


Article

# Direct Phasing of Coiled-Coil Protein Crystals

Ruijiang Fu <sup>1</sup>, Wu-Pei Su <sup>2,\*</sup> and Hongxing He <sup>1,\*</sup> <sup>1</sup> Department of Physics, School of Physical Science and Technology, Ningbo University, Ningbo 315211, China<sup>2</sup> Department of Physics and Texas Center for Superconductivity, University of Houston, Houston, TX 77204, USA

\* Correspondence: wpsu@uh.edu (W.-P.S.); hehongxing@nbu.edu.cn (H.H.)

**Abstract:** Coiled-coil proteins consisting of multiple copies of helices take part in transmembrane transportation and oligomerization, and are used for drug delivery. Cross-alpha amyloid-like coiled-coil structures, in which tens of short helices align perpendicular to the fibril axis, often resist molecular replacement due to the uncertainty to position each helix. Eight coiled-coil structures already solved and posted in the protein data bank are reconstructed ab initio to demonstrate the direct phasing results. Non-crystallographic symmetry and intermediate-resolution diffraction data are considered for direct phasing. The retrieved phases have a mean phase error around 30~40°. The calculated density map is ready for model building, and the reconstructed model agrees with the deposited structure. The results indicate that direct phasing is an efficient approach to construct the protein envelope from scratch, build each helix without model bias which is also used to confirm the prediction of AlphaFold and RosettaFold, and solve the whole structure of coiled-coil proteins.

**Keywords:** coiled coil; ab initio phasing; protein crystallography; hybrid-input output; non-crystallographic symmetry



**Citation:** Fu, R.; Su, W.-P.; He, H. Direct Phasing of Coiled-Coil Protein Crystals. *Crystals* **2022**, *12*, 1674. <https://doi.org/10.3390/cryst12111674>

Academic Editors: Blaine Mooers and Jolanta Prywer

Received: 16 October 2022

Accepted: 18 November 2022

Published: 20 November 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

There is a large number of coiled-coil structures in fibrin proteins, keratin proteins and collagen proteins. Solving the atomic structures of coiled-coil proteins is crucial for understanding their biological functions and in discovering their potential applications. In 2008, the Usón group proposed the Arcimboldo program to solve the structure of coiled-coil motifs [1].

In protein X-ray crystallography, in order to determine the atomic structure, one has to solve the phase problem. In the diffraction experiment, only the amplitudes of the diffracted X-rays are recorded, the phases are lost. In order to obtain the electron density, phases are required. The most common methods to solve the phase problem in protein crystallography are isomorphous replacement (SIR/MIR) [2], anomalous dispersion (SAD/MAD) [3] and molecular replacement (MR) [4]. However, some coiled-coil crystals such as membrane proteins resist the introduction of heavy atoms or selenium replacement. As a result, it is difficult to solve those structures by SIR/MIR or SAD/MAD. Instead, molecular replacement is usually used to solve the atomic structures of coiled-coil proteins.

Molecular replacement method requires a reference structure with high sequence similarity and low structure deviations. Coiled-coil proteins are super-helical twists consisting of at least two  $\alpha$ -helices around each other in the molecule [5]. For long helices, even if the amino acid sequences are similar, their atomic structures might have a large deviation. For short helices, if the target structure contains more than 10 short helices, it is difficult to locate the exact position of each helix in the unit cell during molecular replacement. For completely new coiled coils, a homology reference with high sequence similarity is not available. In a word, molecular replacement could fail in solving coiled-coil structures. Therefore, the direct phasing approach is a necessary supplement to the current phasing methods. It deserves more research and development.

For small molecule crystals, the direct phasing method is a standard technique, which makes use of the Sayre's equation [6], triplet relation [7], tangent formula [8], etc. For protein crystals, the direct phasing method is different, which makes use of iterative projection algorithms [9–12]. It retrieves the lost phases directly from the diffraction data without any prior information about the protein structure. It does not require any heavy-atom derivative crystals or reference structures. Starting from random numbers, the lost phases are retrieved iteratively through thousands of iterations. In each iteration, the calculated density is modified to satisfy density constraints in real space and intensity constraints in reciprocal space. The modified density improves the calculated phases.

Iterative projection algorithms are often used for density modification techniques, such as hybrid input-output (HIO) and difference map (DM). Iterative projection algorithms (IPA) have been developed for decades. In 1978, Fienup proposed that the phase problem in optics could be solved by HIO [13–15], which is an IPA with negative feedback. In 1993, Millane used IPA to solve the phase problem of two-dimensional periodic objects [9]. In 1997, Millane used the same algorithm to reconstruct the structure of a virus from its simulated diffraction data [16]. In 2003, Elser proposed an iterative projection algorithm named difference map [10] and successfully solved several structures of small proteins with around 400 non-hydrogen atoms [17]. In the same year, Marchesini reconstructed the non-periodic image of a small amount of gold nano particles by the HIO algorithm and shrink-wrap technique without any prior information [18]. In 2012, Liu & Dong used the HIO algorithm to solve protein structures with known protein masks [11], which first demonstrated the capability of a direct method to solve the atomic structures of biological macromolecules. In 2015, He & Su developed a complete ab initio phasing method with HIO algorithm and solved several protein structures with high solvent content starting from random numbers [12]. Lo & Millane proposed a direct phasing method (DM) and solved several protein and virus structures starting with approximate protein envelopes and information on non-crystallographic symmetry [19,20]. In 2019, He et al. further developed the HIO phasing method with non-crystallographic symmetry (NCS) constraints to deal with protein crystals with intermediate or low solvent content [21]. In 2022, Kingston and Millane proposed a general method for direct phasing high-solvent-content protein crystals [22].

In the following sections, as an illustration of direct phasing coiled-coil proteins, six structures previously determined by conventional crystallographic analysis were chosen from the Protein Data Bank (PDB). The diffraction resolution of those structures ranges from 2.0 to 3.3 Å. The solvent content is from 35.5 to 80%. The number of  $\alpha$ -helices ranges from 2 to 18 in a single molecule. Real diffraction data were used to reconstruct the electron density. The calculated density was modeled by *ARP/wARP* [23]. More than 60% of amino acids in the sequence are successfully placed. The rebuilt models match the PDB posted structures with a root mean square deviation less than 1 Å.

We first describe the solution of a known structure with PDB code 6c4y, which is an amyloid-like cross- $\alpha$  structure protein [24]. There are 18 copies in the asymmetric unit. The results of 1uii, 6g6b and 1no4 will be briefly described. A structure with five-fold symmetry is used as an example to demonstrate direct phasing of protein crystals with non-crystallographic symmetry. Two other examples of 6eik and 5ez8 with seven-fold rotational NCS will be briefly described. Another amyloid-like cross- $\alpha$  protein with PDB code 6c4z [24] will be used to illustrate the phasing of intermediate- and low-resolution data. The resolution is 3.3 Å.

## 2. Methodology

### 2.1. Direct Phasing Method with Hybrid-Input Output Algorithm

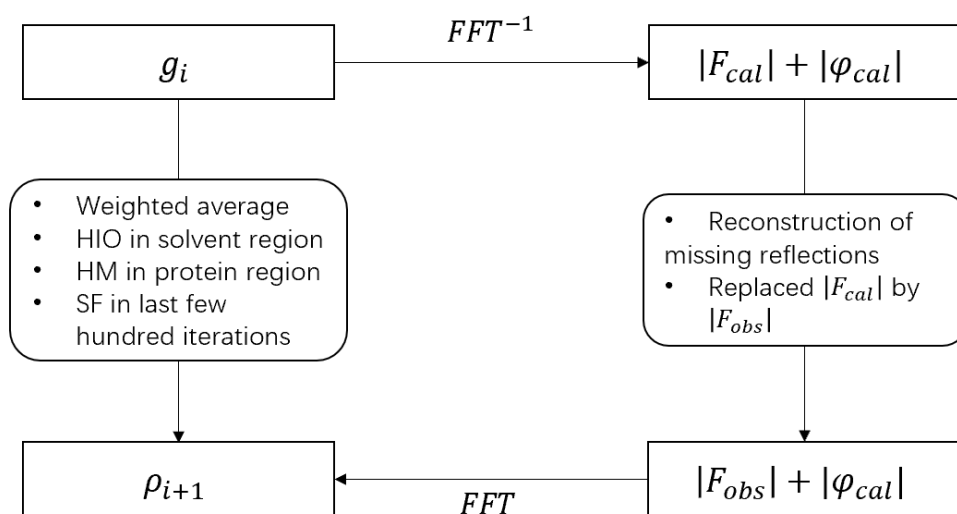
The requirement of a unique solution is important for ab initio phasing. The electron density  $\rho$  is defined on a grid that covers an asymmetric unit containing  $N$  grid points. The diffracted intensities correspond to the square of the Fourier amplitude  $|F(h, k, l)|$ . The fast Fourier transform of densities on  $N$  grid points in real space leads to  $N$  Fourier coefficients

in reciprocal space (Equation (1)). Since densities are real numbers, the Fourier coefficients are conjugate pairs. Only  $N/2$  Fourier coefficients are independent. In that case, for given  $N/2$  Fourier coefficients, the density is uniquely determined. However, the phases of  $N/2$  Fourier coefficients can't not be measured in experiment. Half of the diffraction information are lost. In order to determine a unique density, half of the density must be known or constant. In protein crystals, the solvent between the crystallized macromolecules has a constant density. Therefore, the uniqueness of the solution for ab initio phasing requires a solvent region greater than 50% in order to determine a unique density ab initio from the diffraction data [10,19,25].

$$\rho(x, y, z) = \frac{1}{V} \sum_{h,k,l=-\infty}^{\infty} |F(h, k, l)| e^{i\varphi(h,k,l)} e^{-i2\pi(hx+ky+lz)} \quad (1)$$

An iterative projection method is used to solve the phase problem ab initio when the uniqueness of the solution is satisfied. The density has to satisfy the reconstructed protein-mask constraints in real space and the measured modulus constraints in reciprocal space. In each iteration, IPA method projects the calculated density to satisfy those constraints. After thousands of iterations, the final density satisfies all constraints at the same time. Since the solution is unique, the resulted density must be the solution. Several density modification techniques are often used to modify the calculated density. One is Fienup's hybrid-input output (HIO) algorithm, the other is Elser's difference map (DM) algorithm. HIO introduces a negative feedback term to update the calculated density in the solvent region. The update rule of DM introduces a negative feedback term of the difference between the projected densities of different constraints. Both of them work well in solving the phase problem. In this paper, HIO is used to update the calculated density.

Direct phasing of protein crystals requires thousands of dual-space iterations. Figure 1 shows the flowchart of each iteration. Starting from a random density, the density at the beginning of the  $i$ th iteration is  $g_i$ . After a backward fast Fourier transform [26], the calculated amplitudes  $|F_{cal}|$  are replaced by the observed ones  $|F_{obs}|$  which are downloaded from the protein data bank (PDB), and the calculated phases are retained. After a forward fast Fourier transform, a new density  $\rho_{i+1}$  is obtained. The calculated density has to satisfy several constraints, such as protein boundary, protein density histogram, non-crystallography symmetry, etc. A weighted average density map  $w_i$  is calculated from density  $\rho_{i+1}$ . A cutoff value on  $w_i$  is searched in accordance with an approximate solvent content of the crystal. Grid points with a  $w_i$  value greater than the cutoff form the protein mask. The remaining grid points form the solvent region. HIO is used to update the density in the solvent region. Histogram matching [27] is used to adjust the density in the protein region. After density modification, the next iteration begins. When the error metrics drop below a threshold, the iteration stops. The resulted density satisfies all constraints at the same time and it must be the unique solution.



**Figure 1.** The flowchart for an iteration cycle of direct phasing with hybrid-input-output (HIO). The initial density is random.  $g_i$  is the input density of the  $i$ th iteration.  $\rho_{i+1}$  is the output density before density modification of the  $i$ th iteration.  $|F_{obs}|$  is the diffraction data downloaded from the protein data bank (PDB).

Density modifications in real space include several techniques, such as weighed average density [12], HIO [13], histogram matching(HM) [27], solvent flattening(SF) [28], and so on. A weighted average density is calculated from  $\rho_{i+1}$  to distinguish the protein region from solvent region in the crystal. The weighting function is a Gaussian function (Equation (2)).

$$W_i = \sum_j \exp[-d_{ij}^2 / (2\sigma^2)] \rho_j \quad (2)$$

where  $d_{ij}$  is the distance between two grid points  $i$  and  $j$ .  $\sigma$  is the Gaussian radius with an initial value of 4.0 and it linearly decreases to 3.0 at the end of the iterations. Bigger  $\sigma$  helps to locate an estimate of the protein mask. Smaller  $\sigma$  is good for locating a refined protein mask. Histogram matching is used to adjust the calculated density of the protein region. Because the density-frequency distribution of different molecules at the same resolution is very similar, the histogram of known proteins can be used as a reference histogram to modify the calculated density of an unknown protein. HIO is used to update the density in the solvent region (Equation (3)).

$$g_{i+1} = \begin{cases} \rho_{i+1}, & \text{protein region} \\ g_{i+1} - \epsilon \rho_{i+1}, & \text{solvent region} \end{cases} \quad (3)$$

where  $g_i$  and  $g_{i+1}$  are the input densities of the current and the next iteration.  $\rho_{i+1}$  is the output density after Fourier refinement.  $\epsilon$  is a negative feedback factor with an empirical value of 0.7.

Error metrics, such as  $R_{free}$ ,  $R_{work}$ ,  $\Delta\varphi$  and  $CC$ , are computed to monitor the calculated density (Equations (4)~(7)).  $R_{work}$  measures the difference between the calculated amplitudes and the diffraction data. About 1% diffraction data are randomly selected as a free data set. They are not involved in the phasing process.  $R_{free}$  is calculated using the free data set to avoid over-fitting [29]. For the initial random density,  $R$  values are big. As  $R_{work}$  and  $R_{free}$  exhibit a sudden drop at the same time, the calculated density converges to the solution.  $\Delta\varphi$  measures the difference between the retrieved phases and the true phases. The true phases are computed from the atomic model deposited in the protein data bank.  $CC$  measures the difference between the calculated structure factor and the true structure factor. For an unknown structure,  $R$  values can be computed, but  $\Delta\varphi$  and  $CC$

are not available. For initial random phases,  $\Delta\varphi$  is close to 90 degrees, and CC is around 0. When the calculated density converges to the solution, R values drops to about 0.3.  $\Delta\varphi$  decreases to a value around 30~40 degrees. CC increases to a value close to 1.0.

$$R_{work} = \frac{\sum_{\mathbf{h} \in work} ||F_{obs}(\mathbf{h})| - \lambda|F_{cal}(\mathbf{h})||}{\sum_{\mathbf{h} \in work} |F_{obs}(\mathbf{h})|} \quad (4)$$

$$R_{free} = \frac{\sum_{\mathbf{h} \in free} ||F_{obs}(\mathbf{h})| - \lambda|F_{cal}(\mathbf{h})||}{\sum_{\mathbf{h} \in free} |F_{obs}(\mathbf{h})|} \quad (5)$$

$$\Delta\varphi = \frac{\sum_{\mathbf{h} \in work} \arccos\{\cos[\varphi_{ture}(\mathbf{h}) - \varphi_{cal}(\mathbf{h})]\}}{\sum_{\mathbf{h} \in work} 1} \quad (6)$$

$$CC = \frac{\sum_{\mathbf{h} \in work} |F_{obs}(\mathbf{h})||F_{cal}(\mathbf{h})|\cos[\varphi_{ture}(\mathbf{h}) - \varphi_{cal}(\mathbf{h})]}{[\sum_{\mathbf{h} \in work} |F_{obs}(\mathbf{h})|^2 \sum_{\mathbf{h} \in work} |F_{cal}(\mathbf{h})|^2]^{1/2}} \quad (7)$$

Missing diffraction data are reconstructed during the iterations. Due to the beam stop, tens of small-angle diffraction data are lost. Due to experimental errors, some diffraction data at high resolution are not complete. All missing data are rebuilt according to Equation 8 from the calculated amplitudes with a scale factor [11].

$$|F_{missing}(\mathbf{h})| = \frac{\sum_{\mathbf{h} \in work} |F_{obs}(\mathbf{h})|}{\sum_{\mathbf{h} \in work} |F_{cal}(\mathbf{h})|} |F_{cal}(\mathbf{h})| \quad (8)$$

A data weighting technique is used to improve the phasing speed [20,30] (Equations (9) and (10)). In order to reconstruct the detailed density, a protein profile is determined from the low-resolution data. At the beginning of the iteration, only the low-resolution data are involved in the phasing process. When the iteration proceeds, more high-resolution data get involved. Since the number of low-resolution data is much less than the number of high-resolution diffraction data, data weighting makes it easier to reconstruct the protein profile and speeds up the convergence [30].

$$W_1(S_h) = e^{-2(\pi\sigma_1 S_h)^2} \quad (9)$$

$$|F_{obs, \mathbf{h} \in work}(\mathbf{h})| = w_1 |F_{obs}(\mathbf{h})| \quad (10)$$

## 2.2. Direct Phasing with Non-Crystallographic Symmetry Density Averaging

Non-crystallographic symmetry is a strong density constraint [30]. A coiled-coil protein crystal usually consists of multiple copies related by non-crystallographic symmetry (NCS) [21]. An average density map of NCS related copies reduces the number of unknown densities and helps retrieving the lost phases. Improper NCS often consists of rotational and translational operations. Proper NCS usually consists of rotational operations. In this paper, proper NCS is considered. The direction of the rotational axis is obtained from the self-rotation map. The position of the rotational axis in the unit cell is assumed to be known. Take 4uot and 6eik as examples, the symmetry involves five-fold and seven-fold rotational NCS, respectively.

In order to generate an average density map, an NCS mask is required to cover the correct volume of all equivalent copies [21]. In order to reconstruct the NCS mask, an estimate of the center of the NCS mask is determined in the unit cell. Near the NCS center, a core of the NCS mask is able to indicate each copy of the molecule. The core is isolated from their crystallographic equivalents. Starting from random density, the center of the NCS mask is searched and a core of NCS mask is reconstructed and updated. Starting from the surface of the core, an NCS mask is reconstructed. Following this procedure, the NCS mask has a high probability to cover the correct volume. Three weighted average density maps  $w_1$ ,  $w_2$ , and  $w_3$  are computed from the calculated density  $\rho$ , with three empirical

values of sigma 15 Å, 5 Å, and 3 Å.  $w_1$  is used to locate the center of the NCS mask in the unit cell.  $w_2$  is used to locate a core of the NCS mask around the NCS center.  $w_3$  is used to expand the NCS core into a complete NCS mask. Grid points are added to the surface of the NCS core according to their values of  $w_3$ . For more details, please refer to reference [21].

The existence of non-crystallographic symmetry increases the redundancy of the phase problem [11,21]. Since the NCS related copies have the same density, the number of unknown densities decreases. The uniqueness of the solution is easily satisfied [11,19,21,25,31]. A crystal with a solvent content less than 50% still have a unique solution. For example, if the crystal has a five-fold NCS, only 1/5 of the protein density is unknown. A unique solution only requires the solvent region greater than 1/5 of the protein region, which corresponds to a solvent content of 16.67%. In practice, the protein boundary cannot be accurately reconstructed starting from random phases. Therefore, a slightly higher solvent content is needed in order to retrieve the lost phases in ab initio phasing.

### 2.3. Direct Phasing of Intermediate- and Low-Resolution Data

Protein crystals often diffract to intermediate or low resolution, such as membrane crystals. The capability to direct phasing low resolution data is very useful. The direct phasing method for high-resolution data are modified in accordance with the decrease of the resolution. When phasing intermediate- and low-resolution data, the boundary between protein and solvent regions are not quite clear. The electron density of the solvent region is not a constant any more. The solvent density near the blurred boundary obviously deviates from a constant. Moreover, the highest density of the protein region also becomes smaller which requires slightly different density-modification techniques.

A density-limited HIO has been proposed for direct phasing intermediate- and low-resolution data. Fienup's HIO algorithm modifies the density in the solvent region without any limitations. The modified density could reach a large positive or negative value. The extreme value in the solvent region destroys the calculated density of the unitcell. Therefore, the HIO-modified density is limited into a proper range in accordance with the resolution of the diffraction data. Empirically,  $\pm 1.0 \text{ e}^-/\text{\AA}^3$  is used for phasing 3.0 Å data. At the end of iteration, HIO is turned off by decreasing the limit to  $\pm 0.1 \text{ e}^-/\text{\AA}^3$ . Due to the low resolution, solvent density is not flat any more. Solvent flattening is not suitable and will increase the R value. Therefore, when phasing low-resolution data, at the end of the iteration, a limit of  $\pm 0.1 \text{ e}^-/\text{\AA}^3$  is used for density-limited HIO to replace solvent flattening.

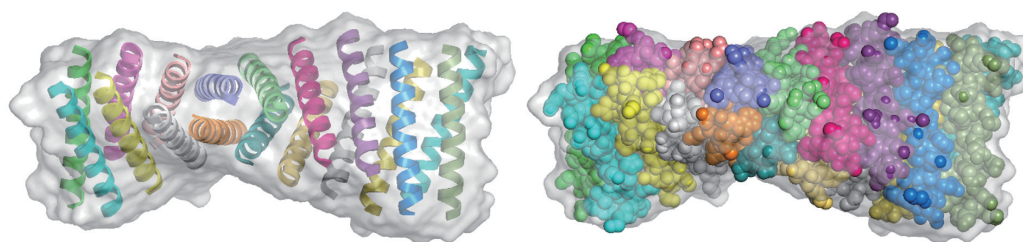
## 3. Results

### 3.1. Direct Phasing of AlphaFold- or MR-Difficult Structures

For a coiled-coil protein containing dozens of short helices, it is difficult to determine the structure by AlphaFold [32]/RosettaFold [33] or molecular replacement. AlphaFold focuses on predicting the structure of a single protein chain. A coiled-coil structure usually contains multiple chains. AlphaFold is capable of predicting an accurate structure for each chain of the coiled-coil protein. If the chain is long and does not contain intrinsic disorder, it is possible to reconstruct the whole structure by molecular replacement. However, if the chain is short or contains regions that are intrinsically disordered, it is difficult to solve the whole structure of the protein with AlphaFold/AlphaFold-Multimer or MR. In this case, direct phasing method provides an alternative approach.

The first trial structure is an artificially designed peptide that assembles into a cross-alpha amyloid-like structure with 18 helices running perpendicular to the fibril axis. It is an analog of the membrane-spanning  $\text{Zn}^{2+}$ -transporting peptide. Their structural fold is useful for directing in vivo protein assemblies with predicted spacing and stabilities. It was designed and solved by Zhang et al. using molecular replacement with great effort [24]. The structure has a PDB code 6c4y. The space group is  $P4_322$  with cell dimensions  $a = b = 125.241 \text{ \AA}$ ,  $c = 119.026 \text{ \AA}$ . The crystal diffracts to 2.5 Å. There are 3747 non-hydrogen atoms in the asymmetric unit. The solvent fraction is about 72%. There are 18 copies of a short  $\alpha$ -helix in the asymmetric unit (Figure 2).



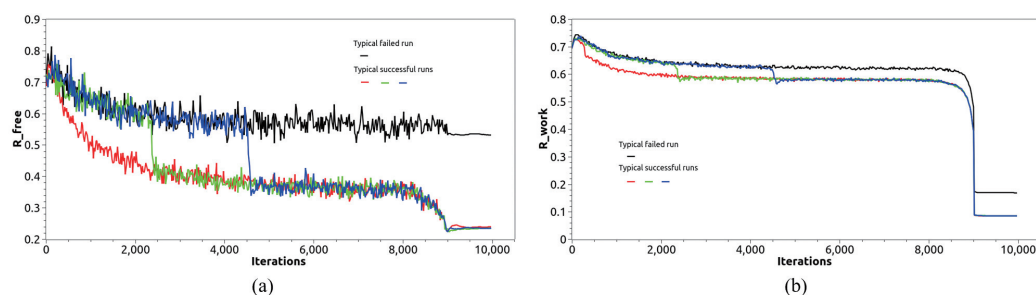


**Figure 2.** The structure of a coiled-coil protein with PDB code 6c4y, consisting of 18 copies of a short helix [24]. AlphaFold predicts an accurate model of each helix, but it is difficult to obtain the model of the whole molecule with MR. The molecule is shown as cartoon on the left and as sphere on the right. The true protein envelope computed from the PDB atomic model is shown as gray profile.

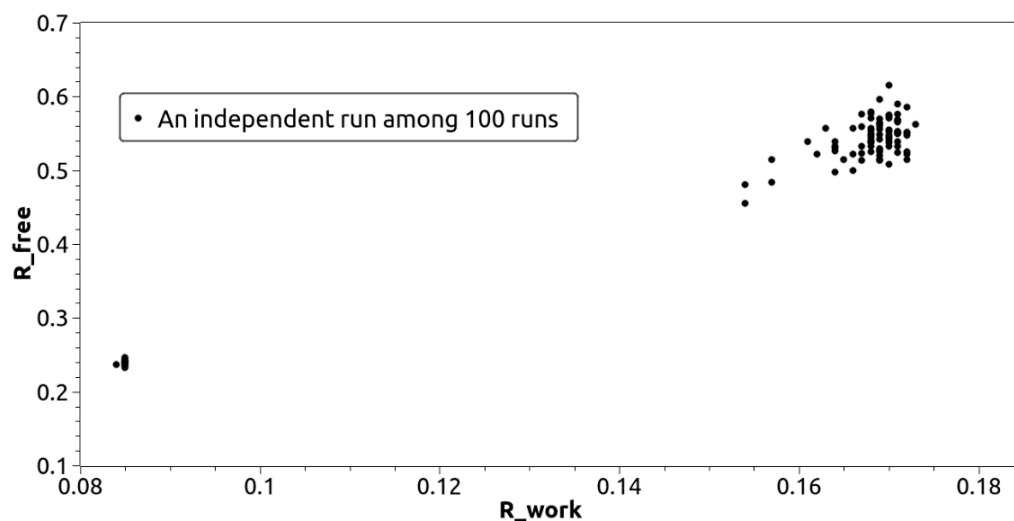
Direct phasing parameters are chosen as follows. The unit cell is divided into a grid with a spacing of 1.0 Å. The Matthews coefficient of 6c4y corresponds to a solvent content of 72.13%. After some trial and error, the solvent content was set to 68%. A loose protein mask helps to locate an approximate protein region. The protein mask is constructed from the calculated density by Equation (2). The initial value of sigma is set to 4.0 which helps to estimate the protein envelope. Sigma decreases linearly with iteration to make the reconstructed envelope more accurate. The final sigma is set to 3.0 at the 8000th iteration. The weighted average density is computed as a Gaussian convolution of the calculated density. The convolution is done in reciprocal space. A cutoff value on the weighted average density map is searched in accordance with the approximate solvent content of the crystal. Grid points with a weighted average density above the cutoff value make up the protein mask. Outside of the protein mask is the solvent region. Fienup's hybrid-input output algorithm is used to modify the computed electron density in the solvent region. Histogram matching is used to modify the calculated electron density in the protein region. The reference histogram is the histogram of 2uxj [34] at 2.5 Å. After 8000 iterations, HIO is gradually turned off. For the last 1000 iterations, Wang's solvent flattening is applied to set the electron density in the solvent region to zero. Error metrics defined by Equations (4)–(7) are computed to monitor the calculated density. In each iteration, the amplitudes of missing reflections including  $F_{000}$  are reconstructed according to Equation (8). Since small-angle reflections often contain large measurement errors, the cutoff value of low resolution is set to 15 Å. The amplitudes of the reflections with a resolution below 15 Å are reconstructed too.

A data weighting technique is applied to help locating the protein envelope [30]. In the early iterations, only low-resolution data are used to retrieve the phases. High resolution data get involved when iterations proceed. The data weighting parameters are determined as follows. There are about 33,367 unique reflections. In the first iteration, the sigma in Equation (9) is set to 1.2. About 3500 low-resolution reflections get involved into the phasing process. The equivalent resolution is around 5.4 Å. In each of the following iterations, three more reflections get involved into the phasing process. At the end of 8000th iteration, all reflections get involved.

Starting from random phases, 100 independent runs were carried out. After 10,000 iterations, we got 16 converged density maps among 100 runs. We ran the program on a Dell R740 server with two Intel Xeon 6230R cpu. Each cpu has 26 cores corresponding to 52 threads. The base frequency is 2.1 GHz. There are 128 GB memory. Figure 3 depicts the evolution of error metrics  $R_{free}$  and  $R_{work}$  of several typical successful runs and a typical failed run. The sudden drop of  $R_{free}$  and  $R_{work}$  indicates a successful run. As shown in Figure 4, the final values of  $R_{free}$  and  $R_{work}$  clearly identify the successful runs from the failed ones.



**Figure 3.** The evolution of (a)  $R_{free}$ , (b)  $R_{work}$  for the ab initio phasing of 6c4y. The initial phases are random. When a solution is achieved, R values indicates a sudden drop. After the 9000th iteration, solvent flattening is applied.



**Figure 4.** R values clearly separate the successful runs from the failed ones [35]. Each point on the figure corresponds to one of the 100 runs. Successful runs have smaller  $R_{work}$  and  $R_{free}$  at the same time. They are located at bottom left on the figure. Failed runs have larger R values. They are located at top right on the figure.

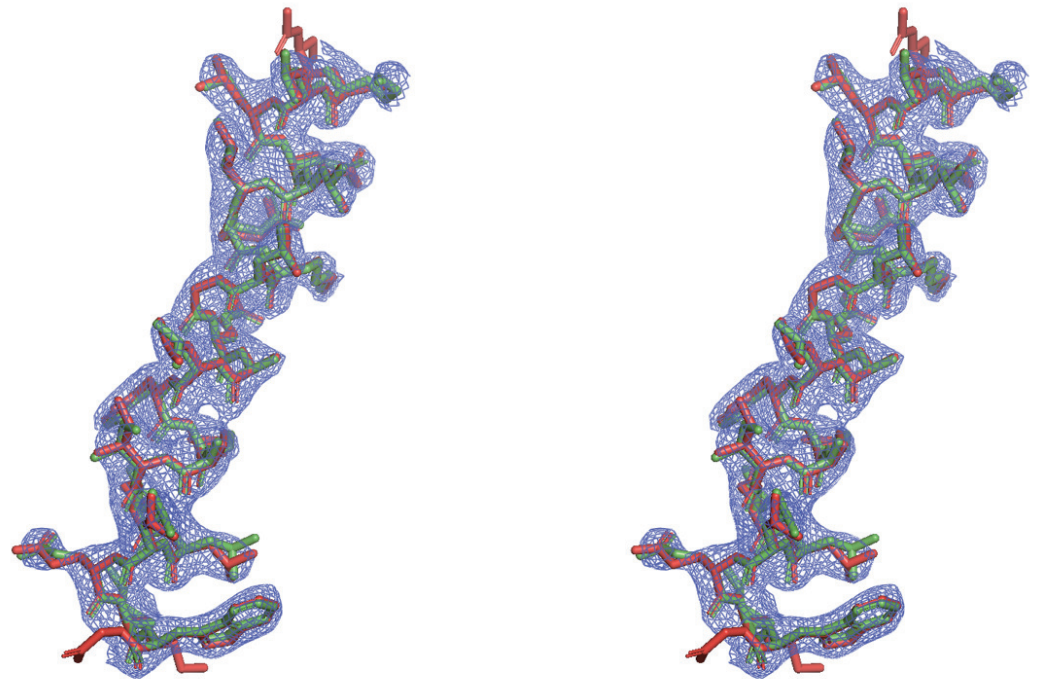
The sudden drop of R values indicates a successful run. For all successful runs,  $R_{free}$  suddenly drops from 0.6 to 0.38. For unsuccessful runs,  $R_{free}$  remains at 0.6. After turning off HIO and applying solvent flattening,  $R_{free}$  for successful runs further decreases to 0.23, while for failed runs  $R_{free}$  stays at 0.58. For a successful run, the mean phase error is about 39 degrees and the calculated electron density is interpretable. An atomic model is automatically rebuilt by *ARP/wARP*. The reconstructed model agrees with the PDB deposited model (Figure 5).

### 3.2. Direct Phasing of Protein Crystals with Non-Crystallographic Symmetry

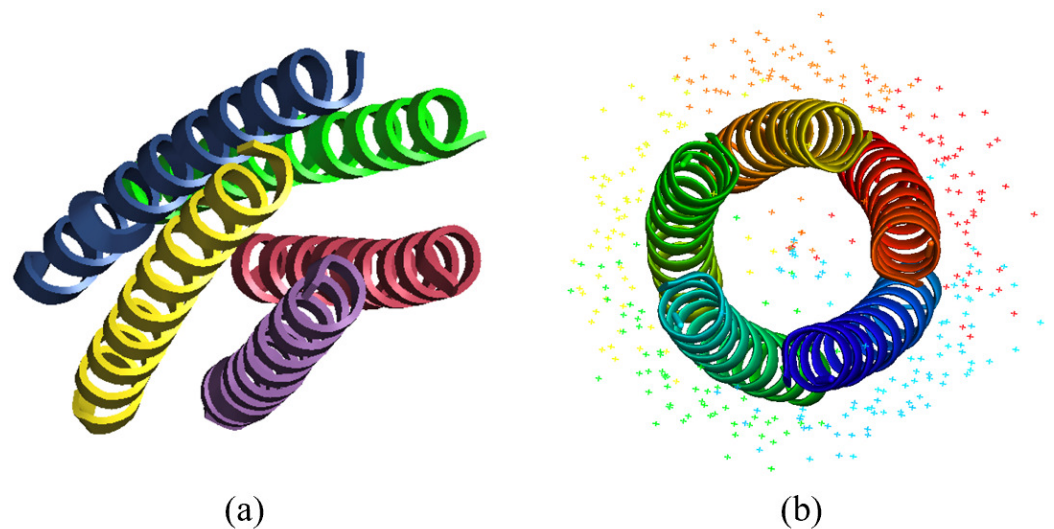
The existence of NCS increases the redundancy of the phase problem. It makes protein crystals with small solvent content solvable with direct phasing methods. Take a parametrically designed helical bundle as an example (Figure 6a). The PDB code is 4uot [36]. The protein was designed by Baker's group and indicated high thermodynamic stability. It has a five-fold rotational symmetry (Figure 6b). The cell parameters are  $a = 55.400 \text{ \AA}$ ,  $b = 88.020 \text{ \AA}$ , and  $c = 103.710 \text{ \AA}$ . There are 1484 non-hydrogen atoms in the asymmetric unit. The crystal diffracts to  $1.69 \text{ \AA}$ . The number of reflections observed in the experiment is 28,784. The Matthews coefficient is 2.48 corresponding to a solvent content of 50%. The solvent content was set to 50% in phase-retrieval iterations. In the trial calculation, the reference histogram is computed from another known structure 6g6e with a resolution of



1.69 Å. Other phasing parameters are the same as 6c6y. More information about the protein crystal is showed in Table 1.



**Figure 5.** Stereograms of the calculated density map (blue mesh) of an  $\alpha$ -helix of 6c4y, reconstructed model (green sticks) by ARP/wAPR [23] and PDB posted model (red sticks) [24].



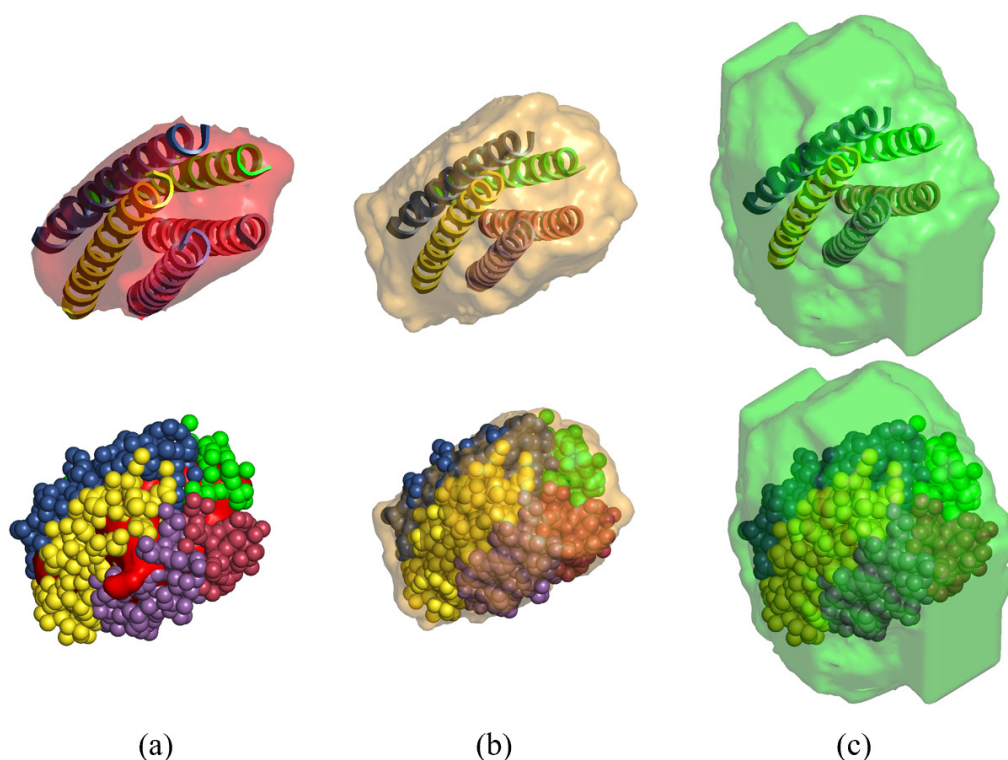
**Figure 6.** The protein molecule of 4uot has a five-fold rotational symmetry. Five copies have the same density related by a rotation symmetry. (a) A side view of the PDB posted model is shown as cartoon. (b) The rotation axis is perpendicular to the paper. The fixed water molecules on the protein surface are displayed. Starting from random phases, the reconstructed NCS mask evolves and finally covers the whole protein including most of the fixed water molecules.

It is crucial to reconstruct an NCS mask covering all equivalent copies. 4uot has a five-fold rotational NCS. The approximate direction of the NCS axis is determined from the self-rotation Patterson map. The position of the rotation axis in the asymmetric unit is assumed to be given. How to find the position of the NCS axis without prior information of the structure will be discussed in Section 4. At the early iterations, the NCS mask is reconstructed frequently. Since it is time-consuming to grow the NCS mask, as the

iteration proceeds, the reconstruction frequency slows down. In each independent run, NCS mask is updated once every iteration in the first three iterations, and updated once every 10 iterations in the first thirty iterations, and updated once every 500 iterations later. Reconstructing the NCS mask includes three steps: searching for the NCS center on the NCS axis, grow an NCS core around the NCS center, and grow a complete NCS mask from the surface of the NCS core. Three weighted average density maps are computed as described in Section 2.2. The construction seems to be complicated but it is capable of generating an NCS mask. No other approaches work when starting from random phases [37]. Figure 7 depicts the reconstructed NCS core, NCS mask covering a complete molecule, and asymmetric unit of a successful run. Although the NCS mask is not updated every iteration, NCS density averaging is performed every iteration. More details could be found in our previous work [21].

**Table 1.** Parameters of known protein structures.

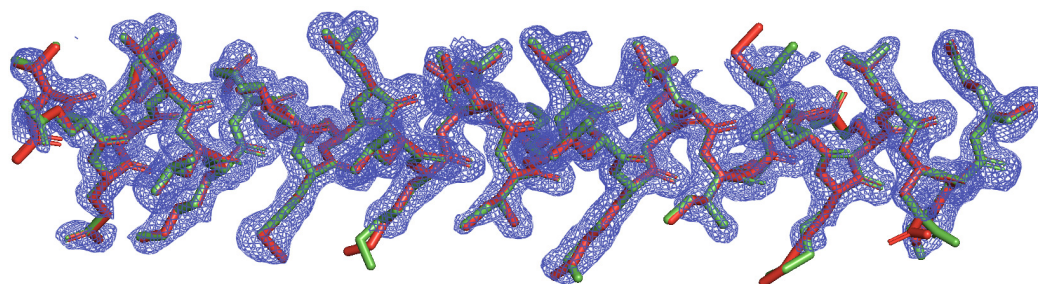
PDB Code	Space Group	Copied in ASU	Resolution (Å)	Solvent Content (%)	Weight (kDa)	Original Method	Success Rate	Speed of Convergence	Final $\Delta\varphi$ (°)
6c4y	P4322	18	2.5	72.31	52.53	MR	17	308	39.36
1uii	P212121	2	2.0	65.30	19.53	MAD	62	260	40.29
6g6b	P4132	3	2.3	74.99	9.75	MR	35	397	39.9
1no4	P21212	4	2.2	77.7	44.66	MAD	4	3341	42.48
6c4z	P6122	18	3.3	60.38	52.53	MR	30	105	58.42
4uot	C2221	5	1.69	63.9	20.68	MR	6	152	29.8
6eik	P42212	7	1.52	44.89	23.97	MR	1	346	42
5ez8	P22121	7	1.95	46	22.59	MR	3	114	42.7



**Figure 7.** In order to apply NCS density averaging, an NCS mask covering the whole protein must be reconstructed from scratch. Figure above demonstrates a reconstructed (a) NCS core, (b) NCS mask, and (c) asymmetric unit of a successful run starting from random phases. The PDB posted structure is superimposed here as a reference.

Direct phasing with non-crystallographic symmetry averaging easily solves the structure of 4uot. Starting from random phases, 100 independent runs are performed. 6 runs out of 100 successfully converge to the correct density. Each run includes 10,000 iterations. After 500 iterations,  $R_{free}$  value drops suddenly from 0.6 to 0.3. The mean phase error

suddenly decreases from 90 to 50 degrees implying a solution has been obtained. After turning off HIO, the final mean phase error further drops to 30 degrees. Since the crystal diffracts to 1.69 Å, the final density map is quite interpretable and ready for automatic model building. The deviation is small between the reconstructed model and the PDB deposited model (Figure 8).



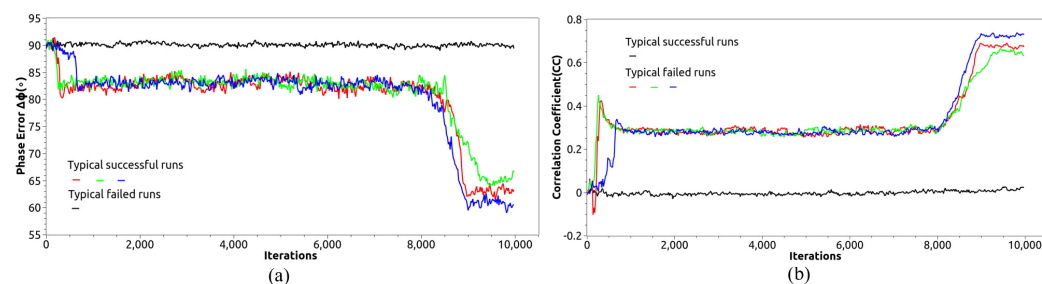
**Figure 8.** The calculated density map (blue mesh) of an  $\alpha$ -helix of 4uot, reconstructed model (green sticks) by ARP/*w*APR and PDB posted model (red sticks).

### 3.3. Direct Phasing of Intermediate- and Low-Resolution Diffraction Data

Macromolecule crystals often diffract to intermediate and low resolutions, such as membrane proteins which often contains coiled-coil domains. Due to the missing high-resolution data, density variance in protein region decreases and density is not flat in solvent region. Limit-density HIO described in Section 2.3 was used to modify the calculated density. Another cross-alpha amyloid-like structure 6c4z [24] is taken as an example to demonstrate the direct phasing results. The crystal diffracts to 3.3 Å with cell parameters  $a = b = 172.279$  Å,  $c = 153.758$  Å. The space group is  $P6_122$ . The solvent content is about 80%. There are 18  $\alpha$ -helices in the asymmetric unit containing 3562 non-hydrogen atoms. The number of reflections observed in experiment is 20,836.

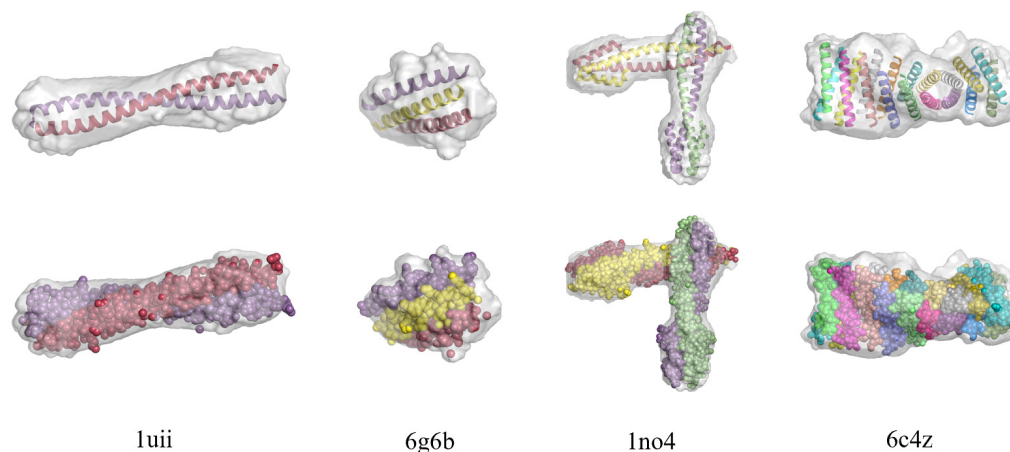
Direct phasing parameters were configured as follows. HIO modified density in the solvent region was limited to within a range of  $-1.0 \sim 1.0 \text{ e}^- / \text{Å}^3$  in accordance with the fact that density variance becomes smaller for low-resolution diffraction data. At 8000th iteration, HIO was turned off by decreasing the limit of density to  $-0.1 \sim 0.1 \text{ e}^- / \text{Å}^3$ . For the final 1000 iterations, instead of solvent flattening, a limit-density HIO was performed. Since the crystal diffracts to low resolution, the solvent density is not quite flat especially in the regions near the protein boundary. Using traditional solvent flattening to force the solvent density to zero, would visibly increase the R values. Additionally, a smaller grid size helps phasing low-resolution diffraction data. Generally, the size of the grid is set to half of the cutoff resolution. In practice, we found a dense grid helps to improve the calculated density for low-resolution diffraction data, although computing density on a dense grid is time-consuming. In this case, the size of the grid was set to 1.0 Å. Moreover, the solvent content was set to 73% implying a loose protein mask was used in phase retrieval. The histogram of 6c4y at 3.3 Å was used as a reference histogram.

Starting from random phases, 11 out of 100 independent runs converged to the correct density. Each run consists of 10,000 iterations. Figure 9 describes the evolution of the error metrics  $\Delta\phi$  and CC for several typical successful runs and one typical failed run. Successful run has lower  $\Delta\phi$  and higher CC. Just like  $\Delta\phi$ , R values also experience a sudden drop when a solution is reached.



**Figure 9.** Ab initio phasing of 6c4z. Evolution of error metrics starting from random phase: (a)  $\Delta\phi$ , (b) CC. When a solution is reached,  $\Delta\phi$  indicates a sudden drop, and CC shows a sudden rise. After 9000th iteration, solvent flattening is applied.

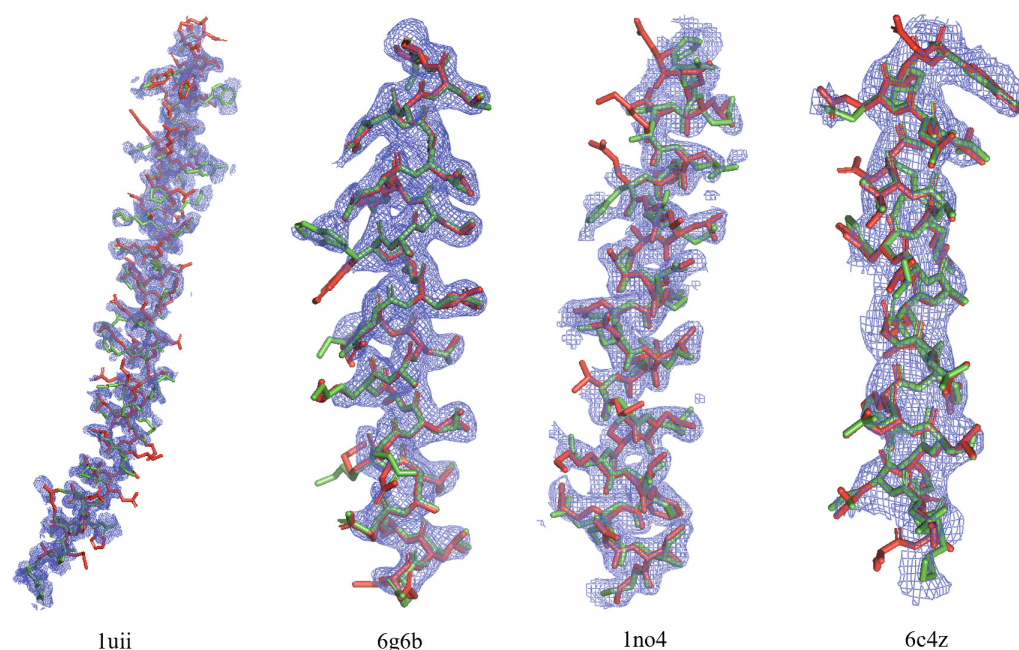
In all successful runs,  $\Delta\phi$  suddenly dropped from 90 to 82° (Figure 9a). At the end of all iterations, due to turning off HIO which flattens the solvent density,  $\Delta\phi$  further dropped from 82 to 60°. In unsuccessful runs  $\Delta\phi$  remained at 90°. For a successful run, the calculated electron density was used for model building by *ARP/wARP*. The comparison of the resulting model with the PDB posted model is shown in Figure 10.



**Figure 10.** The reconstructed molecular envelopes of 1uii, 6g6b, 1no4 and 6c4z. The PDB posted model is shown as cartoon in the first row and as sphere in the second row.

We have also tested our direct phasing method on five other coiled-coil proteins: 1uii, 6g6b, 1no4, 6eik and 5ez8. 1uii is a fragment of cellular protein Geminin [38], originally solved by MAD, consisting of two helices. 6g6b is part of the  $\alpha$ -helical barrel designed by human [39], originally solved by MR, consisting of three helices. 1no4 is a scaffold protein that promotes capsid assembly [40], originally solved by MAD, consisting of four helices. Since the solvent content of those three crystals are greater than 60%, they are easily solved by direct phasing method. The final mean phase error are around 40 degrees (Table 1). Moreover, it takes only around 400 iteration cycles or less than 10 min on a Dell R740 server to retrieve the phases of 6g6b by direct method, indicating that the phasing speed of direct method is much faster than MR. Compared with molecular replacement, another benefit is that the retrieved phases of direct method do not have any model bias, which is quite important during model reconstruction. 6eik [41] and 5ez8 [42] contain seven helices related by a seven-fold rotation symmetry. Since the solvent content of these two crystals are about 40%, their structures couldn't be reconstructed without NCS. Therefore, NCS density averaging was applied in direct phasing. The final mean phase error was around 40 degrees (Table 1). The true structure and mask of some of the proteins are shown in Figure 10. There are 100 independent runs for each structure. The results of the reconstruction are shown in Figure 11.





**Figure 11.** The calculated density maps (blue mesh) of one  $\alpha$ -helix, reconstructed models (green sticks) and the PDB posted models (red sticks) of 1uui, 6g6b, 1no4 and 6c4z.

#### 4. Discussion

In order to achieve an interpretable electron density maps, one has to be cautious when applying symmetry operations of the space group during density modification. In real space, the density of a unit cell is defined on a grid. The density value is placed at the center of each grid point. Otherwise, after applying symmetry operations on an asymmetric unit, the resulted density will not fill a complete unit cell. If applying symmetry operations on unique reflections in reciprocal space, one needs to make sure the resulted reflections fill a complete reciprocal space. In addition, missing reflections including  $F_{000}$  are updated according to Equation (8). A space groups often has several origin choices. The final calculated density map could fall into any of the allowable origin choices. If one wants to compute the average of several density maps, a proper origin translation is often required.

The results of trial calculations on eight coiled-coil structures have clearly proved that direct method is a powerful phasing tool. It could be seen from Table 1 that except for 6c4z, the phase errors of other coiled-coil proteins were reduced to 30~40°. Due to the low resolution of the diffraction data, the mean phase error of 6c4z is around 58°. The obtained density map was interpretable and could be directly used for model building via tools such as *ARP/wARP* [23]. After model building, more than 60% of amino acids were correctly placed. Although the reconstructed model has little deviation from the PDB posted structure, a model refinement process is still necessary.

Theoretically, the uniqueness of the solution requires the solvent content to be greater than 50% [10,19,25]. In practice, due to missing reflections and measurement errors, a little bigger solvent content is preferred such as 65%. The results in Table 1 indicate that direct method works well on phasing crystals with a solvent content greater than 60% without using NCS density averaging. Since the NCS related density are the same, the unique density of the protein region becomes much smaller which increases the redundancy of the unique solution. By applying NCS density averaging, direct method phases crystals with a solvent content much lower than 50% as shown in Section 2.3 and in Table 1.

Compared with AlphaFold and Molecular Replacement, direct method is unique and outstanding. AlphaFold is excellent at predicting single domain proteins and multimers from the sequence when good pairwise sequence alignment(PSA) information is available. It accurately predicts a single helix of a coiled coil if the helix doesn't have flexible region.



However, it is still difficult for AlphaFold to predict a structure consisting of tens of short chains such as 6c4y and 6c4z. The predicted models deviate too much from the true structures. Without a good predicted or homology model as reference, molecular replacement can't be carried out. One of our previous papers has also demonstrated that starting from a reference model direct method retrieves the lost phases within 10 minutes which is much faster than MR [30]. The density reconstructed by direct method doesn't contain model bias which is unavoidable in MR.

When applying NCS density averaging, we have considered only proper NCS, such as rotation symmetry. Our previous paper [21] has shown that non-proper NCS density averaging could also be applied in direct method if the translation and rotation operations are available. Starting from random phases, the translation operation is not easy to determine from the diffraction data. The direction of the rotation axis is retrieved from the Patterson self-rotation map. In order to apply NCS density averaging, we need to correctly position the NCS axis in the unit cell. The position can be located by a try-and-error method. Since an approximate center of mass of the NCS region is retrieved during the phasing iterations, the axis is placed on the approximate center of mass of the NCS region. Although the position has more or less deviation, it is refined by maximizing the local symmetry related density. This has been proved in our previous work [21].

Real diffraction data were used in our trial calculations. The eight structures tested in this paper have been solved by traditional phasing methods and posted in the protein data bank. There are more diffraction data which have not been phased due to low resolution diffraction, big measurement errors, or lack of a good predicted or homology model as reference. The source code is written from scratch and is available on github [https://github.com/Ruijiang-Fu/direct\\_phasing\\_of\\_coiled\\_coil\\_protein\\_crystals.git](https://github.com/Ruijiang-Fu/direct_phasing_of_coiled_coil_protein_crystals.git) (accessed 7 November 2022).

## 5. Conclusions

This paper proposes a direct phasing approach to solve the crystal structures of coiled-coil proteins. The results of trial calculations on eight coiled-coil proteins prove that direct method is a good alternative phasing tool when AlphaFold and molecular replacement fail. Direct phasing does not depend on any reference model. Starting from random numbers, it retrieves the lost phases directly from the diffraction data. Additionally, direct method does not require high resolution diffraction data. It works well on intermediate and low-resolution data. Moreover, it is not sensitive to missing reflections. Those missing reflections will be simultaneously reconstructed during the phasing process. Although the uniqueness of the solution requires a high solvent content for direct phasing, direct method works well on phasing crystals with small solvent content with NCS density averaging. Therefore, direct method provides a fast and efficient phasing approach for protein crystallography. It takes only a few hours, sometimes less than 10 minutes, to retrieve the lost phases on a Dell R740 server.

**Author Contributions:** H.H. and W.-P.S. conceived the concepts; R.F. and H.H. carried out the calculations; R.F., H.H. and W.-P.S. wrote the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The diffraction data were downloaded from the Protein Data Bank at <https://www.rcsb.org> (accessed on 7 November 2022).

**Acknowledgments:** We thank Shaoqing Zhang for providing the information about coiled-coil structures and for the suggestion to solve those structures with direct phasing method. This work was supported by the initial grants of Ningbo University.

**Conflicts of Interest:** The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

## References

1. Caballero, I.; Sammito, M.; Millán, C.; Lebedev, A.; Soler, N.; Usón, I. ARCIMBOLDO on coiled coils. *Acta Cryst. D* **2008**, *74*, 194–204. [[CrossRef](#)] [[PubMed](#)]
2. Kendrew, J.C.; Bodo, G.; Dintzis, H.M.; Parrish, G.R.; Wyckoff, H.; Phillips, D.C. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* **1958**, *181*, 662–666. [[CrossRef](#)] [[PubMed](#)]
3. Hendrickson, W.A.; Smith, J.L.; Phizackerley, R.P.; Merritt, E.A. Crystallographic structure analysis of lamprey hemoglobin from anomalous dispersion of synchrotron radiation. *Proteins Struct. Funct. Bioinform.* **1988**, *4*, 77–88. [[CrossRef](#)]
4. Rossmann, M.G.; Blow, D.M. The detection of sub-units within the crystallographic asymmetric unit. *Acta Cryst.* **1962**, *15*, 24–31. [[CrossRef](#)]
5. Gillingham, A.K.; Munro, S. Long coiled-coil proteins and membrane traffic. *Biochim. Biophys. Acta (BBA)-Mol. Cell Res.* **2003**, *1641*, 71–85. [[CrossRef](#)]
6. Sayre, D. The squaring method: A new method for phase determination. *Acta Cryst.* **1952**, *5*, 60–60. [[CrossRef](#)]
7. Hauptman, H. A minimal principle in X-ray crystallography: Starting in a small way. *Proc. R. Soc. Lond. Ser. A Math. Phys. Sci.* **1993**, *442*, 3–12. [[CrossRef](#)]
8. Zhang, K.Y.J.; Main, P. The use of Sayre's equation with solvent flattening and histogram matching for phase extension and refinement of protein structures. *Acta Cryst. A* **1990**, *46*, 377–381. [[CrossRef](#)]
9. Millane, R.P. Phase retrieval in crystallography and optics. *J. Opt. Soc. Am.* **1990**, *7*, 394–411. [[CrossRef](#)]
10. Elser, V. Phase retrieval by iterated projections. *JOSA A* **2003**, *20*, 40–55. [[CrossRef](#)]
11. Liu, Z.C.; Xu, R.; Dong, Y.H. Phase retrieval in protein crystallography. *Acta Cryst. A* **2012**, *68*, 256–265. [[CrossRef](#)] [[PubMed](#)]
12. He, H.; Su, W.-P. Direct phasing of protein crystals with high solvent content. *Acta Cryst. A* **2015**, *71*, 92–98. [[CrossRef](#)] [[PubMed](#)]
13. Fienup, J.R. Reconstruction of an object from the modulus of its Fourier transform. *Opt. Lett.* **1978**, *3*, 27–29. [[CrossRef](#)] [[PubMed](#)]
14. Fienup, J.R. Phase retrieval algorithms: A comparison. *Appl. Opt.* **1982**, *21*, 2758–2769. [[CrossRef](#)]
15. Fienup, J.R. Phase retrieval algorithms: A personal tour. *Appl. Opt.* **2013**, *21*, 45–56. [[CrossRef](#)]
16. Millane, R.P.; Stroud, W.J. Reconstructing symmetric images from their undersampled Fourier intensities. *J. Opt. Soc. Am. A* **1997**, *14*, 568–579. [[CrossRef](#)]
17. Elser, V. Solution of the crystallographic phase problem by iterated projections. *Acta Cryst. A* **2003**, *59*, 201–209. [[CrossRef](#)]
18. Marchesini, S.; He, H.; Chapman, H.N.; Hau-Riege, S.P.; Noy, A.; Howells, M.R.; Weierstall, U.; Spence, J.C.H. X-ray image reconstruction from a diffraction pattern alone. *Phys. Rev. B* **2003**, *68*, 140101. [[CrossRef](#)]
19. Millane, R.P.; Lo, V.L. Iterative projection algorithms in protein crystallography. I. Theory. *Acta Cryst. A* **2013**, *69*, 517–527. [[CrossRef](#)]
20. Lo, V.L.; Kingston, R.L.; Millane, R.P. Iterative projection algorithms in protein crystallography. II. Application. *Acta Cryst. A* **2015**, *71*, 451–459. [[CrossRef](#)]
21. He, H.; Jiang, M.; Su, W.P. Direct Phasing of Protein Crystals with Non-Crystallographic Symmetry. *Crystals* **2019**, *9*, 55. [[CrossRef](#)]
22. Kingston, R.L.; Millane, R.P. A general method for directly phasing diffraction data from high-solvent-content protein crystals. *IUCrJ* **2022**, *9*, 648–665. [[CrossRef](#)] [[PubMed](#)]
23. Langer, G.; Cohen, S.X.; Lamzin, V.S.; Perrakis, A. Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7. *Nat. Protoc.* **2008**, *3*, 870–875. [[CrossRef](#)] [[PubMed](#)]
24. Zhang, S.Q.; Huang, H.; Yang, J.; Kratochvil, H.T.; Lolicato, M.; Liu, Y.; Shu, X.; Liu, L.; DeGrado, W.F. Designed peptides that assemble into cross- $\alpha$  amyloid-like structures. *Nat. Chem. Biol.* **2018**, *14*, 1171–1179. [[CrossRef](#)]
25. Miao, J.; Sayer, D.; Chapman, H.N. Phase retrieval from the magnitude of the Fourier transforms of non-periodic objects. *J. Opt. Soc. Am.* **1998**, *15*, 1662–1669. [[CrossRef](#)]
26. Fourier Transform Functions. *Intel R Math Kernel Library 11.3 Reference Manual*; Intel Corporation: Santa Clara, CA, USA, 2015; pp.1911–1962.
27. Zhang, K.Y.J.; Main, P. Histogram matching as a new density modification technique for phase refinement and extension of protein molecules. *Acta Cryst. A* **1990**, *46*, 41–46. [[CrossRef](#)]
28. Wang, B.C. Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **1985**, *115*, 90–112. [[CrossRef](#)]
29. Brünger, A.T. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **1992**, *355*, 472–475. [[CrossRef](#)]
30. He, H.; Su, W.-P. Improving the convergence rate of a hybrid input-output phasing algorithm by varying the reflection data weight. *Acta Cryst. A* **2018**, *74*, 36–43. [[CrossRef](#)]
31. Miao, J.; Sayre, D. On possible extensions of X-ray crystallography through diffraction-pattern oversampling. *Acta Cryst. A* **2000**, *56*, 596–605. [[CrossRef](#)]
32. Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Hassabis, D. Highly accurate protein structure prediction with AlphaFold. *Nature* **2021**, *596*, 583–589. [[CrossRef](#)] [[PubMed](#)]

33. Baek, M.; DiMaio, F.; Anishchenko, I.; Dauparas, J.; Ovchinnikov, S.; Lee, G.R.; Baker, D. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **2021**, *373*, 871–876. [[CrossRef](#)] [[PubMed](#)]
34. Koepke, J.; Krammer, E.M.; Klingen, A.R.; Sebban, P.; Ullmann, G.M.; Fritzsche, G. pH modulates the quinone position in the photosynthetic reaction center from *Rhodobacter sphaeroides* in the neutral and charge separated states. *J. Mole. Biol.* **2007**, *371*, 396–409. [[CrossRef](#)] [[PubMed](#)]
35. Jiang, M.; He, H.; Cheng, Y.; Su, W.P. Resolution dependence of an ab initio phasing method in protein X-ray crystallography. *Crystals* **2018**, *8*, 156. [[CrossRef](#)]
36. Huang, P.S.; Oberdorfer, G.; Xu, C.; Pei, X.Y.; Nannenga, B.L.; Rogers, J.M.; Baker, D. High thermodynamic stability of parametrically designed helical bundles. *Science* **2014**, *346*, 481–485. [[CrossRef](#)]
37. Terwilliger, T.C. Finding non-crystallographic symmetry in density maps of macromolecular structures. *J. Struct. Funct. Genom.* **2013**, *14*, 91–95. [[CrossRef](#)]
38. Saxena, S.; Yuan, P.; Dhar, S.K.; Senga, T.; Takeda, D.; Robinson, H.; Kornbluth, S.; Swaminathan, K.; Dutta, A. A dimerized coiled-coil domain and an adjoining part of geminin interact with two sites on Cdt1 for replication inhibition. *Mol. Cell* **2004**, *15*, 245–245. [[CrossRef](#)]
39. Rhys, G.G.; Wood, C.W.; Lang, E.J.; Mulholland, A.J.; Brady, R.L.; Thomson, A.R.; Woolfson, D.N. Maintaining and breaking symmetry in homomeric coiled-coil assemblies. *Nat. Comm.* **2004**, *9*, 1–12. [[CrossRef](#)]
40. Morais, M.C.; Kanamaru, S.; Badasso, M.O.; Koti, J.S.; Owen, B.A.; McMurray, C.T.; Anderson, D.L.; Rossmann, M.G. Bacteriophage */phi29* scaffolding protein gp7 before and after prohead assembly. *Nat. Struct. Mol. Biol.* **2003**, *10*, 572–576. [[CrossRef](#)]
41. Thomas, F.; Dawson, W.M.; Lang, E.J.; Burton, A.J.; Bartlett, G.J.; Rhys, G.G.; Woolfson, D.N. De novo-designed  $\alpha$ -helical barrels as receptors for small molecules. *ACS Syn. Biol.* **2018**, *7*, 1808–1816. [[CrossRef](#)]
42. Burton, A.J.; Thomson, A.R.; Dawson, W.M.; Brady, R.L.; Woolfson, D.N. Installing hydrolytic activity into a completely de novo protein framework *Nat. Chem.* **2016**, *8*, 837–844. [[CrossRef](#)]