crystals

MDPI

Article

Resolution Dependence of an *Ab Initio* Phasing Method in Protein X-ray Crystallography

Mengchao Jiang, Hongxing He, Yunpeng Cheng and Wu-Pei Su *

Department of Physics and Texas Center for Superconductivity, University of Houston, Houston, TX 77204, USA; mch.jiang.nju@gmail.com (M.J.); hellohehongxing@gmail.com (H.H.); ycheng12@uh.edu (Y.C.)

* Correspondence: wpsu@uh.edu

Received: 17 February 2018; Accepted: 31 March 2018; Published: 3 April 2018

Abstract: For direct phasing of protein crystals, a method based on the hybrid-input-output (HIO) algorithm has been proposed and tested on a variety of structures. So far, however, the diffraction data have been limited to high-resolution ones, i.e., higher than 2 Å. In principle, the methodology can be applied to data of lower resolutions, which might be particularly useful for phasing membrane protein crystals. For resolutions higher than 3.5 Å, it seems the atomic structure is solvable. For data of lower resolutions, information of the secondary structures and the protein boundary can still be obtained. Examples are given to support the conclusions. Real experimental data are used. Two aspects of the observed data have been discussed: removal of the measured low-resolution reflections and involvement of the unmeasured high-resolution reflections. The *ab initio* phasing employs histogram matching for density modification. A question arises whether the reference histogram used should match the resolution of the diffraction data or not. It seems that there is an optimal histogram which is good to use for data at various resolutions.

Keywords: resolution dependence; hybrid input-output; iterative projection algorithm; *ab initio* phasing; membrane protein; X-ray crystallography

1. Introduction

Despite the existence of good physical methods for phasing the X-ray reflections of a protein crystal, direct phasing remains a challenging theoretical problem. Iterative projection algorithms have been widely used for phase retrieval [1–17].

For crystals with high solvent content or with adequate non-crystallographic symmetry (NCS), it has been demonstrated that direct phasing is possible [10,12,13]. Completely *ab initio* phasing using an iterative algorithm has been reported [13]. A crucial component of the algorithm is the hybrid-input-output (HIO) scheme employed in enforcing the solvent constraint [1–3]. Instead of requiring the solvent density to be strictly constant, the HIO method uses a negative feedback mechanism to gradually modify the solvent density so that it tends to become more constant.

It has been argued that for HIO to work properly, oversampling is required [5]. For a given number of protein atoms, availability of higher resolution data would therefore seem more favorable. It turns out that the oversampling condition is independent of the resolution and depends only on the structural redundancy [5,11,13], i.e., the low- and medium-resolution data are sufficient for the determination of the corresponding phases as long as there is enough redundancy (high solvent content or NCS).

Ab initio phasing at very low resolution has been reported with a generalized likelihood based approach [18,19]. In this paper, we describe a series of trial calculations using the HIO algorithm involving data at various resolutions (from 2.85 Å to 7 Å). At each resolution, the HIO method is

capable of yielding useful structural information. As expected, only 3.5 Å or higher resolution data can lead to atomic modeling. In addition, at lower resolution partial information such as the secondary structures or the protein boundary can still be obtained.

Membrane protein crystals usually have large solvent content and do not diffract to high resolution. It could be challenging to retrieve the phase using conventional methods [20]. The results of our trial calculations indicate a potential new phasing approach for membrane proteins.

When collecting the experimental data, some low-resolution reflections are missing due to the beam stop. Those measured low-resolution reflections with very small diffraction angles are usually not accurate and deviate a lot from their expected values. They should be replaced by the calculated values during the phasing process. At the same time, we find the HIO phasing method benefits from including some unmeasured high-resolution reflections [21–24]. This is because using the calculated values of those reflections makes the computed density in real space more smooth.

Another resolution dependence of the direct phasing method is histogram matching. A reference histogram is well-known to be very helpful for density modification inside the protein region [25,26]. It is also very much resolution dependent. It would seem only natural that the resolution of the reference histogram should match that of the diffraction data. However, trial calculations show that is not always the case.

It is well-known that the density histogram is universal, i.e., independent of detailed structures. A question naturally arises, namely, what exactly does the histogram encode? Is that the average density of the protein or something else? Although this question is not directly related to the resolution dependence, we find this to be an interesting question to look at.

2. *Ab Initio* Phasing at Various Resolutions

Ab initio phasing using the HIO method has been described in previous articles [13–15,27,28]. In this paper, we made 200 independent runs for each trial calculation and each run has 10,000 iterations. Each iteration consists of density modification in real space and using measured amplitudes in forward Fourier transform to improve phase values in reciprocal space. Before presenting the results of our trial calculations, let us briefly review the HIO phasing algorithm.

At the beginning of each iteration, data weighting [12,15] is used to speed up the convergence of the iterations and to increase the success rate. We give the diffraction data a weight defined in Equation (1) where σ_1 varies with the iteration number. $S_{\mathbf{h}}$ is the reciprocal of the resolution (wavelength) of that reflection. The weighted data defined in Equation (2) are used in the phasing process and updated at the beginning of each iteration. In the first iteration, σ_1 starts from a value of around 1.0 Å. The weight of high-resolution reflections are close to zero and only low- and medium-resolution reflections are involved in the phasing process. Due to a lower number of reflections involved, it helps the calculated protein boundary evolve to the correct shape which speeds up the phasing process and increases the success rate. Then σ_1 decreases smoothly in the following iterations which allows equal number of higher-resolution reflections be incorporated into the phasing procedure at each iteration until all reflections are involved at the 8000th iteration. Finally, σ_1 drops smoothly to zero from the 8000th to the 9000th iteration, so that observed reflections recover their original magnitudes. More details about how σ_1 varies have been described in our previous article [15].

$$w_1(S_{\mathbf{h}}) = e^{-2(\pi\sigma_1 S_{\mathbf{h}})^2} \tag{1}$$

$$|F_w^{obs}(\mathbf{h})| = w_1|F^{obs}(\mathbf{h})| \tag{2}$$

Missing reflections need to be filled with the calculated ones in each iteration. The beam stop used in the diffraction experiment often results in unmeasured reflections at very low resolution. The magnitudes of the missing reflections are replaced by the calculated values according to Equation (3). This replacement is required in order to obtain a good electron density map in real space. About 1% of the observed data were randomly chosen and set aside as a test data set, T, while the

remainder were used as a work data set, W [29]. The reflections in the test data set should also be treated as missing reflections and replaced by the calculated values according to Equation (3).

$$|F^{miss}(\mathbf{h})| = \frac{\sum_{\mathbf{h}\in W}|F_w^{obs}(\mathbf{h})|}{\sum_{\mathbf{h}\in W}|F^{cal}(\mathbf{h})|}|F^{cal}(\mathbf{h})| \tag{3}$$

The electron density in a unit cell is defined on a grid. The grid size is chosen to be half of the high resolution limit of the phasing data. For example, the grid size is 1.43 Å for phasing 2.85 Å data, and 3.5 Å for phasing 7 Å data. Apparently, a bigger grid size leads to less computing time. However, a proper grid size is necessary in order to make sure all reflections to a given resolution have been involved in the computation. *INTEL* forward and backward discrete fast Fourier transform [30] is used to compute the electron density on each grid point in real space and the structure factors in reciprocal space.

The first iteration starts from random electron density in real space. A backward fast Fourier transform is performed to get the calculated structure factors in reciprocal space. The calculated magnitude of each reflection is replaced by the weighted observed magnitude defined in Equation (2). The missing reflections are substituted with the calculated ones according to Equation (3). The new magnitudes and the calculated phases are assembled to form new structure factors. A forward fast Fourier transform is performed to get the calculated electron density $\rho^{(n)}$ in real space. The superscript n denotes the nth iteration.

In order to locate the protein boundary, a weighted average density on each grid point is calculated. Positive density constraint is applied during the calculation of the weighted average density, i.e., negative density is replaced by zero in the averaging. The density weighting function is defined in Equation (4) [7].

$$w_2(d_{ij}) = exp\left(-\frac{d_{ij}^2}{2\sigma_2^2}\right) \tag{4}$$

The subscript i or j represents a grid point in the asymmetric unit. d_{ij} is the distance between the two grid points. The parameter σ_2 measures the width of a Gaussian function which can be used to control the convergence of the solvent region. σ_2 is chosen to be 4.0 Å at the first iteration and it decreases linearly in the following iterations. At the 9000th iteration, σ_2 is reduced to 2.5 Å, and it keeps that value when solvent flattening is applied during the last 1000 iterations. In practice, the weighted average density is calculated in reciprocal space according to the convolution theorem. More information about the calculation of the weighted average density can be found in our previous articles [13,14].

A cutoff value of the weighted average density is used to divide the asymmetric unit into the protein region and solvent region. The cutoff value can be found by adjusting it such that the calculated solvent content agrees with the expected solvent fraction. Since the average density of the protein is greater than the average density of the solvent, if a grid point has a weighted average density greater than the cutoff value, it is assumed to be inside the protein region. Otherwise, it is assumed to be part of the solvent.

After the protein boundary is determined, different density-modification techniques are employed to modify the calculated density in the solvent region and in the protein region, separately. In the solvent region, hybrid input-output introduces a negative feedback density according to Equation (5) [1,2].

$$g^{(n)} = \begin{cases} \rho^{(n)} & \text{in the protein region;} \\ g^{(n-1)} - \varepsilon\rho^{(n)} & \text{in the solvent region.} \end{cases} \tag{5}$$

$g^{(n)}$ denotes the modified density of the nth iteration. $\rho^{(n)}$ is the density of the nth iteration before modification. ε is a feedback parameter which can be used to optimize the convergence of the algorithm. Empirically, ε is chosen to be 0.7. HIO does not change the calculated density of the protein

region. Instead, a standard histogram-matching method is applied to make the calculated density in the protein region satisfy the density distribution of a reference histogram. After the density modification in real space, a backward fast Fourier transform is performed to get the calculated structure factors for the next iteration.

Since the HIO-modified density does not satisfy the solvent constraint, solvent flattening [31–34] is applied in real space during the last 1000 iterations according to Equation (6).

$$g^{(n)} = \begin{cases} \rho^{(n)} & \text{in the protein region;} \\ 0 & \text{in the solvent region.} \end{cases} \tag{6}$$

Having reviewed the HIO phasing algorithm, we now proceed to describe the results of our trial calculations carried out for a membrane protein structure (PDB code 2JLN [35]). 2JLN is a nucleobase-cation-symport-1 benzylhydantoin transporter which is an essential component of salvage pathways for nucleobases and related metabolites. The space group is $P2_12_12_1$. The cell dimensions are $a = 79.70$, $b = 109.14$, and $c = 113.82$ Å. The sequence includes 501 amino acids. Only 464 amino acids have been identified in the refined model. There are 3571 non-hydrogen atoms in the asymmetric unit. The crystal diffracts to 2.85 Å, with a low resolution cutoff at 29 Å. The completeness of the measured data is 88%. The overall R value after model refinement is about 0.24.

A proper value of solvent content is important for HIO phasing. The solvent content listed in PDB is 69% for 2JLN. After checking the model with *sfcheck* [36] in *CCP4* [37], the volume not occupied by model is 68%. In our trial calculations, we have tested several values for the solvent content from 62% to 72%. Although all of them lead to successfully phasing, 68% turns out to be an optimal value with a high success rate and a low phase error.

For histogram matching, a protein structure (PDB code 4W6V [38] of similar size is selected for the computation of a reference density histogram. 4W6V is a peptide transporter which mediates the cellular uptake of di- and tripeptides, and of peptidomimetic drugs. There are about 500 amino acids in the structure. Reference histograms from other structures of similar size also work. Since the histogram of a protein structure depends highly on the average temperature factor of the atomic model, the average B-factor of the reference structure should be rescaled to match the Wilson B-factor computed directly from the measured data of 2JLN. The reference histogram is computed at 2 Å resolution.

Several quantities are defined in Equations (7)–(10) to monitor the progress of the iterative phasing procedure. R_{free} and R_{work} measure the discrepancy between the calculated magnitudes and the observed magnitudes of the reflections in the test data set and the work data set, respectively. Since the phases of an unknown structure are not available, R_{free} and R_{work} should be used to identify whether good phases have been achieved. $\Delta\phi$ is a measure of the difference between the calculated phases and the true phases which are computed from the PDB deposited model with bulk solvent correction. The correlation coefficient, CC, is a measure of both calculated magnitudes and phases. The value of CC is close to one when good phases are achieved. Otherwise, it stays around zero.

$$R_{free} = \frac{\sum_{\mathbf{h}\in T} ||F^{obs}(\mathbf{h})| - |F^{cal}(\mathbf{h})||}{\sum_{\mathbf{h}\in T} |F^{obs}(\mathbf{h})|} \tag{7}$$

$$R_{work} = \frac{\sum_{\mathbf{h}\in W} ||F^{obs}(\mathbf{h})| - |F^{cal}(\mathbf{h})||}{\sum_{\mathbf{h}\in W} |F^{obs}(\mathbf{h})|} \tag{8}$$

$$\Delta\phi = \frac{\sum_{\mathbf{h}\in W} arccos\left\{cos\left[\phi^{true}(\mathbf{h}) - \phi^{cal}(\mathbf{h})\right]\right\}}{\sum_{\mathbf{h}\in W} 1} \tag{9}$$

$$CC = \frac{\sum_{\mathbf{h}\in W} |F^{obs}(\mathbf{h})||F^{cal}(\mathbf{h})| cos\left[\phi^{true}(\mathbf{h}) - \phi^{cal}(\mathbf{h})\right]}{\left[\sum_{\mathbf{h}\in W} |F^{obs}(\mathbf{h})|^2 \sum_{\mathbf{h}\in W} |F^{cal}(\mathbf{h})|^2\right]^{1/2}} \tag{10}$$

Figure 1 shows the evolution of the error metrics defined in Equations (7)–(10). One typical failed run and six typical successful runs are presented. Initially, the phases are random and $\Delta\phi$ is about 90°. *CC* is close to zero. In a failed run, the phase error almost does not change significantly. The value of CC is close to zero. However, in a successful run, the phase error develops a sudden drop when good phases are achieved. The value of CC exhibits a sudden increase. For all runs, when the iterations proceed both R_{work} and R_{free} slowly decrease due to the progressively uniform data weighting. When solvent flattening is applied in the last 1000 iterations, the phase error further drops by several degrees and the CC value further increases. R_{free} also decreases but it can not discriminate between the failed and the successful runs due to the intermediate resolution of the measured data. However, R_{work} reached an obviously smaller value for those successful runs as shown in Figures 1d and 2a. Therefore, R_{work} is still a good indicator when the resolution of the measured data is intermediate.

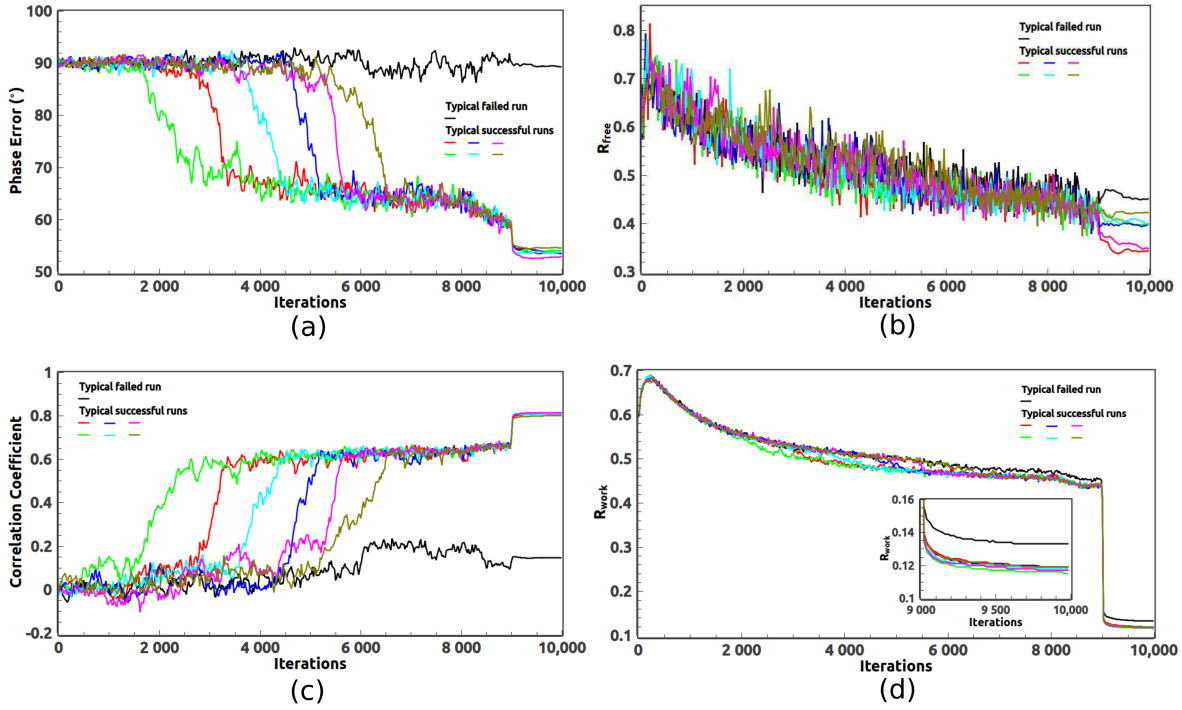

Figure 1. The evolution of (**a**) $\Delta\phi$, (**b**) R_{free}, (**c**) *CC*, and (**d**) R_{work} for the *ab initio* phasing of 2JLN using 2.85 Å observed data. For clarity, one typical failed run and six successful runs are presented.

In order to identify those successful runs, we sorted R_{work} of the 200 runs in ascending order. We checked those runs with low R_{work} (less than 0.125 in Figure 2a) and found that all of those runs are successful runs with small phase errors. We also found the lower the value of R_{work}, generally the smaller the phase error. On the other hand, those runs with high R_{work} correspond to failed runs with phase errors around 90°. There is a clear gap on the distribution of R_{work} in Figure 2a which separates the failed group from the successful group. R_{work} of the failed group are not exactly the same. Some failed runs still contain some correct information about the protein boundary.

The successful runs form a set of low R_{work} (Figure 2a) and they all correspond to a mean phase error around 54°. Since they started from different random phases, they approached the true phases from various directions. In other words, they are not identical due to statistical fluctuations. It is well known that averaging can reduce the fluctuations. As indicated in Figure 2b, the mean phase error can be significantly reduced by averaging over those successful runs starting from the one with the lowest R_{work} and proceeding in the ascending order of R_{work}. Meanwhile, the CC value apparently increases. Therefore, averaging over those successful runs is a powerful phase improvement tool for the iterative phasing method.

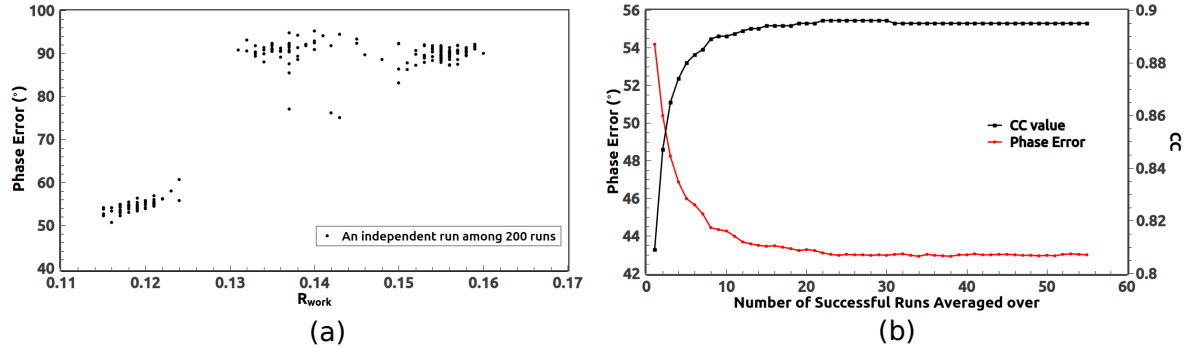

Figure 2. (**a**) A smaller value of R_{work} indicates a successful run. (**b**) After averaging over those successful runs, the phase error further decreases by about 10° and the CC value increases by about 0.1.

The calculated protein boundary of a typical successful run is shown in Figure 3a. Evolving to a good boundary is crucial for the HIO method. If the protein density was not protected by a good boundary, HIO would destroy the calculated protein density during density modification and good phases would not be reached. A good comparison of the averaged density with the PDB deposited model is shown in Figure 3b. Only one major helix of 2JLN is shown for clarity. The atomic model is well traced on the contour map of the averaged density. A model reconstructed from the averaged density using *ARP/wARP* [39] and *AutoBuild* [40] in *PHENIX* [41] is shown in Figure 3c. The reconstructed model is quite close to the deposited model. About 81% of the 501 amino acids are positioned in the model. Further refinement is possible.

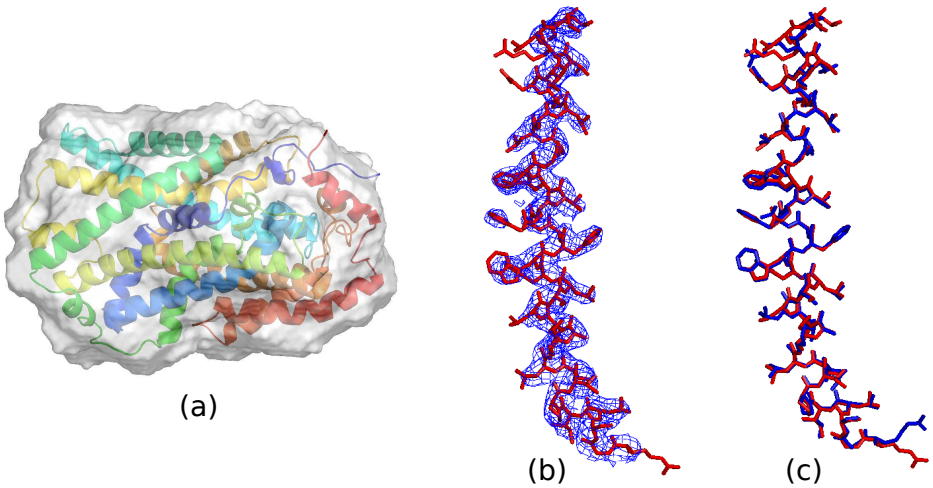

Figure 3. (**a**) The calculated protein boundary, (**b**) density map, and (**c**) reconstructed model (shown as blue sticks) of 2JLN using 2.85 Å observed data. The PDB deposited model is superimposed and shown as cartoons in (**a**) and as red sticks in (**b**,**c**). Only one major helix is shown in (**b**,**c**) for clarity.

As the high-resolution data are not always available, *ab initio* phasing of the medium- and low-resolution data is quite useful. In addition to the 2.85 Å data, we have also tried HIO phasing on the 3, 3.5, 4, 5, 6, and 7 Å data. For example, when phasing the 3 Å data, we pretend the measured data are limited to 3 Å. The evolution of the phase errors are displayed in Figure 4. When phasing low-resolution data such as the 7 Å data, the expected density in the solvent region deviates a lot from a constant which weakens the power of the HIO method. As a result, a very low success rate is expected. A sudden drop of the phase error is no longer expected for those successful runs. If the resolution of the data goes much lower, such as 8 Å, it becomes difficult for the HIO method to achieve good phases for 2JLN.

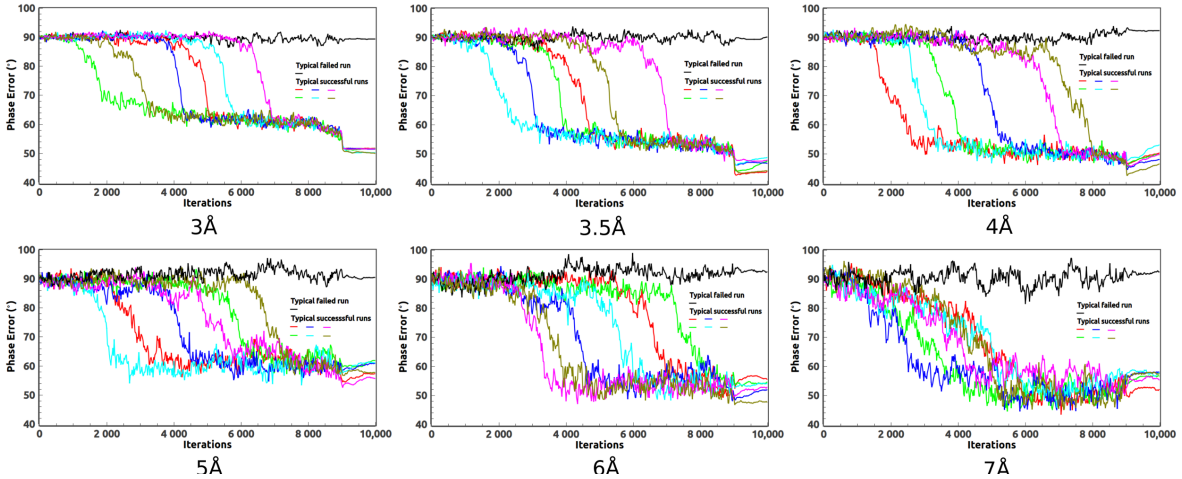

Figure 4. The evolution of the phase errors of 2JLN using the 3, 3.5, 4, 5, 6, and 7 Å observed data. For clarity, one failed run and six successful runs are presented. The total number of successful runs are listed in Table 1.

Like the 2.85 Å data, phase averaging over those successful runs can significantly improve the calculated phases of the medium- and low-resolution data as shown in Figure 5. For the 3, 3.5, 4, and 5 Å data, there are around 50 successful runs among 200. After averaging over about 10 successful runs, the phase error becomes flat. It implies that less than 200 runs are needed in order to get the best phases. For the 6 and 7 Å data, since the success rate is low, more than 200 runs should be carried out. In fact, when averaging over more successful runs, better phases are obtained for the 6 and 7 Å data. On the other hand, averaging over too many runs may not be beneficial. In Figure 5, the phase error of the 5 Å data slightly increases after averaging over too many runs.

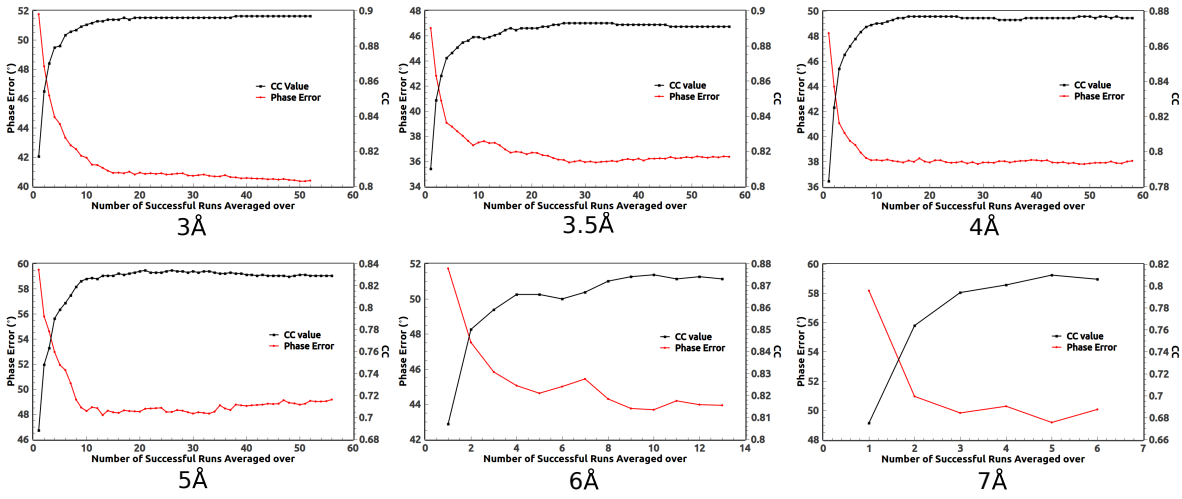

Figure 5. After averaging over those successful runs presented in Figure 4, the phase errors further decrease by about 10° and the CC values further increase by about 0.1.

The calculated protein boundaries of typical successful runs for data of various resolutions are displayed in Figure 6. For the 3, 3.5, 4, and 5 Å data, the reconstructed boundaries are smooth and well match the protein region. For the 6 and 7 Å data, small parts of some side chains get located outside of the calculated boundaries. The surfaces of the boundaries are rough due to the large grid size.

The averaged densities of those successful runs for data of various resolutions are shown in Figure 7. The PDB deposited model is superimposed as a reference. Side chains can be traced on the density maps of the 3 and 3.5 Å data. For the 4 Å data, side chains are not very interpretable but the

secondary structures are clearly visible. For the 5 and 6 Å data, secondary structures can be traced. For the 7 Å data, only partial secondary structures can be traced.

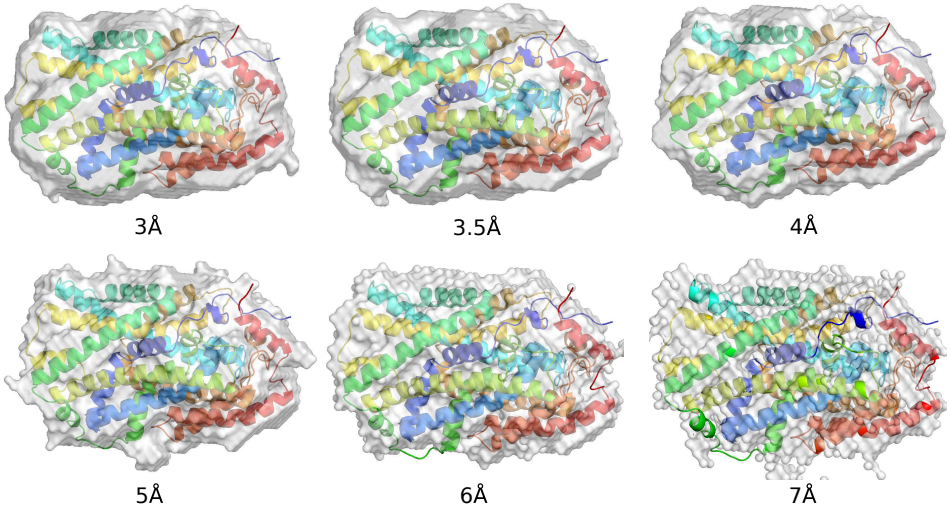

Figure 6. The calculated protein boundaries of 2JLN using the 3, 3.5, 4, 5, 6, and 7 Å observed data. The PDB deposited model is displayed as cartoons.

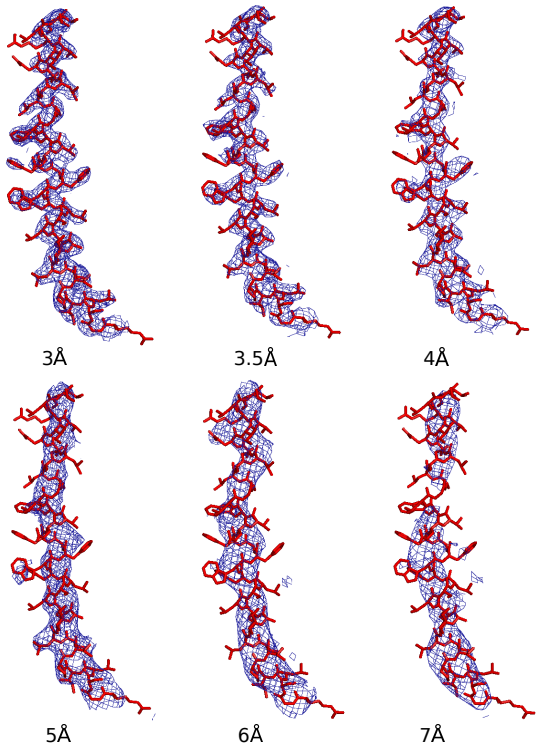

Figure 7. The calculated density maps of 2JLN using the 3, 3.5, 4, 5, 6, and 7 Å observed data. One major helix is presented for clarity. The PDB deposited model is displayed as red sticks.

The reconstructed models of the averaged densities for data of various resolutions are displayed in Figure 8. Side chains can be rebuilt for the 3 and 3.5 Å data using *ARP/wARP* [39]. Secondary structures can be clearly traced for the 4 Å data. They can be located for the 5 and 6 Å data, but not completely. For the 7 Å data, only partial helix can be traced.

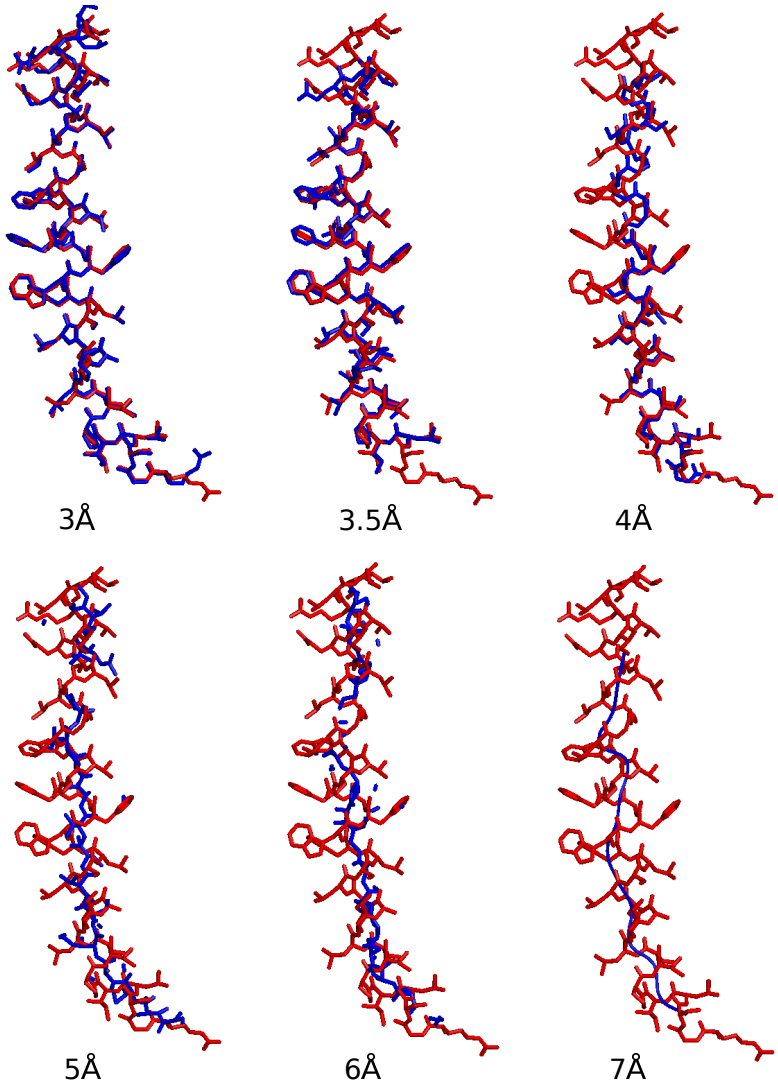

Figure 8. The reconstructed models of 2JLN using the calculated phases of the 3, 3.5, 4, 5, 6, and 7 Å observed data. One major helix is presented for clarity. The PDB deposited model is displayed as red sticks. The reconstructed model is colored in blue.

In summary, Table 1 lists the success rate, the final phase error, the final CC value, and the completeness of the reconstructed models for data of various resolutions. For the 2.85, 3, 3.5, 4, and 5 Å data, the number of successful runs are around 55 among 200. For the 6 and 7 Å data, the success rates decrease a lot. Overall, the final phase errors are less than 50° and the final CC values are above 0.8. The completeness of the reconstructed model declines when the resolution of the data decreases, as the expected density in the solvent region is not flat for low-resolution data, which is not favorable to the HIO method. For the 8 Å data, no successful runs have been reached. The 3 Å data seems better than the 2.85 Å data in rebuilding the model. That is probably because the measured data in the resolution shell from 2.85 to 3 Å contain more errors which can be seen in Figure 9a.

Table 1. The success rate, error metrics, and completeness of the reconstructed model for the *ab initio* phasing of 2JLN using data at various resolutions.

Data (Å)	Successful Runs in 200	Final Phase Error	Final CC Value	Model Completeness %
2.85	55	43	0.89	81
3.0	52	40	0.90	84
3.5	57	36	0.89	74
4.0	58	38	0.88	56
5.0	56	48	0.83	34
6.0	13	44	0.87	34
7.0	6	49	0.81	24
8.0	0	na	na	na

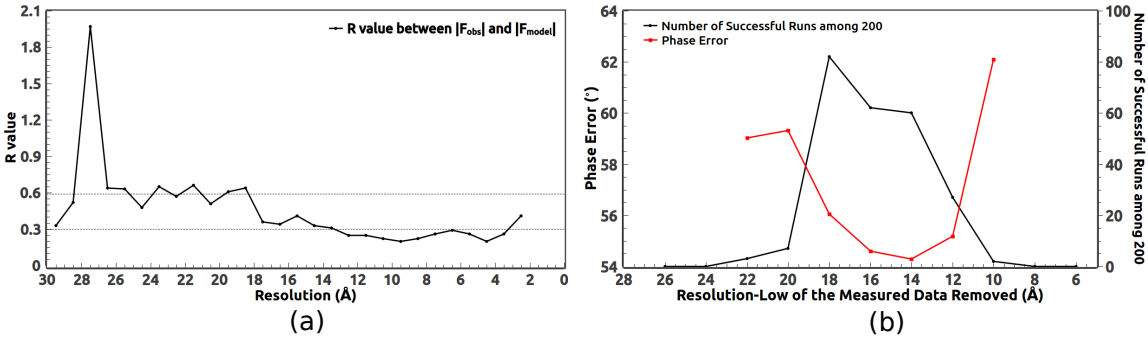

Figure 9. (**a**) R value between $|F_{obs}|$ and $|F_{model}|$. $|F_{obs}|$ are the diffraction data of 2JLN measured down to 29 Å. $|F_{model}|$ are the synthetic data of 2JLN calculated from the PDB deposited model with bulk solvent correction. (**b**) Removal of very low resolution data is necessary for 2JLN in order to increase the success rate. When the measured low-resolution reflections are ignored up to 14 Å, a lower phase error can be achieved.

3. Removal of Very Low Resolution Data

In an actual diffraction data set, there are always missing reflections at very low resolution due to the beam stop. They have to be calculated from Equation (3). For those very low resolution reflections which have been measured, the errors usually are larger than those of the intermediate or high resolutions (Figure 9a). R value indicates the agreement between the measured data and the expected data computed from the deposited model with bulk solvent correction. Measured reflections with a R-value less than 0.3 are assumed to be very good. If a measured reflection has a R-value greater than 0.59, it deviates too much from the expected value. In Figure 9a, the measured data lower than 18 Å have a R-value around 0.59. Inclusion of those measured reflections into the HIO iteration will compromise the accuracy of the calculation. When including the measured data lower than 24 Å, no successful runs are reached among 200 attempts. Therefore, it is actually advantageous to ignore the corresponding data. As shown in Figure 9b, when removing the measured data lower than 18 Å, we get the highest success rate with 82 successful runs among 200. Removing diffraction data of 2JLN up to 14 Å actually further reduces the phase error. On the other hand, if too many measured reflections are removed, it becomes difficult to find proper values for all of them during the calculation. The success rate goes to zero when the measured data lower than 8 Å are ignored.

Empirically, when the measured data have a high resolution limit such as 2 Å, the very low resolution reflections usually contain less errors. Removing the measured data up to 20 Å or 18 Å is a proper choice. When the measured data have an intermediate resolution such as 2.85 Å, removing the measured data up to 16 Å is good. In our trial calculations on 2JLN, the measured data up to 16 Å were ignored and they were filled with the calculated values according to Equation (3).

4. Inclusion of the Unmeasured High-Resolution Reflections

The diffraction data of 2JLN are measured up to 2.85 Å. The expected density in the solvent region is not flat at 2.85 Å resolution. The calculated data at high resolution need to be incorporated into the HIO phasing iterations. It has been reported that it is possible to improve the phases by extrapolating to a high resolution that was not actually measured [21–24]. Figure 10 indicates the phase error and the success rate when more and more unmeasured high-resolution reflections are involved. The phase error obviously decreases when more unmeasured reflections are involved and becomes stable after including up to 2.0 Å reflections. At the same time, the success rate slightly decreases due to more unmeasured reflections to be filled. As a result, filling the unmeasured reflections up to 2.0 Å seems to be a proper choice when phasing the 2.85 Å data of 2JLN.

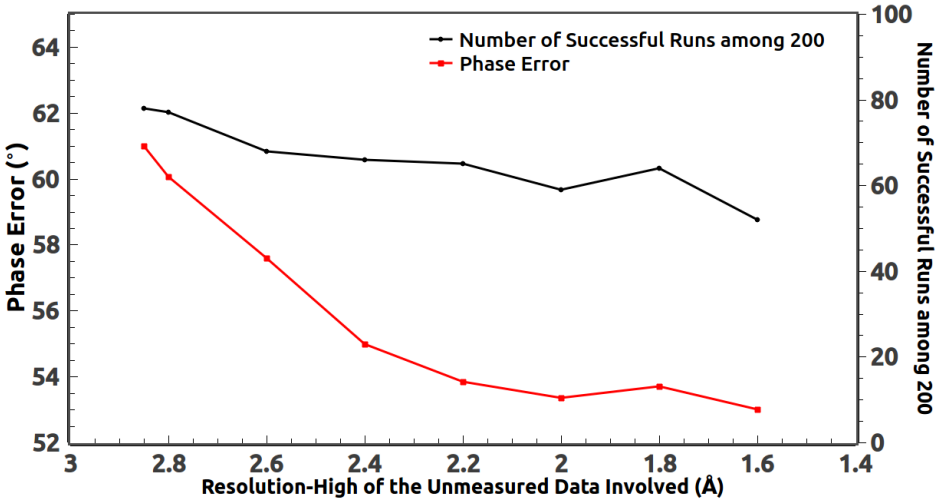

Figure 10. The diffraction data of 2JLN are measured up to 2.85 Å. When some unmeasured high-resolution reflections are involved into the HIO phasing, a lower phase error can be achieved with a slightly decreased success rate.

There is an alternative method to include the unmeasured high-resolution reflections. In order to include all 2.0 Å reflections, the size of the grid in real space has to be at least 1.0 Å. When the grid size in real space is very small, it is time-consuming to compute the densities and the corresponding structure factors during thousands of iterations. Hence a bigger size of grid is always preferred. We find a grid size of 1.43 Å works well for phasing the 2.85 Å data. Instead of including the unmeasured reflections up to a certain resolution such as 2 Å, we include all unmeasured reflections on the grid in reciprocal space with a certain grid size. When the grid size is 1.43 Å, some reflections up to 1.7 Å are already incorporated in reciprocal space. The resultant phase error is about 54° which is quite close to the one we obtained by including the unmeasured reflections up to 2.0 Å.

In our trial calculations, the grid size in real space is chosen to be half of the high resolution limit of the data. For example, when phasing 7 Å data, we choose a grid size of 3.5 Å. All unmeasured reflections on the reciprocal grid are considered in the phasing procedure. Although the expected density in solvent is still not a constant, the HIO method is quite capable of automatically adjusting the calculated magnitudes of those floating reflections to offset this effect. In other words, the HIO method needs to have a way to absorb the side effect caused by measurement errors, uneven solvent density, missing reflections, and so on. The existence of those floating reflections provides such a means.

5. Optimal Resolution of the Reference Histogram

As mentioned in the introduction, when solving for a diffraction data set at a certain resolution, a reference histogram of the same resolution is not always the optimal one. Instead of 4W6V,

the histogram of 2JLN itself is used in this section. It helps to make the conclusions clear and straightforward. In order to find an optimal reference histogram, histograms of 2JLN computed at 2.0, 2.85, 3.0, 3.5, 4.0, 5.0, 6.0, and 7.0 Å resolutions have been tested. For example, the histogram at 2.0 Å resolution is calculated from 2.0 Å diffraction data and true phases of 2JLN. The optimal histogram should lead to a low phase error and a high success rate.

The protein density histograms of 2JLN at various resolutions are illustrated in Figure 11 and the area under each curve equals one. All histograms are calculated from the deposited atomic model with bulk solvent correction. The average density in the solvent region is zero which can be achieved by adjusting $|F_{000}|$. Since the deposited model of 2JLN has an average temperature factor of 60 Å^2, density histograms with a resolution higher than 2 Å vary little which are not displayed in Figure 11. When the resolution goes from 2 to 7 Å, the density distribution becomes less and less positively skewed. At the same time, the maximum density in the protein region decreases. When the resolution is lower than 7 Å, the density histogram gets close to a Gaussian distribution. Since the average density of those histograms in Figure 11 are the same, when a histogram gets less positively skewed, the position of its peak moves close to its average density.

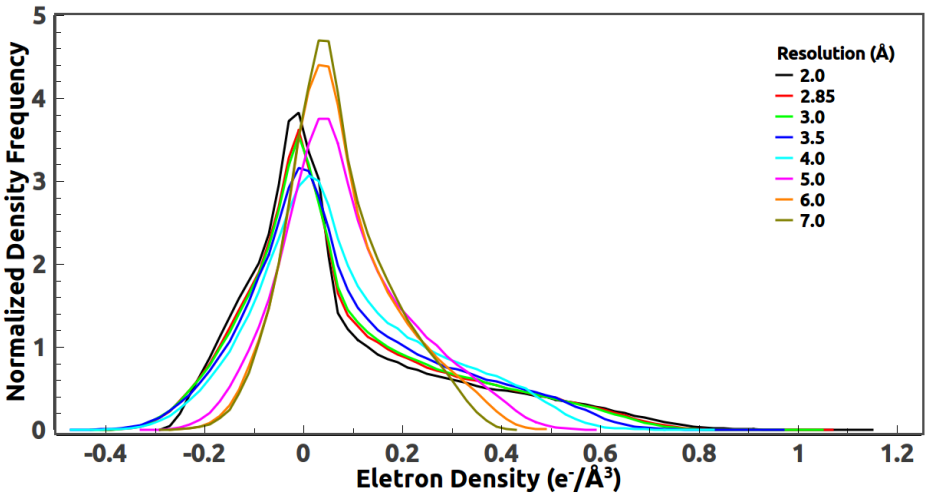

Figure 11. The density histograms of 2JLN computed from the atomic model with bulk-solvent correction, at various resolutions. For clarity, each histogram is displayed as a curve. The area under each curve equals one. All histograms have the same average density.

In order to find an optimal reference histogram for phasing data at various resolutions, the histograms shown in Figure 11 will be tested on both measured and synthetic diffraction data at various resolutions in the following two subsections.

5.1. Optimal Resolution of the Reference Histogram Tested on the Observed Data

Firstly, the reference histograms shown in Figure 11 are tested on the observed diffraction data at various resolutions. The phase error and the success rate have been displayed in Figure 12. It can be seen in Figure 12 that the resolution of the optimal histogram does not need to match the resolution of the phasing data. In conclusion, the 2 Å histogram generally leads to a lower phase error and a higher success rate for any resolution data of 2JLN.

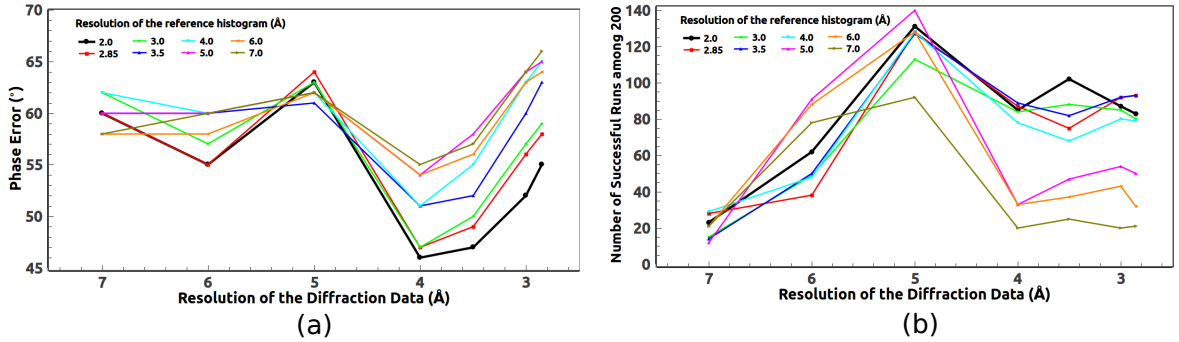

Figure 12. (**a**) The phase error and (**b**) the number of successful runs of 2JLN when the reference histograms in Figure 11 are tested on the observed data at various resolutions.

5.2. Optimal Resolution of the Reference Histogram Tested on the Synthetic Data

Since the measured data always include errors which could make the previous conclusion untenable. Testing those histograms shown in Figure 11 on a synthetic data set is necessary. The synthetic data are computed from the atomic model of 2JLN with bulk solvent correction. We suppose the synthetic reflections with a resolution lower than 18 Å are missing due to the beam stop. The phase error and the success rate have been displayed in Figure 13. It is clear from Figure 13 that for ideal data of 2JLN the reference histogram of 2 Å is the optimal one to use for data at any resolution.

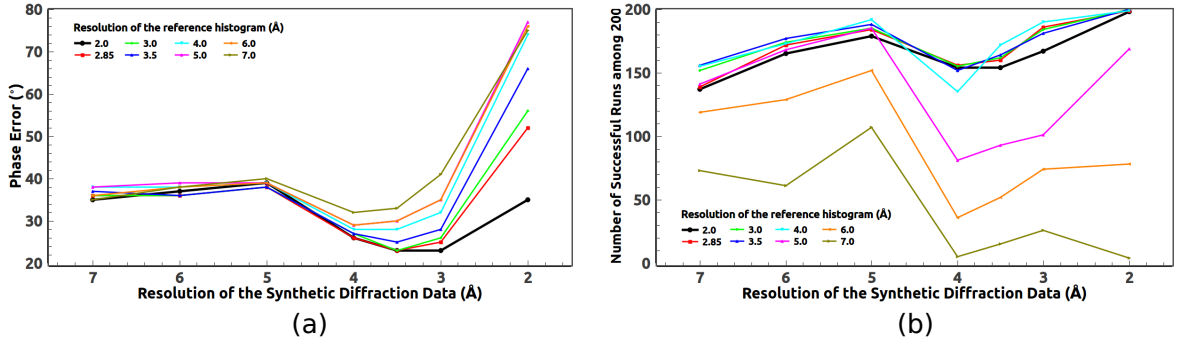

Figure 13. (**a**) The phase error and (**b**) the number of successful runs of 2JLN when the reference histograms in Figure 11 are tested on the synthetic data at various resolutions.

5.3. Reference Histogram Encodes the Information about Atomic Distance

The reference histogram at 2 Å in Figure 11 exhibits peculiar features such as a steep peak near the solvent density together with a long tail at high density. To better understand those features, we magnified the atomic distances by a certain factor and recalculated the histograms at 2 Å. As shown in Figure 14, the new histograms are significantly less positively skewed than the original one, indicating that the actual bond distances are encoded in the original density histogram. When the atomic distances are magnified by 1.5 times, atoms are greatly separated and the densities of two neighbor atoms have a less overlap. The overall density histogram looks like a Gaussian distribution.

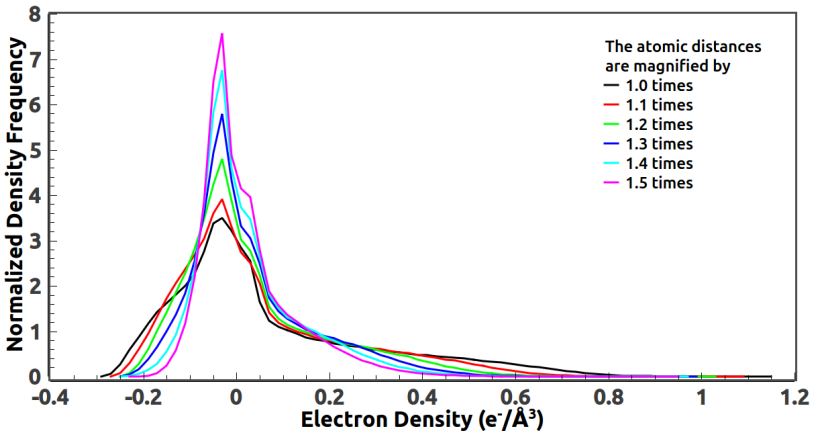

Figure 14. The density histogram of 2JLN at 2 Å is positively skewed (shown as a black line). When the atomic distances are magnified by certain times, the density histograms become less and less positively skewed. The difference between histograms reflects the change of atomic distances. All histograms are calculated from the 2 Å synthetic data of a structure with magnified atomic distances.

6. Discussion

The HIO phasing algorithm has been successfully demonstrated for high-resolution structures [13–15]. However, it is not entirely clear that it will yield anything useful when the effective resolution is less than 3 Å.

As we have shown in this paper, the structure can still be solved with the 3 Å data, and even at 5 Å useful information such as the secondary structures and the protein boundary can still be extracted from the data. Such results extend the usefulness of the HIO phasing algorithm. Since membrane protein crystals often have high solvent content and diffract to moderate resolution, HIO phasing algorithm is a potentially useful alternative approach for solving those structures.

Along the way, we have discussed several important extensions of the algorithm. First, the average density of a number of successful runs can significantly reduce the phase error. This could make a crucial difference in the solution of a new structure. Secondly, the removal of less accurate low-resolution reflections could also improve the success rate and the accuracy of the results. Thirdly, filling unmeasured high-resolution reflections with calculated values has similar effects. This is somewhat surprising in view of the fact that the degrees of freedom are increased but the data set stays the same, i.e., effectively the redundancy is reduced, yet the phase error is also reduced. Finally, we have shown that the 2 Å reference histogram seems to be optimal for data of various resolutions of 2JLN. All those small enhancements when taken together, could constitute a substantial methodological improvement over the previous version of the HIO phasing. Source code is available from the authors upon request.

The variable data resolution offers an alternate approach to a difficult structure. A high resolution structure may not be straightforwardly solvable. In that case, it might be better to truncate the data to obtain the protein boundary and the secondary structures first, and then gradually extending them to high resolution structures by including more data [42]. This resembles the variable weighting scheme we have proposed [15].

7. Conclusions

The HIO phasing method proposed previously for direct phasing of high-solvent protein crystals is employed to solve protein structures with data at intermediate and low resolutions. In the previous paper, we mainly focused on phasing high-resolution data (around 2 Å). In this paper, we further exploit the phasing method with data ranging from 2.85 Å to 7 Å resolutions. Real experimental data were used during the test and the calculated phases were directly used for model building. The atomic model can be automatically reconstructed from the calculated phases when the resolution of the data

is greater than 3.5 Å. When the resolution of the data drops below 4 Å, the protein boundary and the secondary structures can be retrieved. As the data goes to lower resolution (8 Å), fewer reflections are involved and the phase retrieval using HIO becomes difficult. In addition, we also presented four approaches to improve the calculated phases including averaging the calculated density over a number of successful runs, removing very low resolution reflections, including some unmeasured high-resolution reflections, and choosing an optimal reference histogram. The results of our trial calculations show that the HIO phasing method is still a very good choice for crystallographers even though high-resolution data are not available.

Acknowledgments: This work was supported by the Texas Center for Superconductivity and the Robert A. Welch Foundation (E-1070). The authors acknowledge the use of the Maxwell/Opuntia Cluster and the advanced support from the Center for Advanced Computing and Data Systems at the University of Houston to carry out the research presented here.

Author Contributions: Mengchao Jiang and Hongxing He conceived the concepts; Mengchao Jiang, Hongxing He and Yunpeng Cheng carried out the calculations; Hongxing He, Mengchao Jiang and Wu-Pei Su wrote the paper.

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. Fienup, J.R. Reconstruction of an object from the modulus of its Fourier transform. *Opt. Lett.* **1978**, *3*, 27–29, doi:10.1364/OL.3.000027.
2. Fienup, J.R. Phase retrieval algorithms: A comparison. *Appl. Opt.* **1982**, *21*, 2758–2769, doi:10.1364/AO.21.002758.
3. Fienup, J.R. Reconstruction of a complex-valued object from the modulus of its Fourier transform using a support constraint. *J. Opt. Soc. Am. A* **1987**, *4*, 118–123, doi:10.1364/JOSAA.4.000118.
4. Millane, R.P. Phase retrieval in crystallography and optics. *J. Opt. Soc. Am.* **1990**, *7*, 394–411, doi:10.1364/JOSAA.7.000394.
5. 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, doi:10.1364/JOSAA.15.001662.
6. Elser, V. Phase retrieval by iterated projections. *Acta Cryst. A* **2003**, *59*, 201–209, doi:10.1364/JOSAA.20.000040.
7. 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, doi:10.1103/PhysRevB.68.140101.
8. Wu, J.S.; Weierstall, U.; Spence, J.C.H.; Koch, C.T. Iterative phase retrieval without support. *Opt. Lett.* **2004**, *29*, 2737, doi:10.1364/OL.29.002737.
9. Marchesini, S. Invited Article: A unified evaluation of iterative projection algorithms for phase retrieval. *Rev. Sci. Instrum.* **2007**, *78*, 011301, doi:10.1063/1.2403783.
10. Liu, Z.C.; Xu, R.; Dong, Y.H. Phase retrieval in protein crystallography. *Acta Cryst. A* **2012**, *68*, 256–265, doi:10.1107/S0108767311053815.
11. Millane, R.P.; Lo, V.L. Iterative projection algorithms in protein crystallography. I. Theory. *Acta Cryst. A* **2013**, *69*, 517–527, doi:10.1107/S0108767313015249.
12. Lo, V.L.; Kingston, R.L.; Millane, R.P. Iterative projection algorithms in protein crystallography. II. Application. *Acta Cryst. A* **2015**, *71*, 451–459, doi:10.1107/S2053273315005574.
13. He, H.; Su, W.-P. Direct phasing of protein crystals with high solvent content. *Acta Cryst. A* **2015**, *71*, 92–98, doi:10.1107/S2053273314024097.
14. He, H.; Fang, H.; Miller, M.D.; Phillips, G.N., Jr.; Su, W.-P. Improving the efficiency of molecular replacement by utilizing a new iterative transform phasing algorithm. *Acta Cryst. A* **2016**, *72*, 539–547, doi:10.1107/S2053273316010731.
15. 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, doi:10.1107/S205327331701436X.
16. Rossmann, M.G. *Ab initio* phase determination and phase extension using non-crystallographic symmetry. *Curr. Opin. Struct. Biol.* **1995**, *5*, 650–655, doi:10.1016/0959-440X(95)80058-1.

17. Giacovazzo, C.; Siliqi, D.; Zanotti, G. The *ab initio* crystal structure solution of proteins by direct methods. III. The phase extension process. *Acta Cryst. A* **1995**, *51*, 177–188, doi:10.1107/S0108767394010305.
18. Lunin, V.Y.; Lunina, N.L.; Petrova, T.E.; Urzhumtsev, A.G.; Podjarny, A.D. On the *ab initio* solution of the phase problem for macromolecules at very low resolution. II. Generalized likelihood based approach to cluster discrimination. *Acta Cryst. D* **1998**, *54*, 726–734, doi:10.1107/S0907444997012456.
19. Lunin, V.Y.; Lunina, N.L.; Petrova, T.E.; Skovoroda, T.P.; Urzhumtsev, A.G.; Podjarny, A.D. Low-resolution *ab initio* phasing: Problems and advances. *Acta Cryst. D* **2000**, *56*, 1223–1232, doi:10.1107/S0907444900010088.
20. Carpenter, E.P.; Beis, K.; Cameron, A.D.; Iwata, S. Overcoming the challenges of membrane protein crystallography. *Curr. Opin. Struct. Biol.* **2008**, *18*, 581–586, doi:10.1016/j.sbi.2008.07.001.
21. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; De Caro, L.; Giacovazzo, C.; Siliqi, D. Phasing at resolution higher than the experimental resolution. *Acta Cryst. D* **2005**, *61*, 556–565, doi:10.1107/S090744490500404X.
22. Sheldrick, G.M. A short history of SHELX. *Acta Cryst. A* **2008**, *64*, 112–122, doi:10.1107/S0108767307043930.
23. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; De Caro, L.; Giacovazzo, C.; Siliqi, D. *Ab initio* phasing at resolution higher than experimental resolution. *Acta Cryst. D* **2005**, *61*, 1080–1087, doi:10.1107/S0907444905015519.
24. Caliandro, R.; Carrozzini, B.; Cascarano, G.L.; De Caro, L.; Giacovazzo, C.; Siliqi, D. Advances in the free lunch method. *J. Appl. Cryst.* **2007**, *40*, 931–937, doi:10.1107/S0021889807034073.
25. 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, doi:10.1107/S0108767389009311.
26. 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, doi:10.1107/S0108767389012158.
27. Millane, R.P.; Stroud, W.J. Reconstructing symmetric images from their undersampled Fourier intensities. *J. Opt. Soc. Am. A* **1997**, *14*, 568–579, doi:10.1364/JOSAA.14.000568.
28. Van der Plas, J.L.; Millane, R.P. Ab-initio phasing in protein crystallography. *Proc. SPIE* **2000**, *4123*, 249–260, doi:10.1117/12.409276.
29. Brünger, A.T. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. *Nature* **1992**, *355*, 472–475, doi:10.1038/355472a0.
30. Fourier Transform Functions. *Intel® Math Kernel Library 11.3 Reference Manual*; Intel Corporation: Santa Clara, CA, USA, 2015; pp. 1911–1962.
31. Wang, B.C. Resolution of phase ambiguity in macromolecular crystallography. *Methods Enzymol.* **1985**, *115*, 90–112, doi:10.1016/0076-6879(85)15009-3.
32. Leslie, A.G.W. A reciprocal-space method for calculating a molecular envelope using the algorithm of B.C. Wang. *Acta Cryst. A* **1987**, *43*, 134–136, doi:10.1107/S0108767387099720.
33. Terwilliger, T.C. Reciprocal-space solvent flattening. *Acta Cryst. D* **1999**, *55*, 1863–1871, doi:10.1107/S0907444999010033.
34. Abrahams, J.P.; Leslie, A.G.W. Methods used in the structure determination of bovine mitochondrial F1 ATPase. *Acta Cryst. D* **1996**, *52*, 30–42, doi:10.1107/S0907444995008754.
35. Weyand, S.; Shimamura, T.; Yajima, S.; Suzuki, S.; Mirza, O.; Krusong, K.; Carpenter, E.P.; Rutherford, N.G.; Hadden, J.M.; O'Reilly, J.; et al. Structure and molecular mechanism of a nucleobase-cation-symport-1 family transporter. *Science* **2008**, *322*, 709–713, doi:10.1126/science.1164440.
36. Vaguine, A.A.; Richelle, J.; Wodak, S.J. SFCHECK: A unified set of procedure for evaluating the quality of macromolecular structure-factor data and their agreement with atomic model. *Acta Cryst. D* **1999**, *55*, 191–205, doi:10.1107/S0907444998006684.
37. Winn, M.D.; Ballard, C.C.; Cowtan, K.D.; Dodson, E.J.; Emsley, P.; Evans, P.R.; Keegan, R.M.; Krissinel, E.B.; Leslie, A.G.W.; McCoy, A.; et al. Overview of the CCP4 suite and current developments. *Acta. Cryst. D* **2011**, *67*, 235–242, doi:10.1107/S0907444910045749.
38. Boggavarapu, R.; Jeckelmann, J.M.; Harder, D.; Ucurum, Z.; Fotiadis, D. Role of electrostatic interactions for ligand recognition and specificity of peptide transporters. *BMC Biol.* **2015**, *13*, 58, doi:10.1186/s12915-015-0167-8.
39. Langer, G.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*, 1171–1179, doi:10.1038/nprot.2008.91.
40. Terwilliger, T.C.; Grosse-Kunstleve, R.W.; Afonine, P.V.; Moriarty, N.W.; Zwart, P.H.; Hung, L.-W.; Read, R.J.; Adams, P.D. Iterative model building, structure refinement and density modification with the PHENIX AutoBuild wizard. *Acta Cryst. D* **2008**, *64*, 61–69, doi:10.1107/S090744490705024X.

41. Adams, P.D.; Afonine, P.V.; Bunkóczi, G.; Chen, V.B.; Davis, I.W.; Echols, N.; Headd, J.J.; Hung, L.-W.; Kapral, G.J.; Grosse-Kunstleve, R.W.; et al. PHENIX: A comprehensive Python-based system for macromolecular structure solution. *Acta Cryst. D* **2010**, *66*, 213–221, doi:10.1107/S0907444909052925.
42. Su, W.-P. 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* **2008**, *64*, 625–630, doi:10.1107/S0108767308027554.