Phase refinement through density modification

Abrahams, J.P.; Plaisier, J.R.; Ness, S.R.; Pannu, N.S.; Sanderson, M.R.; Skelly, J.V.

No cover image available

Macromolecular Crystallography: conventional and high-throughput methods

Mark R. Sanderson (ed.), Jane V. Skelly (ed.)

https://doi.org/10.1093/acprof:oso/9780198520979.001.0001

Published: 2007 **Online ISBN:** 9780191706295 **Print ISBN:** 9780198520979

CHAPTER

CHAPTER 10 Phase refinement through density modification

Jan Pieter Abrahams, Jasper R. Plaisier, Steven Ness, Navraj S. Pannu

https://doi.org/10.1093/acprof:oso/9780198520979.003.0010 Pages 143–154

Published: August 2007

Abstract

This chapter discusses the concept of density modification, where a priori information about the structure of the macromolecule being studied is used to improve phase estimates. These concepts of density modification for the improvement of phases from crystallography have been implemented in many different programs in the past. Some of the popular density modification programs implementing the real space restraints are SOLOMON (Abrahams, 1997), DM (Cowtan, 1999), and RESOLVE (Terwilliger, 2003).

Keywords: X-ray diffraction, phase measurement, electron density, density modification

Subject: Biochemistry, Genetics and Genomics, Proteins

Collection: Oxford Scholarship Online

10.1 Introduction

It is impossible to directly measure phases of diffracted X-rays. Since phases determine how the measured diffraction intensities are to be recombined into a three-dimensional electron density, phase information is required to calculate an electron density map of a crystal structure. In this chapter we discuss how prior knowledge of the statistical distribution of the electron density within a crystal can be used to extract phase information. The information can take various forms, for example:

> *Solvent flatness*. On average, protein crystals contain about 50% solvent, which on an atomic scale usually adopts a random, non-periodic structure within the crystal and hence is featureless within the averaged unit cell. Therefore, if we know the location of the solvent regions within a macromolecular crystal, we already know a considerable part of the electron density (i.e. the part that is flat and featureless), and 'flattening' the electron density of the solvent region can improve the density of our macromolecule of interest.

Non-crystallographic symmetry. Many protein crystals contain multiple copies of one or more molecules within the asymmetric unit. Often the conformations of such chemically indistinguishable but crystallographically non-equivalent molecules are sufficiently alike to treat them as identical. In this case, we can improve the signal to noise ratio of the electron density of our molecule of interest by averaging the density of the multiple copies in the asymmetric unit.

Electron density statistics. At high resolution we know the shape of the electron density of an atom, in which case we only need to know its exact location to reconstruct the electron density in its immediate vicinity. At lower resolution we can impose an expected shape on the uni- or multivariate distributions of electron density within the protein region in a procedure that is known as histogram matching.

The problem is not so much in understanding the restrictions these types of prior knowledge impose on (suboptimal) electron density, but rather in using these restraints in reciprocal space. In practice an iterative procedure is followed. First the electron density of an initial model is calculated, which is then modified to satisfy the expected, previously determined restraints. From the modified map, the diffraction data are recalculated. The resulting phases are combined with the measured data and their associated phase probability distributions. On this basis the currently most probable phase set is calculated. The procedure is repeated until convergence. Below, we briefly describe the mathematical background of these procedures and discuss some of their essential aspects. We pay special attention to visual, geometric concepts, as we believe them important for developing an intuitive grasp for the process of density modification in phase refinement. To this end, we illustrate the concepts on a one-dimensional, centrosymmetric map, as this allows us to depict phases simply by their sign.

p. 144

10.2 Fourier transforms and the phase problem

We want to know the electron density that is determined by the measured structure factor amplitudes and their phases. The electron density at point $\boldsymbol{x}$ is calculated by a Fourier summation:

$$\rho\left(\mathrm{x}\right) = \frac{1}{V}\sum_{j=0}^{2N}\left|F\left(\mathrm{h_j}\right)\right|e^{\mathrm{i}\varphi(\mathrm{h_j})-2i\pi\mathrm{x}\cdot\mathrm{h}_j} \tag{1}$$

In the above equation, $\rho(\mathbf{x})$ is the electron density at $\mathbf{x}$, while V is the volume of the unit cell, $2N$ is the number of relevant structure factors, and $|F(\mathbf{h}_j)|$ is the amplitude of the structure factor with Miller indices $\mathbf{h}_j=(h_j, k_j, l_j)$ and a phase of $\varphi(\mathbf{h}_j)$. Note that $2N$ is determined by the size of the unit cell and the resolution.

Equation 1 is a discrete Fourier transform. It is discrete rather than continuous because the crystalline lattice allows us to sum over a limited set of indices, rather than integrate over structure factor space. The discrete Fourier transform is of fundamental importance in crystallography—it is the mathematical relationship that allows us to convert structure factors (i.e. amplitudes and phases) into the electron density of the crystal, and (through its inverse) to convert periodic electron density into a discrete set of structure factors.

Even though the Fourier transform in Eq. 1 is discrete, $\rho(\mathbf{x})$ is continuous, as it can be calculated for any grid point $\mathbf{x}$. Obviously one could calculate Eq. 1 on an arbitrary fine grid, but in that case the density at any grid point is correlated to that of its neighbours through interpolation. Since Fourier transforms neither create nor destroy information, the maximum number of uncorrelated, independent density grid points is limited

to $2N$, the number of structure factors going into the summation of Eq. 1. To conclude, if there are $2N$ structure factors, the corresponding electron density map has $2N$ independent grid points and there are $2N$ independent equations of type 1 relating the former to the latter.

Intuitively, the correct application of restraints and constraints to electron density should improve the phases. To quantify this notion, it is useful, though unconventional, to proceed as if we are solving a system of non-linear equations. A solution of such a system requires at least as many independent equations as there are unknowns. Let us therefore count the number of unknowns: there are $2N$ unknown electron densities at the independent grid points and $2N$ unknown phases. These $4N$ unknowns are inter-related by $2N$ Fourier summations, so the system is underdetermined. Its solution clearly requires at least $2N$ additional equations.

However, in view of experimental errors, $2N$ additional equations are unlikely to be sufficient to solve the phase problem. In practice, we can only expect a statistically meaningful solution if we include many more equations and identify the solution that agrees most with all equations simultaneously. Furthermore, since Eq. 1 is non-linear in $\varphi(\mathbf{h}_j)$, we cannot expect to find an analytic solution. Hence, we have to make initial guesses for the unknowns and improve from there.

Constraints effectively reduce the number of unknowns while restraints add to the number of equations in the system, without changing the number of unknowns. The effectiveness of a restraint is partially determined by the number of independent equations the restraint introduces in the minimization. Also the discriminating potential of the individual terms between right and wrong models affects the scale of the improvement. Clearly, a robust term with a sharp minimum contributes much more to phasing than a permissive term that hardly distinguishes a wrong model from a correct one. Below we summarize the general form of the additional information, constraints and restraints, which in practice leads to a system of equations that is no longer underdetermined.

10.3 Reciprocal space constraints

10.3.1 Friedel's law

In protein crystallography we assume that all electron density is real, and does not have an imaginary component. In reciprocal space this observation is known as Friedel's law, which states that a structure factor $F(\mathbf{h})$ and its Friedel mate $F(-\mathbf{h})$ have equal amplitudes, but opposite phases. The correspondence of these two assumptions follows straight from Fourier theory and, in consequence, explicitly constraining all
p. 145 electron density to be real is entirely equivalent to introducing N additional equalities of ↳ the following type:

$$\varphi(\mathrm{h_j}) = -\varphi(-\mathrm{h_j}) \tag{2}$$

Straight substitution of these equalities into Eq. 1 reduces the magnitude of the problem; rather than $4N$ unknowns, we are left with $3N$ unknowns.

10.3.2 Differences between corresponding $|F(\mathbf{h}_j)|$'s

In order to obtain initial phase estimates, crystallographers typically use either experimental phasing techniques or molecular replacement. Obtaining initial phase estimates and their associated phase probability distributions are treated in other chapters of this book, and for the remainder of this chapter we assume that the initial phases and the associated phase probabilities have been calculated. These phase probability distributions are conveniently described by Hendrickson–Lattman coefficients.

$$P\left(\varphi\left(\mathrm{h}_j\right)\right) = K_j e^{A_j \cos\left(\varphi\left(\mathrm{h}_j\right)\right) + B_j \sin\left(\varphi\left(\mathrm{h}_j\right)\right) + C_j \cos\left(2\varphi(\mathrm{h}_j)\right) + D_j \sin\left(2\varphi(\mathrm{h}_j)\right)} \tag{3}$$

Here, $P(\varphi(\mathbf{h}_j))$ is the probability function of a phase $\varphi(\mathbf{h}_j)$, whilst A_j, B_j, C_j, and D_j are its Hendrickson–Lattman coefficients and K_j is a normalizing constant. Clearly, $P(\varphi(\mathbf{h}_j))$ cannot be inserted straight into Eq. 1. However, it does provide additional equations, one for each phase for which A_j, B_j, C_j, or D_j are non-zero.

10.4 Real space restraints

10.4.1 Solvent flatness

In disordered solvent regions of the unit cell, the density is featureless and flat. In practice, the location and size of the solvent region are inferred from an initial electron density map and the molecular weight of the molecule. Since the average electron density of the solvent usually is very similar to that of the protein, automated procedures identify the solvent mask by determining the regions which have the smallest *variation* in electron density. If we know which regions of the unit cell are featureless, additional equations of the following type can be inferred:

$$\rho(\mathbf{x}_{\text{solvent}}) = \rho_{\text{solvent}} \tag{4}$$

Here, $\mathbf{x}_{\text{solvent}}$ is a real space coordinate within the solvent region and ρ_{solvent} is the mean electron density of the solvent.

The number of additional terms based on Eq. 4 is determined by the solvent fraction of the crystal. If the solvent content is 50% (which is the average for protein crystals), N independent, additional equations of type 4 are introduced. As these equations can be substituted into the Fourier summations of Eq. 1, they effectively reduce the number of unknowns by $2N$ times the solvent fraction–provided they accurately distinguish disordered solvent from protein.

10.4.2 Non-crystallographic symmetry (NCS)

If the electron density of one area of the asymmetric unit is sufficiently similar to that of another (after a translation and/or a rotation), additional equations of the following type can be inferred:

$$\rho(\mathrm{x}_{\mathrm{unique}}) = \rho(\mathrm{T}\mathrm{x}_{\mathrm{unique}}) \tag{5}$$

Here, $\mathbf{x}_{\mathrm{unique}}$ is a real space coordinate within a region of density that is repeated elsewhere in the asymmetric unit after a rotation and/or translation defined by the transformation $\mathbf{T}$. Also these equations can be substituted in the Fourier summation of Eq. 1, effectively further reducing the number of unknowns in real space down to the number of independent grid points within the fraction of unique density.

10.4.3 Electron density statistics

The overall distribution of randomly phased electron density is Gaussian, whereas a correctly phased map is expected to have a non-Gaussian distribution at resolutions beyond about 2.5 Å. An electron density distribution is described by a histogram, in which for each density value the likelihood is plotted of finding such a value within the unit cell. The shape of this histogram is determined by the resolution of the map (at low resolution, extreme values are less likely) and the chemical composition (heavy atoms will cause more extreme histograms). Proteins share characteristic electron density histograms, provided they do not
p. 146 contain many heavy atoms or large, disordered volumes. This implies that for a given ↳ structure a good guess of the correct histogram can be obtained, resulting in the following equation:

$$H\left(\rho\left(\mathrm{x}_{\mathrm{protein}}\right)\right) = H^{\mathrm{obs}}\left(\frac{1}{V}\sum_{j=0}^{2N}\left|F\left(\mathrm{h}_j\right)\right|e^{i\varphi\left(\mathrm{h}_j\right)-2i\pi\mathrm{x}_{\mathrm{protein}}\cdot\mathrm{h}_{\mathrm{j}}}\right) \tag{6}$$

Here $\mathbf{x}_{\mathrm{protein}}$ is a real space coordinate within the protein region, $H(\rho(\mathbf{x}))$ is the expected, non-Gaussian histogram of the electron density and $H^{\mathrm{obs}}(\rho(\mathbf{x}))$ is the observed histogram of protein density which may or may not have phase errors.

Equation 6 cannot be substituted into Eq. 1 and therefore it does not further reduce the number of unknowns. However, it does provide additional equations, their number being determined by the number of independent grid points within the unique protein region. Its effectiveness is determined by the difference between the theoretical histogram of a protein at a given resolution, and that of randomly phased data.

10.5 The practice of phase refinement: Fourier cycling

In theory, density modification could produce perfect phases, if the Eqs 2 to 6 are sufficiently restrictive. Let us illustrate this by an example; assume a crystal with 50% solvent and three-fold non-crystallographic symmetry. There are $2N$ Fourier summations (Eq. 1) with $4N$ unknowns. After substitution with Friedel's Law (Eq. 2), only N phases remain unknown, so now there are $3N$ unknowns in total. For all phases we have experimental information encoded in Hendrickson–Lattman coefficients (Eq. 3), so we can add N equations to our set. As we know the location of the solvent region we can reduce the number of unknown densities at independent grid points from $2N$ to N upon substituting with Eq. 4. Non-crystallographic symmetry further reduces the number of unknown densities to $N/3$ by substituting with Eq. 5. Histogram matching can further reduce the search space of solutions and improve convergence.

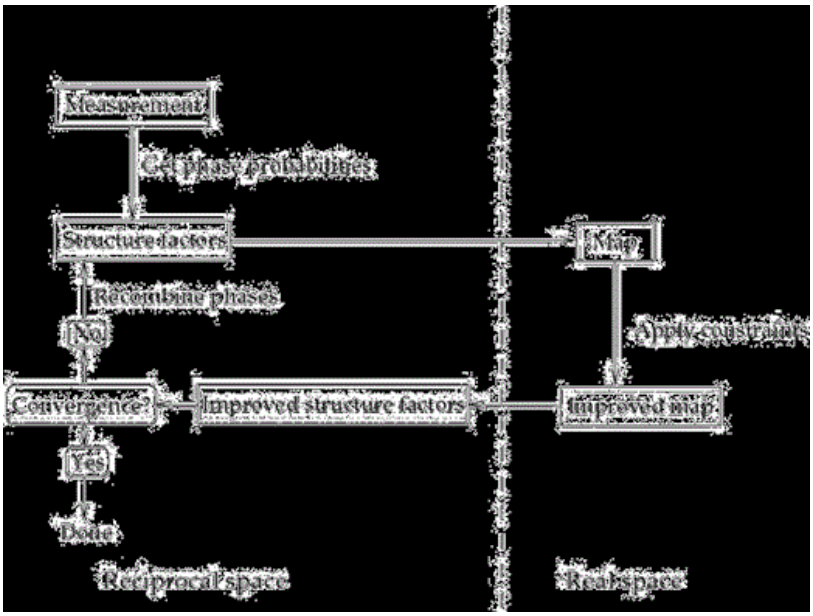
If there are more equations than unknowns, why then can we not determine phases accurately without building an atomic model? Well, the additional equations of type 3 and 6 are of a statistical nature and therefore may be less restrictive. Furthermore, inaccuracies in determining the solvent mask and the non-

crystallographic symmetry operators and masks will compromise the procedure. Nevertheless, the additional information imposed often leads to substantial improvement.

Unfortunately, the relations between the electron density, the restraints we have discussed here, and the structure factors are non-linear. Thus, the only strategy we can adopt is to use the approximate phases we start out with and improve these iteratively. Even this is not straightforward, mainly because Eq. 1 is expensive to compute. However, there exists a powerful and straightforward procedure that is used in virtually all phase refinement programs: Fourier cycling.

In Fourier cycling, the approximate phases we have available at the beginning of the process of density modification are used to calculate an initial map. The real space restraints, solvent flatness, non-crystallographic symmetry averaging, and histogram matching, are imposed on this initial density map. After Fourier transformation, the structure factors obtained typically no longer obey the reciprocal space constraints such as the measured amplitude and the phase probability distribution. Therefore existing reciprocal space restraints are recombined with the phase probability distribution obtained after back transformation of the restrained electron density. These modified structure factors are used to calculate a new map. This new map may, in turn, no longer obey the real space restraints, so these are reimposed. The procedure is repeated until it converges on a density map satisfying the equations available as well as is possible. In Fig. 10.1, a flow chart of the process of Fourier cycling in shown.

Figure 10.1

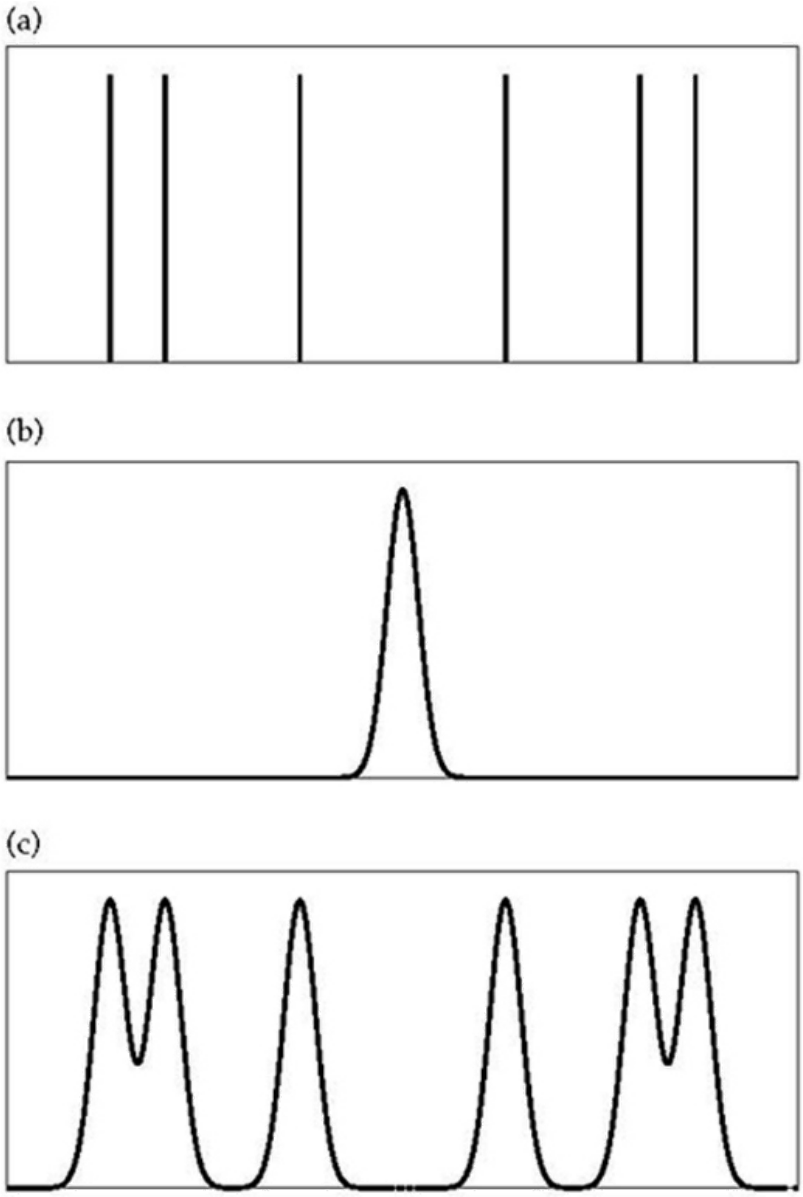


A flow chart demonstrating phase refinement in practice with the iterative recycling between real and reciprocal space.

Before we can understand why Fourier cycling works, we have to deepen our understanding of the Fourier transform. In particular, we need to understand the effects of modifying density on the structure factor amplitudes and phases. The mathematical tool that describes this is the convolution operator.

Convolution is a commonly used mathematical technique that takes as input two functions, say A(x) and
p. 147 B(x). To convolute A(x) with B(x), first take the function A(x) and place it at the origin of the second ↳ function, then multiply the two functions. Now, do this for each point in B, moving A to each point in B, multiplying the functions, and adding all the product functions. The result is the convolution of A with B. See Fig. 10.2 for an example of the convolution operator.

Figure 10.2



The procedure of convolution, represented graphically. (a) A one-dimensional centrosymmetric structure. (b) A Gaussian distribution, which could potentially be an atomic shape function. (c) The convolution of the function in (a) and (b).

Mathematically, the convolution operator $\otimes$ is defined as follows:

$$C(\mathrm{y}) = \left(A \otimes B\right)(\mathrm{y}) = \int_{-\infty}^{\infty} A(\mathrm{y}-\mathrm{x})B(\mathrm{x})\delta\mathrm{x} \tag{7}$$

A useful property of the Fourier transform and convolution is that the Fourier transform of the convolution of two functions is equal to the Fourier transform of the two functions multiplied together. Thus, since the convolution of two functions is typically a time consuming process, this property, together with the Fast Fourier Transform, is used to significantly speed up the process of convolving two functions.

In conclusion, when modifying the density in real space is equivalent to a multiplication with another map, in reciprocal space this results in the convolution of the Fourier transform of both maps (and *vice versa*).

10.6 Phase recombination in Fourier cycling

Clearly, for the procedure outlined in Fig. 10.1 to work, we need to take care of several critical steps; next to reasonable initial phase estimates required to formulate the initial restraints, we need a statistically valid procedure for the combination of the phases obtained by back transformation of the real space restrained map and the initial phase probability distribution. This recombination step is discussed below.

In general, estimates of high resolution phases will be less accurate than the ones at low resolution. The reason is that in the beginning of a structure determination, it is easier to establish low resolution contours than high resolution details. This is because contours are hardly affected by errors at high resolution. On the other hand, the contrast of high resolution features is severely affected by errors at low resolution. Hence, in phase refinement we generally see the improvement progressing from low to high resolution as we cycle through the procedure. It makes sense to weight down structure factors with erroneous phases. Therefore
p. 148 we need to introduce a weighting scheme that typically has a higher ↳ weight or lower fall-off as the phase refinement converges.

In practice, recombination of structure factors involves first weighting of the phases of the modified structure factors in a resolution dependent fashion, according to their estimated accuracy or probability. Every phase also has an experimental probability (determined by experimental phasing techniques and/or molecular replacement). The two distributions are combined by multiplication, and the new phase is calculated from this combined probability distribution. The measured associated structure factor amplitude is then scaled by the probability of the phase, and we have our set of recombined structure factors.

However, there is a problem with the phase recombination approach. Essentially we're combining probability distributions: (1) real space restraints give distributions for modified structure factors, while (2) the phasing experiments give partially independent phase probability distributions. Combining distributions is easy: we just multiply them, *provided we know they are independent*. However, here consecutive distributions are clearly not independent and treating them as independent would inevitably lead to an undesirable bias towards the very first map with which we started the Fourier cycling. How do we deal with this situation? We *separate out* the dependent component, and multiply the independent components. In order to explain how this is done in practice, we give a more quantitative explanation of the reason why Fourier cycling and phase recombination works, first for solvent flattening and subsequently for NCS and histogram matching.

10.7 Why does Fourier cycling improve phases in solvent flattening?

Before we can flatten the solvent, we need to know where it is. One of the implementations to obtain a good approximation of the solvent mask computes the variance of the electron density within a small sphere throughout the entire unit cell. Regions in the unit cell where a low variance is found then are considered to be solvent, whereas a high variance indicates protein. Most density modification programs use a binary solvent mask, with one value representing the protein region and the other value representing the solvent region. Some programs have reported good results by extending this and using real valued numbers between 0 and 1, where the value of the grid point indicates the probability of being in a protein region (Terwilliger, 2003).

Now return to Eq. 3, which describes the process of solvent flattening. As a restraint, it can be written down as follows:

$$\rho_{\mathrm{mod}}(\mathrm{x}) = \rho_{\mathrm{init}}(\mathrm{x})g(\mathrm{x}) + \rho_{\mathrm{solvent}}\hat{g}(\mathrm{x}) \tag{8}$$

where:

$g(\mathbf{x})$ is a mask function which is equal to one in the protein region and is zero in the solvent region.

$\hat{g}(\mathbf{x})$ is a mask function that is zero in the protein region and one in the solvent region.

p. 149 $\rho_{mod}(\mathbf{x})$ is the modified electron density.

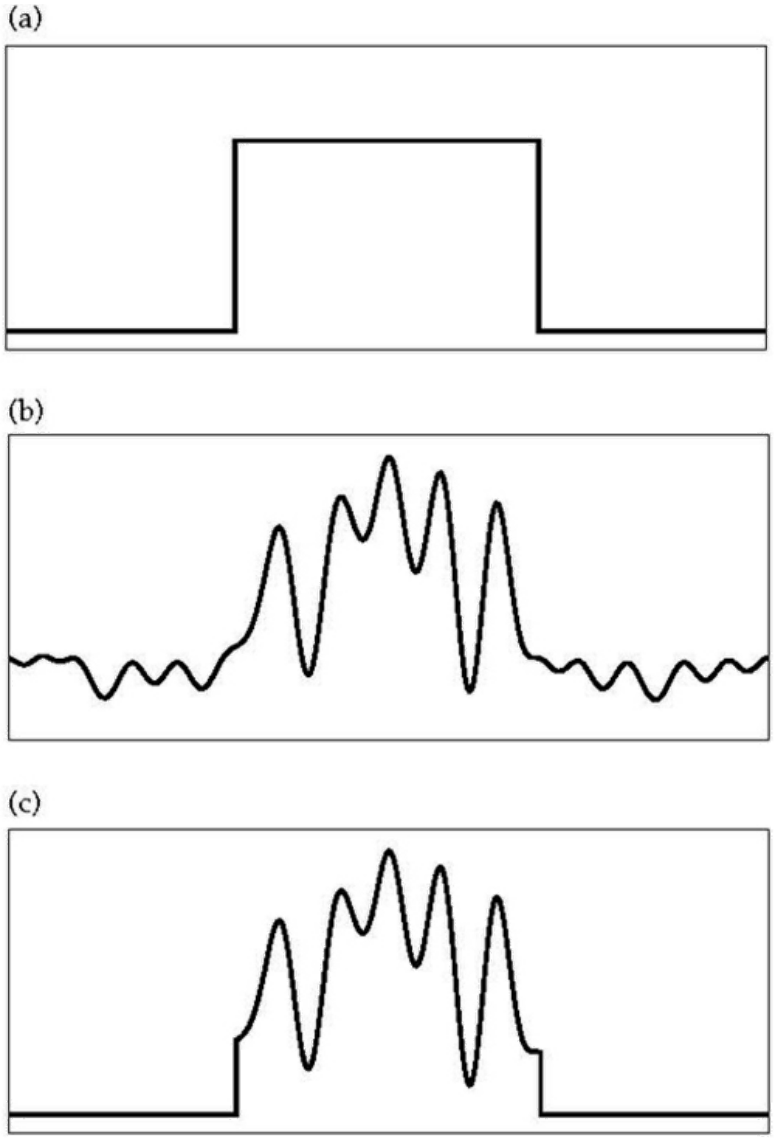
$\rho_{init}(\mathbf{x})$ is the initial electron density.

$\rho_{solvent}$ is the mean density in the solvent region.

A graphical representation of solvent flattening in real space is shown in Fig. 10.3. In Eq. 8, we multiply the two functions $\rho_{init}(\mathbf{x})$ and $g(\mathbf{x})$ as we flatten the density within the solvent region. However, multiplication in real space is equivalent to a convolution in reciprocal space. Therefore, we can rewrite Eq. 8 as follows:

$$\mathrm{F}_{\mathrm{mod}}(\mathrm{h}) = (\mathrm{F}_{\mathrm{init}} \otimes \mathrm{G})(\mathrm{h}) + C(\mathrm{h}) \tag{9}$$

Figure 10.3



A graphical representation of solvent flattening in real space. (a) represents a one-dimensional solvent mask, (b) is a one-dimensional, unflattened electron density, and (c) is the resulting flattened electron density map that imposes the solvent mask.

Where:

$F_{mod}(\mathbf{h})$ is the Fourier transform of the modified density map, $\rho_{mod}(\mathbf{x})$.

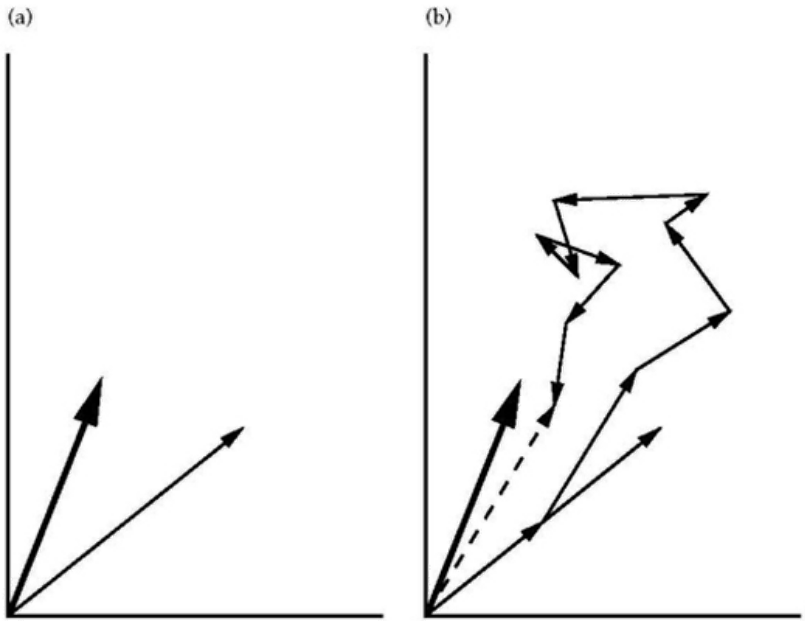
$F_{init}(\mathbf{h})$ is the Fourier transform of the unmodified density map, $\rho_{init}(\mathbf{x})$.

$G(\mathbf{h})$ is the Fourier transform of the protein mask, often referred to as an interference function.

$C(\mathbf{h})$ defines a small correction, mainly peaking at the origin.

Applying a mask function in real space is equivalent to combining many structure factors through a convolution in reciprocal space. This results in an improvement because the random error component of the structure factors will average out, whereas the true values of the structure factors will add up systematically. Fig. 10.4 gives a graphical example showing this phenomenon.

Figure 10.4



Vector diagram of the effect of convolution of many structure factors in reciprocal space. In (a) is shown the true structure factor (thick arrow) and the initial, unmodified structure factor (thin arrow). In (b) the convolution operator applied to the initial structure factor (thin arrow) results in a closer estimate (dashed arrow) of the true structure factor (thick arrow).

As is described in Abrahams (1997), plotting the radial distribution of the intensity of the interference function G(**h**), most of the intensity is around the origin. Thus, when you convolute the structure factors F with the interference function G, each structure factor will mainly be recombined with structure factors that are close by.

However, this procedure is not entirely without problems. Importantly, there is a term in the convolution given in Eq. 9 which cannot be neglected: it is the value of the G(a) function at (h = 0). The magnitude of this term determines how much of the original (partially erroneous) map is unaffected by the convolution. The modified map is actually a scaled down version of the initial map, to which is added a new map containing new information. Now we have effectively identified the bias component: it is defined by the magnitude of G(**h**) (or by the mean value of g(**x**), as follows from Fourier theory). In order to do a proper phase recombination, we have to set G(0) to zero. Several ways of achieving this have been developed: the reflection omit method (Cowtan, 1996); the gamma-correction (Abrahams, 1997), which speeds up the procedure by an order of magnitude; and the perturbation gamma method (Cowtan, 1999), which generalizes the method to any type of bias determination.

10.8 Fourier cycling and NCS averaging

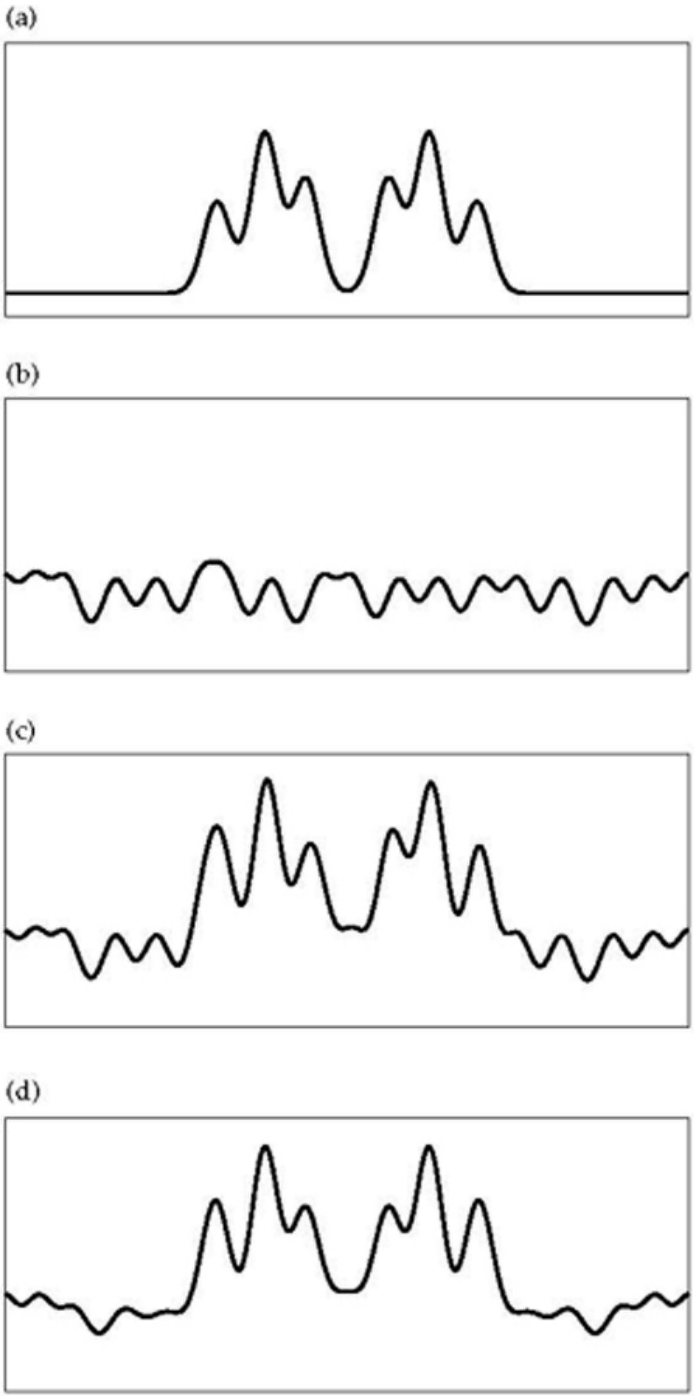
One of the earliest ways of doing density modification is non-crystallographic symmetry averaging.
p. 150 Sometimes, different molecules in the asymmetric ↳ unit are related by rotation and/or translation operators that do not belong to the crystal symmetry. Thus, when crystallographic symmetry operators are used, they superimpose the entire crystal lattice onto itself; non-crystallographic symmetry operators do not have this property.

Another way of looking at this is that non-crystallographic symmetry operators cannot be used to tile three-dimensional space, and thus are not of the class of crystallographic symmetry operators. Because these symmetry related molecules are not related by crystallographic symmetry, extra symmetry is

introduced in reciprocal space over and above the symmetry of the Laue group. Bricogne (1974) gives mathematical relationships necessary to efficiently take advantage of this extra source of information.

In many ways, NCS averaging is the easiest density modification technique to intuitively understand, especially if considered in real space. In the process of NCS averaging, one simply takes all the different NCS related molecules in the asymmetric unit, superimposes them, and then replaces their density with the average density. Because these molecules are in a similar chemical environment, and are of similar shapes, when we superimpose them, regions of similar electron density reinforce each other. We therefore increase the signal from the protein, and as we overlay multiple proteins, their signal increases additively. Likewise, the noise decreases by $1/n^{1/2}$, where n is the number of non-crystallographic symmetry related molecules. This property of signal amplification and noise reduction in NCS averaging is illustrated graphically in Fig. 10.5.

Figure 10.5



Increase of the signal to noise ratio in non-crystallographic symmetry averaging. In (a) is shown a one-dimensional representation of the electron density of a macromolecule. In (b), a graph of the noise that results from the sources of errors in the crystallographic process, including experimental phasing and measurement errors. In (c), the observed density composed of the true electron density with the noise component. In (d), the effect of non-crystallographic symmetry improves the signal from the macromolecule while decreasing the noise level, the dotted lines shows the level of bias.

In NCS-averaging, bias problems occur in Fourier cycling that are similar to the ones we mentioned in solvent flattening. For example, in two-fold averaging the result within the protein region is biased towards the initial map by 50%. Since we calculated the average of the two molecules at each grid point, half of the original density is retained. Therefore, similar treatment of the bias is required. Removing the bias results in

swapping of the densities in two-fold averaging, and replacing the density of a molecule by the average of the others in high non-crystallographic symmetries.

p. 151

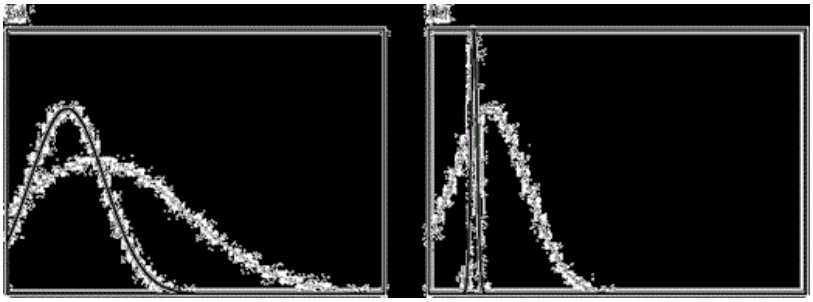
10.9 Histogram matching

Proteins often fold into a compact, globular form, composed of secondary elements such as beta sheets and alpha helices. This regular collection of building blocks leads to interesting similarities between the electron density across protein folds and families. These similarities can be exploited, and can provide a path to improve poor electron density.

Histogram matching is a technique that took its cue from the field of image processing. In various fields of study, one common problem is that of controlling contrast and brightness of an image. One such field is map making using aerial photography. Due to the wide variety of conditions and ground types, the aerial photographs are often badly corrected for contrast and brightness. One way of fixing this problem is that of histogram matching. In histogram matching, one does not look at the image *per se*, but instead, the histogram of the intensity values of each pixel, binning them into their appropriate histogram area.[1] After making a histogram of your model, comparing it to a histogram of a known good image and changing the model so that its histogram resembles the known image, the contrast and brightness of the starting image are dramatically improved.

As was mentioned before, solvent flattening helps to improve the low contrast solvent region, whereas histogram matching helps to improve the high contrast protein region. In addition, histogram matching may be used for phase extension, where by adding and phasing thin shells of data in reciprocal space, good phase estimates for structure factors of previously unphased structure factors can be obtained. One situation where this happens often is when solving a structure by MIR. Often, the diffraction pattern of a protein modified by isomorphous replacement will not diffract as far out into reciprocal space as those of the native crystal. By a judicious application of phase extension, phases that could not be obtained from the MIR experiment may be obtained by phase extension. In addition to the mainly textual description given above, a simple one-dimensional illustration is shown in Fig. 10.6.

Figure 10.6



Histogram matching. In (a) are shown a histogram from a first map in phase refinement (dashed line) and a theoretical protein histogram(solid line). In (b), the protein histogram (dashed line) and a sharp solvent histogram (solid line) are shown.

In contrast to the situation in solvent flattening and non-crystallographic symmetry averaging, in histogram matching, the theoretical gamma correction performs less well than the perturbation gamma.
p. 152 Cowtan (1999) shows that, in the case of histogram matching, there is not a single, ↳ global gamma value because of non-linearities in the histogram matching method. The problem is underdetermined, leading to multiple solutions. The perturbation gamma algorithm runs a single histogram-matching cycle twice, once

normally, and once with a map with a small amount of noise added to it. By comparing the correlation between these, one can determine a gamma correction for any subset of the data.

Technical notes

Some of the popular density modification programs implementing the real space restraints discussed above are Solomon (Abrahams, 1997), DM (Cowtan, 1999), RESOLVE (Terwilliger, 2003), Pirate (Cowtan, http://www.ysbl.york.ac.uk/~cowtan/pirate/pirate.html) and SHELXE (Sheldrick, 2002).

Solomon. This was the first density modification program to use solvent flipping, where density in the solvent region is inverted or 'flipped' to enhance density modification. Can be downloaded with the CCP4, but a superior script implementation is found in the SHARP suite.

Availability: http://www.ccp4.ac.uk

http://www.globalphasing.com

DM. One of the most popular density modification programs, DM comes bundled with the CCP4 suite. It incorporates many different ideas in density modification including histogram matching, NCS averaging, multi-resolution modification, Sayre's equation and skeletonization.

Availability: http://www.ccp4.ac.uk

RESOLVE. An easy to use density modification program that uses statistical density modification, which is an application of the general principles of maximum likelihood to density modification. One of its central algorithms is to iterate through every reflection in turn, examining all of the possible phases for the one that gives the most probable map. This procedure determines the most statistically valid hypothesis for every phase and is designed to help reduce bias. The RESOLVE program is also capable of performing model building and has many other advanced tools for protein structure solution.

Availability: http://solve.lanl.gov

Pirate. A new statistically based density modification program that uses sparseness/denseness and order/disorder in a statistical framework to model a new protein structure from ones that have been previously determined.

Availability: http://www.ysbl.york.ac.uk/~cowtan/pirate/pirate.html

http://www.ccp4.ac.uk

SHELXE. Part of the SHELX suite of programs, SHELXE uses a variety of novel algorithms to help perform density modification. One of it's primary algorithms is the 'sphere-of-influence' method, where a 2.42 Å sphere of 92 (or 272) points is systematically moved in real space through the starting electron density map, regions of high variance are assigned to be protein and regions of low variance are assigned to be solvent.

Availability: http://www.uni-ac.gwdg.de/SHELX

10.10 Conclusions

In this chapter we have presented the concept of density modification, where we take *a priori* information about the structure of the macromolecule we are studying to improve phase estimates. These concepts of density modification for the improvement of phases from crystallography have been implemented in many different programs in the past. Some of the popular density modification programs implementing the real space restraints discussed above are SOLOMON (Abrahams, 1997), DM (Cowtan, 1999), and RESOLVE (Terwilliger, 2003).

Note

1 A histogram plots frequency distributions of observed values. If observations within a certain interval occur frequently, a histogram will plot a high value for this interval, irrespective of where or when these observations occur. A histogram of the values cast by a single perfect die is flat, whereas a histogram for the total value cast by a pair of dice peaks at 7, where it is six times higher than at values of 2 or 12. For frequency distributions of correlated observations, multidimensional histograms may be useful.

References

Abrahams, J. P. (1997). Bias reduction in phase refinement by modified interference functions: introducing the correction. *Acta Crystallogr. D* **53**, 371–376. 10.1107/S0907444996015272
WorldCat Crossref

Cowtan, K. (1999). Error estimation and bias correction in phase-improvement calculations. *Acta Crystallogr. D* **55**, 1555–1567. 10.1107/S0907444999007416
WorldCat Crossref

Sheldrick, G. M. (2002). Macromolecular phasing with SHELXE. *Z. Kristallogr.* **217**, 644–650. 10.1524/zkri.217.12.644.20662
WorldCat Crossref

p. 154 Terwilliger, T. C. (2003). Statistical density modification using local pattern matching. *Acta Crystallogr. D* **59**, 1688–1701. ↳ 10.1107/S0907444903015142
WorldCat Crossref