

## 15. DENSITY MODIFICATION AND PHASE COMBINATION

### 15.1. Phase improvement by iterative density modification

BY K. Y. J. ZHANG, K. D. COWTAN AND P. MAIN

#### 15.1.1. Introduction

Density modification is a technique for improving the quality of an approximate electron-density map based on some conserved features of the correct electron-density map. These conserved features are independent of the unknown fine detail of the structural conformation. They are often expressed as constraints on the electron density in various forms, either in real or reciprocal space. Since the structure-factor amplitudes are known, these constraints restrict the values of phases and can therefore be used for phase improvement.

The structure-factor amplitudes and phases are independent of each other if we know nothing about the electron density. Therefore, the phases are indeterminable given only the amplitudes (Baker, Krukowski & Agard, 1993). The information about the electron density provides the missing link between structure-factor amplitudes and phases. It is only through the knowledge of the chemical or physical properties of the electron density that the phases can be retrieved. Density modification is usually the most straightforward application of the constraints on electron density. However, this is only a matter of convenience in implementation. Sometimes the constraints can be more readily implemented in reciprocal space on structure factors.

Density-modification methods are usually implemented as an iterative procedure that alternates between density modification in real space and phase combination in reciprocal space. This paradigm was first proposed by Hoppe & Gassmann (1968) in their 'phase correction' method. This approach takes advantage of the particular properties of the constraints and uses them in a way that is most convenient to implement.

Density-modification methods usually require an initial map with substantial phase information. In most cases, these phases are obtained from multiple isomorphous replacement (MIR) or multi-wavelength anomalous dispersion (MAD), but it is also possible to improve maps from other sources, such as molecular replacement. The amount of information in the initial map is dependent on phase accuracy, data resolution and completeness. As more powerful constraints are incorporated, the density modification can be initiated from lower-resolution maps with less accurate phases. *Ab initio* phasing would be achieved if a density-modification method could start from a map generated from random phases. Therefore, density modification can potentially lead to *ab initio* phasing methods, although it does not seek direct solution to the phase problem as its immediate goal.

There are two major components in a density-modification procedure. One is the type of electron-density constraints. The other is the way the constraints are exploited. These two components combined determine the phasing power of the procedure. In this chapter, we will review various electron-density constraints and the way they are exploited for phase improvement.

#### 15.1.2. Density-modification methods

The aim of density-modification calculations is to obtain new or improved phase estimates for observed structure-factor amplitudes. Often, this includes calculation of phases for previously unphased reflections, for example, in the case of phase extension. The

calculation of weights, which indicate the degree of confidence in the new phase estimates, is also an important part of the calculation. Improved phase estimates are obtained by bringing the initial phase estimates into consistency with additional sources of structural information.

One difficulty in combining information from various sources is that the amplitudes and phases are represented in reciprocal space and include good estimates of error, whereas the other constraints are in real space and in general, represent expectations about the structure which may be hard to quantify. As a result, the method that has been adopted is iterative and divided into real- and reciprocal-space steps. A weighted map is calculated and used as a basis for applying all the real-space modifications. The modified map is then back-transformed to produce a set of amplitudes and phases. The agreement between the observed amplitudes and the amplitudes calculated from the modified map is then used to estimate weights for the modified phases, which are used to combine the modified phases with experimental phases to produce new phases. This process is shown diagrammatically in Fig. 15.1.2.1.

A broad range of techniques have been applied to electron-density maps to impose chemical or physical information. Some sources of information used in density modification are summarized in Table 15.1.2.1. The list included here is not exhaustive, but covers the most widely used methods. Here, we describe some of the constraints and the techniques through which these constraints are implemented for phase improvement.

##### 15.1.2.1. Solvent flattening

Solvent flattening exploits the fact that the electron density in the solvent region is flat at medium resolution, owing to the high thermal motion and disorder of solvent molecules. The flattening of the solvent region suppresses noise in the map and therefore improves phases.

##### 15.1.2.1.1. Introduction

Biological molecules are typically irregular in shape, often taking roughly globular forms. When they are packed regularly to form a crystal lattice, there are gaps left between them, and these

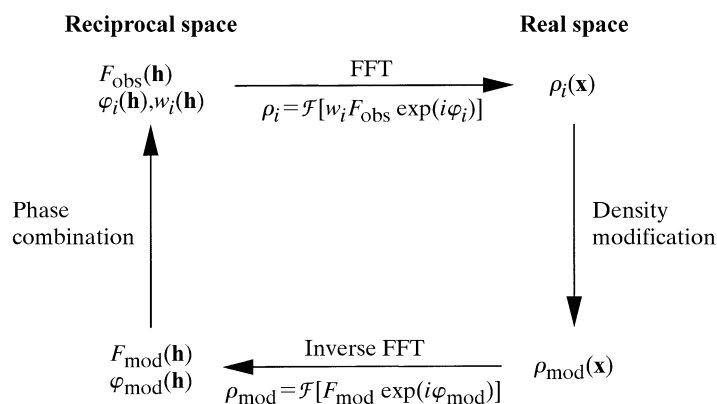


Fig. 15.1.2.1. Density-modification calculation showing iterative application of real-space and reciprocal-space constraints.

## 15. DENSITY MODIFICATION AND PHASE COMBINATION

Table 15.1.2.1. Constraints used in density modification

Constraints	Use	Effectiveness and limitation
(1) Solvent flatness	Solvent flattening	Works best at medium resolution. Relatively resolution insensitive. Good for phase refinement. Weak on phase extension.
(2) Ideal electron-density distribution	Histogram matching	Works at a wide range of resolutions. More effective at higher resolution. Very effective for phase extension.
(3) Equal molecules	Molecular averaging	Works better at low to medium resolution. Its phasing power increases with the number of molecules in the asymmetric unit.
(4) Protein backbone connectivity	Skeletonization	Requires near atomic resolution to work.
(5) Local shape of electron density	Sayre's equation	The equation is exact at atomic resolution. It can be used at non-atomic resolution by choosing an appropriate shape function. Its phasing power increases quickly with resolution. Very powerful for phase extension.
(6) Atomicity	Atomization	If the initial map is good enough, iteration could lead to a final model.
(7) Structure-factor amplitudes	Sim weighting	Can be used to estimate the reliability of the calculated phases after density modification. It assumes the random distribution of errors that caused the discrepancy between the calculated and observed structure-factor amplitudes.
(8) Experimental phases	Phase combination	This can be used to filter out the incorrect component of the estimated phases. Most phase-combination procedures assume independence between the calculated and observed phases.

spaces are filled with the solvent in which the crystallization was performed. This solvent is a disordered liquid, and thus the arrangement of atoms in the solvent regions varies between unit cells, except in those small regions near the surface of the protein. The X-ray image forms an average of electron density over many cells, so the electron density over much of the solvent region appears to be constant to a good approximation.

The existence of a flat solvent region in a crystal places strong constraints on the structure-factor phases. The constraint of solvent flatness is implemented by identifying the molecular boundaries and replacing the densities in the solvent region by their mean density value.

When solving a structure, the contents of the unit cell are usually known, and so an estimate can be formed of how much of the cell volume is taken up by solvent (Matthews, 1968). If the solvent region can be located in the cell, then we can improve an electron-density map by setting the electron density in this region to the expected constant solvent density. Once the resulting modified phases are combined with the experimental data, an improvement can often be seen in the protein regions of the map (Bricogne, 1974).

The solvent region of a unit cell may usually be determined even from a poor MIR map using the following features:

(1) The mean electron density in the solvent region should be lower than that in the protein region. Note that this information will come from the low-resolution data, which dictate long-range density variations over the unit cell.

(2) The variation in density in the flat solvent region should be much smaller than that in the ordered protein region containing isolated clumps of density. The 'peakiness' of the protein region comes from the high-resolution data.

A good method for locating the solvent region therefore takes into account information from both low- and high-resolution structure factors. Many methods have been proposed to locate the protein-solvent boundary. The first of these were the visual identification methods. The boundary was identified by digitizing a mini-map with the aid of a graphic tablet (Hendrickson *et al.*, 1975; Schevitz *et al.*, 1981). The hand-digitizing procedure was very time-consuming and prone to subjective judgmental errors. Nevertheless, these methods demonstrated the potential of solvent flattening and stimulated further improvement on boundary-identification methods. An automated method using a linked, high-density approach was first proposed by Bhat & Blow (1982). Based on the fact that the densities are generally higher in the protein region than in the solvent region, they defined the molecular boundary by locating the protein as a region of linked, high-density points.

Convolution techniques were subsequently adopted as an efficient method of molecular-boundary identification. Reynolds *et al.* (1985) proposed a high mean absolute density value approach. The electron density within the protein region was expected to have greater excursions from the mean density value than the solvent region, which is relatively featureless. The molecular boundary was located based on the value of a smoothed 'modulus' electron

## 15.1. PHASE IMPROVEMENT BY ITERATIVE DENSITY MODIFICATION

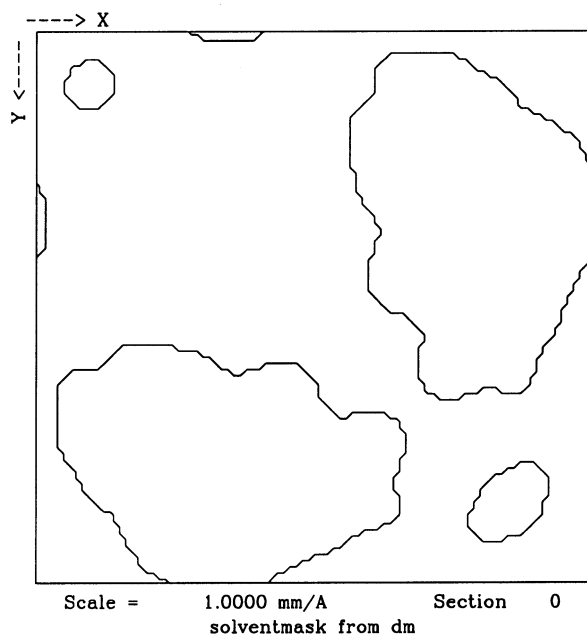
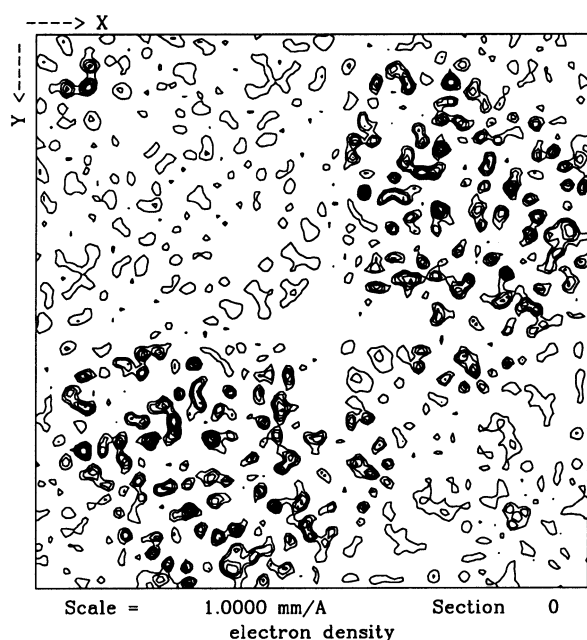


Fig. 15.1.2.2. Solvent mask determined from a map by Wang's method.

density, which is the sum of the absolute values of all density points within a small box.

### 15.1.2.1.2. The automated convolution method for molecular-boundary identification

Wang (1985) suggested an automated convolution method for identifying the solvent region which has achieved widespread use. His method involved first calculating a truncated map:

$$\rho_{\text{trunc}}(\mathbf{x}) = \begin{cases} \rho(\mathbf{x}), & \rho(\mathbf{x}) > \rho_{\text{solv}} \\ 0, & \rho(\mathbf{x}) < \rho_{\text{solv}} \end{cases} \quad (15.1.2.1)$$

The electron density is simply truncated at the expected solvent value,  $\rho_{\text{solv}}$ ; however, since the variations in density in the protein region are much larger than the variations in the solvent region, it is generally only the protein region which will be affected. Thus, the mean density over the protein region is increased. Similar results may be obtained using the mean-squared difference of the density from the expected solvent value.

A smoothed map is then formed by calculating at each point in the map the mean density over a surrounding sphere of radius  $R$ . This operation can be written as a convolution of the truncated map,  $\rho_{\text{trunc}}$ , with a spherical weighting function,  $w(\mathbf{r})$ ,

$$\rho_{\text{ave}}(\mathbf{x}) = \sum_{\mathbf{r}} w(\mathbf{r}) \rho_{\text{trunc}}(\mathbf{x} - \mathbf{r}), \quad (15.1.2.2)$$

where

$$w(\mathbf{r}) = \begin{cases} 1 - |\mathbf{r}|/R, & |\mathbf{r}| < R \\ 0, & |\mathbf{r}| > R \end{cases} \quad (15.1.2.3)$$

Leslie (1987) noted that the convolution operation required in equation (15.1.2.2) can be very efficiently performed in reciprocal space using fast Fourier transforms (FFTs),

$$\rho_{\text{ave}}(\mathbf{x}) = \mathcal{F}^{-1}\{\mathcal{F}[\rho_{\text{trunc}}(\mathbf{x})]\mathcal{F}[w(\mathbf{r})]\}, \quad (15.1.2.4)$$

where  $\mathcal{F}$  denotes a Fourier transform, and  $\mathcal{F}^{-1}$  represents an inverse Fourier transform.

The Fourier transform of the truncated density can be readily calculated using FFTs. The Fourier transform of the weighting function can be calculated analytically by

$$g(s) = \mathcal{F}[w(\mathbf{r})] = \frac{3[\sin(2\pi Rs) - 2\pi Rs \cos(2\pi Rs)]}{(2\pi Rs)^3} - \frac{3\{4\pi Rs \sin(2\pi Rs) - [(2\pi Rs)^2 - 2] \cos(2\pi Rs) - 2\}}{(2\pi Rs)^4}, \quad (15.1.2.5)$$

where

$$s = 2 \sin \theta / \lambda.$$

Therefore, the averaging of the truncated electron density by a spherical weighting function can be achieved by two FFTs. This greatly reduced the time required for calculating the averaged density. Other weighting functions may be implemented by the same approach.

A cutoff value,  $\rho_{\text{cut}}$ , is then calculated, which divides the unit cell into two portions occupying the correct volumes for the protein and solvent regions. All points in the map where  $\rho_{\text{ave}}(\mathbf{x}) < \rho_{\text{cut}}$  can then be assumed to be in the solvent region. A typical mask obtained from an MIR map by this means, and the modified map, are shown in Fig. 15.1.2.2.

The radius of the sphere,  $R$ , used in equation (15.1.2.3) for the averaging of electron densities is generally around 8 Å. The molecular envelope derived from such an averaged map tends to lose details of the protein molecular surface. Paradoxically, a large averaging sphere is required for the identification of the protein-solvent boundary based on the difference between the mean density of the protein and solvent, which is very small and can only be distinguished when a sufficiently large area of the map is averaged. Abrahams & Leslie (1996) proposed an alternative method of molecular-boundary identification that uses the standard deviation of the electron density within a given radius relative to the overall mean at every grid point of a map. The local-standard-deviation map is the square root of a convolution of a sphere and the squared map, which can be calculated in reciprocal space in a similar way to the procedure described in equations (15.1.2.4) and (15.1.2.5) as proposed by Leslie (1987). By integrating the histogram of the local-standard-deviation map, the cutoff value of the local standard deviation corresponding to the solvent fraction can be calculated. Using this procedure, a molecular envelope that contains more details of the protein molecular surface can be obtained, since the radius of the averaging sphere can be as low as 4 Å (Abrahams & Leslie, 1996).

## 15. DENSITY MODIFICATION AND PHASE COMBINATION

### 15.1.2.1.3. The solvent-flattening procedure

Once the envelope has been determined, solvent flattening is performed by simply setting the density in the solvent region to the expected value,  $\rho_{\text{solv}}$ :

$$\rho_{\text{mod}}(\mathbf{x}) = \begin{cases} \rho(\mathbf{x}), & \rho_{\text{ave}}(\mathbf{x}) > \rho_{\text{cut}} \\ \rho_{\text{solv}}, & \rho_{\text{ave}}(\mathbf{x}) < \rho_{\text{cut}} \end{cases} \quad (15.1.2.6)$$

If the electron density has not been calculated on an absolute scale, the solvent density may be set to its mean value.

A related method is solvent flipping, developed by Abrahams & Leslie (1996). In this approach, the flattening operation is modified by the introduction of a relaxation factor,  $\gamma$ , where  $\gamma$  is positive, effectively ‘flipping’ the density in the solvent region.

$$\rho_{\text{mod}}(\mathbf{x}) = \begin{cases} \rho(\mathbf{x}), & \rho_{\text{ave}}(\mathbf{x}) > \rho_{\text{cut}} \\ \rho_{\text{solv}} - [\gamma/(1-\gamma)][\rho(\mathbf{x}) - \rho_{\text{solv}}], & \rho_{\text{ave}}(\mathbf{x}) < \rho_{\text{cut}} \end{cases} \quad (15.1.2.7)$$

The effect of this modification is to correct for the problem of independence in phase combination and is discussed in Section 15.1.4.3.

### 15.1.2.2. Histogram matching

Histogram matching seeks to bring the distribution of electron-density values of a map to that of an ideal map. The density histogram of a map is the probability distribution of electron-density values. It provides a global description of the appearance of the map, and all spatial information is discarded. The comparison of the histogram for a given map with that expected for an ideal map can serve as a measure of quality. Furthermore, the initial map can be improved by adjusting density values in a systematic way to make its histogram match the ideal histogram.

#### 15.1.2.2.1. Introduction

Histogram matching is a standard technique in image processing. It is aimed at bringing the density distribution of an image to an ideal distribution, thereby improving the image quality. The first attempt at modifying the electron-density distribution was that by Hoppe & Gassman (1968), who proposed the ‘3-2’ rule. The electron density was first normalized to a maximum of 1 and modified by imposing positivity. Subsequently, the electron density was modified by  $\rho_{\text{mod}} = 3\rho^2 - 2\rho^3$ . Podjarny & Yonath (1977) used the skewness of the density histogram as a measure of quality of the modified map. Harrison (1988) used a Gaussian function as the ideal histogram in his histogram-specification method for protein phase refinement and extension. The choice of the Gaussian function as the ideal electron-density distribution was based on theoretical arguments instead of experimental evaluation. The Gaussian function was also made independent of resolution. Lunin (1988) used the electron-density distribution to retrieve the values of low-angle structure factors whose amplitudes had not been measured during an X-ray experiment. The electron-density distribution was thought to be structure specific and was derived from a homologous structure. Moreover, the histogram was derived from the entire unit cell, including both the protein and the solvent. Zhang & Main (1988) systematically examined the electron-density histogram of several proteins and found that the ideal density histogram is dependent on resolution, the overall temperature factor and the phase error. It is, however, independent of structural conformation. The sensitivity to phase error suggests that the density histogram could be used for phase improvement. The structural conformation independence made it possible to predict the ideal histogram for unknown structures.

### 15.1.2.2.2. The prediction of the ideal histogram

Polypeptide structures in particular, and biological macromolecules in general, display a broadly similar atomic composition, and the way in which these atoms bond together is also conserved across a wide range of structures. These similarities between different protein structures can be used to predict the ideal histogram even when positional information for individual atoms is not available in a map. If the positional information is removed from an electron-density map, then what remains is an unlabelled list of density values. This list is the histogram of the electron-density distribution, which is independent of the relative disposition of these densities. The shape of the histogram is primarily based on the presence of atoms and their characteristic distances from each other. This is true for all polypeptide structures.

The frequency distribution,  $P(\rho)$ , of electron-density values in a map can be constructed by sampling the map and counting the density values in different ranges. In practice, once the electron-density map has been sampled on a discrete grid, this frequency distribution becomes a histogram, but for convenience, it is treated here as a continuous distribution.

At resolutions of better than 6.0 Å and after exclusion of the solvent region, the frequency distribution of electron-density values for protein density over a wide range of proteins varies only with resolution and overall temperature factor to a good approximation. If the overall temperature factor is artificially adjusted, for example, by sharpening to  $B_{\text{overall}} = 0$ , then the frequency distributions may be treated as a function of resolution only. Therefore, once a good approximation to the molecular envelope is known, the frequency distribution of electron densities in the protein region as a function of resolution may be assumed to be known. Therefore, the ideal density histogram for an unknown map at a given resolution can be taken from any known structure at the same resolution (Zhang & Main, 1988, 1990a).

The ideal electron-density histogram can also be predicted by an analytical formula (Lunin & Skovoroda, 1991; Main, 1990a). The method adopted by Main (1990a) represents the density histogram by components that correspond to three types of electron density in the map. The first component is the region of overlapping densities, which can be represented by a randomly distributed background noise. The second component is the region of partially overlapping densities. The third component is the region of non-overlapping atomic peaks, which can be represented by a Gaussian.

The histogram for the overlapping part of the density can be represented by a Gaussian distribution,

$$P_o(\rho) = N \exp\left[-(\rho - \bar{\rho})^2/2\sigma^2\right], \quad (15.1.2.8)$$

where  $\bar{\rho}$  is the mean density and  $\sigma$  is the standard deviation. The region of partially overlapping densities can be modelled by a cubic polynomial function,

$$P_{po}(\rho) = N(a\rho^3 + b\rho^2 + c\rho + d). \quad (15.1.2.9)$$

The histogram for the non-overlapping part of the density can be derived analytically from a Gaussian atom,

$$P_{no}(\rho) = N(A/\rho)[\ln(\rho_0/\rho)]^{1/2}, \quad (15.1.2.10)$$

where  $\rho_0$  is the maximum density,  $N$  is a normalizing factor and  $A$  is the relative weight of the terms between equation (15.1.2.8) and equation (15.1.2.10).