New developments in phase refinement

Abrahams, J.P.; Graaff, R.A.G. de

New developments in phase refinement

JP Abrahams* and RAG De Graaff

A longstanding problem in X-ray crystallography is that vital information regarding the crystal phases is missing from the experimental data that are gathered in the diffraction experiment. Prior knowledge needs to be introduced in order to resolve phase ambiguities whenever the diffraction data are not sufficient to unequivocally reconstruct the crystal phases through anomalous or isomorphous differences. Very recent developments include progress in the application of direct methods to small proteins and other compounds of a similar small size (Shake 'n' Bake, SHELXD, CRUNCH and SIR96), bias-free refinement through the γ-correction (Solomon), improvements in the determination of phase probability distributions (SHARP) and automated atomic refinement (wARP).

Addresses
Leiden Institute of Chemistry, Gorlaeus Laboratories, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands
*e-mail: Abrahams@chem.leidenuniv.nl

Current Opinion in Structural Biology 1998, **8**:601–605

http://biomednet.com/elecref/0959440X00800601

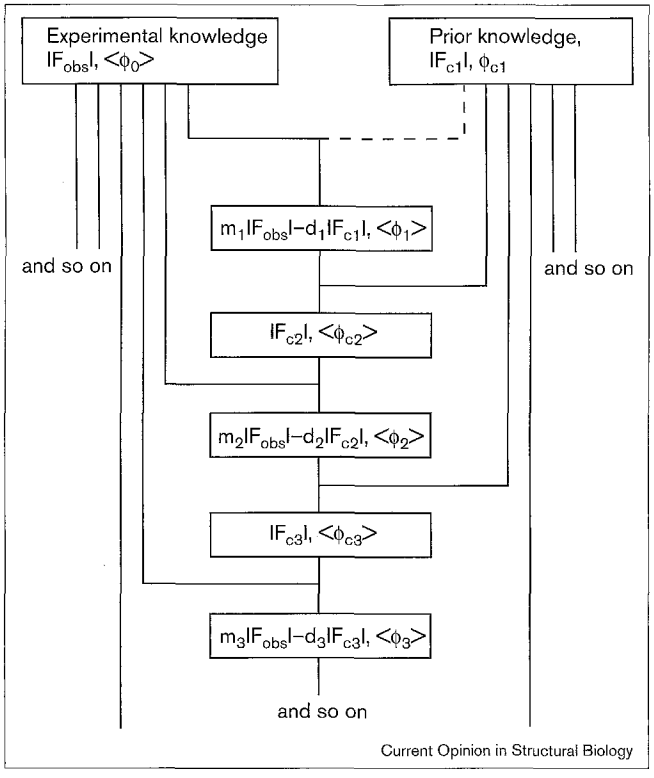
Abbreviation
KH Karle–Hauptmann

Introduction

In X-ray crystallography, the data that are used to reconstruct the electron density of the unit cell of the crystal are gathered through diffraction in reciprocal space. The periodicity of the crystal dictates the discrete sampling intervals in reciprocal space in which the diffracted X-rays do not interfere destructively. The intensity of the X-rays at each of these Bragg positions tells us the magnitude of the integral repeat of the electron density within the unit cell, with a direction and periodicity that are defined by the angles of diffraction. Unfortunately, an essential feature of the crystal structure is lost in the process — one still needs to infer the way in which all the criss-crossing static planar waves of electron density interfere with one another to form the Moire pattern that is the crystal. This is what is known as the phase problem in crystallography. Phases need to be inferred and refined using additional data.

Additional information is required to reconstruct electron density from measured diffraction data on macromolecular crystals. Two classes of such additional information can be distinguished — difference diffraction data, gathered in reciprocal space, and prior knowledge of the physical characteristics of the molecules in the crystal lattice. Since the prior knowledge is usually most easily expressed in real space, whereas the experimental data are reciprocal, most phase refinement techniques manipulate data in real and reciprocal space alternately. Figure 1 summarizes the archetypal phase refinement. Note that it is not a true cyclical process, the observed data and prior knowledge are reintroduced alternately until convergence has been achieved. Going from a set of phased, weighted structure-factor differences ($m_n|F_{obs}|-d_n|F_{cn}|$, ϕ_n) to a set with new phases and modified structure-factor amplitudes ($|F_{cn+1}|$, ϕ_{n+1}) (see legend to Figure 1), the new information is often introduced in real space, requiring two Fourier transforms.

Figure 1



In phase refinement, two types of information are combined – experimental data (which can include information on phases) are measured in reciprocal space, and prior knowledge, which is based on the physical characteristics of the molecules in the unit cell. Prior knowledge is most easily formulated in real space. It can be general (e.g. electron density is never negative, even when the atomic shape is deconvoluted and the density of large solvent cavities will be disordered) or specific to a certain crystal (e.g. the existence of noncrystallographic symmetry, the availability of a molecular replacement model, multiple isomorphous replacement, single isomorphous replacement or multiple anomalous dispersion difference data, etc.). Prior knowledge can improve the phases, leading to a better map, which can then in turn be constrained more specifically. A constrained map needs to be corrected for bias, using the γ-correction, before its structure factors can be recombined with the experimental data using σ_A weighting. d_n, weighting factor resulting in minimum bias; $|F_{cn}|$, structure-factor amplitudes of the 'n'th model; $|F_{obs}|$, observed structure-factor amplitudes; m_n, resolution-dependent weighting factor of the observed structure factors, based on a comparison with a model; $\langle\phi_0\rangle$, phases based on experimental information; $\langle\phi_{cn}\rangle$, phases of the 'n'th model; $\langle\phi_n\rangle$, recombined phases of the 'n'th model.

Each time the prior knowledge is introduced, structure factors with different reciprocal lattice vectors are combined with one another, whereas each time that experimental knowledge is introduced, structure factors with identical reciprocal lattice vectors, but from different data sets are recombined (see also [1,2•]). This is as a direct result of the fact that experimental data is measured in reciprocal space, whereas prior knowledge is based on physical considerations in real space.

Direct methods

Recently, Hauptman has written an excellent review on direct phasing methods for protein crystallography [3•]. These methods do not require experimental phase information or a molecular replacement model and they only work when data to a very high resolution (better than 1.2 Å) are available. Here, we will only deal with some of the most recent developments in the field of *ab initio* phasing.

Direct methods rely on sufficiently general types of prior knowledge being introduced in reciprocal space without major conceptual difficulty. For example, the concept of atomicity can be introduced by deconvoluting the atomic shape from the measured structure-factor amplitudes, transforming these into E-values. A data set of E-values is equivalent to a data set of rescaled structure-factor amplitudes, with a mean value that is the same for every resolution bin of the data set. E-values are therefore also referred to as 'normalised structure-factor amplitudes'. When atomic resolution is available, properly phased E-values give rise to point-like atoms upon a Fourier transform into real space, as is to be expected since the atomic shape was deconvoluted from the structure. The certainty that electron density is never negative and derived statistical relationships between structure factors inspired various other phase restraints in reciprocal space. Examples are the triplets, quartets, and so on, that define phase relationships between three, four or more structure factors. Considerations like these gave rise to the mainstream direct methods programs. In the field of protein crystallography, recent successes have been reported for the Shake 'n' Bake method [4,5•,6] and a different version of a similar procedure, Half Baked, devised by Sheldrick and Gould [7]. These methods have in common the fact that the summed internal consistency of many small sets of inter-related structure factors is optimized, whilst continuously imposing atomicity. This is illustrated elegantly by the function $R(\phi)$, which is minimized in Shake 'n' Bake:

$$R(\phi) = \frac{\sum_{H,K}\left[\kappa_{HK}\left(\cos\phi_{HK} - \frac{I_1(\kappa_{HK})}{I_0(\kappa_{HK})}\right)^2\right]}{\sum_{H,K}\kappa_{HK}}$$

where κ_{HK} is the weight and ϕ_{HK} is the phase associated with the triplet consisting of reflections H, K and HK. Given the value of κ, the quotient of the two Bessel functions, $I_1(\kappa_{HK})/I_0(\kappa_{HK})$, represents the expected value of $\cos\phi_{HK}$, the cosine of the triplet phase. For a given set of structure factors with random or nonrandom phases, $R(\phi)$ is minimized by a parameter shift method. Next, a map is generated from which atoms are picked. These atoms are then used to calculate new structure factors, the phases of which are used as the starting point for a new round of minimizing R. The procedure is iterated until convergence. Many runs, each using a different starting point in reciprocal space, are required in order to find the right solution. The procedure devised by Sheldrick and Gould [7] is comparable, although the conventional tangent formula is used instead of $R(\phi)$ and much more attention is given to picking the best possible atoms from the maps.

The certainty that electron density never is negative dictates that a Karle–Hauptman (KH) matrix is a semipositive definite [8]. It was subsequently demonstrated that the most probable set of phases maximizes the determinant of a KH matrix [9,10]. A practical application of these properties is the structure determination package CRUNCH [11], which has been used to solve structures approaching the size of small proteins. Instead of correlating phases indirectly through triplets and so on, this method imposes many phase relationships of different orders concurrently. Very recent developments include a variation on this theme, in which the driving force in reciprocal space is the minimization of the sum of the square of the negative eigenvalues of large KH matrices (large with respect to the number of atoms in the unit cell) [12•,13]. An eigenvector of a matrix is a special vector that does not change direction when multiplied with this matrix. The corresponding eigenvalue is the scale factor by which the magnitude of the vector changes upon multiplication with the matrix. An important advantage is that the screening of random starts before refinement seems possible.

All three direct methods of phase determination described here have their problems:

1. The minimization of $R(\phi)$ by a parameter shift is rather crude, but more sophisticated methods are frustrated because first and second derivatives of $R(\phi)$ with respect to the phases are not easily available.

2. The tangent formula is a crude instrument for protein structures, since its basic assumption, that all triplet phases are zero, is obviously untrue.

3. The storage requirement of the KH matrices required for the eigenvalue method increases with N^2, the square of the number of atoms present in the unit cell. This currently limits its practicality.

Bias-free refinement

It is essential to the procedure of phase refinement (see Figure 1) that prior knowledge is introduced into $\{m_n|F_{obs}| - d_n|F_{cn}|, \phi_n\}$-type data (the 'weighted experimental map', resulting in $\{|F_{cn+1}|, \phi_{n+1}\}$-type data (the 'new model map'). The new model map, containing the prior knowledge, is then recombined with the experimental data. Prior knowledge and the experimental data can only be combined meaningfully when they are scaled appropriately. More specifically, in order to prevent model bias, prior knowledge should only be introduced into resolution ranges in which it can be expected to be pertinent. In practice, resolution-dependent weighting procedures using Sim or σ_A weighting factors are used. These resolution-dependent weighting factors are calculated using the correspondence between structure-factor amplitudes of the experimental data and the data modified according to prior knowledge [14,15•]. If, in a certain resolution range, these amplitudes match poorly, the information in this range is weighted down.

The statistics upon which the calculation of the Sim or σ_A factors is based assume that the experimental data and new model map, constrained by prior knowledge, are mutually independent. It is usually the case, however, that despite modification of the structure factors through the introduction of prior knowledge, certain aspects of the data remain unaffected or are affected to only a small degree. For example, when the solvent area of a map is flattened (here one introduces prior knowledge that the solvent in large channels and cavities will be disordered [16]), the electron density in the protein region remains the same. As a result, a certain fraction of each individual structure factor of the $\{|F_{cn+1}|, \phi_{n+1}\}$ data will be biased towards a corresponding factor of the $\{m_n|F_{obs}| - d_n|F_{cn}|, \phi_n\}$ data [17••]. It was assumed until recently that it is too difficult to separate the bias component from the new information, so this problem is usually ignored. The result is model bias.

It was recently demonstrated that the degree to which constrained structure factors are individually biased towards unmodified factors can, in many instances, be calculated without major difficulty [17••]. To calculate this bias, one first expresses the prior knowledge as a real-space function, g, that is multiplied at every grid point by the map, f, into which this knowledge is to be introduced:

$$f_{new} = f.g.$$

The bias can be removed by subtracting from f_{new} the scalar product of f and the mean value of g (denoted by γ):

$$f_{(new,\ unbiased)} = f.g - \gamma f.$$

In this equation, γf is the remnant of the original, unconstrained map that still exists after the introduction of prior knowledge. As this remnant doesn't carry any new information, it represents the bias component. Recombination of the experimental data with $f_{(new,\ unbiased)}$ using σ_A weighting is now warranted, as the subtraction of γf makes both types of data independent. As a result, better maps can be obtained [18].

Better phase probability distributions

Experimental phase information can be obtained by measuring the differences in diffraction that are induced by a heavy or anomalous scatterer. Such phase information cannot be exact, but instead must be represented by a probability distribution, usually bimodal. Each individual structure factor has its own associated phase probability distribution.

Recently, it has become clear that proper determination of the phase probability distributions requires a maximum likelihood protocol for the refinement of the heavy or anomalous atom parameters [19••]. The program SHARP maximizes $L(\{g\})$, the sum of the logarithms of the likelihood of each of the complex structure factors of a data set, given the differences in diffraction of a constellation of heavy and/or anomalous scattering atoms described by a set of parameters {g}. Centric reflections (*cen*) and acentric reflections (*acen*) have different types of probability distributions and, therefore, need to be treated separately:

$$L(\{g\}) = \sum_{acen} \log\left[\Lambda^{TOT}_{acen}(\{g\})\right] + \sum_{cen} \log\left[\Lambda^{TOT}_{cen}(\{g\})\right]$$

The likelihood of a structure factor, $\Lambda^{TOT}(\{g\})$, is the product of all the probability distributions associated with a given native structure factor, integrated over all possible phases and over all possible moduli of the structure factor, given {g}. The value of $\Lambda^{TOT}(\{g\})$ depends not only on the heavy-atom parameters, but also on the experimentally determined isomorphous and anomalous differences and on estimates of the variances of these measurements. Nonisomorphism is modeled by two components — one that increases the variance with resolution (modeling random global perturbations of the structure) and one that decreases with resolution (modeling localized differences, such as altered solvent contrast, missing low occupancy sites or locally induced conformational changes). The set of parameters {g} for which the likelihood is maximal is then used to calculate the phase probabilities.

Based on such a set of phase probability distributions, which are a true statistical representation of both the experimental accuracy of the data and the lack of isomorphism, unbiased density-modification procedures can approximate the true phases by combining and redistributing the phase probabilities of all the measured structure factors. To this aim, a γ-corrected solvent-flattening procedure (implemented in Solomon and linked to SHARP in the user interface) plays an important part in generating accurate maps.

Automatic atomic refinement

Instead of visually interpreting the maps obtained using standard methods, such as single isomorphous replacement, multiple isomorphous replacement, multiple anomalous dispersion or molecular replacement, it is also possible to first automatically improve the phases. In an extension of the program ARP [20•], the procedure of wARP [21••,22•] generates a number of equal-atom atomic models automatically. About 20% more atoms than are required in order to simulate the atomic weight of the protein are placed in high density at sensible distances from each other. The models are then refined with a maximum likelihood method. After each refinement step, the models are checked using F_{obs}–F_c and $3F_{obs}$–$2F_c$ difference maps. In an F_{obs}–F_c difference map, positive density appears where the model is incomplete, atoms in negative difference density are wrong and correct parts of the structure have no appreciable density. From a $3F_{obs}$–$2F_c$ difference map, similar information can be deduced, but correct parts of the model now also have associated density. These maps can be easier to interpret. Atoms are added if unexplained density is encountered, whereas others are discarded if they conflict with one another chemically or if their temperature factors become too high. Finally, structure factors based on a number of these models are averaged and the individual phases are weighted on the basis of the consistency between the different models. If the initial phases are of limited quality and the unphased data is available to at least 2.5 Å, this method will lead to much improved maps that are generally easily interpreted.

Conclusions

Clearly, a relationship exists between the recent developments reviewed in this paper. All the methods quoted here rely on a successful admixture of information in real and reciprocal space. Provided the data are of sufficient quality, the program SHARP, followed by Solomon, provides the user with maps that are generally readily interpretable.

If this is not the case or when a poor molecular replacement model is available, wARP provides the user with an automatic means of phase improvement and extension. It is interesting to see how the authors of the wARP method succeed in implementing atomicity in their phase refinement technique, even though the data available are not of atomic resolution.

Current direct methods do need atomic resolution data, which obviously detracts greatly from their usefulness. Procedures such as Shake 'n' Bake may be applicable to more usual protein data if the concept of atomicity could be incorporated into the phase refinement in some way. The wARP method is obviously rather expensive in terms of computer resources. The effectiveness in improving the phases suggests, however, that the ideas involved may be of interest to other authors concerned with the development of methods for structure determination.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Rossmann MG, Blow DM: **The detection of sub-units within the crystallographic asymmetric unit.** *Acta Crystallogr* 1962, **15**:24-31.

2. • Vellieux FDM, Read RJ: **Non crystallographic symmetry averaging in phase refinement and extension.** *Methods Enzymol* 1997, **277**:18-52.

The authors provide a clear explanation of the relationship between the structure factors imposed by noncrystallographic symmetry and disordered solvent.

3. • Hauptman HA: **Phasing methods for protein crystallography.** *Curr Opin Struct Biol* 1997 **7**:672-680.

An excellent review of new developments in direct methods. In particular, the procedure used in the 'Shake 'n' Bake' program is lucidly explained.

4. De Titta GT, Weeks CM, Thuman P, Miller R, Hauptman HA: **Structure solution by minimal-function phase refinement and Fourier filtering I. Theoretical basis.** *Acta Crystallogr A* 1994, **50**:203-221.

5. • Chang CS, Weeks CM, Miller R, Hauptman HA: **Incorporating tangent refinement in the shake and bake formalism.** *Acta Crystallogr A* 1997, **53**:436-444.

The relative merits of tangent refinement and the minimization of the R(ϕ) function are discussed. Tangent refinement can be more efficient for solving small structures, but the minimization of R(ϕ) is preferable for larger structures.

6. Weeks CM, De Titta GT, Hauptman HA, Thuman P, Miller R: **Structure solution by minimal-function phase refinement and Fourier filtering II. Implementation and applications.** *Acta Crystallogr A* 1994, **50**:210-220.

7. Sheldrick GM, Gould RO: **Structure solution by iterative peak list optimization and tangent expansion in space group P1.** *Acta Crystallogr B* 1995, **51**:423-431.

8. Karle J, Hauptman HA: **The phases and magnitudes of structure factors.** *Acta Crystallogr* 1950, **3**:181-187.

9. Tsoucaris G: **A new method for phase determination: the maximum determinant rule.** *Acta Crystallogr A* 1970, **26**:492-499.

10. Heinerman JJL, Kroon J, Krabbendam H: **Conditional phase probability distributions of structure factors in a Karle-Hauptman matrix.** *Acta Crystallogr A* 1979, **35**:105-107.

11. de Gelder R, de Graaff RAG, Schenk H: **Automatic determination of crystal structures using Karle-Hauptman matrices.** *Acta Crystallogr A* 1993, **49**:287-293.

12. • van der Plas JL, de Graaff RAG, Schenk H: **On the use of eigenvalues and eigenvectors in the phase problem.** *Acta Crystallogr A* 1998, **54**:262-266.

The authors discuss recent developments in matrix methods. The use of matrices with orders larger than N, the number of atoms in the unit cell, is investigated.

13. van der Plas JL, de Graaff RAG, Schenk H: **Karle-Hauptman matrices and eigenvalues: a practical approach.** *Acta Crystallogr A* 1998, **54**:267-272.

14. Read RJ: **Improved Fourier coefficients for maps using phases from partial structures with errors.** *Acta Crystallogr A* 1986, **42**:140-149.

15. • Read RJ: **Model phases: probabilities and bias.** *Methods Enzymol* 1997, **277**:110-128.

This is a clear introduction to the problems associated with phase refinement and overcoming these problems using σ_A weighting.

16. Wang BC: **Resolution of phase ambiguity in macromolecular crystallography.** *Methods Enzymol* 1985, **115**:90-112.

17. •• Abrahams JP: **Bias reduction in phase refinement by modified interference functions: introducing the γ correction.** *Acta Crystallogr D* 1997, **53**:371-376.

A general method is described for calculating the magnitude of the bias component in phase refinement. The paper also contains a clear description of the phase relationships introduced by the constraint of solvent flatness.

18. Abrahams JP, Leslie AWG: **Methods used in the structure determination of bovine mitochondrial F_1 ATPase.** *Acta Crystallogr D* 1996, **52**:30-42.

19. De la Fourtelle É, Bricogne G: **Maximum-likelihood heavy atom**
•• **refinement for multiple isomorphous replacement and multiwavelength anomalous diffraction methods.** *Methods Enzymol* 1997, **276**:472-494.
The only readily available description of the methods used in SHARP, a program that uses MIR and/or MAD data for phasing.

20. Lamzin VS, Wilson KS: **Automated refinement for protein**
• **crystallography.** *Methods Enzymol* 1997, **277**:269-305.
The authors introduce a method for macromolecular crystallographic phase refinement in which atomicity is introduced automatically, rather than manually through model building.

21. Perrakis A, Sixma TK, Wilson KS, Lamzin VS: **wARP: improvement**
•• **and extension of crystallographic phases by weighted averaging of multiple refined dummy atomic models.** *Acta Crystallogr D* 1997, **53**:448-455.
A description of a procedure for phase refinement using multiple automatically generated atomic models. The authors show that the procedure is a significant improvement on the one described in [20•].

22. Van Asselt EJ, Perrakis A, Kalk KH, Lamzin VS, Dijkstra BW:
• **Accelerated X-ray structure elucidation of a 36 kDa muraminidase/transglycosylase using wARP.** *Acta Crystallogr D* 1998, **54**:58-73.
An elaborated analysis of the wARP procedure.