From the Institute of Biochemistry
of the University of Lübeck
Director: Prof. Dr. Rolf Hilgenfeld

Genetic algorithms for phase determination in macromolecular crystallography

Dissertation

in Fulfillment of

Requirements

for the Doctoral Degree

of the University of Lübeck

Submitted by

Monarin Uervirojnangkoorn

from Bangkok, Thailand

Lübeck, 2012

First referee: Prof. Dr. Rolf Hilgenfeld
Second referee: Prof. Dr. Jürgen Prestin
Date of oral examination: 13.3.2013

Approved for printing: 13.3.2013

Abstract

Macromolecular X-ray crystallography is a technique used for elucidating macromolecular structures at high resolution. It is an iterative process of the following steps: (1) growing crystals of molecules, (2) performing X-ray diffraction experiments with the crystals and converting diffraction intensities to structure-factor amplitudes (Fourier coefficient amplitudes), (3) obtaining phases by phasing techniques (molecular replacement, experimental phasing, and *ab-initio* phasing, (4) transforming the amplitudes and the phases of the structure factors to an electron density function that reveals molecular structures, (5) building an atomic model into the electron density, (6) refining the model against the structure-factor amplitudes.

While the structure-factor amplitudes are derived directly from the diffraction intensities, phases have to be obtained via an additional step where a variety of computing techniques is used intensively. Two phasing problems are at the focus of this thesis and genetic algorithms were developed to understand and solve the problems. For experimental phasing, a computer program, SISA, was developed to optimize the quality of phases prior to density modification and model building. SISA improved the quality of phases for a few strongest reflections using the electron-density map skewness as the target function. In all test cases, the optimized phases led to improvements of the model building and in one case, a model could be derived where this had been impossible before application of the algorithm. For *ab-initio* phasing, another genetic algorithm was developed to search for solutions for 2-dimensional structures using the structure-factor amplitude correlation as the target function. Three structures with different levels of solvent were artificially generated. Results from the search showed that when information about the structures were lacking, the amplitude correlation was not a useful measure for the quality of phase. With an increasing amount of known information, the usability of the correlation increased and the test structures were recovered by the algorithm. The results also showed that the high-solvent structure required the smallest amount of known information to achieve similar results.

Contents

1 Introduction

Crystallographic studies would cease to exist without computers. The rapid increase of the number of structures solved every year validates this, for such an increase would be impossible by mere handwork. Crystallographers cultivate their structures from collecting diffraction data, constructing electron-density maps, building and refining models, and validating results. All triumph due to the computers complementing the theories.

Macromolecular crystallography has a tremendous merit on biological studies such as enzyme mechanisms and emerging viruses by revealing their 3-dimensional images at high resolution. The concept of Bragg reflection and the Ewald sphere define the geometry of X-ray diffraction. Crystallographers perform the non-trivial process of deducing structure-factor amplitudes from diffracted X-ray intensities. The inverse Fourier transform would reveal the anticipated molecular pictures, if only the phases were present – unfortunately, these phases are lost in the diffraction experiment.

Lack of phases and constraints on the electron-density function permit infinite possibilities for its shape. Direct Methods, based on the relations between the structure factors (Cochran, 1952) and the Tangent Formula (Karle & Hauptman, 1956), celebrated their success due to the recognition of the positive and resolved nature of atoms. Their applications become less efficient with an increasing amount of atoms, hence macromolecules subscribe to other phasing techniques. Crystallographers employ non-crystallographic symmetry and molecular replacement when they can seize comparable structures; if not, they must inevitably rely on experimental phasing.

Of all crystallographic problems, phasing prevails. As computers continue to empower stochastic methods, more possibilities are open for phasing to grasp problems with higher complexity. The emergence of stochastic methods salvages the problems where a less-than-optimum solution is preferable over none. I addressed two

problems in macromolecular crystallography and introduced stochastic algorithms to solve them.

SISA: SIR/SAD phase optimization

Experimental phasing of diffraction data from macromolecular crystals involves deriving phase probability distributions. These distributions are often bimodal, making their weighted average, the centroid phase, improbable, so that electron density maps computed using centroid phases are often uninterpretable. Density modification brings in information about the characteristics of electron density in protein crystals. In successful cases, this allows a choice between the modes in the phase probability distributions, and the electron-density maps can cross the borderline between uninterpretable and interpretable.

Based on the suggestions by Vekhter (2005), I got interested in the impact of assigning low-error phases to a small number of strong reflections prior to the density-modification process, while using the centroid phase as a starting point for the remaining reflections. A genetic algorithm, SISA, was developed to search for optimal phases using the skewness of the density map as a target function. Phases optimized this way are then used in density modification and model building. Experimental data that had failed to give complete structures were selected to demonstrate that SISA could improve the quality of phases. The optimized phases led to greater success in subsequent model building.

Ab-initio phasing: resolving phase ambiguities for 2-dimensional problems

Given a structure consisting of only equal atoms, with information about this structure lacking (coordinates or phases are unknown), the correlation of the observed and the calculated structure-factor amplitudes is not a good measure for the quality of phases. Other measures such as electron-density histograms, connectivity properties, and statistical likelihood also fail to resolve phase ambiguities (Lunin *et al*., 2000). The usages of α-helical polyalanine search fragments cannot solve the problem in the general case (Rodriguez *et al*, 2009). Phase ambiguities are key problems for macromolecular *ab-initio* phasing.

In a 2-dimensional setting, I investigated the usability of structure-factor amplitude correlation as a measure of phase quality in *ab-initio* phasing. The goal was to find out the amount of prior information on the structure needed to rely on structure-factor amplitude correlation. Three structures on 10x10 grids with different levels of solvent were artificially generated. A genetic algorithm that searches for solutions using the structure-factor amplitude correlation as the target function was developed. I demonstrated that usability of the structure-factor amplitude correlation depends on the amount of prior information on the structure and the magnitude of solvent content.

2 Background

The Fourier transform enables us to describe an object in real space (time domain) and reciprocal space (frequency domain). Bragg connected X-ray diffraction to reciprocal lattices, which led to conversion of measured intensities (I) to structure-factor amplitudes (F):

$$F(obs) = kk'\sqrt{I} \qquad (2.1)$$

where k contains all angle-independent corrections such as the Lorentz factor (Pflugrath, 1999) and k' contains all angle-dependent corrections such as the polarization factor (Kabsch, 1988).

For phases, the introduction of isomorphous replacement in proteins (Green *et al*., 1954; Perutz, 1956) pioneered the field of *de-novo* phasing, while the observation of non-crystallographic symmetry (Rossmann & Blow, 1962) led to the development of molecular replacement. Attempts to uncover relations of the structure factors (Karle & Hauptmann, 1950; Sayre, 1952; Cochran, 1952) based on the positive and resolved features of the electron density, were the basis of the success of direct phasing for small molecules. We routinely combine structure-factor amplitudes (F) and phases (α) (eq. 2.2) based on these grounds to obtain an electron-density function (ϱ), which represents the 3-dimensional image of our structures.

$$\varrho\,(\mathbf{x}) = (1/V) \sum \boldsymbol{F}_{\mathrm{h}} \exp\,(-2\pi i \mathbf{hx}), \qquad (2.2)$$
$$\boldsymbol{F}_{\mathrm{h}} = F_{\mathrm{h}} \exp\,(i\alpha_{\mathrm{h}})$$

where $\boldsymbol{h}$ is a vector of reciprocal index, $\boldsymbol{x}$ is a vector of real-space index, and the bold-letter notation of the structure factor represents a complex number with the amplitude and the phase component.

Developments in computing science support us to test our assumptions and gain insights more rapidly. The emergence of stochastic methods, complementary to the deterministic ones, enables us to deal with complex problems. Genetic algorithms (Holland, 1975; Goldberg, 1989) follow stochastic techniques by offering search

procedures based on the concept of natural selection. In crystallography, genetic algorithms were implemented in a variety of computational methods such as small-angle scattering (Svergun & Franke, 2009), powder diffraction (Shankland *et al*., 1997; Harris *et al*., 2004; Feng & Dong, 2007) and *ab-initio* phasing for macromolecular crystals at low resolution (Miller *et al*., 1996; Webster & Hilgenfeld, 2001; Zhou & Su, 2004; Immirzi *et al*., 2009).

2.1 Molecular replacement (MR)

The condensed idea: *Phases from analogous structures*.

Rossmann & Blow (1962) use the Patterson function (Patterson, 1934) to locate similar subunits in crystals of oligomeric protein. The maximum of the function peaks when rotation operators correctly locate the subunits. Recognition of a large number of known structures enhanced this original idea to use other similar structures as search models, which led to current practices of molecular replacement (Rossmann, 2001). Early developments in molecular replacement include the fast rotation and translation functions (Crowther & Blow, 1967; Crowther, 1972). Today, molecular replacement can be carried out using computer programs such as *AMoRe* (Navaza, 1994), *EPMR* (Kissinger *et. al*., 1999), and *Phaser*, which introduced the concept of maximum likelihood in molecular replacement functions (McCoy *et. al*., 2007).

Key requirements for molecular replacement involve obtaining a model with at least 25% sequence identity and finding correct rotation and translation vectors to place the model in the unit cell. We obtain these vectors by identifying a maximum on the Patterson map (autocorrelation map). A 2-dimensional example as illustrated in Fig. 2.1 demonstrates how this method works. Given a known structure ***A***, which shares some similarities to an unknown structure ***B***, a suboptimal orientation and location (Fig 2.1 top) would produce an autocorrelation map with its largest peak (*ccmax*) less than the optimal one (Fig. 2.1 bottom).

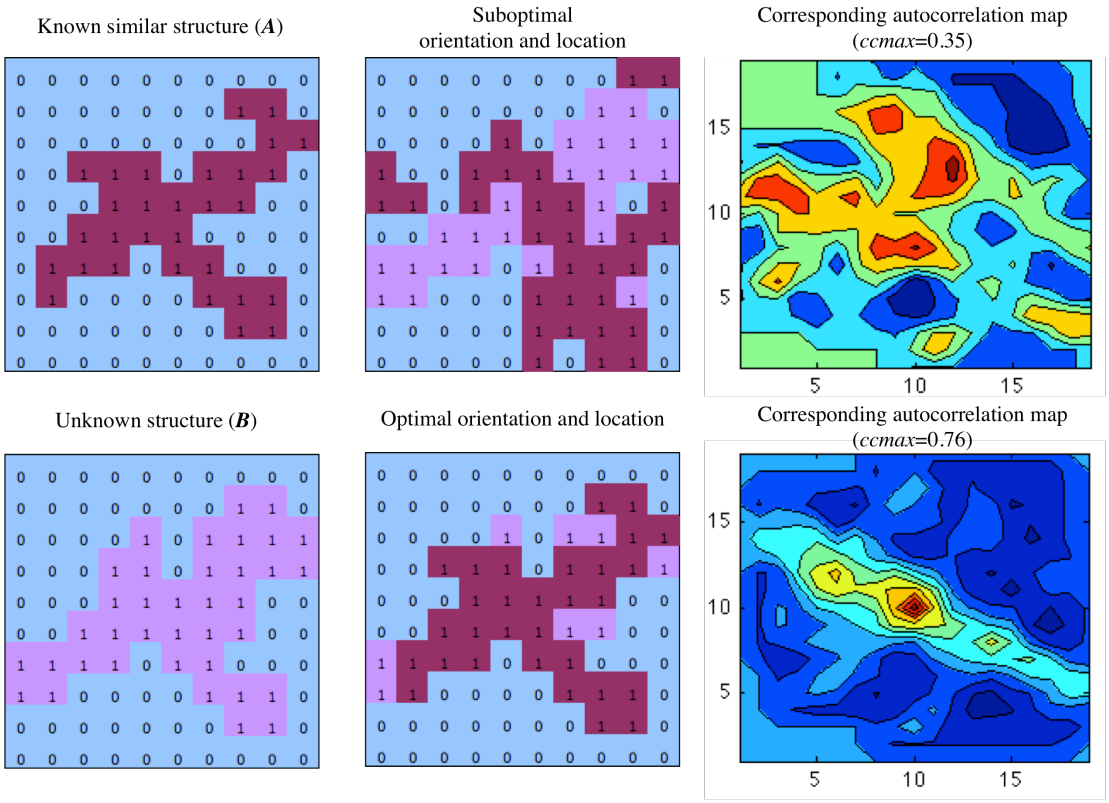


Figure 2.1 An example showing a usage of the Patterson map (the autocorrelation map) to find the optimal orientation and location for superimposing structure *A* onto structure *B*. The suboptimal location results in a smaller Patterson peak (*ccmax*) (top) while the optimal location results in a larger peak (bottom).

Model errors affect the structure-factor amplitudes and consequently the autocorrelation map. *Phaser* treats effects of resolution and these errors by incorporating the maximum likelihood (Read, 2001). The *SIGMAA* function (Read, 1986), employed in *Phaser*, weights a pair of observed and calculated (model) structure-factor amplitudes according to its resolution and the model errors in the calculation of the likelihood function. Since we know the model's sequence identity, we can estimate the value of coordinate errors (root-mean-square or r.m.s. errors) using the relation between sequence identity and r.m.s. error (Chothia & Lesk, 1986). Figure 2.2 shows plots of *SIGMAA* functions calculated with model r.m.s. errors of 0.2, 1.0, and 2.5 Å within a resolution range of 2.0 – 18.0 Å. With the low-error model, the *SIGMAA* function shows large values for high-resolution reflections and begins to fall off in lower resolution ranges. The error-prone model, on the other hand, causes the *SIGMAA* to decrease for both high- and low-resolution reflections. Different behaviors of the *SIGMAA* function affect the summation of the log-likelihood (*LL*).

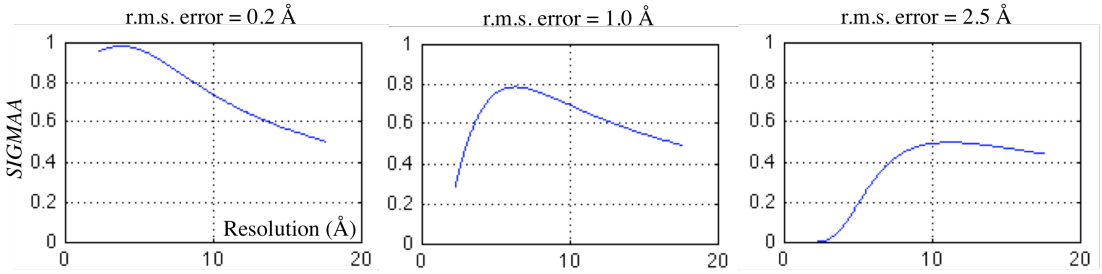


Figure 2.2 SIGMAA functions calculated with different values for model r.m.s. errors.

Fig. 2.3 shows an example of the log-likelihood function calculated from the PDB model of the CcmK1 C-terminal structure (PDB Code: 3DN9; Tanaka *et al.*, 2009) with r.m.s. = 0.2 Å and r.m.s. = 2.5 Å. Since we know that this is the correct model, we would expect that the log-likelihood summation should be very large for an r.m.s. value close to 0. The plots and the calculations of the log-likelihood summation (Fig. 2.3) confirm our expectation; the smallest r.m.s. errors of 0.2 Å yields the largest log-likelihood summation of 1.13 x 10^4.

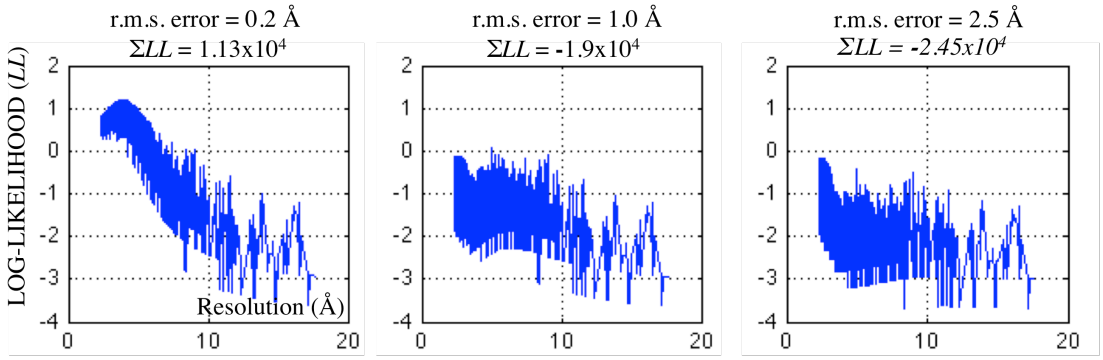


Figure 2.3 Likelihood functions viewed on a logarithmic scale for a comparison of the observed and the calculated normalized structure factors (E value: see eq. 2.18 on page 22) for CcmK1 C-terminal (generated from the PDB coordinates). The calculations were done using r.m.s. values of 0.2, 1.0, and 2.5 Å.

The log-likelihood function affects the search for orientation and location to place the model. Locating a model with high sequence identity would yield a very probable solution. The search still derives some solutions for a low-sequence-identity model; crystallographers can then try to improve these solutions using other techniques.

2.2 Experimental phasing

The condensed idea: *Solving phase equations.*

"Experimental phasing" refers to techniques that obtain phases without structural information (*de-novo* structure determination). Two techniques in experimental phasing are isomorphous replacement and anomalous dispersion. Note that figures and detail explaining the anomalous-dispersion section as presented here are modified after the course on exploiting anomalous scattering in macromolecular structure determination (EMBO'07) at ESRF, France (http://www.esrf.eu/events/conferences/embo2007/weiss_AnomScatt_2007.pdf, with permission of Dr. Manfred Weiss).

2.2.1 Isomorphous replacement

Crystallographers need to obtain isomorphous crystals with heavy atoms or compounds bound to native protein molecules. Obtaining phases from isomorphous crystals involves the following steps:

Step 1. The substructures (heavy atoms/compounds) in the derivative crystals are responsible for providing phase information for the proteins. We obtain structure-factor amplitudes of the protein (F_P) and the derivative (F_{PH}) from diffraction data of the native and the isomorphous crystal respectively. We depend on the isomorphism of the derivative crystal because only under this condition, the summation of the protein structure factors ($\boldsymbol{F_P}$) and the heavy-atom structure factors ($\boldsymbol{F_H}$) will be equal to the derivative structure factors ($\boldsymbol{F_{PH}}$).

$$\boldsymbol{F_{PH}} = \boldsymbol{F_P} + \boldsymbol{F_H} \qquad (2.3)$$

For $\boldsymbol{F_H}$, we derive the amplitudes from the differences between structure factor amplitudes of the protein and of the derivative. These differences are used in the Patterson technique or in Direct Methods to determine phases of the substructures.

Step 2. We substitute $\boldsymbol{F_H}$ and the observed amplitudes of $\boldsymbol{F_P}$ and $\boldsymbol{F_{PH}}$ into equation 2.3 to derive phases for the protein. We use the Harker diagram, which constructs possibilities for the protein phases using two circles. Each circle comprises three

vectors $\boldsymbol{F}_{PH}$, $\boldsymbol{F}_{P}$, and $\boldsymbol{F}_{H}$ (Fig. 2.4*a*). We derive the solutions by determining the angles where the two circles intersect (Fig. 2.4*b*).

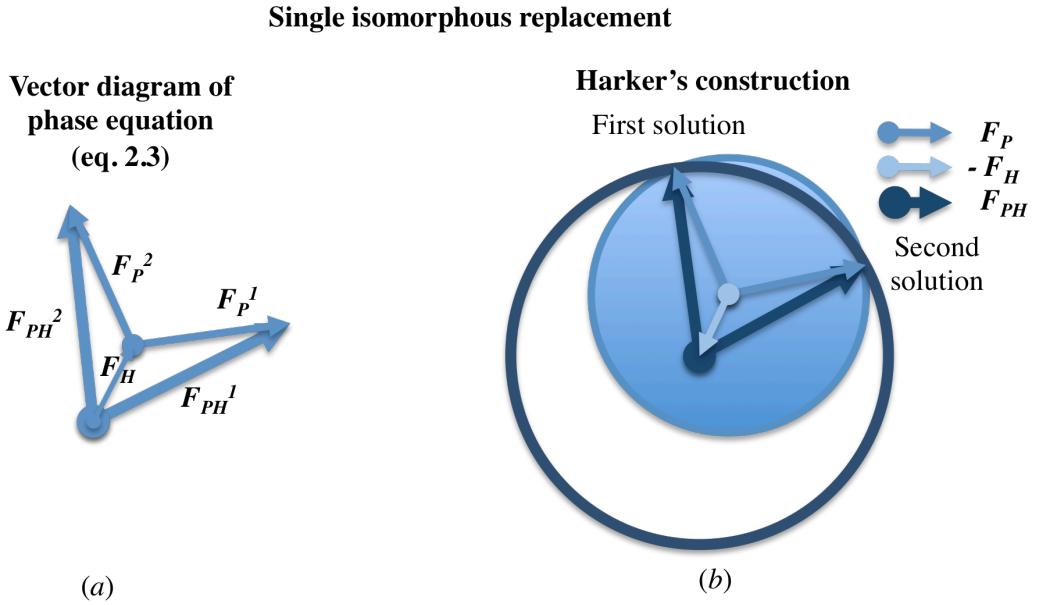


Figure 2.4 Single Isomorphous Replacement (SIR) (a) Phase equations represented by two triangles; each formed by three vectors $\boldsymbol{F}_{PH}$, $\boldsymbol{F}_{P}$, and $\boldsymbol{F}_{H}$. (b) Solving phase equations using the Harker diagram.

Step 3. We convert the solutions in the Harker diagram to a phase probability distribution (Blow & Crick, 1959; Otwinowski, 1991; McCoy *et al*., 2004). Phases derived from the equation expand along the range 0 to 2π, leading to the derivation of a confident level for each phase angle. We calculate the error quantity (ε) from the *lack of closure* of the summation of the three vectors $\boldsymbol{F}_{PH}$, $\boldsymbol{F}_{P}$, and $\boldsymbol{F}_{H}$ (Fig. 2.5*a*) for each phase angle (often with 5° interval). We turn a collection of ε to a phase probability distribution that describes the phases with a degree of confidence known as *figure of merit* (*FOM*).

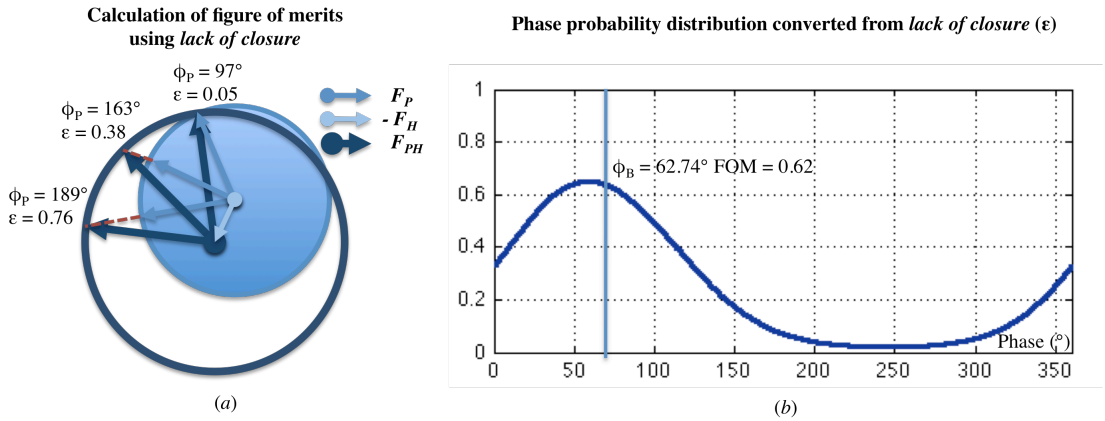


Figure 2.5 Conversion of phases from the Harker diagram to a phase probability distribution. (a) Calculation of figures of merit (*FOM*) using *lack of closure*. (b) The phase probability distribution calculated from (a) with its best phase (ϕ_B) and the corresponding *FOM*.

Crystallographers encapsulate a phase probability distribution in the Hendrickson-Lattmann coefficients (Hendrickson & Lattman, 1970) for subsequent applications. The coefficient consists of four real numbers represented as *A*, *B*, *C*, and *D* in equation 2.4. We extract the centroid phase (ϕ_B) and its corresponding *FOM* from the result calculated from the formula.

$$P(\phi) = N \exp(A \cos(\phi) + B \sin(\phi) + C \cos(2\phi) + D \sin(2\phi)) \qquad (2.4)$$

where N is a normalization constant such that $\int P(\phi)\, d\phi = 1$.

Single isomorphous replacement (SIR) phasing leads to ambiguity in phase solutions, which affects the resulting electron density map to become uninterpretable. Crystallographers employ density-modification techniques that use the characteristics of electron density in protein crystals to improve the quality of phases. Maps that are still uninterpretable will require further treatments to resolve the ambiguity of phases: more derivatives for the multiple isomorphous replacement (MIR) or more wavelengths for the anomalous dispersion.

2.2.2 Anomalous dispersion (scattering)

Prior to solving the phase equation, the anomalous dispersion method requires either derivative crystals such as in the case of the isomorphous replacement or so called "selenomethionine (Se-Met) crystals" produced by replacing methionine residues with selenomethionine. The two approaches produce the same effect; waves with energy close to the transition energy of the electron in heavy atoms or selenium will

be absorbed and eject core electrons. This phenomenon causes the wave to scatter anomalously. Fig. 2.6 shows differences between normal scattering and anomalous scattering; for normal scattering, Friedel's law holds, so structure factors of a Friedel pair (***F*(*h k l*)** and ***F*(-*h* –*k* –*l*)**) are equal, and for anomalous scattering, Friedel's law is broken since anomalously scattering waves add two additional quantities ($\Delta f'$ and $\Delta f''$) to the normal scattering form factor (f^o).

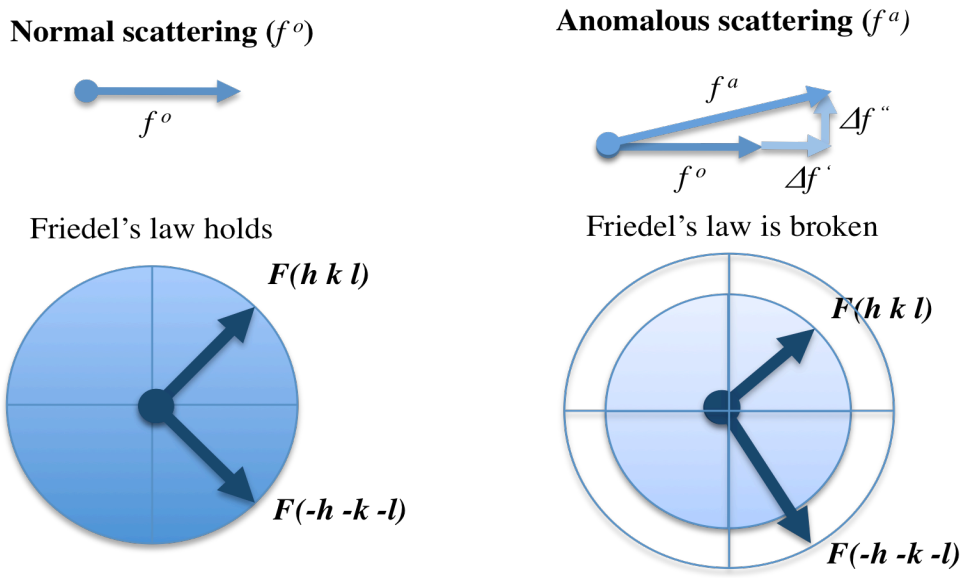


Figure 2.6 Comparison of normal and anomalous scattering. Friedel's law holds for the first but is broken for the latter. In anomalous scattering, the structure factors of the Friedel pair are not equal for both the magnitudes and the phases.

We can resolve the ambiguity of phases resulting from single isomorphous replacement when the derivative crystal diffracts anomalously. This technique is known as single isomorphous replacement with anomalous scattering (SIRAS), which requires:

- one native crystal with normal scattering (F_P)
- one derivative crystal with normal scattering (F_{PH})
- and the same derivative crystal with anomalous scattering (F_{PH}^+ and F_{PH}^-)

The two additional quantities, which are added to the scattering form factor, break the symmetry of the structure factors (F_{PH}^+ and F_{PH}^-) as described by Friedel's law (Fig. 2.6), therefore yielding additional phase information to resolve the phase ambiguity occurring in SIR. We supply $\boldsymbol{\Delta F_H'}$ and $\boldsymbol{\Delta F_H''}$ in the phase equation and arrive at

$$\boldsymbol{F}_{PH}^{+} = \boldsymbol{F}_{P} + \boldsymbol{F}_{H}^{o} + \boldsymbol{\Delta F}_{H}^{'} + \boldsymbol{\Delta F}_{H}^{''} \qquad (2.5)$$

$$\boldsymbol{F}_{PH}^{-} = \boldsymbol{F}_{P} + \boldsymbol{F}_{H}^{o} + \boldsymbol{\Delta F}_{H}^{'} - \boldsymbol{\Delta F}_{H}^{''} \qquad (2.6)$$

where

- $\boldsymbol{F}_{H}^{o}$ is the heavy atom structure factor (from normal scattering), which can be derived using the Patterson technique or Direct Methods in the same way as it is done in SIR.
- $\boldsymbol{\Delta F}_{H}^{'}$ is the dispersive component of the anomalous scattering from the heavy atom. Its phase depends on $\boldsymbol{F}_{H}^{o}$ and its magnitude can be extracted from the absorption curve.
- $\boldsymbol{\Delta F}_{H}^{''}$ is the anomalous component of the anomalous scattering from the heavy atoms. Its phase is orthogonal to $\boldsymbol{\Delta F}_{H}^{'}$ and its magnitude can also be extracted from the absorption curve.

Fig. 2.7*a* shows vector constructions of equations 2.5 and 2.6. To resolve the ambiguity of SIR phases, we construct the Harker diagram as seen in Fig 2.4*b* with two circles representing F_P and F_{PH}, then add two additional circles for F_{PH}^{+} and F_{PH}^{-}. The extra two circles are constructed by adding the negative $\boldsymbol{\Delta F}_{H}^{'}$ to $\boldsymbol{F}_{H}^{o}$ and setting center to draw the first circle F_{PH}^{+} on one side of $\boldsymbol{\Delta F}_{H}^{''}$ and the second circle F_{PH}^{-} on the other side of $\boldsymbol{\Delta F}_{H}^{''}$. The ambiguity of the two solutions from SIR is resolved and one arrives at a single solution as seen at the intersecting point of the four circles.

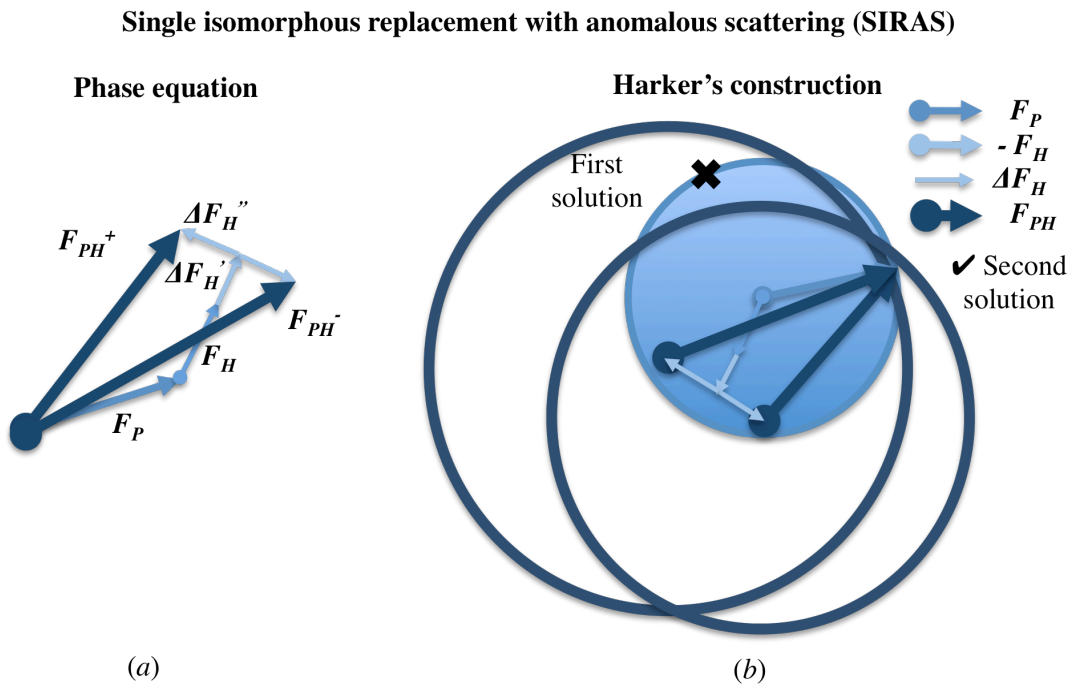


Figure 2.7 Solving the phase equation using the anomalous scattering from a derivative crystal (SIRAS). (a) A representation of equation 2.5 and 2.6 in vector forms. (b) A Harker's construction is added with phase information from the anomalous scattering to break the phase ambiguity in SIR.

As one gains understanding about how the protein phases can be derived using the anomalous scattering technique based on a single derivative crystal (SIRAS) as explained here, it is now easy to understand the phasing technique of single anomalous dispersion (SAD) and why the protein phases obtained from the technique are still ambiguous.

Another technique to resolve phase ambiguity involves a native crystal with either sulfur (Weiss *et al.*, 2001; Liu *et al.*, 2012) (under the condition that a relatively long wavelength can be applied), or selenomethionine. The technique requires only one data set, which contains the anomalous scattering components F_{PH}^{+} and F_{PH}^{-}. To understand the relationship between the two components, Fig. 2.8 shows a version of the vector forms as presented in Fig. 2.7*a* with unavailable parameters represented as dotted line. The three closed vectors seen in bold (Fig 2.7*a*) are comparable to the three vectors formed in SIR (Fig 2.5*a*). We derive $\boldsymbol{\Delta F_H''}$ from the positions of the scatterers (sulfur or selenium) and construct the Harker diagram in a similar way as it is done in SIR; phases from SAD phasing remain ambiguous and crystallographers need to apply density modification to improve the quality of phases.

Phase equation for single anomalous dispersion (SAD)

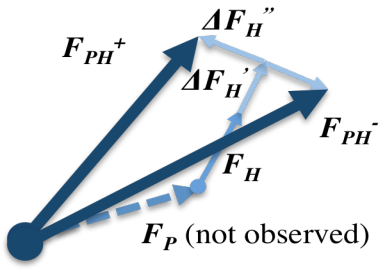


Figure 2.8 The phase equation for single anomalous dispersion (SAD). The protein structure factor is not observed, leaving only the three closed vectors F_{PH}^+, F_{PH}^-, and $\Delta F_H''$.

In case phase ambiguity still exists after SIRAS or SAD, crystallographers need to employ more derivatives or wavelengths. The techniques are known as multiple isomorphous replacement with anomalous scattering (MIRAS), when we use more than one derivative crystal, and as multiple anomalous dispersion (MAD) when we use more than one wavelength. Table 2.1 provides a summary of available experimental phasing techniques.

Method	**Detail**	
SIR	*Single isomorphous replacement*	
	Required data sets • A native crystal (F_P) • A derivative crystal (F_{PH})	**Phase equation** $F_{PH} = F_P + F_H$ Solved by Harker's construction (Fig. 2.4*b*)
SIRAS	*Single isomorphous replacement with anomalous scattering*	
	Required data sets • A native crystal (F_P) • A derivative crystal (F_{PH}) • The same derivative crystal with anomalous scattering (F_{PH}^+ and F_{PH}^-)	**Phase equations** $F_{PH}^+ = F_P + F_H^o + \Delta F_H' + \Delta F_H''$ $F_{PH}^- = F_P + F_H^o + \Delta F_H' - \Delta F_H$ Solved by Harker's construction (Fig 2.7*b*)

Method	Detail	
SAD	*Single anomalous dispersion*	
	Required data sets • A native crystal with sulfur (with long wavelength) or selenium atom(s) with anomalous scattering (F_{PH}^{+} and F_{PH}^{-})	**Phase equations** are solved in the same way as SIR with the three closed vectors being $\boldsymbol{F_{PH}^{+}}$, $\boldsymbol{F_{PH}^{-}}$, and $\boldsymbol{\Delta F_{H}''}$.
MIR	*Multiple isomorphous replacements* Requirements and phase equations are similar to SIR; only more derivative crystals are needed for more experiments.	
MIRAS	*Multiple isomorphous replacements with anomalous scattering* More than one derivative crystal diffracting anomalously.	
MAD	*Multiple anomalous dispersion* A crystal with SAD requirements diffracting anomalously at different wavelengths.	

Table 2.1 A summary of experimental phasing methods.

2.3 *Ab-initio* phasing for macromolecules

The condensed idea: *Finding relations between phases of structure factors.*

Ab-initio phasing refers to techniques that require only structure-factor amplitudes for phase determination. The basis of Direct Methods for small molecules provides a foundation for their adaptation and extension for macromolecules. This basis includes Sayre's equation (Sayre, 1952), Cochran's integral (Cochran, 1952), and the tangent formula (Karle & Hauptman, 1956).

Current progress in *ab-initio* phasing methods for macromolecular crystals diverges into two directions. Development in the first direction inherited the original Direct Methods for small molecules and extended them for macromolecules. Current computer programs include *Shake-and-Bake* (Debaerdemaeker & Woolfson, 1983) and *SHELXD* (Sheldrick & Gould, 1995). These programs limit the size of the

applicable molecules to < 1,000 non-H-atoms in the asymmetric unit, which led to a call for an alternative concept. The second direction drew in macromolecular features to enhance the *ab-initio* phasing. Methods, which explored an imitation of macromolecules, include a series of developments as summarized in Lunin *et al.*, (2000; 2012), spherical scatterers (Subbiah, 1991), binary scatterers (Webster & Hilgenfeld, 2001; Su, 2008), and α-helical fragments as scatterers (Rodríguez *et al.*, 2009).

2.3.1 Basis of Direct Methods

Direct Methods rely on the relations between the structure-factor amplitudes derived from the positive and resolved features of the electron-density function. When reflections up to atomic resolution can be measured, the electron-density function representing the molecules and the solvent will be positive almost everywhere with the locations of atoms resolved from each other (Fig. 2.9). The early pioneers of Direct Methods (Harker & Kasper, 1948; Karle & Hauptmann, 1950; Goedkoop, 1950; Sayre, 1952; Cochran, 1952) used these features to derive the relations between the structure factors.

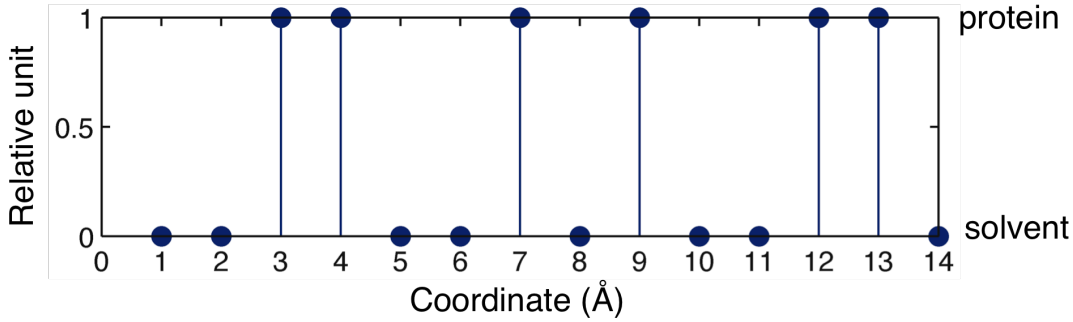


Figure 2.9 Discrete and resolved features of a 1-dimensional electron-density function mimicking an atomic structure with 6 atoms in a unit cell of length 14 Å.

Sayre's equation

The condensed idea: *There are limited choices of signs that can be assigned to the structure factors, to yield a positive and resolved electron-density function.*

David Sayre (1952) presented support for the idea that the signs (phases) of the structure factors in the centrosymmetric case were deducible from the diffraction data. By incorporating the convolution theory, he showed that a set of structure-factor products should follow a set of signs. An electron-density function (ϱ) and its squared function (ϱ^2) are comparable when the structure consists of only positive and resolved atoms; thus, the structure factor of the function ($\boldsymbol{F}$) and its squared function ($\boldsymbol{F}^{sq}$) should be almost identical (Fig 2.10*a*). Another contrasting example demonstrates that $\boldsymbol{F}$ and $\boldsymbol{F}^{sq}$ are different when the electron density function (ϱ) fails to follow the conditions (Fig 2.10*b*).

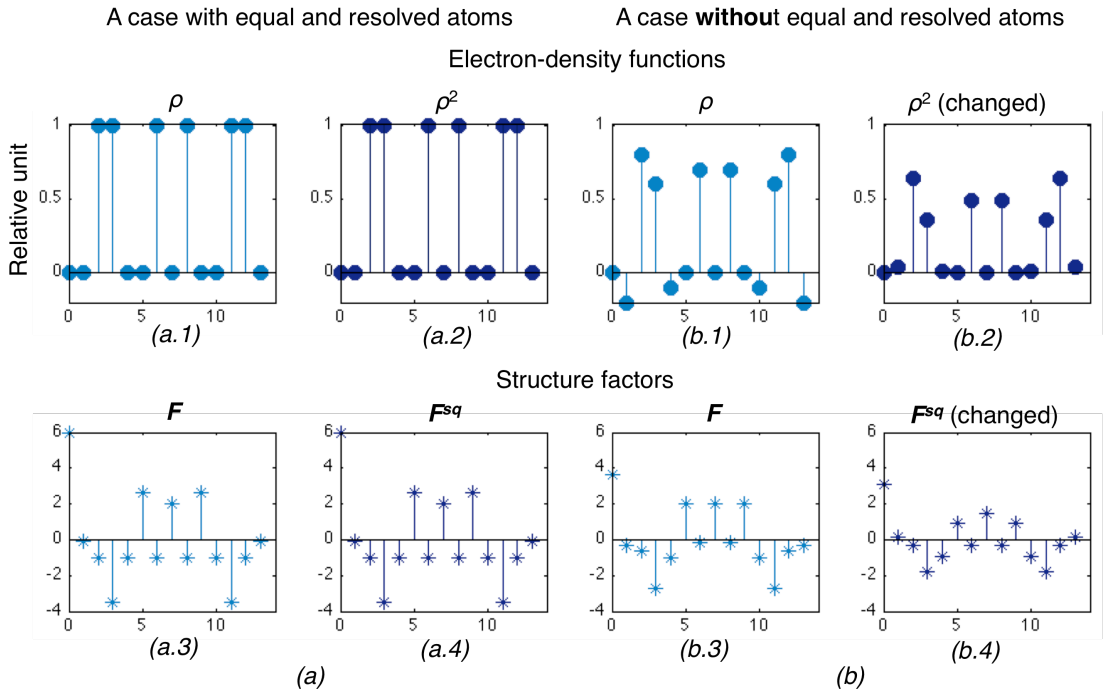


Figure 2.10 Fourier transforms of an electron-density function (ϱ) and its squared fuction (ϱ^2) for (a) a case with equal and resolved atoms, ϱ and ϱ^2 are equal (a.1 and a.2); they have the same structure factors (a.3 and a.4) (b) a case without equal and resolved atoms, ϱ and ϱ^2 are not equal (b.1 and b.2); they have different structure factors (b.3 and b.4).

Sayre derived relations between the structure factors based on the convolution theorem for the case when $\boldsymbol{F}^{sq}$ is nearly identical to $\boldsymbol{F}$. From the convolution theorem, the Fourier transformation of the squared electron-density function (ϱ^2) is equal to a convolution of its structure factors

$$F^{sq}(h,k,l) = \frac{1}{V}\sum_{p}\sum_{q}\sum_{r}F(p,q,r)F(h-p,k-q,l-r) \quad (2.7)$$

where the right-hand side is a self-convoluting process for all reciprocal indexes p, q, and r, and V is the volume of the unit cell. Since $\boldsymbol{F}^{sq} \cong \boldsymbol{F}$ for the equal- and resolved-atoms function, we can substitute $\boldsymbol{F}^{sq}$ with $\boldsymbol{F}$, move the volume (V) to the left-hand side, and arrive at

$$V \cdot F(h,k,l) = \sum_{p}\sum_{q}\sum_{r}F(p,q,r)F(h-p,k-q,l-r) \quad (2.8)$$

Equation 2.8 is known as Sayre's equation, from which signs or phases of the structure factors can be determined. In order to see how this relation can be used to derive the signs of the structure factors, consider the convolution of a selected index $p = 3$ for a Fourier series of a 1-dimensional function with length 5 Å

$$\begin{aligned} 5\,\boldsymbol{F}(3) &= \boldsymbol{F}(3)\boldsymbol{F}(3-0) + \boldsymbol{F}(3)\boldsymbol{F}(3-1) + \boldsymbol{F}(3)\boldsymbol{F}(3-2) + \boldsymbol{F}(3)\boldsymbol{F}(3-3) + \boldsymbol{F}(3)\boldsymbol{F}(3-4) \\ &= \boldsymbol{F}(3)\boldsymbol{F}(3) + \boldsymbol{F}(3)\boldsymbol{F}(2) + \boldsymbol{F}(3)\boldsymbol{F}(1) + \boldsymbol{F}(3)\boldsymbol{F}(0) + \boldsymbol{F}(3)\boldsymbol{F}(-1) \qquad (2.9) \end{aligned}$$

For a centrosymmetric structure, structure factors can only have a phase value of either 0 or π. Substituting 0 for the phase of $\boldsymbol{F}(3)$ in equation 2.10 results in

$$\begin{aligned} 5\,|F(3)| = {} & |F(3)|^2 + |F(3)|\,|F(2)|\exp(i\alpha(2)) + \\ & |F(3)|\,|F(1)|\exp(i\alpha(1)) + |F(3)|\,|F(0)| + |F(3)|\,|F(-1)|\exp(i\alpha(-1)) \quad (2.10) \end{aligned}$$

We can only select choices of sign for $\alpha(2)$, $\alpha(1)$, and $\alpha(-1)$ to make the two sides of equation 2.10 equal. We know that the choices of sign are correct when we substitute them to other convolution terms (e.g. for $p = 1$ or $p = 2$) and they also fulfill those equations.

It can also be realized now that those terms with large amplitude would likely be the ones that determine the sign of the whole summation and that when the number of the Fourier terms is large (in the case of large structures), finding the correct signs that meet all the equations can become very difficult.

Cochran's integral and distribution

The condensed idea: *For a positive electron-density function, its cubed function will result in another positive function. The cubed function can be interpreted as the double convolution of the structure factors, which defines the relation for three reflections. To yield a positive function to the extent that the integral of the cubed function is maximum, there should be as many cases as possible where the signs of the two structure factors in the first convolution are equal to the third structure factor.*

Cochran (1952) provided another view of *ab-initio* phase determination based on relations of the structure factors. He pointed out that structure factors were redundant and the inverse Fourier transform using some of them could already regenerate a recognizable electron-density function. The selected Fourier terms should fulfill

$$\{\sum_{1-p} F_s^2\}^{1/2} \geq F(0) \qquad (2.11)$$

where p is the number of selected Fourier terms and $\boldsymbol{F}_\mathbf{s}$ denotes the structure factor of a point atom.

Examples of this construction for a 1-dimensional function of a centric structure using a partial series with two and four terms are shown in Fig. 2.11*b-c*. The structure consists of 6 atoms, making $F(0) = 6$. The sums of the left side of 10 for these chosen two and four Fourier terms are 4.94 and 6.16 accordingly. The inverse Fourier transforms show a recognizable electron-density function (Fig. 2.11*c*) only for the second case where $\{\Sigma_{1\text{-}p} \boldsymbol{F}_\mathbf{s}^2\}^{1/2} > 6$.

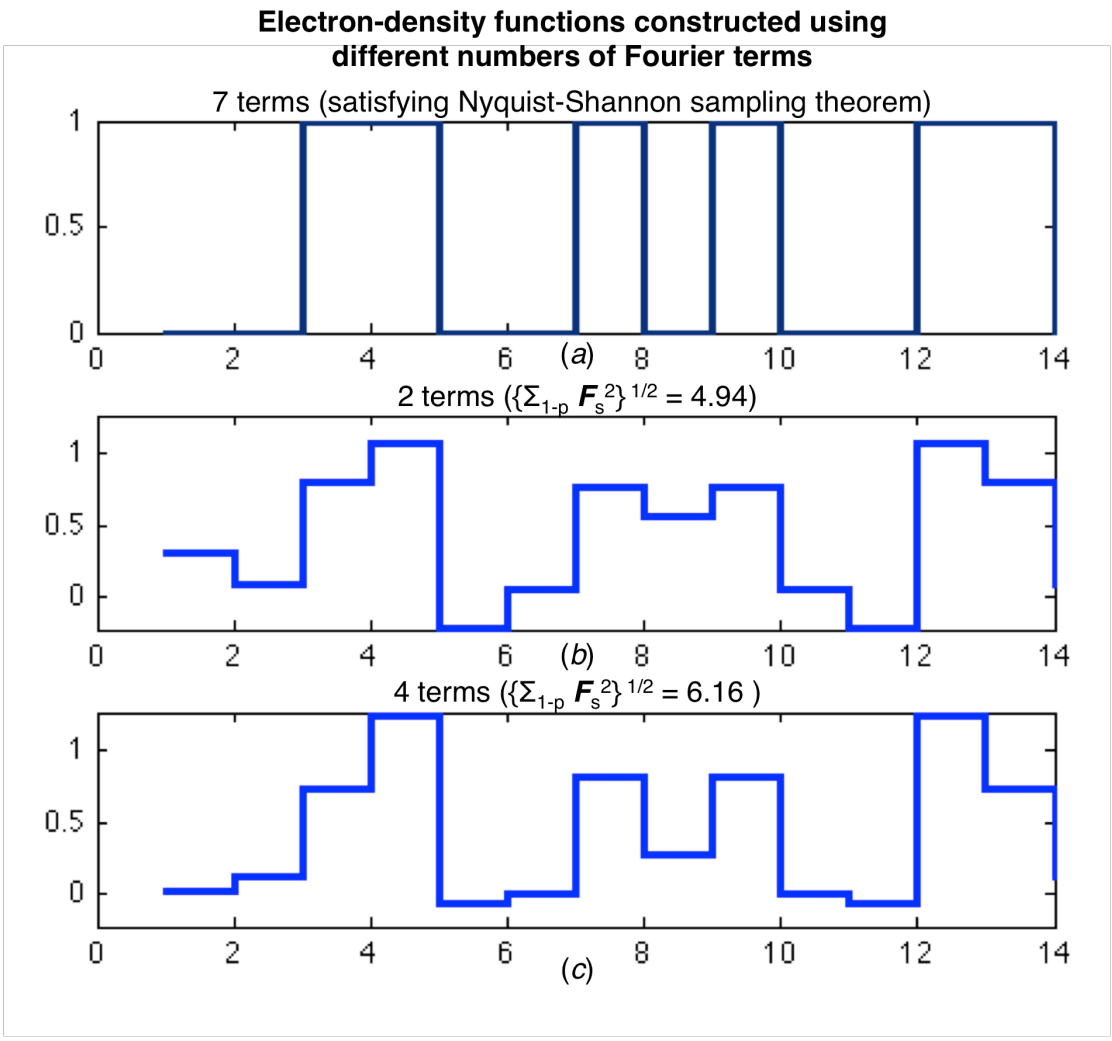


Figure 2.11 Numbers of Fourier terms required to construct a recognizable electron-density function illustrated by using constructions of the function based on different numbers of Fourier terms. (a) Original structure. (b) Two terms with $\{\Sigma_{1\text{-p}} \boldsymbol{F}_s^2\}^{1/2} < \boldsymbol{F}(0)$. (b) Four terms with $\{\Sigma_{1\text{-p}} \boldsymbol{F}_s^2\}^{1/2} > \boldsymbol{F}(0)$.

Cochran suggested that an integral of a cubed function (eq. 2.12) could differentiate a positive function from a real function; therefore, it may be used as a measure for correct phases. Fig. 2.12 demonstrates two 1-dimensional functions calculated from four Fourier terms. The function with four correct signs (Fig. 2.12*a*) yields a value of the integral larger than the value calculated from the function with only two correct signs (Fig. 2.12*b*).

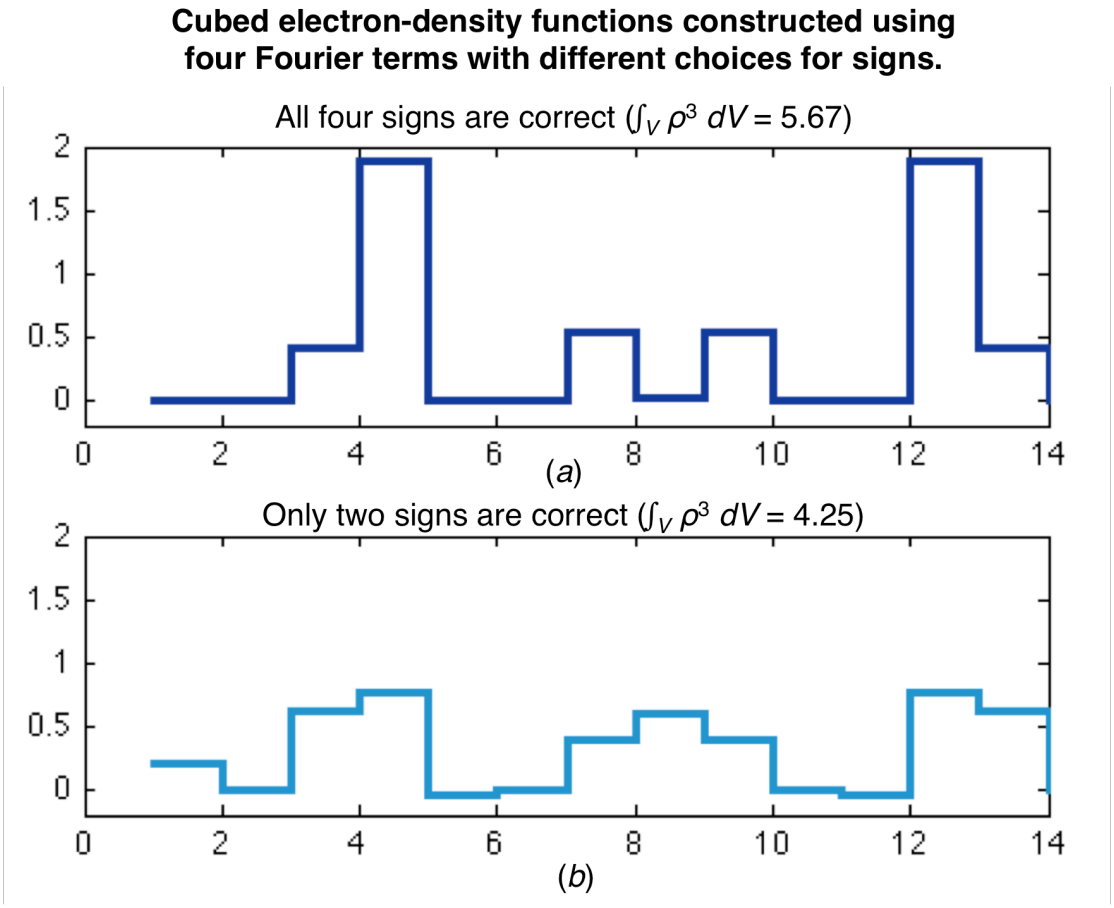


Figure 2.12 Using integrals of the cubed electron-density functions to determine the correct signs for the structure factors. The two plots show the inverse Fourier transform of four Fourier terms with (a) four correct signs. (b) only two correct signs.

$$\int_V \varrho^3\, dV \qquad (2.12)$$

An integral of a cubed function in real space is comparable to a summation of a double convolution in reciprocal space (eq. 2.13), which implies another relation, in addition to the one by Sayre, for three structure factors (eq. 2.14). Following the sign (s) derivation in equation 2.14 results in a maximum of the integral (eq. 2.12), which may lead to a correct recovery of the structure.

$$\int_V \varrho^3\, dV = \Sigma_{1\text{-}p}\, \boldsymbol{F}_s\,(\mathrm{h})\, \boldsymbol{G}_s\,(\mathrm{h}), \text{ where}$$
$$\boldsymbol{G}_s\,(\mathrm{h}) = 1/\mathrm{V}\, \Sigma_{\mathrm{h}'}\, \boldsymbol{F}_s\,(\mathrm{h}')\, \boldsymbol{F}_s\,(\mathrm{h}+\mathrm{h}') \qquad (2.13)$$

$$s(\mathrm{h}) = s(\mathrm{h}')\, s(\mathrm{h}+\mathrm{h}') \quad (2.14)$$

For non-centrosymmetric structures, phases of the structure factors can be any value from 0 to 2π. Cochran (1955) applied the same idea onto the non-centrosymmetric cases and introduced a relation between the phases of structure factors as

$$\alpha(\mathrm{h}) = \alpha(\mathrm{h'}) + \alpha(\mathrm{h} - \mathrm{h'}) \quad (2.15)$$

The relations (eq. 2.14 and eq. 2.15) return the maximum for the integral of the cubed function; however, they fail to assure correct phase determination for large structures. Cochran (1955) estimated a probability for phase determination of a triplet phase (ϕ_3) (eq. 2.16) given three normalized structure factors (E_h, E_k, and E_{h-k}) and the number of non-H atoms (N) (eq. 2.17).

$$\phi_3 = \alpha_h + \alpha_k + \alpha_{h-k} \approx 0 \quad (2.16)$$

$$P(\phi_3 \mid \frac{2 \mid E_h E_k E_{h-k} \mid}{\sqrt{N}}) = \frac{1}{2\pi I_0(\frac{2 \mid E_h E_k E_{h-k} \mid}{\sqrt{N}})} \exp[(\frac{2 \mid E_h E_k E_{h-k} \mid}{\sqrt{N}}) \cos\phi_3] \quad (2.17)$$

where I_0 is a zeroth-order Bessel function of the first kind and E is a normalized structure factor that can be calculated as

$$E = \sqrt{\frac{F^2}{\varepsilon \sum_{i=1}^{N} f_i^2}} \quad (2.18)$$

where ε (epsilon-factor) is a measure of how often a reflection is superimposed onto the same diffraction spot due to crystallographic symmetry, f is the atomic-scattering factor, and N is the number of atoms in the asymmetric unit.

The probability calculated from equation 2.17 is very small when N (number of atoms) is large - this explains the limitation of the method for macromolecules.

The tangent formula and its practical applications in Direct Methods

The condensed idea: *Extending sets of phases by triplet relations.*

Equations 2.14 and 2.15 indicate that the phase of the third reflection can be derived from phases of the other two reflections involved in the relation. From the cosine law and the normalized structure factors (E), this third phase can be calculated using

$$\tan\varphi_h = \frac{\sum_k |E_k E_{h-k}| \sin(\varphi_k + \varphi_{h-k})}{\sum_k |E_k E_{h-k}| \cos(\varphi_k + \varphi_{h-k})} \qquad (2.19)$$

where φ_h is involved in a series of triplet relations:

$$\varphi_h = \varphi_{h2} - \varphi_{h-h2}$$
$$\varphi_h = \varphi_{h3} - \varphi_{h-h3}$$
$$\varphi_h = \varphi_{h4} - \varphi_{h-h4} \, etc.$$

as proposed by Karle & Hauptman (1956). Equation 2.19 is the simplest form of the Tangent Formula, which is the main tool used in conventional Direct Methods.

A complete implementation of Direct Methods requires an understanding of diffraction and symmetry. The theory and practice of Direct Methods are covered briefly by Gilmore (2000) and comprehensively by Giacovazzo (2006).

2.4 Genetic algorithms

The condensed idea: *Finding solutions based on the "survival of the fittest".*

The primary concern of the work described in this thesis is about finding solutions for the phase problem under the scopes of different phasing methods (experimental SIR/SAD phasing and *ab-initio* phasing). These tasks involve finding protein phases in the range of 0 to 2π; thus, the size of the problem could exponentially grow with increasing choices for phases and amount of reflections selected. Finding an exact set of phases for 1,000 reflections, given that the phase can only be π/4, 3π/4, -π/4, or -3π/4, would require testing $4^{1,000}$ combinations of phases, a number that is too large to be computed. Yet in most cases, the number of reflections could be up to a few ten thousands or more and choices of phase can be any real number from 0 to 2π. To obtain some answers, stochastic methods may be more suitable for this type of problem.

The problem of finding the correct phases is also highly complex since its solution landscape is prone to local minima. Types of stochastic search such as hill climbing

become less efficient when dealing with this kind of problem. To avoid local minima, other types of algorithms such as genetic algorithms or simulated annealing are more suitable.

Genetic algorithms were pioneered by Holland (1975) and have been used as search, optimization, and machine-learning tools. They encapsulate and perturb problem settings through chromosomes and genetic operators. For example, we can use the algorithms to solve a problem of finding a binary string that produces the maximum value from a function

$$\sum_{i=1}^{N} X_i \quad (2.20)$$

where X_i is the value stored on bit i of a chromosome with length N (Fig. 2.13).

The genetic algorithms encode real values representing candidates for the solution in a binary-type chromosome. The search evaluates a fitness value for each chromosome using the given target function (eq. 2.20). At each generation, the algorithms select a pair of chromosomes and apply the genetic operators on them to produce offspring that represents two more chromosomes. These operators play a crucial role by allowing binary bits in the chromosomes to be perturbed through crossover, mutation, or inversion. The offspring would get propagated to the next generation or not, depending on their fitness calculated from the target function (eq. 2.20). The procedure terminates either when one of the chromosomes demonstrates acceptable fitness values or when the predefined maximum number of generations is reached.

No.	Chromosome (Length = 20)	*Fitness*	*Percent of total*
1	0 1 1 0 1 1 1 0 0 0 1 0 1 1 0 1 0 0 0 0	9.00	20.93
2	1 1 0 1 1 1 1 1 0 1 0 1 0 0 1 1 0 1 0 0	12.00	27.91
3	0 1 0 0 0 1 0 0 1 0 0 0 0 0 1 1 1 0 1 0	7.00	16.28
4	0 1 0 0 0 0 1 0 1 1 1 0 0 1 1 0 0 1 1 1	10.00	23.26
5	1 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 1	5.00	11.63
Total		43.00	100.00

(*a*)

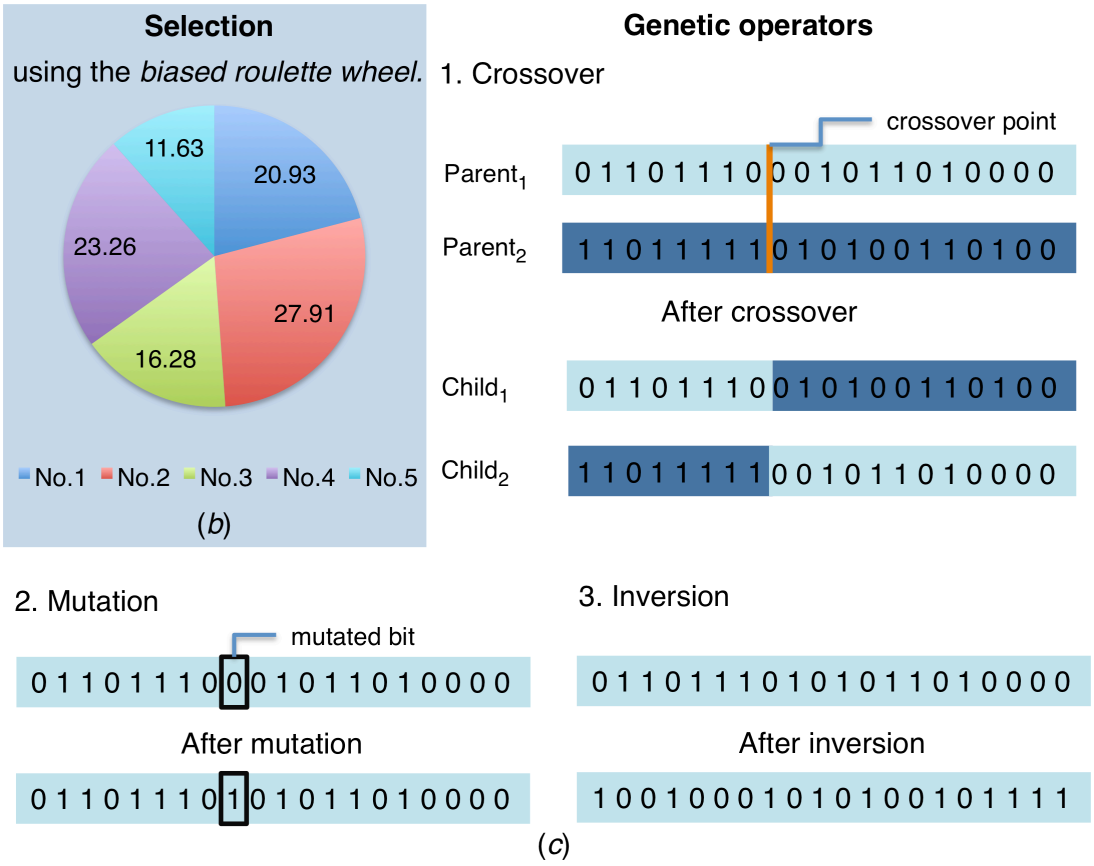


Figure 2.13 An example of genetic algorithms for solving equation 2.20. (a) A representation of chromosomes, fitness values (calculated using equation 2.20), and percent-of-total values. (b) Selection process using the *biased roulette wheel* – each slot, which represents a chromosome, has its size proportionate to its fitness value or its percent of total. (c) Three genetic operators: a crossover operator with 1 crossover point swaps chromosome sections specified by the crossover point to produce two offspring; a mutation operator with 1 mutation bit inverts the value of the selected bit; an inversion operator inverts all the bits in a chromosome.

There are a number of ways to implement genetic algorithms. The four operations, selection, crossover, mutation, and inversion can be performed in different ways depending on the complexity of the problem. The example as shown in Fig. 2.13(*b*) illustrates the selection process using the *biased roulette wheel*. Each slot on the

roulette wheel, which represents a chromosome, has its size proportionate to its fitness value (eq. 2.20) or its percent of total. Each time a roulette ball is rolled, it is likely going to get into the hole with the bigger size. However, the smaller-size holes would still have a chance to get selected but only with less probability. This is an example of how the genetic algorithms use probability to avoid local minima. Allowing some less fit chromosomes to remain in the population helps retaining variants, which might emerge to be good solutions in later stages of the search process. For more complex problems, there are also other more sophisticated selection techniques, which are usually designed to prevent the *crowding* problem (Mitchell, 1997). This problem occurs when fitter populations reproduce themselves too fast, leaving fewer variants in the genetic pool.

The genetic algorithms perturb binary bits in each chromosome by applying the crossover, mutation, and the inversion operators. The simplest forms of crossing genes (series of bits) on a pair of chromosomes are *1- or 2- point crossover* operators, where 1 or 2 positions are selected and the two chromosomes can exchange their genes based on the slices defined by these positions. These crossover operators are usually effective on schematized problems, e.g. the first 5 bits represent variable *X* and the next 7 bits represent variable *Y*. Another type of crossover is the *uniform crossover* (Syswerda, 1989). The uniform crossover generates a cross template by randomly selecting locations for the exchange across the chromosome. The method may perform better for some problems with large search space, giving variants in a population the potential to improve the performance of the search (De Jong & Spears, 1991).

Another genetic operator, which has a minor role in the search process, is the mutation operator. The genetic algorithms select bits on a chromosome randomly and adjust their values (‘0’ → ‘1’ and ‘1’ → ‘0’ for a binary chromosome). The mutation operator is usually set to occur infrequently, leaving most of the manipulation tasks to the crossover operator. Through the process of natural selection, the genetic algorithms derive the fittest solution as the output.

Applications of genetic algorithms involve transforming problem context to these operations provided in the algorithms. Choices of problem encapsulation and genetic operators determine the level of success at the end of the optimization.

3 SISA: SIR/SAD phase optimization

3.1 Introduction

Experimental SAD phasing allows us to obtain phasing information by solving equations based on differences between Friedel pairs of structure factors. The possible solutions for a reflection are represented in the form of a probability distribution (Blow & Crick, 1959; Otwinowski, 1991; McCoy *et al*., 2004). Toward solving a structure, this phasing information is passed onto density modification, which exploits expected features of molecular maps to break the ambiguity existing in the initial distribution (Wang, 1985). In the case where many reflections have accurate phases, obtaining an interpretable map is straightforward. In contrast, when the majority of the reflections are poorly determined, resolving the ambiguity remains a difficult task.

A SAD dataset of Gene V protein (Skinner *et al*., 1994) was selected as an example of this situation. Solving this structure from just the peak wavelength of SAD data is challenging due to the low quality of the electron density map obtained after density modification. The structure could however be solved from a MAD dataset. This is a common situation when experimental phases result in a poor map.

Vekhter (2005) presented an interesting study where it was shown that by assigning low-error phases to a few of the strongest reflections, the entire set of phases could become significantly improved after density modification. There were five structures with 5,000 to 17,000 reflections in that test and it was very encouraging to see that datasets so large could be improved by having only the 124 strongest reflections assigned the correct phase. Vekhter (2005) assigned correct phases calculated from the model and proposed that, in practice, phases could be measured experimentally by a three-beam diffraction experiment. The study performed as part of the thesis follows up this analysis by exploring computational methods to select improved phases for a few of the strongest reflections before feeding them into density modification. The following points were addressed to pursue the goal.

An analysis was performed to test if the map skewness (Podjarny & Yonath, 1977), which describes the extent to which the extreme values in a map tend to be systematically positive or negative, could be used to identify the correct phases for a few of the strongest reflections. The test was done by implementing an algorithm that searched for combinations of phases for the strongest reflections that, in the presence of the entire data set, led to better values of skewness. Observed results showed that correct phases for the strongest reflections correlate with increasing values of the map skewness.

Another analysis was carried out to test the efficiency of having the skewness as the target function to implement an algorithm and protocols that optimize the quality of phases for strongest reflections for four structures. In order to observe the effect of this improvement, the optimized dataset was passed to density modification and model building to obtain a resulting model, which can be compared with the model obtained using the original data.

The datasets presented in Table 3.1 were selected because they represented borderline cases where density-modified phases were not good enough to generate an interpretable map.

Structure	PDB entry	Space group	Resolu-tion for phase optimizat ion (Å)	No. of non-H atoms (a.u.)	Unit cell	
					axes (Å)	angles [°]
I. Gene V protein (single-stranded DNA-binding protein; Skinner *et al*., 1994)	1VQB	*C*2	2.6	682	*a*=75.81 *b*=27.92 *c*=42.4	β=103.1
II. Heterogeneous ribonucleoprotein A1 (Shamoo *et al*., 1997)	1HA1	$P2_1$	3.0	1,338	*a*=38.1 *b*=44.0 *c*=56.1	β=94.8
III. Cytosolic C2A-C2B domains of synap-totagmin III (Sutton *et al*., 1999)	1DQV	$P6_222$	3.2	2,191	*a*=*b*=125.96 *c*=118.44	
IV. RNA molecule containing domains 5 and 6 of the yeast ai5g group II self-splicing intron (Zhang & Doudna, 2002)	1KXK	$P6_122$	3.5	1,497	*a*=*b*=91.68 *c*=241.65	

Table 3.1 Summary of data for test proteins

There are two key ideas exploited here in order to improve the quality of the experimental maps. The first key idea involves the role of the strongest reflections. Considering that only a few strongest reflections can have an impact on density modification (Fig 3.1*a*), it is possible to implement algorithms that search for phase combinations in this compact solution space. Together with the fact that knowledge about phases is obtainable from experimental phasing, the choice of phases for a reflection based on its probability distribution can also be limited.

The second key idea is based on measures of molecular map quality. Note that we are going to choose alternative phases for only a few of the strongest reflections. The rest of the reflections will be used with their original centroid phase and any new map will be calculated using the complete set of reflections. In this way, the phases for the reflections that are not varied provide a background of known information used for the map calculation, and the phases that are varied are being tested for consistency with the other phases. The phase choices for the strongest reflections are not generated from random sources but from the probability distribution obtained from the experiment, therefore the prior knowledge about phase is still preserved. The newly generated maps are assumed to have some molecular features as a starting point that can be used to calculate a measure of map quality. The skewness of the density values in an electron-density map was chosen in this work as it was pointed out in Terwilliger *et al*. (2009) that it was the most accurate one for estimating map quality out of ten measures tested. The skew function (eq. 3.1) as the target function for the search algorithm is

$$skew = \frac{< \rho^3 >}{< \rho^2 >^{3/2}} \qquad (3.1)$$

Fig. 3.1*b* shows a comparison of the electron-density histogram generated from phases from the SAD data (Φ_B) and phases from the solved structure (Φ_C) for Gene V protein. Electron-density maps for the two sources of phase were generated accordingly and a threshold of ±5σ was applied for the density cutoff on the maps. The skewness was calculated using equation 3.1 and values of about 0.22 and 1.11 were obtained for the first and the second case accordingly. It is necessary to apply the threshold cutoff to truncate the density map since most of the starting experimental maps tend to have some highly positive and negative values. The truncation helps prevent extreme values of map skewness, resulting from a few very large peaks.

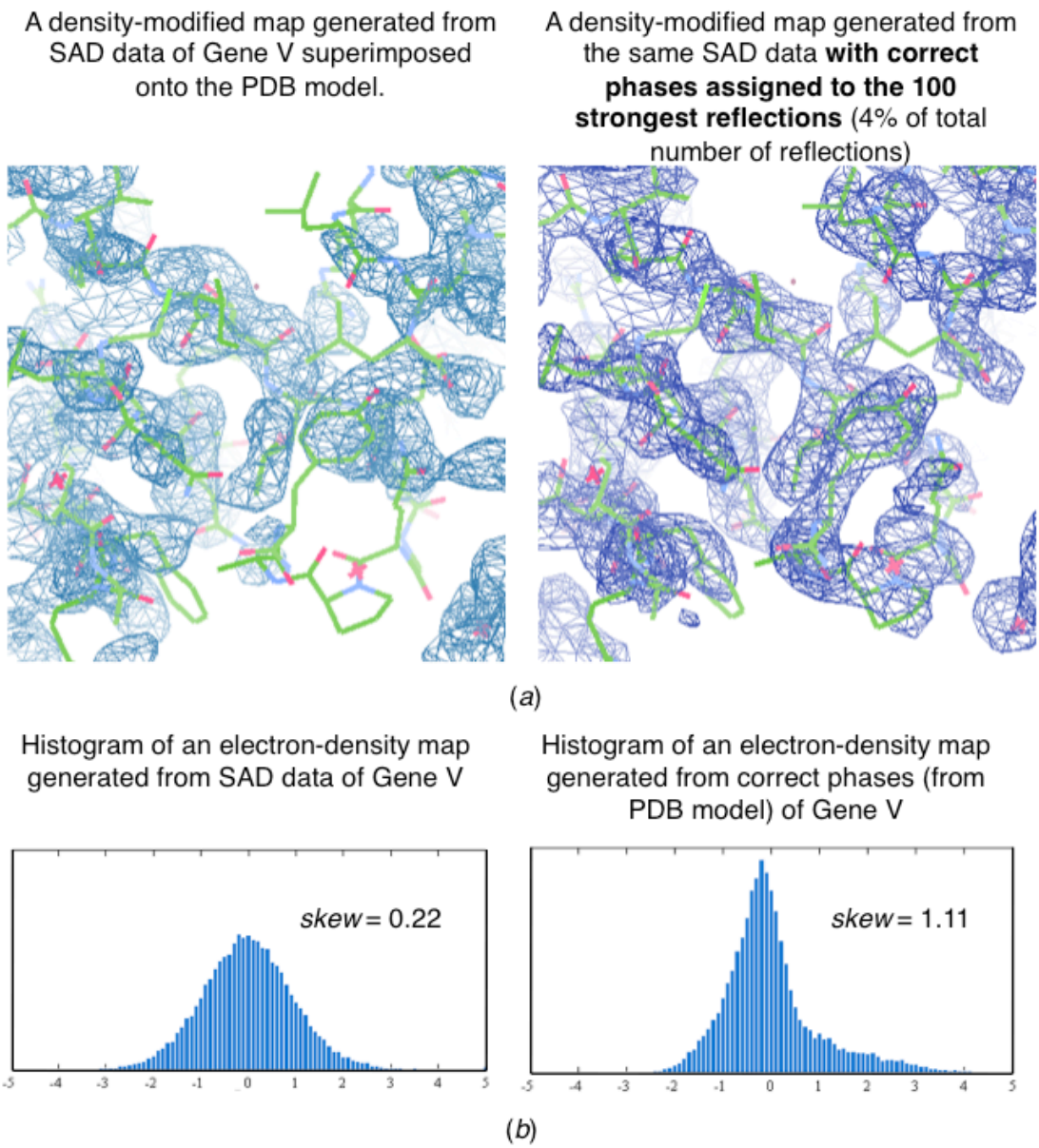


Figure 3.1 Two key ideas exploited in the implementation of the method. (a) A comparison of two density-modified maps generated from SAD data of Gene V: the first map was derived from a reflection set with the original centroid phases (ϕ_B) while the second map was derived by assigning correct phases (from PDB model) to 100 strongest reflections of the same reflection set. Map correlation of the second map was significantly improved from 0.46 to 0.77. (b) A comparison of two electron-density histograms: the histogram on the left was generated from the electron-density map calculated using the centroid phases (ϕ_B) resulting in a small value of map skewness (*skew* = 0.22) (eq. 3.1) while the histogram on the right side was generated from the map calculated using the correct phases (ϕ_C) resulting in a large value of map skewness (*skew* = 1.11).

3.2 Materials and methods

Genetic algorithms (Holland, 1975) were chosen as the optimizing tools because of their useful features in problem representation and search space exploration. A variety of methods in crystallography such as small angle scattering (Svergun & Franke,

2009), powder diffraction (Shankland *et al*., 1997; Harris *et al*., 2004; Feng & Dong, 2007) and *ab-initio* phasing for macromolecules at low resolution (Miller *et al*., 1996; Webster & Hilgenfeld, 2001; Zhou & Su, 2004; Immirzi *et al*., 2009) have implemented the genetic algorithms to solve the phase problem. The implementation takes phase probability distributions of the strongest reflections selected as input, creates a data structure analogous to chromosomes to store these phases, manipulates each chromosome by genetic operators, selects only those with higher skew value, and outputs the solution with a high value for the target function (Fig. 3.2). At the end of each run, phase improvement is determined by calculating the map correlation coefficient (Read, 1986; Lunin & Woolfson, 1993) between the solution phases (Φ_S) and the calculated phases from the correct model (Φ_C)

$$CP\{\rho_1,\rho_2\} = (\sum_{i=1}^{N} F_{obs}^2 \cos[PHI_{C,i} - PHI_{S,i}]) / \sum_{i=1}^{N} F_{obs}^2 \quad (3.2),$$

where N is the number of the selected reflections.

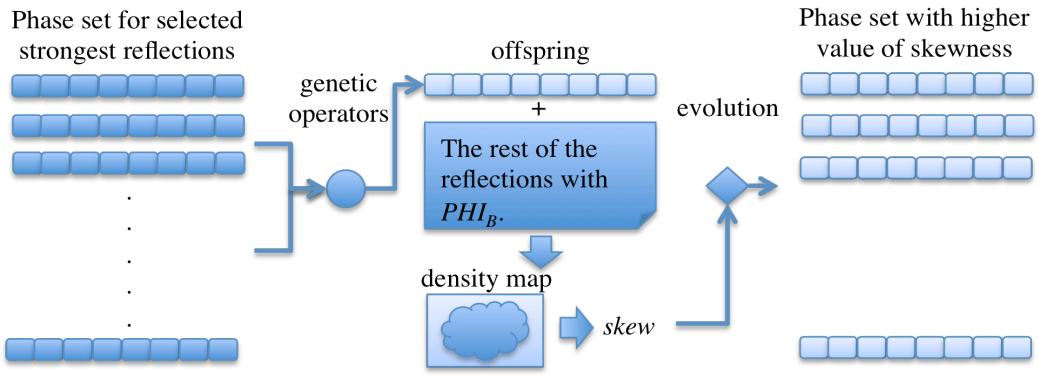


Figure 3.2 Implementation of the genetic algorithm.

The implementation was divided into three parts. In the first part, phase choices are generated from the phase probability distribution function. The second part involves the construction of genetic algorithm and genetic operators with the target function being the skewness of the density map. The last part deals with selection of the best solution, treats them with new figures of merit, and passes them to density modification and model building procedures. All parts of the algorithm were written in Python together with the usage of the *cctbx* libraries (Grosse-Kunstleve *et al*., 2002).

In the first part, phase choices were generated for a reflection according to its phase probability distribution function encoded in Hendrickson-Lattman coefficients (Hendrickson & Lattman, 1970). An example of selecting a phase for a reflection is shown in Fig. 3.3. In the case of this bimodal distribution, traditionally, a centroid phase (Φ_B) is selected. In this method here, other phases were allowed for selection according to their probability distribution. In practice, the phase probability distribution (Fig. 3.3*a*) was converted to the cumulative one (Fig. 3.3*b*). At a time, a random number in the range of 0-1 was picked, then a line was drawn horizontally to intersect with the cumulative function. The phase that met this point vertically was selected out. By doing this many times, all possible choices of phase could be sampled out for that reflection. It is also clear that those phases with higher probability are most likely to be selected because of the high slope of the cumulative function. At the end, a number of phase choices was generated according to the desired number of density maps and the same was done for the rest of the selected reflections.

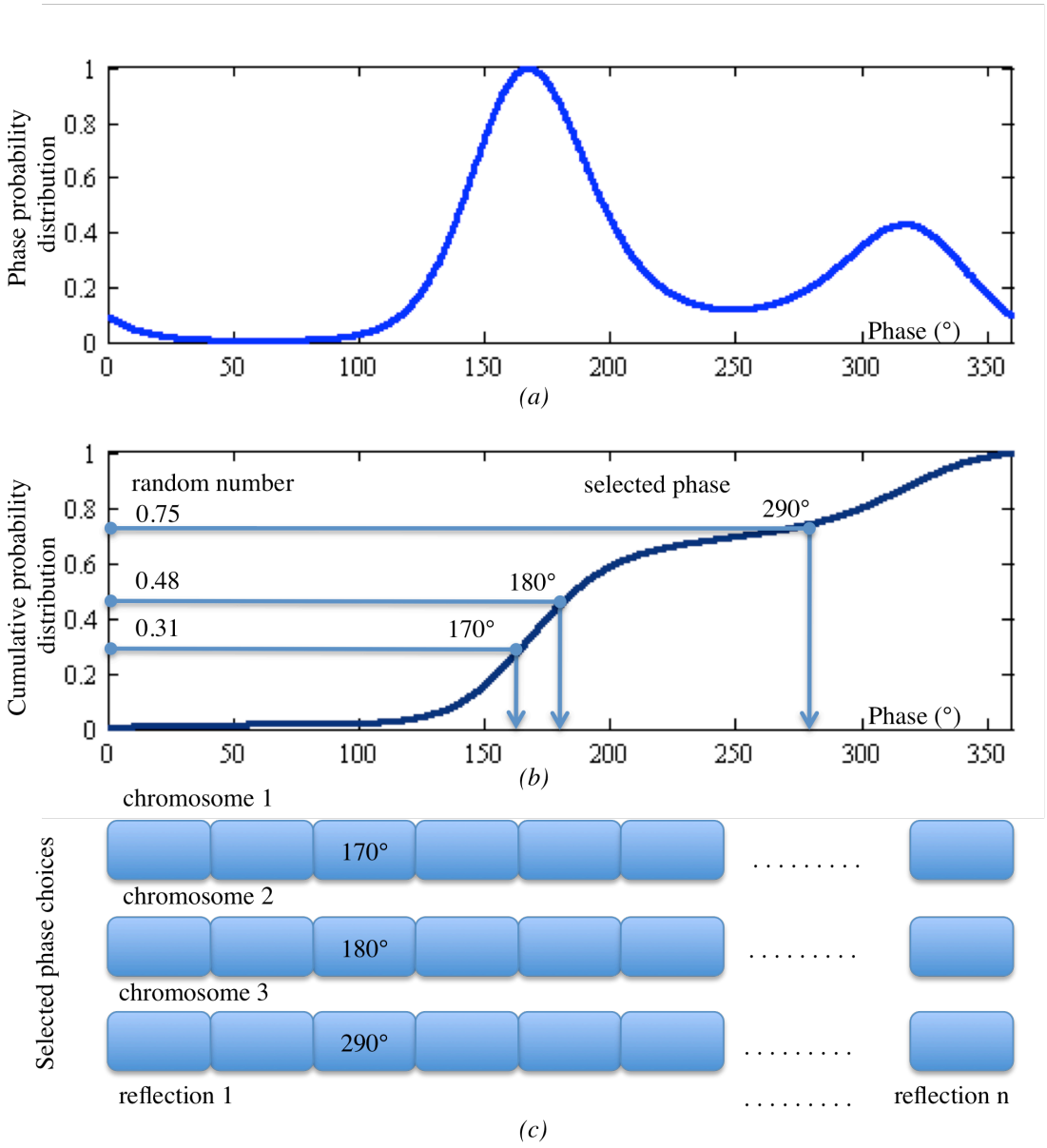


Figure 3.3 Selection of phase choices other than Φ_B for a reflection. (a) Phase- probability distribution function. (b) Cumulative distribution function calculated from (a). (c) Chromosomes storing phase choices for the genetic algorithm.

Note that these alternative phase choices were applied only on the strongest reflections. The rest of the reflections, which comprised the majority, maintained the centroid phases (Φ_B). Even though the phases of the remaining reflections were not perturbed, they play an important role in interacting with the varied reflections to determine the skew value. It will be shown below that phase improvements could be obtained only when the varied reflections are used with the other reflections to calculate the map skewness.

The second part is the implementation of the genetic algorithm, SISA. This kind of stochastic search algorithm has two important features. The first feature is the way information representing the possible solution to the problem is stored. The genetic algorithm treats each set of answers as a chromosome, which looks like the output constructed from the first part (Fig. 3.3*c*) where each value of phase is a possible answer for a reflection. Note that the values stored in the chromosomes are not represented by binary strings but by the set of non-negative integers from 0 to 359. SISA treats these many combinations of phases that had been created as a starting pool of chromosomes. The second feature comprises the selection and recombination process. In order to increase search performance, the geographical-restraint technique (Connor, 1994) was chosen over the probability- weighted (also known as roulette-wheel) method (Bäck et al., 1997) for this selection process. To compare the search performance, tests to search for phases for 100 reflections were attempted for one of the selected test cases, Gene V protein, using both selection techniques. For the geographical-restraint method, the solution phases were found after around 9 – 11 generations. This performance was seen as a significant improvement over the roulette-wheel technique where around 95 – 97 generations were needed before the algorithm could terminate. Note that the SISA was designed to terminate when all chromosomes in the same generation yielded phase differences less than 2°. Both selection techniques resulted in a similar quality of the 100 solution phases with map correlation of 0.53 for the geographical-restraint and 0.54 for the roulette-wheel (the 100 original centroid phases yielded a map correlation of 0.4).

Fig. 3.4 illustrates how the geographical-restraint technique was implemented for the selection and recombination process. At any time, a parent chromosome is selected from a random location on a map where another smaller map is drawn to cover the selected position (Fig 3.4*a*). The algorithm performs random walks on this smaller map to select candidates for recombination and choose the one with the highest fitness value. In comparison to the roulette-wheel, where only the fittest chromosomes determined globally are likely to be selected at any given time, the geographical-restraint method also allows other fittest chromosomes determined locally on the fitness landscape to be selected for the recombination.

The evolution process is triggered by the recombination of the parents, which depends on the crossover and mutation operators. These two mechanisms are controlled by the probability of crossover and mutation accordingly, so that many of the fitter solutions and some non-fit solutions would get selected for the next generation.

The uniform crossover, which allows randomly selected segments from the parents' chromosome to be exchanged (Syswerda, 1989), was selected for the recombination process. It was suggested as a suitable operator for problems with complex search spaces where the practical population size could not meet the necessary sampling accuracy (De Jong & Spears, 1991), which might be the case for this work. In the problem settings here, one way to imagine the size of the solution space is to consider the number of phase sets that must be tested for 1,000 reflections. If each reflection has 2 choices for the phase (like in the case of the bimodal distribution), there are $2^{1,000}$ combinations of phases to be tested in order to obtain the best answer. In order to still be able to compute the answer, the approach applied here only generates around 400 combinations of phase per each test run and this number is much smaller than the number required to obtain an accurate answer. Note that the 1-point and 2-point crossover operators were also tested in this work; however, the results showed that all chromosomes turned homogeneous after a small number of generations without deriving a significantly higher value for the target function.

An example of how the recombination process works for the method is illustrated in Fig. 3.4*b*. From the population pool, a pair of phase sets is selected. In order to recombine their chromosomes, a random template is generated indicating locations where the genes will be swapped. This template is newly created every time crossover occurs. With a certain probability, some of the genes of these two new offspring chromosomes are mutated as well. When mutation happens, the algorithm randomly selects a new phase from the reflection's original phase probability distribution. The target function is then recalculated from the new phase combination. The parent pair is replaced with their offspring only when the latter has a higher value for the target function.

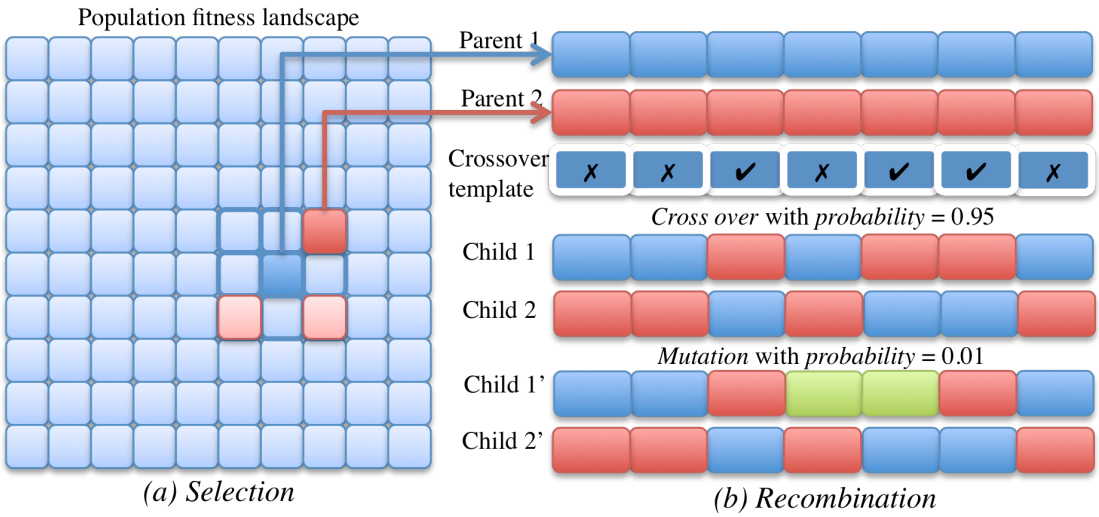


Figure 3.4 Geographical-restraint technique used in the selection and recombination process for SISA. (a) A parent is selected (dark blue location) from a random location on the fitness landscape where a local map is drawn around it. By performing random walks on this local map, more chromosomes are selected as candidates (red locations) and the fittest one (dark red) is chosen for the recombination process. (b) A pair of selected chromosomes is chosen for the recombination under controls of probability of crossover and mutation. The uniform crossover technique was used for the crossover operation where only locations indicated on the crossover template were exchanged between the parents. The mutation operator occurred on randomly selected locations on the child chromosomes where their phases were replaced by new phases redrawn from the phase probability distribution.

The last part of the process concerns the selection of the best solution from the optimization process. To sample solution space, several independent microruns were carried out so that many solutions from different starting points could be obtained. Once all runs were completed, the results show that there were different solutions that could produce similar values of map skewness. This means that for a selected value of map skewness, the value of phase difference between the best solution and the worse solution can be up to around 15°. In order to avoid selecting the worse solution, those solutions for which the value of the fitness was higher than the average value were selected and their centroid phases were calculated as the best solution. This composite best solution is the output from each run of the search process.

Additionally, although the original figure of merit could be used for density modification, the results from different runs show that setting the figure of merit to 1.0 for all reflections selected in the search resulted in more complete models after the density modification and the model building. These solution phases (Φ_S) and figures of merit from the optimization were combined with the rest of the reflections. The

impact of optimizing the strongest reflections was measured by feeding this new set of reflections to the density modification and model building process.

Throughout each run, SISA processes were controlled by the following parameters.
$N_{chromosomes}$: number of chromosome
$N_{generations}$ number of generations
P_{cross}: probability for crossover operator (0.0 – 1.0)
P_{mutate}: probability for mutation (0.0 – 1.0)
$R_{crosspoints}$: number of crossover points represented by a fraction of chromosome size
$N_{mutatepoints}$: number of mutation points

These parameters determine the size of the solution space that each run can represent and the amount of computing time required. Note that *CCP4* Suite (Collaborative Computational Project, Number 4, 1994) and *PHENIX* (Adams *et al*., 2010) were also used during this work.

3.3 Results and discussion

3.3.1 Case I - Gene V Protein (PDB Code: 1VQB)

The SAD dataset from this crystal delivered phases with a mean figure of merit of 0.42 for the entire set of reflections. By supplying the dataset with the sequence of the molecule to an automatic model building program, *PHENIX AutoBuild*, a model was obtained at the end of the run with 42 out of 87 residues built with R value = 0.46. The dataset was collected from a crystal with unit cell parameters as shown in Table 1. The crystal packing is in spacegroup C2 with about 2,500 reflections.

There are two points that direct the test procedures here. The first goal involves examining if the skew function can be used to improve the phases of a few strongest reflections and, if so, to determine if the new phases could make an impact on the density modification and model building process. To meet the first goal, the optimization algorithm was set to run on varying numbers of strongest reflections selected. Apart from the different number of reflections, the same parameters ($N_{chromosomes}$= 400, $N_{generations}$= 100, P_{cross}= 0.95, P_{mutate}= 0.01, $R_{crosspoints}$= 0.2, and $N_{mutatepoints}$= 1) were assigned to SISA for all the runs and the procedure was terminated

when every chromosome had phase differences < 2°. To observe changes in phase quality, the map correlation coefficient (eq. 3.2) was calculated for a particular chromosome, which stored phase choices for the selected reflections, in comparison to the known Φ_C. A scatter plot (Fig. 3.5) between the map correlation (vertical axis) and the skew value (horizontal axis) having one particular point representing a set of phases for selected reflections was generated. The color, which goes from light green to dark blue, represents an increase in number of generations during the optimization process. The square and diamond markers represent Φ_B and the solution phases Φ_S accordingly. Note that Φ_S is the new centroid phase calculated from the selected chromosomes that have a skew value greater than the average. These plots also reveal the variation of overall phase quality during the optimization process as can be seen from the series of filled dots. Each filled dot represents the phase quality of the centroid phases computed from a collection of phase sets with similar skew values. These centroid phases tend to have higher phase quality than the individual samples, as evident in particular when larger numbers of reflections are varied.

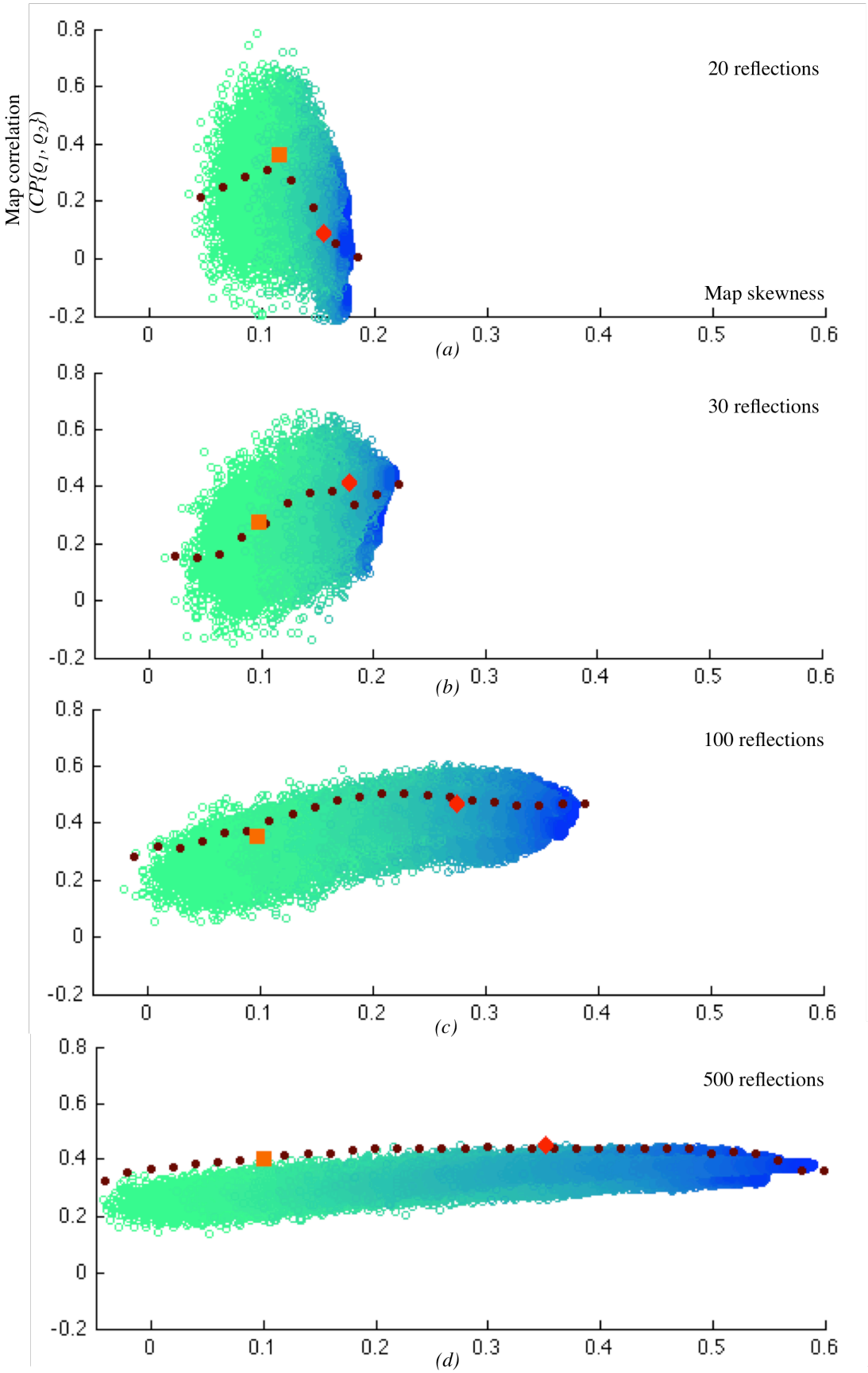

Map correlation
(CP{ϱ1, ϱ2})
20 reflections
Map skewness
(a)
30 reflections
(b)
100 reflections
(c)
500 reflections
(d)
0.8
0.6
0.4
0.2
0
-0.2
0
0.1
0.2
0.3
0.4
0.5
0.6

Figure 3.5 Measures of the quality of the solution phases for the strongest reflections selected in the search. The measures were calculated using map correlation coefficients (eq. 3.2) of the solution phases (ϕ_S) and the known phases (ϕ_C). Each plot is displayed with the skew value of a density map that each set of phases represents. The filled dots show map correlation coefficients of the centroid phases calculated from a group of phases with similar skew value. All plots show the results from 10 independent runs with a square marker representing the centroid phases $\boldsymbol{\phi}_B$ and a diamond marker representing the solution phases $\boldsymbol{\phi}_S$ selected as output of the search process for: (a) the 20 strongest reflections. (b) the 30 strongest reflections. (c) the 100 strongest reflections. (d) the 500 strongest reflections.

These plots reveal that at least 30 strongest reflections should be selected in order to obtain phase improvements because with at least this many reflections chosen, the solution phases $\boldsymbol{\Phi}_S$ with better map correlations than the one calculated using $\boldsymbol{\Phi}_B$ could be obtained. As the number of selected reflections was increased, the algorithm achieved higher values of map skewness with less overall average improvement in phase quality for the varied reflections. The same procedures were used to run 10 independent trials for cases with 100, 300, 500, and 1,000 strongest reflections selected in the search and the map correlation coefficients were calculated. The results from the calculations are shown in the plot in Fig. 3.6(*a*). The plot shows the maximum (top end of the line), the minimum (bottom end), and the median (connected horizontal line) of the calculated map correlations for the results from the 10 runs. The results shown are grouped according to the number of the strongest reflections used in the search process. For each group, the quality of $\boldsymbol{\Phi}_B$ for the selected amount of reflections is shown using the square marker.

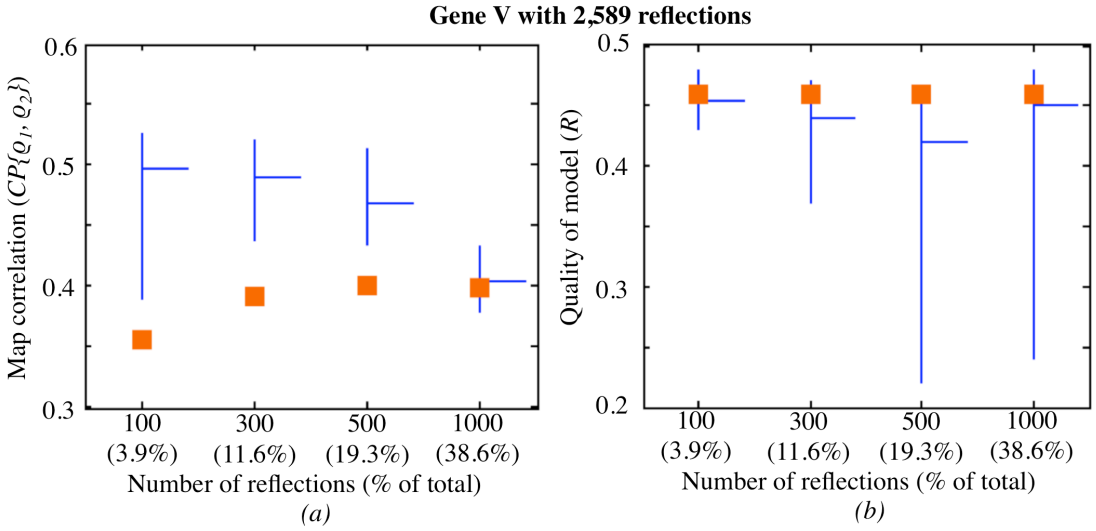


Figure 3.6 Results of running 10 independent trials to search for phases for 100, 300, 500, and 1,000 selected strongest reflections showing (a) Phase quality calculated using map correlation coefficient (eq. 3.2) for phases found after the search and (b) Quality (R factor) of

the final models generated by assigning the selected strongest reflections in the original dataset with the new phases and passing them to the model building program.

Fig. 3.6(*a*) shows that SISA could improve the quality of phases of the original Φ_B for up to around 500 strongest reflections. The highest degree of improvement was achieved in the case of 100 strongest reflections, where the map correlation coefficient of Φ_B was improved from 0.35 to 0.50 (the median case of the solution phases Φ_S). Some improvements could also be observed in the 500 strongest reflection tests with less degrees of improvement (from 0.40 to 0.46). Significant improvements were not observed in the 1,000 strongest reflection case and the median case for the solution phases Φ_S had similar map correlation coefficient to Φ_B.

Two other tests were performed to see if phase improvements obtained from the search depend on other reflections incorporated with the varied ones to calculate map skewness. In the first test, the map skewness (the target function of the search) was calculated using both the varied reflections and the other reflections while in the second test, only the varied reflections were used. Both tests were performed for the 100 and the 300 strongest reflections with 10 independent runs in each case. Fig. 3.7 shows comparisons of map correlations of the solution phases (φ_S) from the first test with map correlations from the second test for the 100 (*a*) and the 300 strongest reflections (*b*). Both plots show that while improvements of map correlations could be observed for the tests with all reflections, this is not the case for the tests using only the varied reflections. For the 100 strongest reflections, map correlation coefficients of 0.36 (φ_B) increased to 0.5 (the median value) when using all reflections but decreased to 0.34 when using only the varied reflections. Similar results were obtained for the 300 strongest reflections when the map correlation coefficient of 0.39 (φ_B) increased to 0.49 when using all reflections but decreased to 0.38 when using only the varied reflections. Note that map skewness reached at the end of all the runs for the two tests increased from 0.1 (φ_B) to similar values of around 0.3 for the 100 strongest reflections and from 0.1 (φ_B) to 0.35 for the 300 strongest reflections.

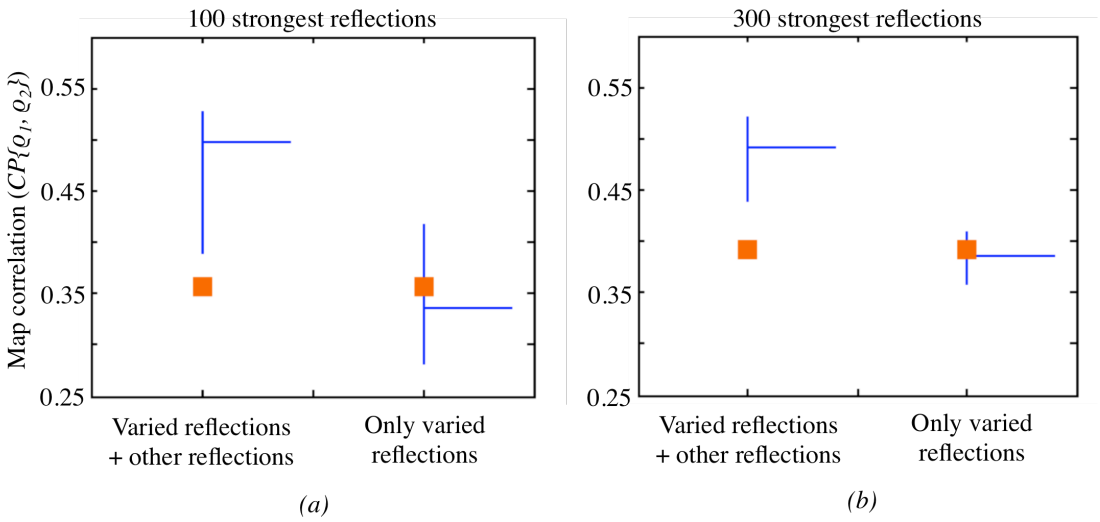


Figure 3.7 Comparisons of map correlations of the solution phases (φ_S) from the tests using the varied and the other reflections with map correlations from the tests using only the varied reflections for the calculation of map skewness as the target function for the search. Each vertical line on the plots displays map correlations calculated using only the varied reflections of 10 independent runs. The square marker indicates map correlation calculated using φ_B. (a) The 100 strongest reflections case. (b) The 300 strongest reflections case.

Another task for this work is to investigate if increasing the population size in the genetic algorithm could help improve the results when searching for more than 300 reflections. A total of 10 runs were performed with an increase of the population size from 400 to 2,500 to search for phases for the 500 strongest reflections. Leaving other parameters for the search to the same values as used previously, map correlation coefficients were obtained in the range of 0.4 to 0.5 for 10 independent runs. These resulting values were similar to the values obtained when the population size of 400 was used in the tests.

In order to see if the new phases as obtained from the search could improve the results from density modification and model building, a new reflection file was created by combining the strongest reflections (assigning their phases to the solution phases Φ_S and figures of merit to 1.0) with the rest of the reflections (using their original phases (Φ_B) and figures of merit) and passed the file to the model building program (*PHENIX AutoBuild* was chosen in the test). The results were evaluated by comparing the *R* factor of the atomic model generated at the end of the runs using the solution phases (Φ_S) and the original phases (Φ_B). The tests were done for the cases of the 100, 300, 500, and 1,000 strongest reflections selected in the search.

Fig. 3.6(*b*) shows that model quality is strongly influenced by the numbers of strongest reflections selected for the search. In Fig. 3.6(*b*), each vertical line represents the range of *R* factors of the atomic models generated from *PHENIX AutoBuild* using the 10 new reflection files, with the connected horizontal line representing the median value. The square marker displays the *R* factor of the model built from the same program using the original Φ_B and the figures of merit. For each case with 10 independent runs, there were 6 runs for the 100 strongest reflections, 9 for the 300 strongest reflections, all 10 runs for the 500 strongest reflections, and 8 runs for the 1,000 reflections where the *R* factors obtained were better than or equal to the *R* factor of the model built from Φ_B. As a comparison to the results obtained here, note that the model obtained from the program using Φ_B had 42 residues (the structure has 87 residues) with *R* factor = 0.46. In comparison to this result, it can be seen that most of the significant improvements were generated from the 500 strongest reflection cases where three of the runs resulted in *R* factors of around 0.31 with 64, 65, and 69 residues built. The best result was also obtained in one of the tests varying the 500 strongest reflections; the resulting model had 84 residues with *R* factor = 0.22.

Two uses of the figure of merit for the solution phases were tested for the density modification: setting the value to the original figure of merit and setting it to 1.0. It was possible to obtain improvements after the density modification with the original figures of merit, but setting these values to 1.0, led to even better results. For gene V, with 500 strongest reflections optimized, an average of map correlation was increased from 0.53 to 0.57.

3.3.2 Cases II - IV

The improvement after the density modification and model building for the SAD dataset of Gene V protein shows that map skewness could be used as a target function to search for more accurate phases than Φ_B. In order to investigate if the same method can be applied to other datasets, three more datasets (Cases II-IV in Table 1), which had failed to give complete structures after density modification and model building were selected for the test here.

The same protocol was applied to these three datasets as for the Gene V protein. Firstly, SISA was set to perform 10 independent runs in order to search for phases for 100, 300, 500, and 1,000 strongest reflections in each test case. The map correlation coefficients were calculated for all the solution phases (Φ_S) in comparison to the known structure (Φ_C). After the search operations were complete, the new set of phases with figures merit of 1.0 was recombined with the original centroid phases (Φ_B) of the other reflections and the figures of merit and passed to *PHENIX AutoBuild*. The *R* factors of the models generated from the runs were collected to investigate the impact of the new phase sets.

The quality of the solution phases (Φ_S) for the three test cases is shown in Fig. 3.8*a*, *c*, and *e*. Resulting phases from the search procedures for these three datasets have relationships to the map skewness similar to the results obtained for the Gene V protein. For the Heterogeneous ribonucleoprotein A1 (Case II) and the Cytosolic C2A C2B domains of synaptotagmin III (Case III), using 100 strongest reflections was enough to obtain phase improvements for the search procedures (Fig. 3.8*a* and *c*). With these 100 strongest reflections, the map correlation coefficients were improved from 0.45 (Φ_B) to 0.54 (Φ_S) for the second case and from 0.49 to 0.57 for the third case. However, for the RNA molecule of the yeast ai5g group II self-splicing intron (Case IV), at least 300 reflections were necessary for the search in order to obtain some improvements (Fig. 3.8*e*). The obtained map correlation coefficient for this fourth case increased from 0.40 to 0.43. As the numbers of the selected strongest reflections increased, the degrees of phase improvements also decreased. In all three cases, for 1,000 strongest reflections, phase improvements were not observed and the median cases for the three datasets resulted in map correlation coefficients less than or equal to the value calculated from the original Φ_B.

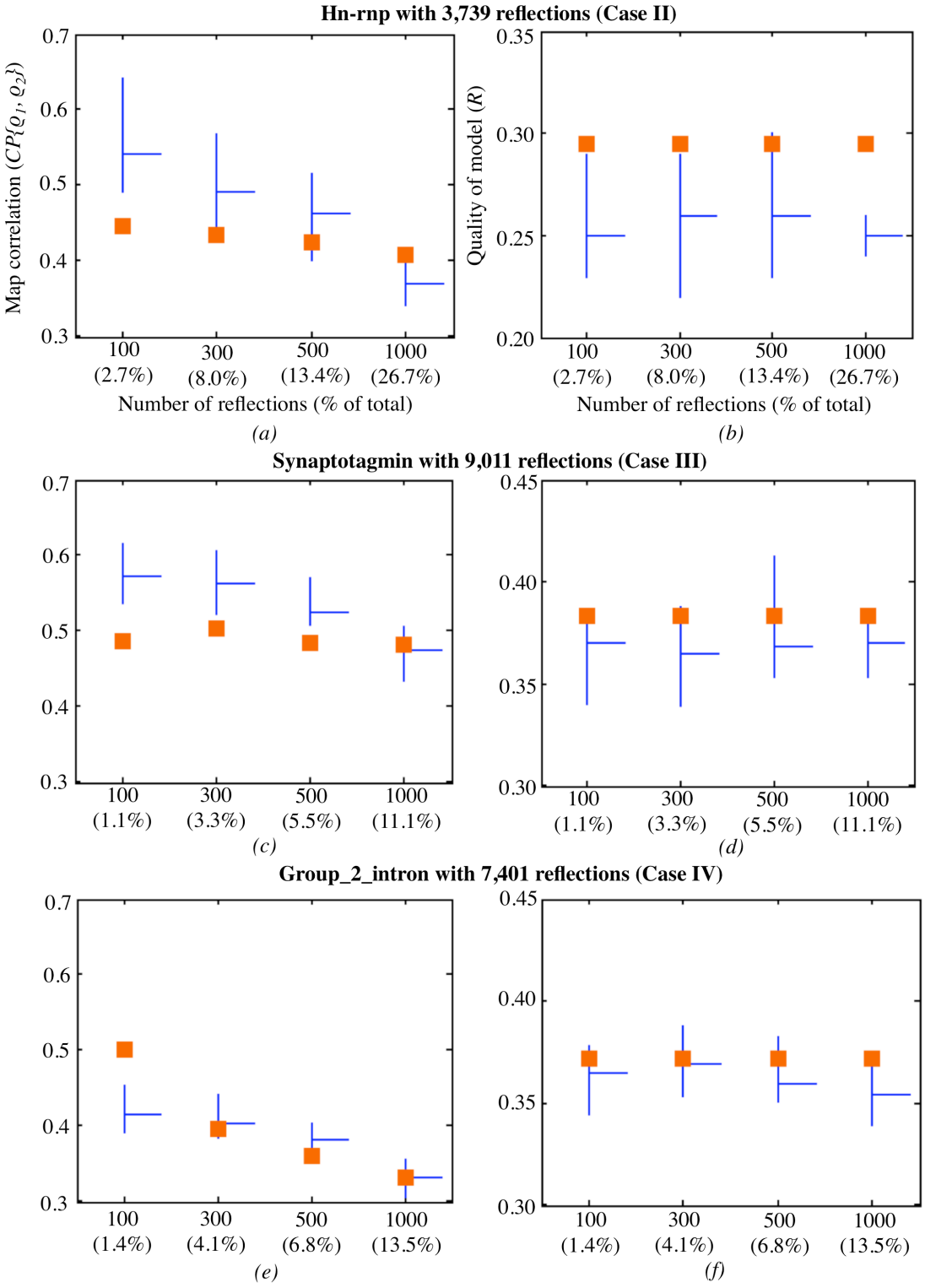


Figure 3.8 Results of running 10 independent trials to search for phases for 100, 300, 500, and 1,000 selected strongest reflections showing map correlation coefficients on the left side and map quality (*R* factor) on the right side for the three datasets.

Fig. 3.8*b*, *d*, and *f* show *R* factors of the models built from *PHENIX AutoBuild* grouped by numbers of reflections used in the search. These plots reveal that *R* factors

for models generated from the new phases are in general better than the values produced from the original phases (Φ_B). There is only one test run from the Cytosolic C2A C2B domains of the synaptotagmin III test case (Fig. 3.8*d*) where *R* of the model generated from the new phases is significantly higher than the *R* value obtained starting from Φ_B ($R\ \Phi_B$=0.38 and $R\ \Phi_S$=0.41). For the three datasets, the results of the most successful test run are shown in Table 3.2.

Test structures	**No. of resi-dues**	**Phases**	**No. of reflec-tions used in search**	**Results from *PHENIX AutoBuild***				
				R	**Rfree**	**Resi-dues traced**	**Side chains**	**Frag-ments**
II. Hn-rnp	184	Φ_B	-	0.29	0.40	140	89	5
		Φ_S	**300**	**0.22**	**0.30**	**153**	**143**	**1**
III. Synaptotagmin	296	Φ_B	-	0.38	0.43	167	30	14
		Φ_S	**300**	**0.34**	**0.39**	**159**	**91**	**8**
IV. Group2intron	70	Φ_B	-	0.37	0.41	61	0	4
		Φ_S	**500**	**0.36**	**0.42**	**58**	**48**	**1**

Table 3.2 Comparisons of results from the model building program using Φ_B and the solution phases Φ_S.

3.4 Conclusions

There are two key ideas explored in this work: first, reducing the phase errors in a small set of the strongest reflections can have a large impact; second, map skewness is a highly effective measure of phase quality. These ideas were combined using the genetic algorithm, SISA, to improve the quality of the density map after density modification, leading to greater success in subsequent model building. Results from the four test cases show that the phases of around 100 – 500 selected strongest reflections could be improved through a search using map skewness as the target function. Based on tests using, variously, the 100, 300, 500, and 1,000 strongest reflections in the search, it can be seen that greater average phase improvement occurred when smaller numbers of reflections were selected. The greatest improvements were observed for 3 test cases (I - III) when only the 100 strongest reflections were varied, or the 300 strongest reflections for the remaining test case (IV). Significant phase improvements were not observed in any of the cases when the

1,000 strongest reflections were varied. For the Gene V protein, phase quality measured by map correlation coefficient did not change significantly when 1000 reflections were varied, whereas the map correlation coefficients actually dropped by about 0.05 for the other three test cases.

When 100-500 phases were varied and combined with the original centroid phases, Φ_B, for the remaining reflections, a large majority of test runs showed a substantial improvement in the quality of the map after density modification and the success of the subsequent model building.

The calculation time for the search depends on the size of the structures and the numbers of the selected reflections. From the four test cases, the smallest structure, the Gene V protein, has 682 non-H atoms with around 2,500 reflections in space group *C*2. Calculations took about 0.5 hours for the 100 strongest reflections and 1.5 hours for the 1,000 strongest reflections. The largest structure, the structure of the Cytosolic C2A C2B domains of synaptotagmin III (Case III), has 2,191 non-H atoms with around 9,000 reflections in space group $P6_222$. Calculation times of 4.4 hours and 16 hours were recorded for the 100 and 1,000 selected reflections respectively.

4 *Ab-initio* phasing: resolving phase ambiguities for 2-dimensional problems

4.1 Introduction

The uniqueness of solutions in *ab-intio* phasing depends on molecular size and data resolution. Atomic-resolution features (positive, equal, and resolved atoms) may help derive a unique solution from the structure factor amplitudes when the molecular size is restricted to < 1,000 atoms (Sheldrick *et al*., 2001). Lower-resolution data and a large number of atoms usually prevent this. The usages of electron density histograms, connectivity properties, and statistical likelihood usually do not resolve phase ambiguities (Lunin *et al*., 2000). Also, the usages of α-helical polyalanine search fragments do not solve the problem in the general case (Rodriguez *et al*, 2009). Thus, uniqueness is the key problem for macromolecular *ab-initio* phasing.

Ab-initio phasing for macromolecules relies on fitting search models with the observed structure factor amplitudes. Usually, the method tries to locate atoms or protein fragments in the unit cell, then identifies solutions by comparing their calculated structure factor amplitudes with the observed ones. When errors (measurement and model errors) are included, we use the maximum likelihood for the comparison; otherwise, we may use the least-square methods or the correlation coefficients.

The amount of prior knowledge about the structure may impact the identification of a unique solution that relies only on the structure-factor amplitudes. Although these amplitudes are less useful for *ab-initio* phasing, methods such as molecular replacement and model building use them to indicate the quality of the models or the phases: molecular replacement uses the maximum likelihood to locate and orient the starting models, model building uses least-squares (crystallographic *R*-factor) to modify and extend the models. The difference between these methods and *ab-initio*

phasing is that both molecular replacement and model building use prior knowledge about the structure.

This work seeks to understand relations between the amount of known information about the structure and the structure-factor amplitudes. Three test structures were artificially generated based on these settings: the structure is in 2-dimensonal grids of binary values ('1' for protein and '0' for solvent), the sampling size is 10x10, the solvent content is varied from low to high, and the pseudo-atoms are all equal. With different amounts of known coordinates given, a genetic algorithm was implemented to search for solutions using the correlation coefficient of the observed and the calculated (from the solutions) structure-factor amplitudes as the target function. At the end of the search, the solutions with the largest amplitude correlation were extracted and their quality of phases were determined – these helped quantify the relations between the amount of known information about the structure and the structure-factor amplitude correlation based on structures with different levels of solvent.

4.2 An example of non-unique solutions

Given a structure consisting of only equal atoms, with information about the structure lacking (coordinates or phases are unknown), the correlation of the observed and the calculated structure-factor amplitudes fails to measure the quality of phases. An example of this problem in 10x10 sampling grids is shown in Fig. 4.1. The structure-factor amplitude correlations (CC_F) between the first structure and two additional structures (Structure B and C) were calculated and a value as high as 0.87 and 0.88 were obtained respectively (a perfect match will yield a correlation of 1); however, their map correlations (CP) (eq. 4.1) (Read, 1986; Lunin & Woolfson, 1993) resulted in a value as low as ~0.3 after applying a different origin, enantiomorph, and Babinet's principle (an opaque body and a hole of the same size and shape) to match the two structures.

$$CP\{\rho_1,\rho_2\} = (\sum_{i=1}^{N} F_{obs}^2 \cos[PHI_{C,i} - PHI_{S,i}]) / \sum_{i=1}^{N} F_{obs}^2 \qquad (4.1)$$

where N is the number of Fourier terms.

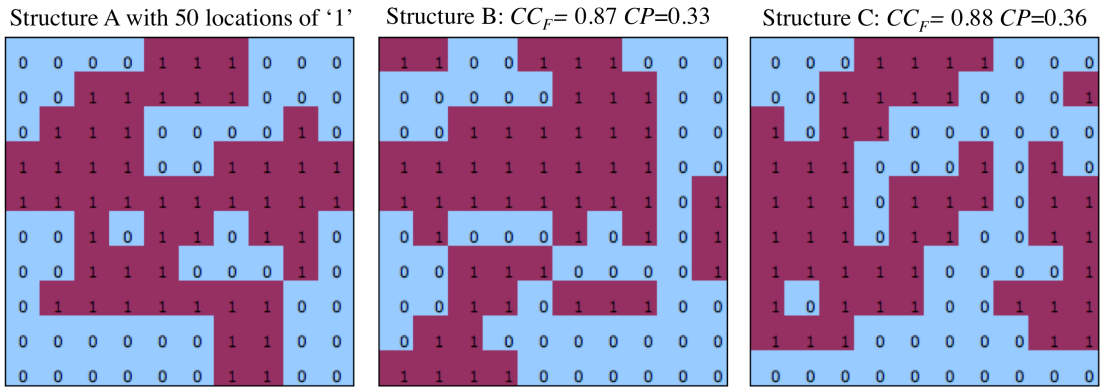


Figure 4.1 An example of non-unique solutions shows a structure used as a comparison, structure A, and two additional structures, structure B and C, that yielded a map correlation of only around 0.3 in comparison with structure A. Nonetheless, structure factor amplitudes of Structure B and C were highly correlated with Structure A (CC_F = 0.87 and 0.88 respectively).

4.3 Materials and methods

Three test structures were generated with solvent content of 70%, 50%, and 30% (Fig. 4.2). These structures consisted of cells containing '1' for the molecular region and '0' for the solvent. The structure factors of these structures were calculated with 10x10 sampling grids and only the amplitudes were used in the search process. Genetic algorithms were selected as search tools and an initial set of solutions was created by randomly locating positions of '1' on the grids. The number of '1'-locations was set to be equal to the amount found in the test structures. Since the goal was to measure usability of the structure-factor amplitude correlation in determining solutions according to the amount of known coordinates, the initial random solutions were generated with 0%, 25%, 50%, and 75% of known coordinates. The algorithm applied genetic operators to perturb the positions of '1' and used the structure-factor amplitude correlation as the target function. For the test with >0% known coordinates supplied, these coordinates stayed unperturbed during the search. The algorithm terminated when the structure-factor amplitude correlations among the solutions were > 0.9 or when the maximum number of generations was reached.

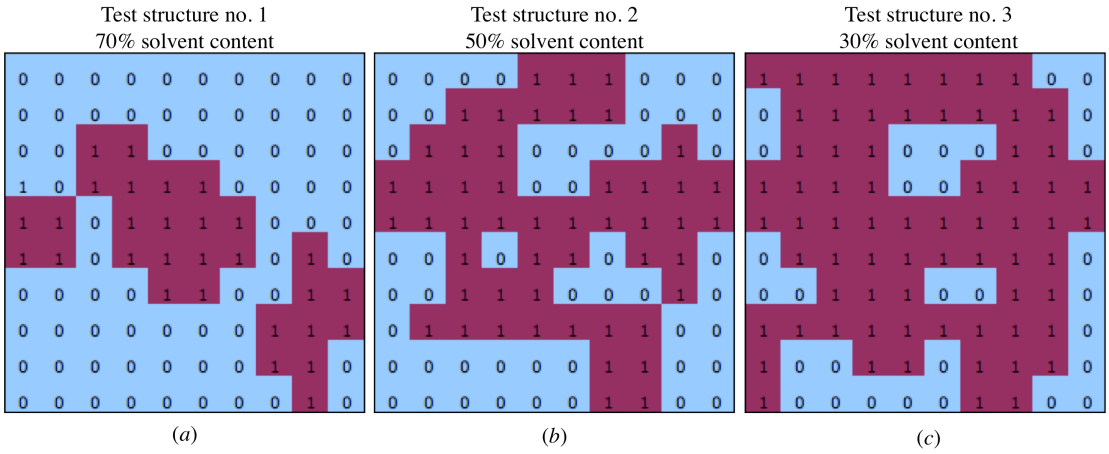


Figure 4.2 Test structures artificially constructed using an equal-atom binary model on the 10x10 sampling grids for (a) 70% solvent content. (b) 50% solvent content. (c) 30% solvent content.

All Fourier terms generated from the 10x10 sampling grids were used to calculate the structure-factor amplitude correlation. These terms provide enough information only for the structures with 50% and 70% solvent content, since the inverse Fourier transforms (the structure) will be recognizable when the selected Fourier terms fulfill

$$\sqrt{\sum_{P} F^2} > F(0) \qquad (4.2)$$

where P is the number of Fourier terms (Cochran, 1952).

An example, which illustrates (eq. 4.2) can be shown by a calculation of an inverse transform using different numbers of strong Fourier terms generated from a 2-dimensional structure. Fig. 4.3 shows structures constructed from the inverse Fourier transforms using 10, 20, and 44 Fourier terms (Fig. 4.3*a*, *b*, and *c*). This example shows that when $\sqrt{\Sigma \boldsymbol{F}^2}$ is larger than $\boldsymbol{F}(0)$ (Fig. 4.3*b*), the inversed transform function already looks similar to the transform using all terms (Fig. 4.3*c*), especially when the structure is discretized to increase the contrast between positive and negative regions.

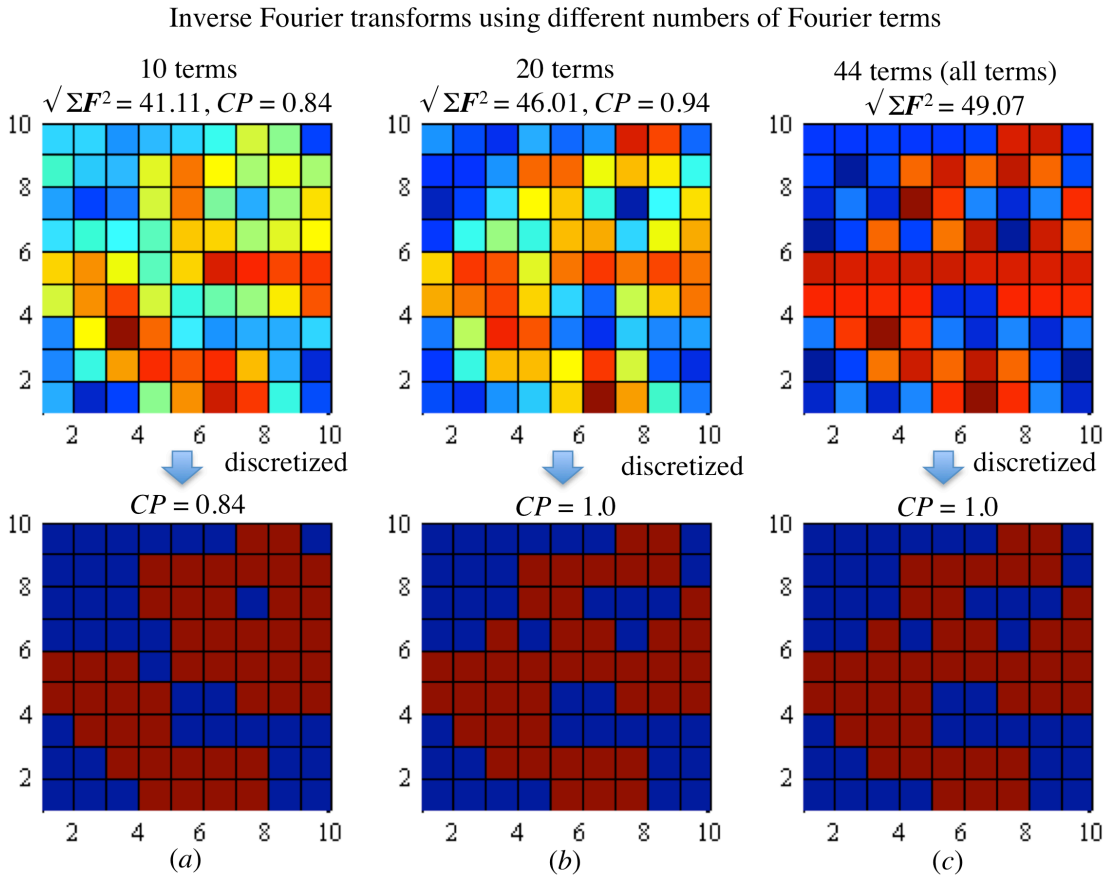


Figure 4.3 Structures calculated from the inverse Fourier transforms using different numbers of the largest Fourier terms. (a) For the 10 selected largest terms; (b) for the 20 selected largest terms; (c) for the 44 largest terms.

The number of selected Fourier terms required to meet eq. 4.2 depends on the solvent content, the 50%-solvent structure required 26 while the 70%-solvent structure required only the 17 largest structure-factor amplitudes. The $\sqrt{\Sigma \boldsymbol{F}^2}$ values were larger than $F(0)$ terms for both cases; therefore, the use of 44 terms generated from 10x10 sampling grids was more than enough to provide information about the structure after the inverse Fourier transform. However, when the solvent was as low as 30%, these 10x10 sampling grids with 44 Fourier terms could only produce $\sqrt{\Sigma \boldsymbol{F}^2} = 63.5$. Note that this value was lower than the $\boldsymbol{F}(0)$ term, which was 70, and inadequate to provide a recognizable structure after the inverse Fourier transform. The test on this case with 30% solvent content was still performed to examine the effect when the available Fourier terms were inadequate to explain the complete structure.

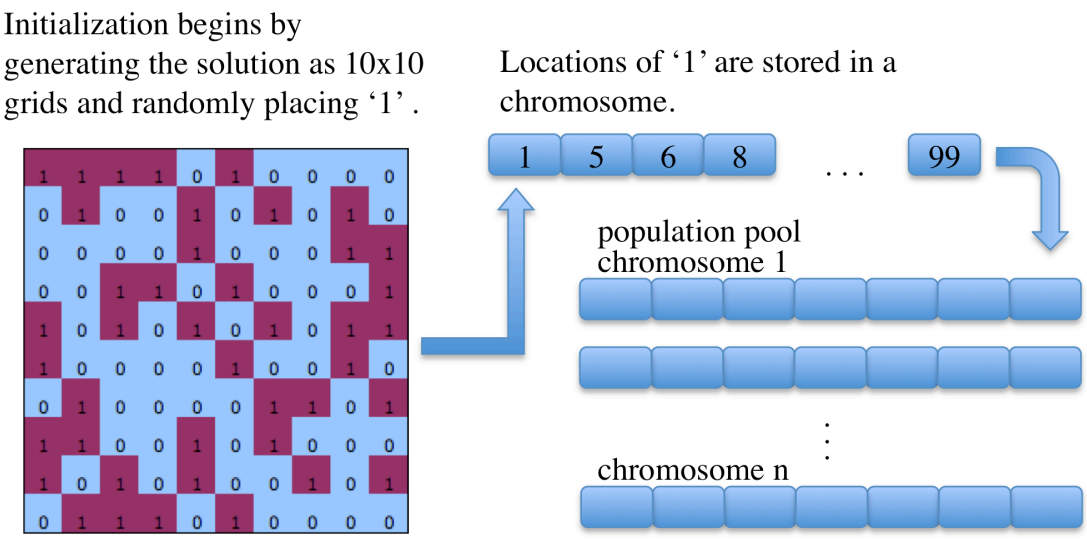
Once the required number of the Fourier terms was known, the genetic algorithm started by initializing binary models and placing '1' randomly on the 10x10 grids.

The amount of '1' placed was set to be equal to the amount of '1' in the selected test structure. Note that for each test case, different amounts of known coordinates were supplied and kept fixed throughout the search process. The genetic algorithm stored a solution as a chromosome (Fig. 4.4*a*); only the location numbers of '1' were recorded in the chromosome to minimize the size of computing memory. This allowed the size of the population to be larger if necessary. At any time when the 10x10-grid models were needed, they could be regenerated from these location numbers stored on the chromosomes.

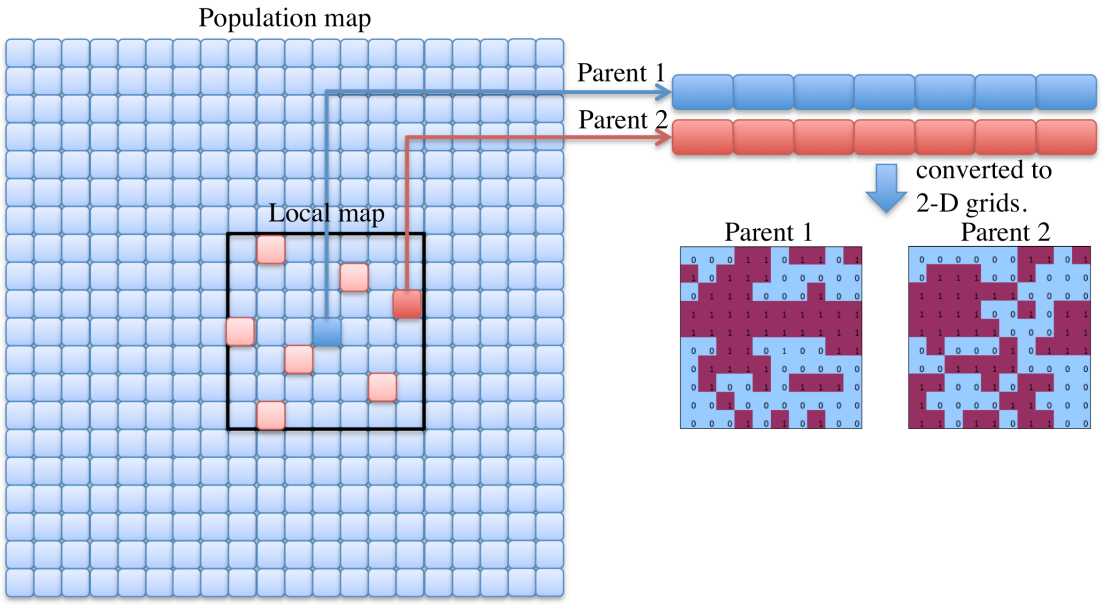
At the beginning of each run, the genetic algorithm generated these chromosomes and stored them on a population landscape where selection and recombination processes took place. The selection process was done using methods similar to the geographical-restraint technique (Connor, 1994) as shown in Fig. 4.4*b* to prevent a crowding problem, where some highly fit solutions quickly reproduce themselves (Mitchell, 1997). These selection techniques involve choosing two chromosomes from a local map, which is generated randomly for a number of times at a particular generation. These two chromosomes are candidates for the recombination processes, which consist of the crossover and the mutation operations. For the crossover, the two chromosomes (parent 1 and 2) will have their genes (binary bits) swapped at locations specified by the crossover template. Note that the two chromosomes were reconstructed back to the 10x10 grids, so they could be aligned to the same origin, enantiomorph, and Babinet structure before gene swapping. Crossover genes were done using a template, which was designed under the uniform crossover scheme (Syswerda, 1989). The scheme selected crossover points randomly along the positions on the chromosomes and regenerated the template every time crossover occurred. At the end of these processes, two offspring chromosomes would be produced and subjected for the mutation operation according to the chosen probability. Mutation also occurred on randomly selected locations where values stored on those bits would be flipped to the opposite state (1→ 0 and 0→1).

The evolution process is triggered by the recombination of the parents, which depends on the crossover and the mutation operators. These two mechanisms are controlled by the probability of crossover and mutation accordingly so that not only the fitter

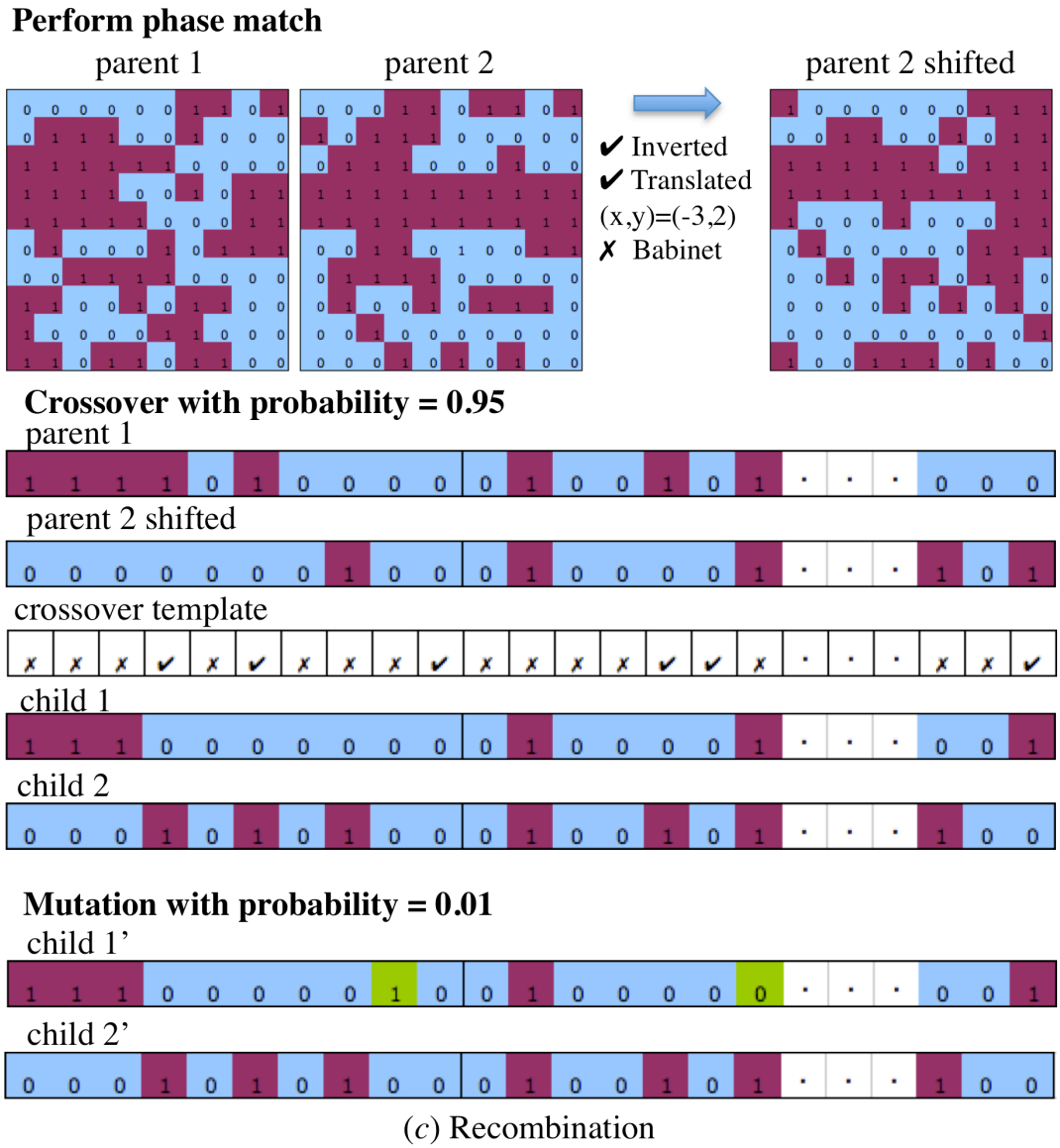
solutions would get selected to the next generation. When crossover or mutation processes occur, the newly generated offspring would be calculated for their fitness values; the aim of the search process is to derive the solutions by maximizing these values.



(*a*) Initialization



(*b*) Selection



(*c*) Recombination

Figure 4.4 An implementation of the genetic algorithm: (a) Initialization of the chromosomes by placing '1' randomly on the 10x10 grids and storing the location numbers in each chromosome. (b) Selection of parental chromosomes for the recombination process via the use of a local map. (c) Recombination of the parental chromosomes by applying genetic operators (crossover and mutation) according to the chosen probabilities.

The chromosome fitness was obtained from the calculation of the correlation between the observed (from the selected test structure) (F_O) and calculated (from the solution) (F_C) structure factor amplitudes (eq. 4.4). The algorithm performed selection, recombination, and fitness evaluation iteratively until all chromosomes in the population pool became homogeneous (averaged map correlation among the chromosomes > 0.9) or when maximum number of generations was reached. When the algorithm terminated, the solutions with the highest fitness values were selected as output.

$$CC_F = \frac{\sum_P (F_O - < F_O >)(F_C - < F_C >)}{\sqrt{\sum_P (F_O - < F_O >)^2} \sqrt{\sum_P (F_C - < F_C >)^2}} \qquad (4.4)$$

where P is the number of Fourier terms.

To measure the quality of the solutions, it was necessary to identify different origin, enantiomorph, and Babinet equivalents. An example to demonstrate these problems in 2-dimensional structures is shown in Fig. 4.5. A structure (Fig. 4.5*a*) can be generated using a set of structure factor-amplitudes and phases. The same set of amplitudes with all the phase angles shifted by a constant, inversed by sign, and shifted by π generates the structure's translated copy (Fig. 4.5*b*), inverted copy (Fig. 4.5*c*), and Babinet copy (Fig. 4.5*d*) respectively. The structure-factor amplitude correlations between these copies and the original structure, in the same order, were 1.0, 1.0, and 0.89 (for the Babinet copy, only the $\boldsymbol{F}(0)$ is different).

These ambiguities must be taken into account when measuring quality of the solutions. We can identify these ambiguities by analyzing the cross-correlation map (the Patterson map) and the peak found on the map (*CMAX*). For the translated structure, the highest peak on the map corresponds to the translation vector and moving the structure according to this vector will superimpose the two structures (Fig. 4.5*b*). For the inverted structure, the height of the peak on the cross-correlation map can be used to determine whether the structure is an inverted version of the other (Fig. 4.5*c*); by inverting the structure in Fig. 4.5*c*, *CMAX* value will increase from 0.48 to 1.0. Lastly for the Babinet structure, the sign of the peak on the cross-correlation map can be checked. By looking at the highest peak on the map, if the sign of the *CMAX* is negative, it means that the structure could be better matched when it is overturned (Fig. 4.5*d*) (Lunin *et al.*, 1993). Once the solutions were aligned with the test structure, map correlations (eq. 4.1) could be calculated.

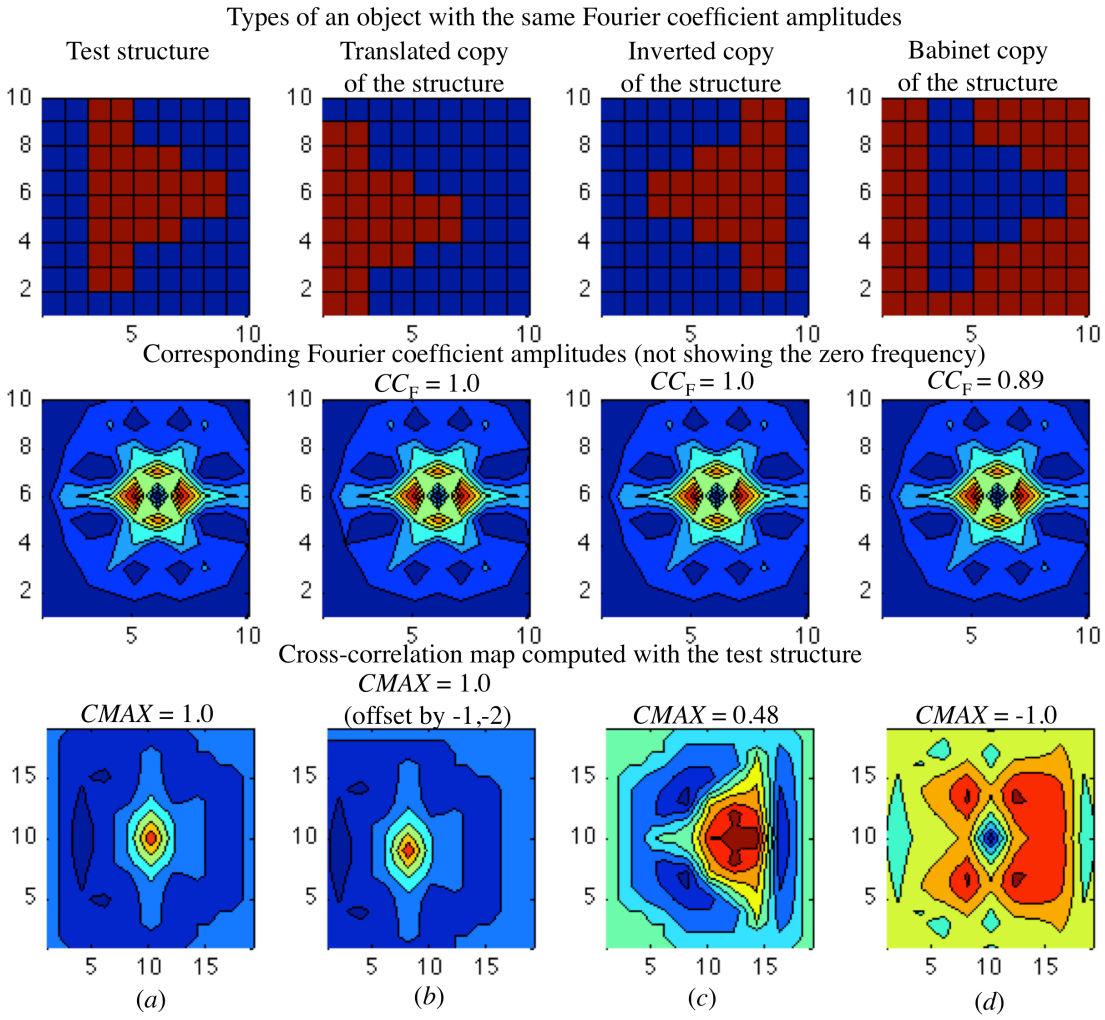


Figure 4.5 Scattering objects displayed with their corresponding Fourier coefficient amplitudes and the cross-correlation maps compared with the test structure. (a) A test structure showing the same amplitudes with the other three objects. (b) A translated copy of the test structure with its cross-correlation map showing the peak shifted by the translation vector (-1, 2). (c) An inverted copy having the maximum value on its cross-correlation map being smaller than it should be (the correct structure would have *CMAX*=1.0). (d) A Babinet copy with its peak on the cross-correlation map being a negative value.

Throughout each run, the genetic algorithm was controlled by the following parameters.

$N_{\text{chromosomes}}$: number of chromosomes

$N_{\text{generations}}$ number of generations

P_{cross}: probability for crossover (0.0 – 1.0)

P_{mutate}: probability for mutation (0.0 – 1.0)

$R_{\text{crosspoints}}$: number of crossover points represented by a fraction of chromosome size

$N_{\text{mutatepoints}}$: number of mutation points

4.4 Results and discussion

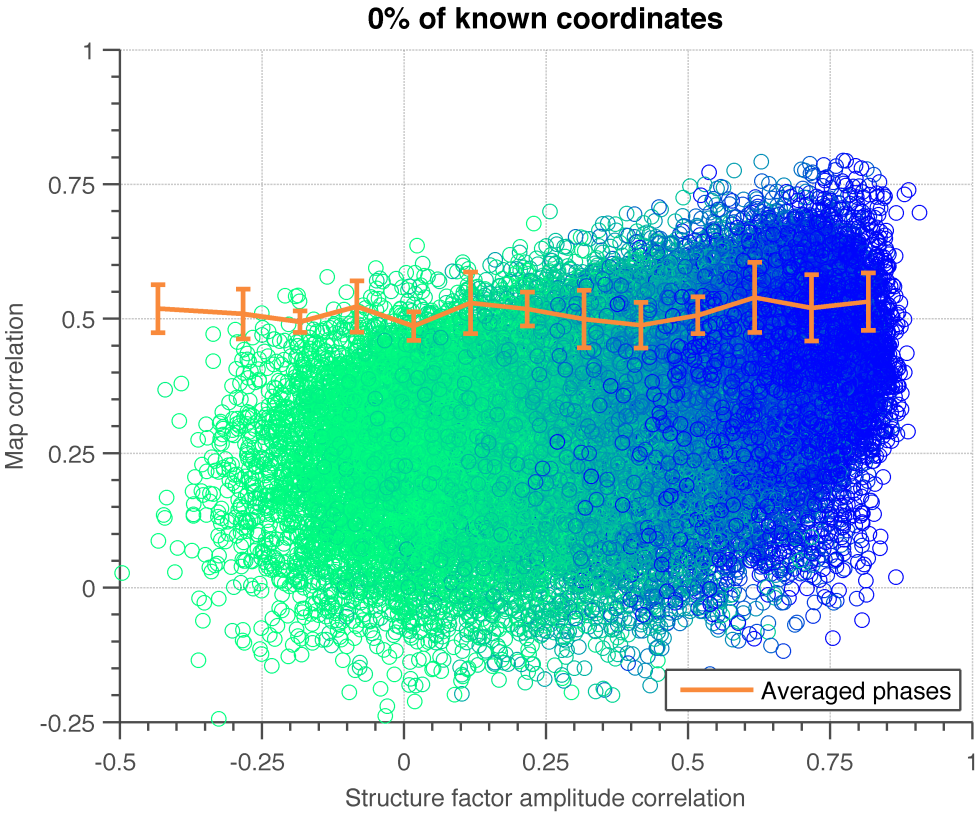
4.4.1 A test structure with 50% solvent content

The first test structure (Fig. 4.2*b*) demonstrates a case where scatterers ('1') occupy the same amount of cells as the background ('0'). The goal of the test involves identifying the amount of known information about the structure needed to rely on structure-factor amplitude correlation as a measure for the quality of phases. The algorithm was set to perform 10 independent runs on four types of test: given amount of known coordinates were 0%, 25%, 50%, and 75%. The genetic algorithm used the same parameters ($N_{\text{chromosomes}}$= 400, $N_{\text{generations}}$= 50, P_{cross}= 0.95, P_{mutate}= 0.01, $R_{\text{crosspoints}}$= 0.2, and $N_{\text{mutatepoints}}$= 1) for all the runs. Termination of the search occurred when every chromosome had map correlations > 0.9 or when the maximum number of generations was reached.

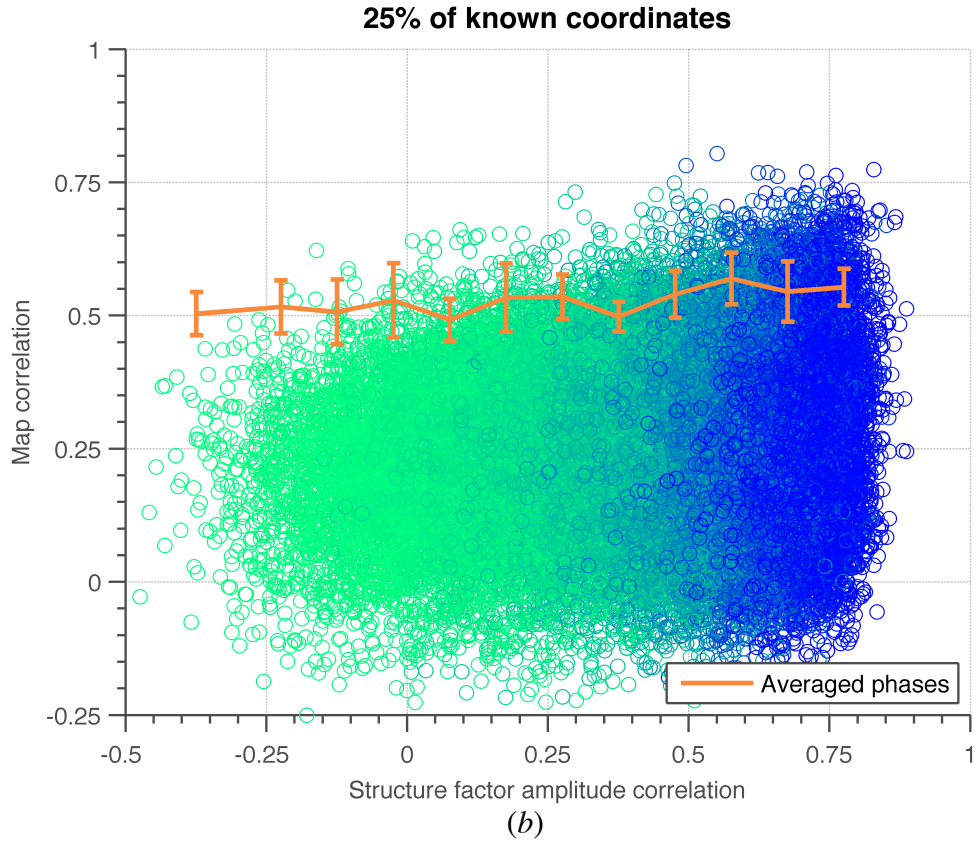
To observe changes in phase quality, we calculated a map correlation (eq. 4.1) by transforming a particular chromosome, which stored locations of '1' back to the 10x10 grids. Phases of the Fourier coefficients were used for the calculation of the map correlation in comparison with the known calculated phase. A scatter plot (Fig. 4.6) between the map correlation (vertical axis) and the structure-factor amplitude correlation (horizontal axis) having one particular point representing a set of phases was generated. The color, which goes from light green to dark blue, represents an increase in number of generations during the optimization process.

Each error bar represents the phase quality of the centroid phases computed from a collection of phase sets with similar structure-factor amplitude correlations. The quality of the centroid phase set is represented as a range (the error bar), because a group of around 100 chromosomes was selected as a representative set. The other chromosomes (with similar values of structure-factor amplitude correlation) had to be matched for different origin, enantiomorph, and Babinet copy to the selected group. The centroid phases were calculated after the matching procedures; therefore, the quality of phase was approximated from the representative chromosome set. We filled each plot in Fig. 4.6: the error bar on the plot represents the range of map correlations, the line that connects each bar represents the mean value, and the width of the bar represents 1σ above and below the mean value. These centroid phase sets tend to

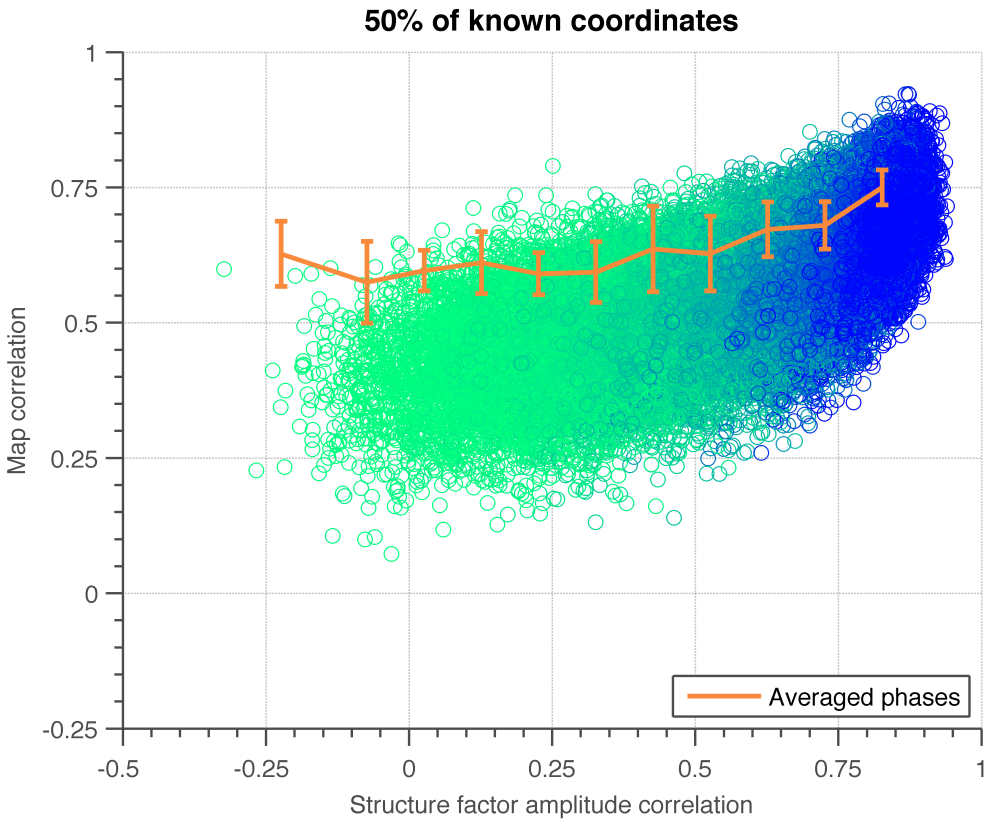
have higher phase quality than the individual samples, as evident in particular when larger amounts of known coordinates are supplied.



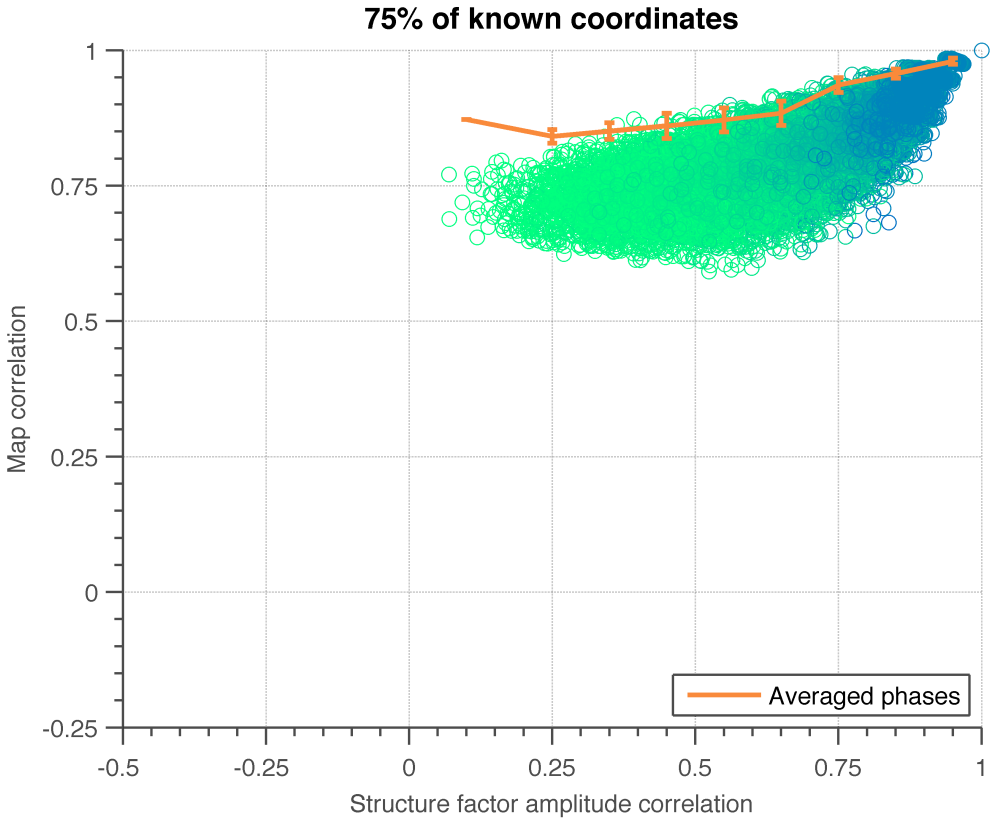
(*a*)



(*b*)


50% of known coordinates
Map correlation
Structure factor amplitude correlation
Averaged phases
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.25
0
0.25
0.5
0.75
1

(*c*)


75% of known coordinates
Map correlation
Structure factor amplitude correlation
Averaged phases
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.25
0
0.25
0.5
0.75
1

(*d*)

Figure 4.6 Measures of solution-phase quality for Fourier terms calculated from the solutions (10x10 sampling grids) for the test structure with 50% solvent content. The measures were calculated using map correlations (eq. 4.1) between the solution phases and the known phases. Each plot shows relations of the structure factor amplitude correlation and the map correlation calculated from chromosomes in different generations. The error bars show map correlations of the centroid phases calculated from a group of solutions with similar structure-factor amplitude correlations. The quality of centroid phases is represented as a range, since 100 phase sets were selected and the rest were used to perform phase match with the selected sets. The line connecting each error bar represents mean values and the width of the error bars show 1σ above and below the mean values. All plots show the results from 10 independent runs for the four tests with a different amount of known coordinates of: (a) 0% (b) 25%. (c) 50%. (d) 75%.

For cases of 0% (Fig. 4.6*a*) and 25% (Fig. 4.6*b*) supplied known coordinates, solutions with increasing values of structure-factor amplitude correlation failed to improve phase quality (map correlation). Mean values of map correlations for both cases stayed almost constant at around 0.5 from the range of structure-factor amplitude correlations from -0.5 to 0.9. The algorithm was terminated without reaching a homogeneous population and the calculated map correlations among the chromosomes in the last generation (generation 50^{th}) that yielded a value of ~0.4, show that the chromosomes in the population still remained different.

With the amount of known coordinates increased to 50% (Fig. 4.6*c*), the correlation between observed and calculated structure-factor amplitudes could be used to distinguish the solution from the non-solutions. Note that all solutions had these 50% of correct locations of '1' fixed throughout the optimization process and only the locations of the other 50% of '1' were varied. At a range of structure-factor amplitude correlations of -0.5 to 0.25, the averaged map correlation was 0.59. The range of amplitude correlation of 0.25 to 0.75 produced an averaged map correlation of 0.69 and the largest value of amplitude correlation above 0.75 produced the largest value of an averaged map correlation of 0.78. This result shows that for this test case when 50% of correct coordinates were known, increasing values of structure factor amplitudes would result in phase improvement for all the structure factors used in the calculation. Note that the algorithm failed to converge with homogeneous population; however, the averaged map correlation among chromosomes in the last generation was as high as 0.73.

When the amount of known coordinates is as large as 75%, the test structure could be completely recovered (solutions with map correlation = 1). An initial set of random solutions (with 70% of correct coordinates) yielded an averaged map correlation of 0.87 when structure-factor amplitude correlations were less than 0.2. Averaged map correlations increased to 0.98 when structure factor amplitude correlations reached the value of 0.9.

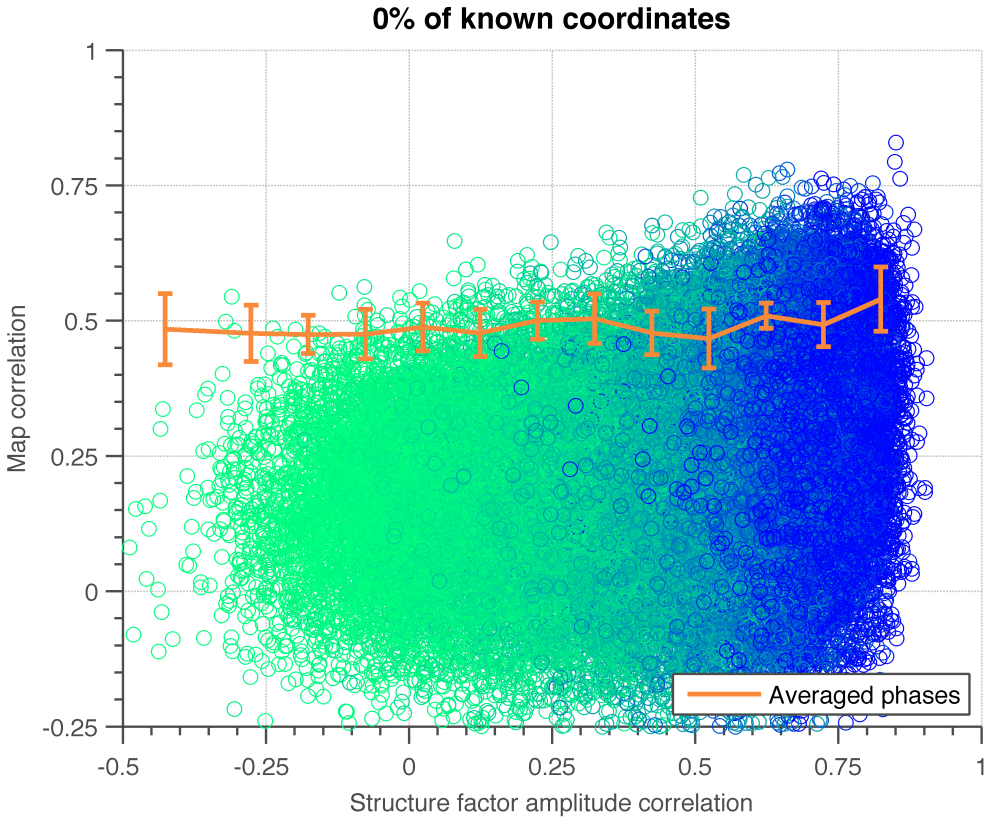
4.4.2 Structures with low and high solvent content

Another two structures with solvent content of 70% (Fig. 4.2*a*) and 30% (Fig. 4.2*c*) were artificially generated to test if solvent content affects the amount of known information about the structure needed to distinguish the solutions from the non-solutions using only the structure factor amplitude correlation. The second structure with low solvent content had 30 of '1'-locations while the last structure had 70 of '1'-locations.

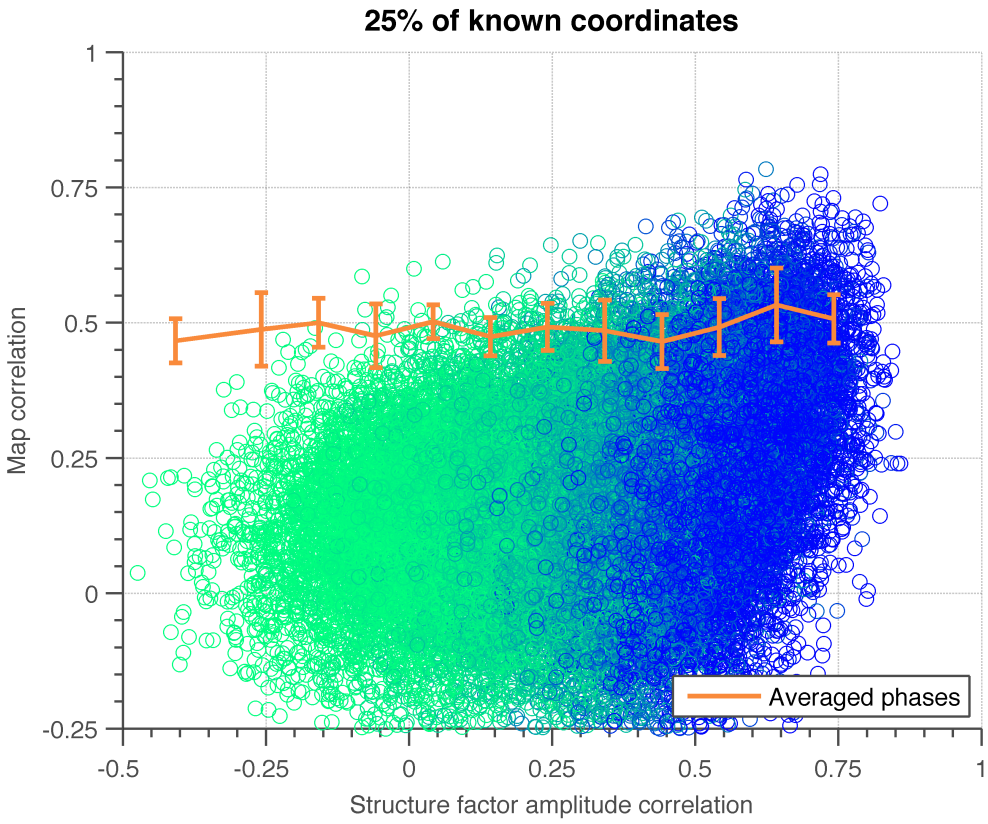
Tests were done using the same procedures as performed in 4.4.1. The genetic algorithm generated the first population pool using the known amount of coordinates of 0%, 25%, 50%, and 75%. For the cases with >0% known information, locations of '1' stayed the same throughout the search process and only the rest of '1'-locations were varied by the genetic operators. The algorithm terminated the run when all chromosomes were homogeneous or when the maximum number of generations was reached. Each chromosome, which stored the locations of '1', was transformed back to the 10x10 grids where a map correlation could be calculated.

Fig. 4.7 shows a measure of phase quality for the low-solvent structure (30%) for the 0%, 25%, 50%, and 75% known coordinate cases. Note that since the $\sqrt{(\Sigma \boldsymbol{F}^2)}$ of all Fourier terms calculated using 10x10 sampling grids (44 terms in total) yielded a value of only 63.5, these terms failed to provide enough information to describe the structure with $\boldsymbol{F}(0) = 70$. Similar results were observed for the case when 0% of known coordinates were supplied. An increase in structure-factor amplitude correlations failed to improve phase quality generated from the solutions. Even with as many as 50% of known coordinates supplied, significant improvement of map correlations could not be observed. It required the amount of known coordinates of 75% such that improvements in the quality of phases could be obtained. The averaged

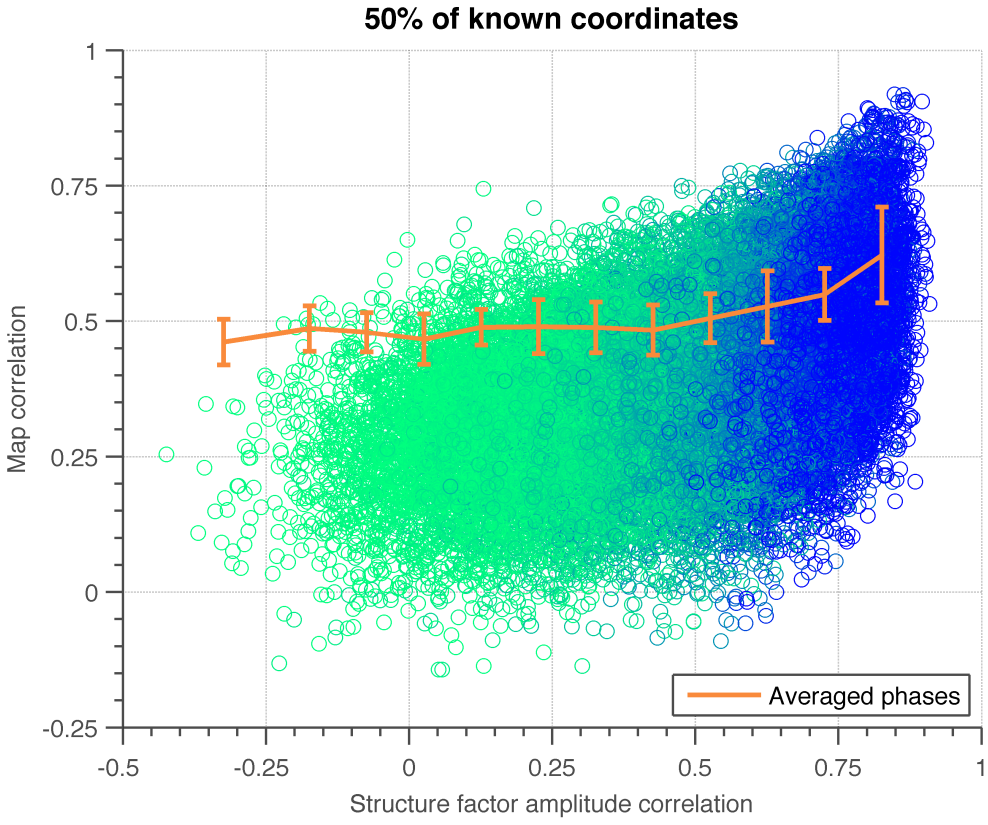
phases calculated from the chromosomes with structure-factor amplitude correlations > 0.88 yielded averaged map correlations of around 0.94. Note that the initial value of map correlation was already as large as 0.76.



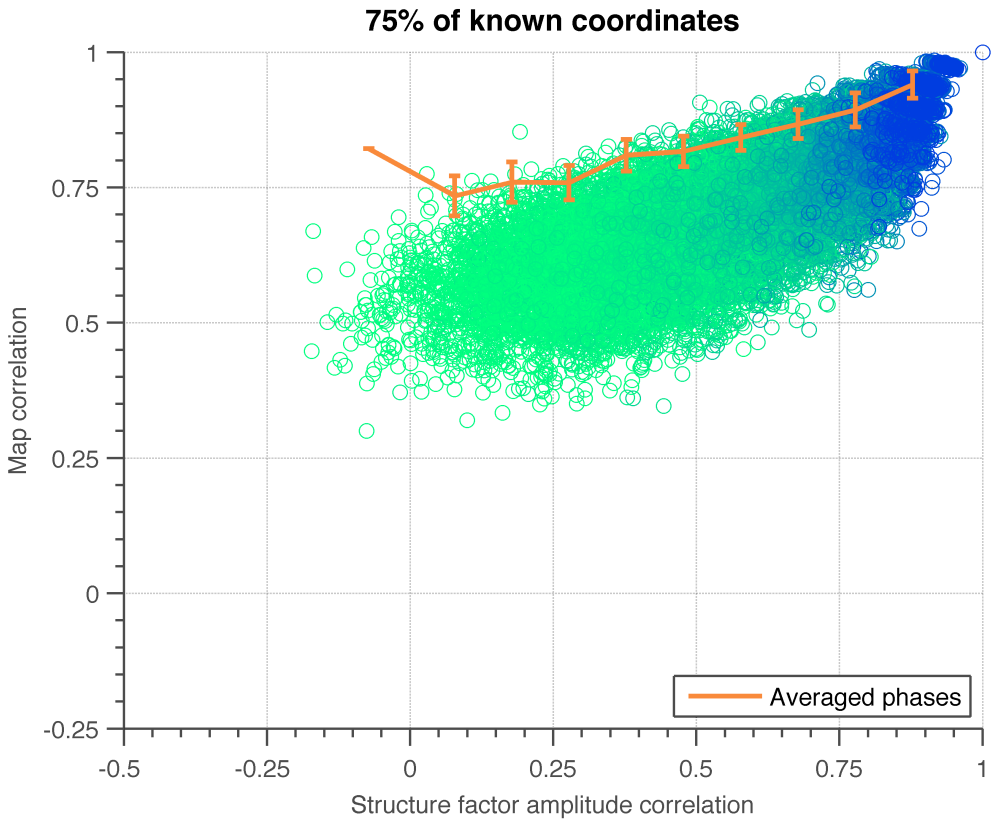
(*a*)


25% of known coordinates
Map correlation
Structure factor amplitude correlation
Averaged phases
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.25
0
0.25
0.5
0.75
1

(*b*)


50% of known coordinates
Map correlation
Structure factor amplitude correlation
Averaged phases
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.25
0
0.25
0.5
0.75
1

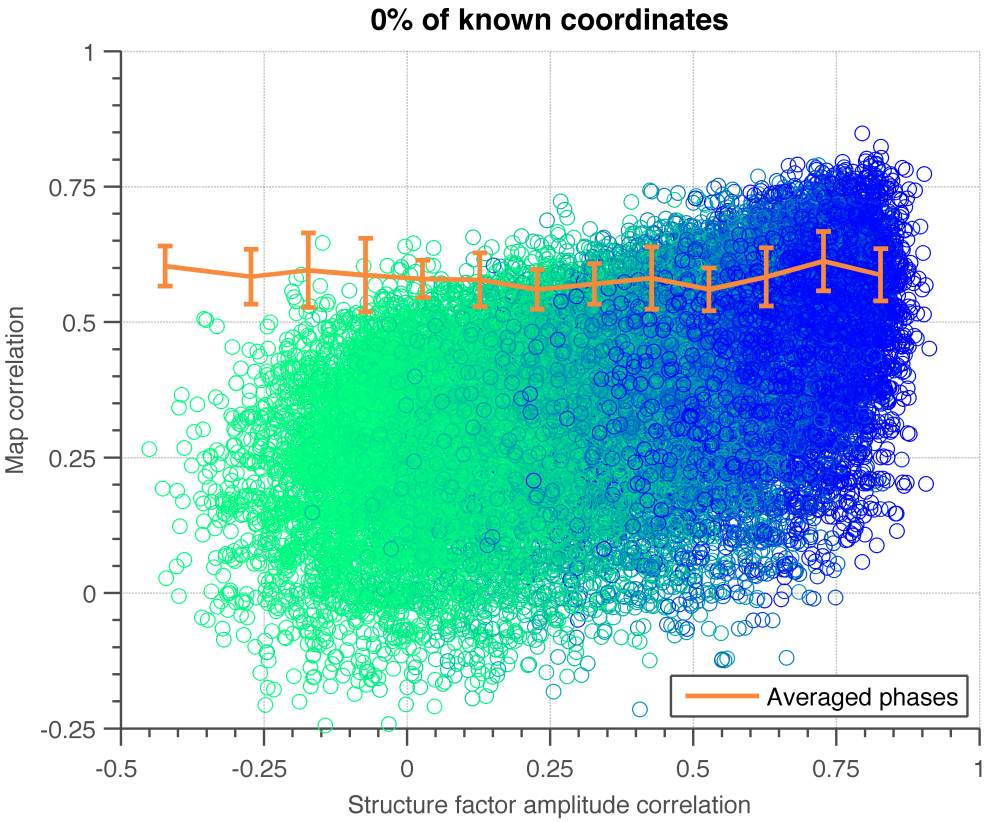
(*c*)



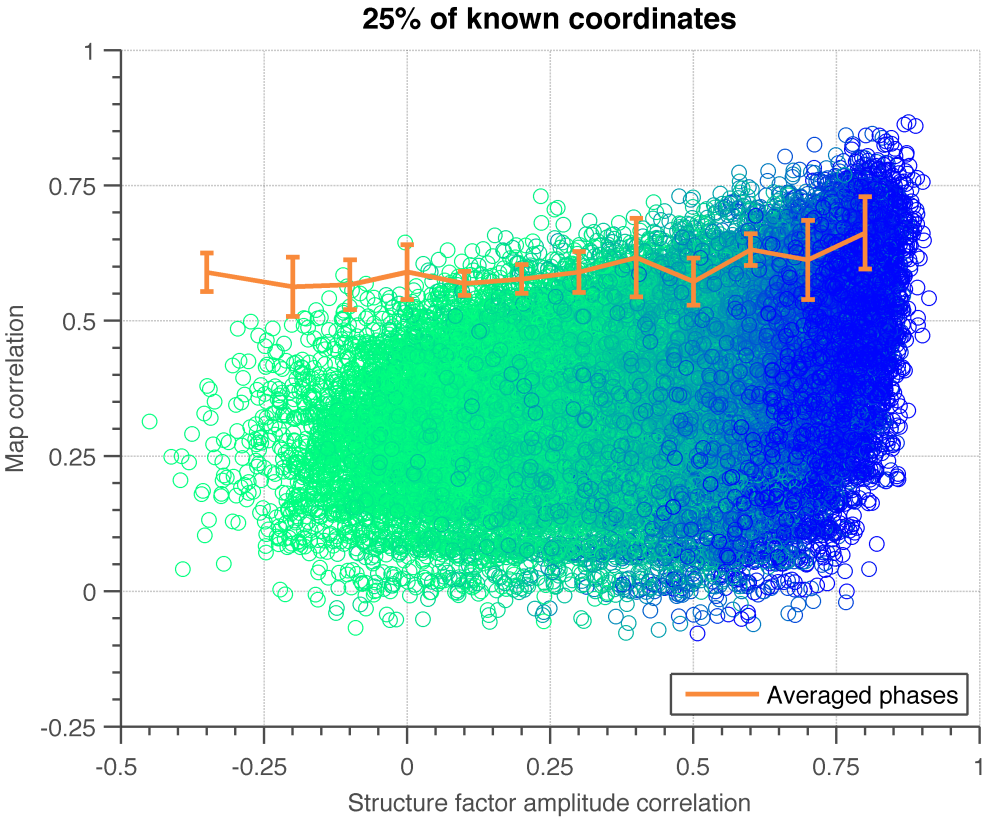
(*d*)

Figure 4.7 Measures of solution-phase quality for Fourier terms calculated from the solutions (10x10 sampling grids) in the case of 30% solvent content. The four test cases were set with different amounts of known coordinates of (a) 0%, (b) 25%, (c) 50%, (d) 75%.

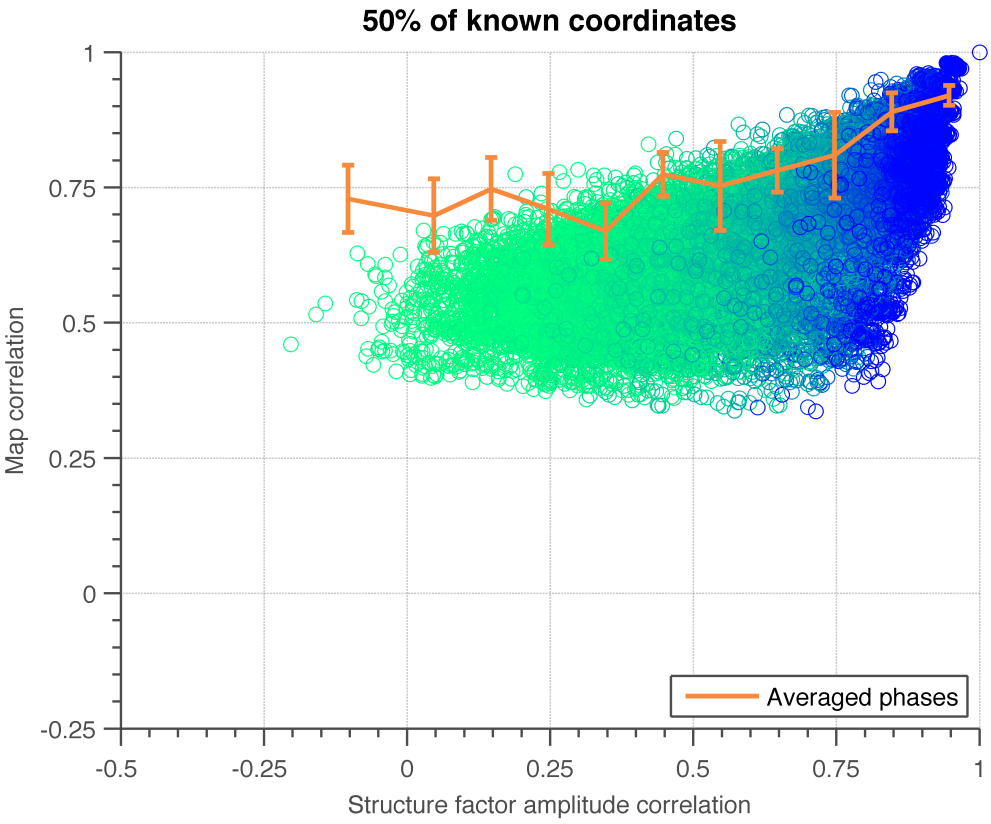
For the high-solvent content structure, the results with 0% known coordinates supplied show that map correlations stayed unchanged at a value of ~0.6 independent of the increase of the structure-factor amplitude correlations (Fig. 4.8*a*). Note that for this test structure with 30 locations of '1' ($\boldsymbol{F}(0) = 30$), only the 17 largest structure-factor amplitudes were adequate to generate a recognizable inverse Fourier transform for an equal-atom binary structure. Therefore, the use of 44 Fourier terms in the search provided more than enough information required to generate the structure. The test structure was completely recovered when 50% of known coordinates were used in the search. The set of chromosomes with the structure-factor amplitude correlation > 0.8 produced an averaged map correlation of 0.93. The algorithms were terminated with an averaged map correlation among chromosomes in the last population of 0.79. With as high as 75% known coordinates supplied (Fig. 4.8*d*), it took only 19 generations on average for the algorithms to converge. The solutions were found with an averaged map correlation of 0.97.


0% of known coordinates
Map correlation
Structure factor amplitude correlation
Averaged phases
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.25
0
0.25
0.5
0.75
1

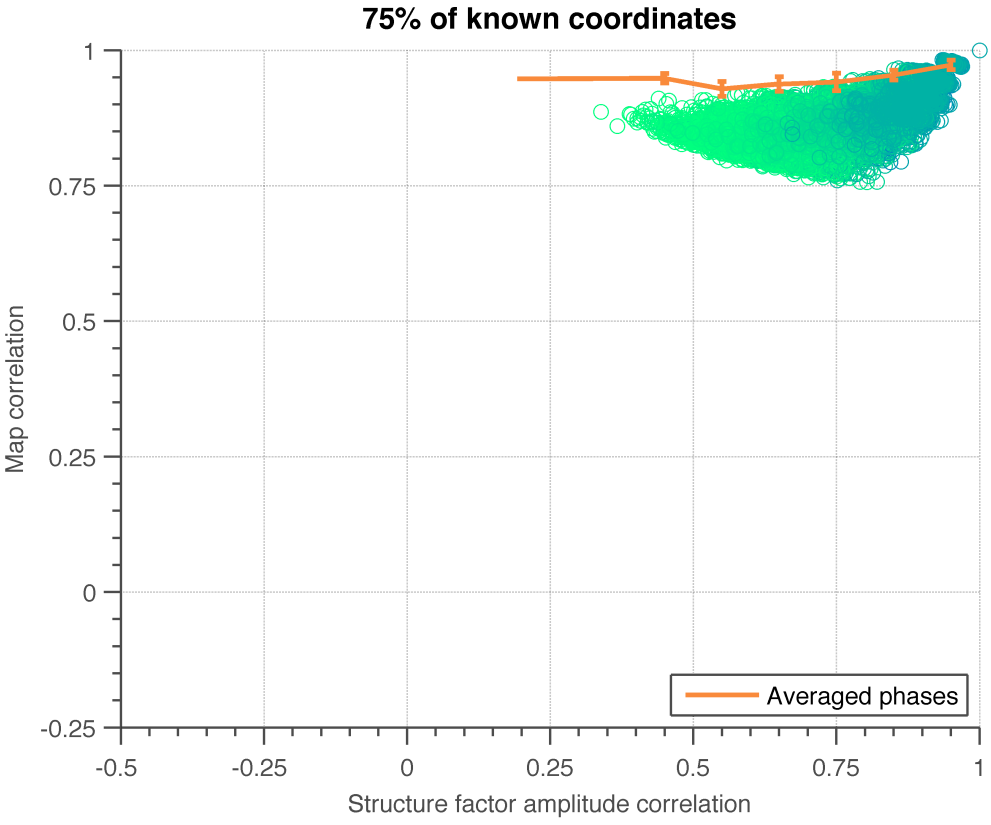
(*a*)


25% of known coordinates
Map correlation
Structure factor amplitude correlation
Averaged phases
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.25
0
0.25
0.5
0.75
1

(*b*)


50% of known coordinates
Map correlation
Structure factor amplitude correlation
Averaged phases
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.25
0
0.25
0.5
0.75
1

(*c*)


75% of known coordinates
Map correlation
Structure factor amplitude correlation
Averaged phases
1
0.75
0.5
0.25
0
-0.25
-0.5
-0.25
0
0.25
0.5
0.75
1

(*d*)

Figure 4.8 Measures of solution phase quality for Fourier terms calculated from the solutions (10x10 sampling grids) in the case of 70% solvent content. The four test cases were set with different amounts of known coordinates of (a) 0%, (b) 25%, (c) 50%, (d) 75%.

To differentiate the amount of known information about the structure needed to rely on structure-factor amplitude correlation as a measure of phase quality for the structures with different level of solvent, a summary of results from the three test structures grouped by 0%, 25%, 50%, and 75% of known-coordinate tests is given in Fig. 4.9. The results shown here were generated from a group of chromosomes with structure-factor amplitude correlations < 0.2 (circle markers) and > 0.8 (upward-pointing triangle markers). The dark, medium, and light colors of the markers show the results of the structures with solvent content of 70%, 50%, and 30% respectively. With 0% of known coordinates given, no significant phase improvement could be observed from the increases of structure-factor amplitude correlations from -0.5 to around 0.9. The lowest solvent content structure (30%) required as high as 75% of known coordinates to be supplied for the search to obtain a significant increase of map correlation. The higher solvent content structures (50% and 70%) required around 50% of known coordinates to obtain a similar improvement with more improvement observed for the structure with the highest solvent content. Both test structures with 30% and 50% solvent content were completely recovered when 75% of known coordinates were given. The highest solvent-content structure, the 70%, was recovered with fewer amounts of known coordinates supplied (50% of known coordinates).

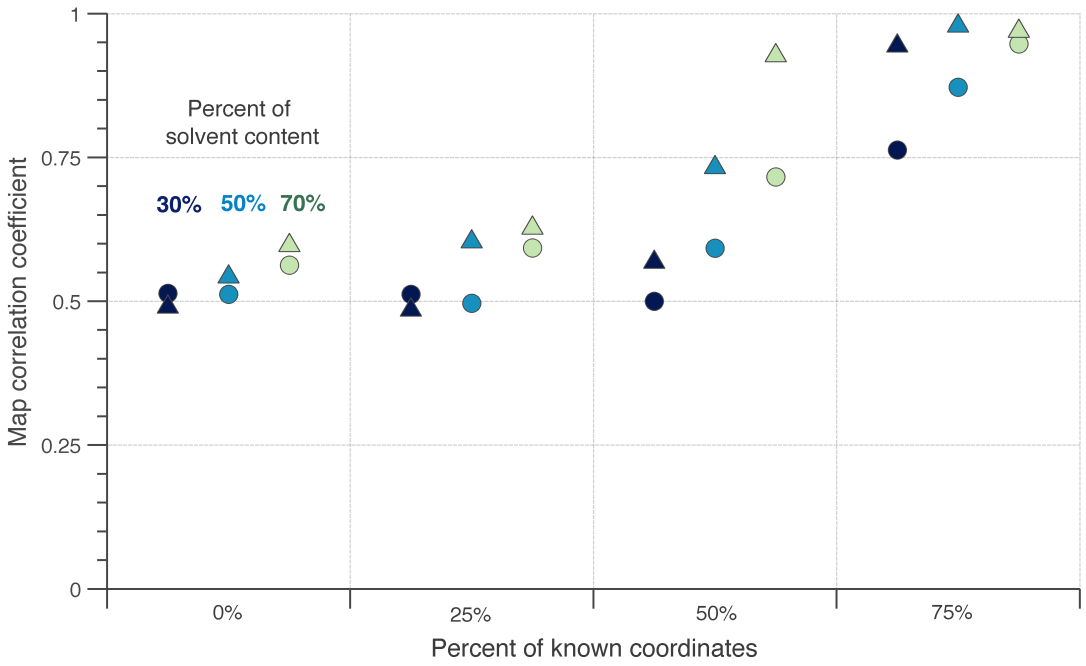


Figure 4.9 Required amounts of known coordinates to obtain increases of map correlation with the increases of the structure-factor amplitude correlation for the three test structures with solvent content of 30% (dark color), 50% (medium dark color), and 70% (light color). The circle marker and the upward triangular marker show map correlations calculated from averaged solutions with structure-factor amplitude correlation <0.2 and >0.8 respectively. Map correlation improvements were displayed in separated columns according to the percent of known coordinates supplied for the search. The plot reveals that significant map correlation improvement could be obtained only when at least 50% of known coordinates were given with the highest improvement occurring in the test structure with the highest solvent content (70%).

4.5 Conclusions

This work explored if structure-factor amplitude correlations could be used to determine the quality of phases in *ab-initio* phasing. The focus was on an equal-atom binary structure in 2-dimensional layout with 10x10 sampling grids. Three types of structures with 30%, 50%, and 70% solvent content were artificially generated for the tests. Their structure-factor amplitudes were used with the genetic algorithm to search for solutions with high values of the correlation between the observed (generated from the test structure) and the calculated (generated from the solutions) structure-factor amplitudes. The three structures were tested with an initial set of solutions generated using different amount of known coordinates of 0%, 25%, 50%, and 75%. A measure of phase quality was calculated from a map correlation (eq. 4.1) between the solutions found during the search and the test structures.

The results for all test structures show that large values of structure-factor amplitudes generated from the solutions, which lacked information about known coordinates were uncorrelated with the structures' map correlations. When structure-factor amplitude correlation increased from 0 to ~0.8, map correlations calculated from the solutions stayed unchanged at a value of ~0.5. Results of the structure with 50% solvent content show that at least 50% of known coordinates should be supplied, because with at least this amount of information, the structure-factor amplitude correlation could be used to distinguish the solutions from the non-solutions.

The amount of solvent content had some impact on the amount of known coordinates needed for the search to rely on the structure-factor amplitude correlation to determine the quality of phases. Higher solvent content required fewer amounts of known coordinates to be supplied for the search to obtain solutions. The test structure with 70% solvent content needed 50% of known coordinates for the search to be fully recovered. The lower-solvent content structures, the 50% and 70% solvent structures, required 75% of known coordinates to yield the same result.

The algorithm was designed to terminate under two conditions: 1) the solutions in the population pool had map correlation > 0.9 or 2) the maximum number of generations was reached. Note that only when the algorithm found a completely recovered test structure, the first condition was met. In other test runs, when the chromosomes reached map correlations < 1, these chromosomes in the population pool still remained different. These runs were set to terminate at the limit number of generations of 50 due to the observation that with this value, all chromosomes could already reach a structure factor-amplitude correlation of around 0.9.

5 Summary and outlook

This thesis combines crystallographic knowledge and computational algorithms to tackle the phase problem. I associated the role of the strongest reflections and map skewness, a measure of the quality of density map, to improve the quality of phases from experimental phasing. The developed genetic algorithm, SISA, reconstructed optimized phases that can be used to improve density modification. For *ab-inito* phasing studies, I shed light onto using the structure-factor amplitude correlation as a measure of map quality.

A computer program, SISA (SIR/ SAD phase optimization), was created to optimize the quality of phases for a few strongest reflections using map skewness as a target function. For the tests, I selected experimental data that had failed to give complete structures after density modification and model building. The program reduced phase errors in the optimized phases compared to the original centroid phases, leading to a greater success in the subsequent model building. In one of the cases, this new method enabled successful model building where SAD phasing had failed to do so. Additionally, results from the tests also showed that integrating the rest of the reflections with the varied ones during the search played a significant role in phase improvements. SISA calculation times depend on the size of the experimental data and the number of the selected strongest reflections.

The role of the strongest reflections and the map skewness are the key ideas for the success of phase optimization in SISA. The algorithm used the structure-factor amplitudes to sort and select these strongest reflections; the normalized structure factors (E) can also be used instead. Phases optimized this way might have fewer errors. However, the density modification and the model building based on an electron-density map calculated from the normalized structure factors must be used to measure the impact of the optimized phases. Additional measures of the quality of an electron-density map such as the local r.m.s error can be used to improve the efficiency of the target function in optimizing the quality of phases. Fewer phase

errors for the strongest reflections help decrease errors in an electron-density map, leading to a greater improvement in density modification and model building.

I also constructed another genetic algorithm to study the usability of structure-factor amplitude correlation for *ab-initio* phasing. The problem was configured in a 2-dimensional setting using 10x10 grids. The focus was on finding out the amount of known coordinates needed to rely on the structure-factor amplitude correlation as a measure of phase quality. Given a structure consisting of only equal atoms, when information about the structure was lacking (coordinates or phases unknown), solutions with large values of structure-factor amplitude correlation still yielded incorrect structures. The test structures were categorized according to their solvent content; as a result, I observed that the low-solvent structure required a higher amount of known coordinates to obtain phase improvement with the increase of structure-factor amplitude correlations. The medium- and high-solvent structures required fewer amounts of known coordinates to achieve similar results; with 50% of known coordinates supplied, both structures were fully recovered.

The study of structure-factor amplitude correlation for 2-dimensional structures helps in understanding their usability for *ab-initio* phasing. Although structure-factor amplitude correlation is not useful for *ab-initio* phasing in the complete absence of prior structural information, the results in this thesis show that phases can be recovered using the genetic algorithm when some coordinates of the structures are known. The high-solvent structure yielded the best results (in terms of the completeness of the solutions and the calculation times); this work can be extended to 3-dimensional data for macromolecules with an inclusion of finding the translated, inverted, and Babinet copy of a structure in a 3-dimensional setting.

Bibliography

Adams, P.D., Afonine, P.V., Chen, V., Echols, N., Headd, J. J., Hung, L.-W., Grosse-Kunstleve, R.W., McCoy, A.J., Moriarty, N.W., Read, R.J., Richardson, D.C., Richardson, J.S., Terwilliger, T.C. and Zwart, P. H. (2010). *PHENIX*: a comprehensive Python-based system for macromolecular structure solution. *Acta Cryst*. **D66**, 213–221.

Bäck, T., de Graaf, J. M., Kok, J. N. & Kosters, W. A. (1997). Theory of genetic algorithms. *Bull. Eur. Assoc. Theor. Comput. Sci*. **63**, 161-192.

Blow, D. M. & Crick, F. H. C. (1959). The treatment of errors in the isomorphous replacement method. *Acta Cryst*. **12**, 794-802.

Chothia, C. & Lesk, A.M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO J*. **5**, 823-826.

Cochran, W. (1952). A relation between the signs of structure factors. *Acta Cryst*. **5**, 65-67.

Cochran, W. (1955). Relations between the phases of structure factors. *Acta Cryst*. **8**, 473-478.

Collaborative Computational Project, Number 4 (1994). The *CCP4* suite: programs for protein crystallography. *Acta Cryst*. **D50**, 760-763.

Connor, R. (1994). *Practical Handbook of Genetic Algorithms: Applications*, edited by L. D. Chambers, pp. 57-74. Boca Raton: CRC Press.

Crowther, R. (1972). The fast rotation function. In *Rossmann M (Ed.) The Molecular Replacement Method*. New York: Gordon and Breach Science Publishers.

Crowther, R. & Blow, D.M. (1967). A method of positioning a known molecule in an unknown crystal structure. *Acta Cryst*. **23**, 544-548.

De Jong, K. A. & Spears W.M. (1991). An analysis of the interacting roles of population size and crossover in genetic algorithms. *Proceedings of the 1st workshop on parallel problem solving from nature*. Springer-Verlag, UK.

Debaerdemaeker, T. & Woolfson, M. M. (1983). On the application of phase relationships to complex structures. XXII. Techniques for random phase refinement. *Acta Cryst*. **A39**, 193-196.

Feng, Z. J. & Dong, C. (2007). GEST: a program for structure determination from powder diffraction data using a genetic algorithm. *J. Appl. Cryst*. **40**, 583-588.

Giacovazzo, C. (2006). Direct methods. *International Tables for Crystallography. Vol. B*, ch. 2.2, 210-234

Gilmore, C.J. (2000). Direct methods and protein crystallography at low resolution. *Acta Cryst*. D**56**, 1205-1214.

Goedkoop, J.A. (1950). Remarks on the theory of phase limiting inequalities and equalities. *Acta Cryst*. **3**, 374-378.

Goldberg, D.E. (1989). Genetic algorithms in search, optimization, & machine learning. *Addison Wesley Longman, Inc*. Crawfordsville, IN, USA.

Green, D. W., Ingram, V. M. & Perutz, M. F. (1954). The structure of haemoglobin. IV. Sign determination by the isomorphous replacement method. *Proc. R. Soc. London Ser*. A**225**, 287-307.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W. & Adams, P. D. (2002). The *Computational Crystallography Toolbox*: crystallographic algorithms in a reusable software framework. *J. Appl. Cryst*. **35**, 126-136.

Harker, D. & Kasper, J.S. (1948). Phases of Fourier coefficients directly from crystal diffraction data. *Acta Cryst*. **1**, 70-75.

Harris, K. D. M., Habershon, S., Cheung, E. Y. & Johnston, R. L. (2004). Developments in genetic algorithm techniques for structure solution from powder diffraction data. *Z. Kristallogr*. **219**, 838-846.

Hendrickson, W. A. & Lattman, E. E. (1970). Representation of phase probability distributions for simplified combination of independent phase information. *Acta Cryst*. **B26**, 136-143.

Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems*. Pittsburg: University of Michigan Press.

Immirzi, A., Erra, L., & Tedesco, C. (2009). Finding crystal structures of peptides by random search and evolutionary algorithms. *J. Appl. Cryst*. **42**, 810-814.

Kabsch, W. (1988). Evaluation of single-crystal X-ray diffraction data from a position-sensitive detector. *J. Appl. Cryst*. **21**, 916-924.

Karle, J. & Hauptman, H. (1950). The phases and magnitudes of the structure factors. *Acta Cryst*. **3**, 181-187.

Karle, J. & Hauptman, H. (1956). A theory of phase determination for the four types of non-centrosymmetric space groups 1P222, 2P22, 3P12, 3P22. *Acta Cryst*. **9**, 635-651.

Kissinger, C.R., Gelhaar, D.K., & Fogel, D.B. (1999). Rapid automated molecular replacement by evolutionary search. *Acta Cryst*. D**55**, 484-491.

Liu, Q., Dahmane, T., Zhang, Z., Assur, Z., Brasch, J., Shapiro, L., Mancia, F., & Hendrickson, W.A. (2012). Structures from anomalous diffraction of native biological macromolecules. *Science* **336**, 1033-1037.

Lunin, V. Y. & Woolfson, M. M. (1993). Mean phase error and the map-correlation coefficient. *Acta Cryst*. D**49**, 530-533.

Lunin, V.Y., Lunina, N.L., Petrova, T.E., Skovoroda, T.P., Urzhumtsev, A.G., & Podjarny, A.D. (2000). Low-resolution *ab initio* phasing: problems and advances. *Acta Cryst*. D**56**, 1223-1232.

Lunin, V. Y., Lunina, N.L., Casutt, M.S., Knoops, K., Schaffitzel, C., Steuber, J. Fritz, G. & Baumstark, M.W. (2012). Low-resolution structure determination of Na+-translocating NADH:ubiquinone oxidoreductase from *Vibrio cholerae* by *ab initio* phasing and electron microscopy. *Acta Cryst*. D**68**, 724-731

McCoy, A.J., Storoni, L.C. & Read, R.J. (2004). Simple algorithm for a maximum-likelihood SAD function. *Acta Cryst*. D**60**, 1220-1228.

McCoy, A.J., Grosse-Kunstleve, R.W., Adams, P.D., Winn, M.D., Storoni, L.C., & Read, R.J. (2007). *Phaser* crystallographic software. *J. Appl. Cryst*. **40**, 658-674.

Miller, S.T., Hogle, J.M. & Filman, D.J. (1996). A genetic algorithm for the ab initio phasing of icosahedral viruses. *Acta Cryst*. D**52**, 235-251.

Mitchell, T.M. (1997). *Machine Learning*. Singapore: McGraw-Hill Companies, Inc.

Navaza, J. (1994). AMoRe: an automated package for molecular replacement. *Acta Cryst*. A**50**, 157-163.

Otwinowski, Z. (1991). Maximum likelihood refinement of heavy atom parameters. *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80-86. Warrington: Daresbury Laboratory.

Patterson, A.L. (1934). A Fourier series method for the determination of the components of interatomic distances in crystals. *Phys. Rev.* **46**(5), 372-376.

Perutz, M.F. (1956). Isomorphous replacement and phase determination in non-centrosymmetric space groups. *Acta Cryst.* **9**, 867-873.

Pflugrath, W. (1999). The finer things in X-ray diffraction data collection. *Acta Cryst.* D**55**, 1718-1725.

Podjarny, A. D. & Yonath, A. (1977). Use of matrix direct methods for low-resolution phase extension for tRNA. *Acta Cryst.* A**33**, 655-661.

Read, R. J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Cryst.* A**42**, 140-149.

Read, R. J. (2001). Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Cryst.* **D57**, 1373-1382.

Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M., & Usón, I. (2009). Crystallographic *ab-initio* protein structure solution below atomic resolution. *Nat. Methods* **6(9)**, 651–653.

Rossmann, M.G. (2001). Molecular replacement – historical background. *Acta Cryst.* D**57**. 1360-1366.

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

Sayre, D. (1952). The squaring method: a new method for phase determination. *Acta Cryst.* **5**, 60-65.

Shamoo, Y., Krueger, U., Rice, L.M., Williams, K.R., Steitz, T.A. (1997). Crystal structure of the two RNA binding domains of human hnRNP A1 at 1.75 A resolution. *Nat. Struct. Biol.* 4(3), 215-222.

Shankland, K., David, W. I. F. & Csoka, T. (1997). Crystal structure determination from powder diffraction data by the application of a genetic algorithm. *Z. Kristallogr.* **212**, 550-552.

Sheldrick, G.M. & Gould, R.O. (1995). Structure solution by iterative peaklist optimization and tangent expansion in space group *P*1. *Acta Cryst.* **B51**, 423-431.

Sheldrick, G. M., Hauptman, H. A., Weeks, C. M., Miller, M. & Usón, I. (2001). *Ab-initio* phasing. *International Tables for Macromolecular Crystallography, Vol. F*, ch. 16.1, 333–345.

Skinner, M. M., Zhang, H., Leschnitzer, D. H., Bellamy, H., Sweet, R. M., Gray, C. M., Konings, R. N. H., Wang, A. H.-J. & Terwilliger, T. C. (1994). Structure of the gene V protein of bacteriophage f1 determined by multiwavelength x-ray diffraction on the selenomethionyl protein. *Proc. Natl Acad. Sci. USA*, **91**, 2071-2075.

Su, W.-P. (2008). Retrieving low- and medium-resolution structural features of macromolecules directly from the diffraction intensities - a real-space approach to the X-ray phase problem. *Acta Cryst.* A**64**, 625-630.

Subbiah, S. (1991). Low-resolution real-space envelopes: an approach to the *ab-initio* macromolecular phase problem. *Science*. 252, 128-133.

Sutton, R.B., Ernst, J.A., & Brunger, A.T. (1999). Crystal structure of the cytosolic C2A-C2B domains of synaptotagmin III. Implications for Ca(+2)-independent snare complex interaction. *J Cell Biol*. 147(3), 589-98.

Svergun, D.I. & Franke, D. (2009). DAMMIF, a program for rapid *ab-initio* shape determination in small-angle scattering. *J. Appl. Cryst.* **42**, 342-346

Syswerda, G. (1989). Uniform crossover in genetic algorithms. *Proceedings of the 3rd International Conference on Genetic Algorithms*. Morgan Kaufman Publishing, USA.

Tanaka, S., Sawaya, M.R., Phillips, M. & Yeates, T.O. (2009). Insights from multiple structures of the shell proteins from the beta-carboxysome. *Protein Sci*. **18**, 108-120.

Terwilliger, T. C., Adams, P. D., Read, R. J., McCoy, A. J., Moriarty, N. W., Grosse-Kunstleve, R. W., Afonine, P. V., Zwart, P. H. & Hung, L.-W. (2009). Decision-making in structure solution using Bayesian estimates of map quality: the *PHENIX AutoSol* wizard. *Acta Cryst*. D**65**, 582-601.

Vekhter, Y. (2005). Improving experimental phasing: the role of strongest reflections. *Acta Cryst*. D**61**, 899-902.

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

Webster, G. & Hilgenfeld, R. (2001). An evolutionary computational approach to the phase problem in macromolecular X-ray crystallography. *Acta Cryst*. A**57**, 351-358.

Weiss, M. S., Sicker, T., Djinovic-Carugo, K. & Hilgenfeld, R. (2001). On the routine use of soft X-rays in macromolecular crystallography. *Acta Cryst*. **D57**, 689-695.

Zhang, L. & Doudna, J.A. (2002). Structural insights into group II intron catalysis and branch-site selection. *Science*. 295, 2084-2088.

Zhou, Y. & Su, W.-P. (2004). Solving the Sayre equations for centrosymmetric structures with a genetic algorithm. *Acta Cryst*. A**60**, 306-310.

Acknowledgements

Thank you.

Prof. Hilgenfeld for your initiative to start the computational work on *ab-initio* phasing; for giving me useful resources to tackle the problem; for your support on SISA work; for the stimulating discussions; for an extensive review of this thesis, and most of all, for the research opportunity.

Dr. Mesters for your teaching in crystallography course in my first year; for suggestions on my *ab-initio* phasing work – it turned out to be a very interesting result!

Prof. Read from the University of Cambridge for giving the talk about the maximum likelihood at ECM26, Darmstadt in 2010; for initiating ideas on the role of the strongest reflections and experimental phase optimization; for a careful review on the manuscript of SISA work, which is part of this thesis; and most of all for the internship opportunity – all of these led to the birth of SISA.

Dr. Terwilliger from the Los Alamos National Laboratory for your light-speed replies to my emails; for a careful review on SISA manuscript; for writing the map-likelihood function for me to test; and for very interesting discussions.

Prof. Martinetz and Dr. Mamlouk for interesting discussions on the algorithms; for the neural-network course; for MATLAB server, which I used intensively to build SISA and *ab-initio* phasing search algorithm prototypes.

Helgo, Raspudin, Friedrich, and Gabor (University of Cambridge) for useful discussions on the phase problem. Helgo provided SIRAS data from *WaaA* to test with SISA. Friedrich provided several datasets from *Tsp*.

Sarah and Claudia for your friendship.

Achim for your helps on IT.

Krishna, Karin, Anja, Walter, Aditi, Silke, Doris, and other staffs at the institute for making the institute an interesting place to be.

and my family for your love and support.