CROATICA CHEMICA ACTA
CCACAA, ISSN-0011-1643, ISSN-1334-417X

Croat. Chem. Acta **82** (2) (2009) 421–432
CCA-3331

Review

Iterative Methods for the Solution of the Phase Problem in Protein Crystallography*

Anton Thumiger[a] and Giussepe Zanotti[a,b,]**

[a]*Department of Biological Chemistry, University of Padua, Viale G. Colombo 3, 35131 Padua, Italy*
[b]*Venetian Institute of Molecular Medicine (VIMM), Via Orus 2, 35129 Padua, Italy*

RECEIVED APRIL 7, 2008; REVISED JUNE 17, 2008; ACCEPTED JULY 15, 2008

Abstract. The phase problem is a major challenge when using X-ray crystallography for structure determination. This is especially true when the objects studied are macromolecular crystals, which contain many atoms and diffract quite poorly. For this reason, conventional direct methods, which are very successful for small and medium-sized molecule crystals, generally fail with protein crystals that do not diffract to atomic resolution. In this paper, we review some of the iterative phase retrieval methods used in optics, and present our own results obtained while trying to extend these methods to the field of macromolecular crystallography. A binary constraint on density has been incorporated in a new iterative algorithm, as well as into an existing Difference Map, in order to attempt crystallographic phase retrieval. Another existing algorithm, Charge Flipping, has been modified to test a connectivity-based phasing approach. While the results on binary densities could not be extended to realistic cases, the connectivity criterion has shown to possess some phase extension power.

Keywords: phase problem, iterative methods, flipping algorithm, binary approximation, density modification

INTRODUCTION

Many iterative methods exist for non-periodic object reconstruction. From a general point of view, all these methods operate by creating some succession of points in phase (or density) space, *i.e.*, in the space where possible solutions are defined. Each point represents a set of phases $\{\phi_h\}$, or, equivalently, the corresponding density function $\rho(x)$. Usually, a starting point is chosen at random and the succession is constructed in such a way that, almost for an appreciable percentage of starting points, convergence to the solution occurs. This solution, satisfying all the constraints simultaneously, must lie at the intersection between two constraint subsets: one defined by the experimental moduli and the other determined by *a priori* constraints (which are often easier to express in real space). The generator of the succession is a map

$$\Gamma : \rho_n \rightarrow \rho_{n+1} \quad \text{(a)} \qquad (1)$$

usually devised in such a way that the solution $\hat{\rho}$ is a *fixed point attractor* for the iterations:

$$\Gamma(\hat{\rho}) = \hat{\rho} \qquad (2)$$

(in some cases, the attractor can be a *limiting cycle* $\Gamma^n(\hat{\rho}) = \hat{\rho}$). A fixed point is left unchanged by the application of the map, so that once the iterations have converged to it, no further evolution occurs. Nevertheless, the existence of fixed points does not suffice *per se* to ensure convergence, and it is not possible to set an upper bound to the number of iterations needed to reach the solution. In this sense, a completely satisfactory phase retrieval algorithm has not been proposed yet.

Given an N-point sampling, a generic density is represented by a vector in $\Re^N$. If we call C_R and C_{MOD} the two subsets corresponding to the densities consistent respectively with real-space constraints and observed moduli, the solution must belong to their intersection

* Dedicated to Professor Emeritus Drago Grdenić, Fellow of the Croatian Academy of Sciences and Arts, on the occasion of his 90th birthday.

** Author to whom correspondence should be addressed. (E-mail: giuseppe.zanotti@unipd.it)

(a) Iterating a map gives rise to a memory-less trajectory or *Markov chain*, because only the last point determines the next. It can be argued that, in this way, some useful information from the whole past trajectory remains unexploited. This does not hold, for instance, in protein crystallography when new phases are combined with a previous set of phases.

$C^* = C_R \cap C_{MOD}$. In absence of supplementary data, the starting point is a randomly chosen element in C_{MOD}, which is generated simply by Fourier transforming the known moduli with random phases. A repeated application of the map Γ generates a trajectory in phase space, which in favorable conditions is likely to end in the intersection. When the origin is not fixed from the beginning (for example, by specifying some region in which the object density has known values) the intersection is not represented by a point, but rather by a continuous or a discrete set of points according to the space group symmetry (a three-dimensional submanifold of $\Re^N$), since all the possible choices for origin and enantiomorph are equally valid. The trajectory can be thought to evolve in real space (object density) as well in phase space, since for a given set of moduli there is a one-to-one correspondence between points in the two spaces.

Usually, the map used in iterative phasing can be constructed by composing elementary operations known as *vectorial subset projections*. The projection of an element $x \in U$ on a subset of $U, Y \subset U$ is written as $\Pi_Y : x \to \{\tilde{y}\}$ and associates to x the set $\{\tilde{y}\}$ of its nearest elements in Y:

$$\Pi_Y(x) = \{\tilde{y} \in Y : \|\tilde{y} - x\| = \inf_{y \in Y} \|y - x\|\} \qquad (3)$$

The set $\{\tilde{y}\}$ always contains a single element when the subset Y is convex. A set Y is said to be convex when, for every arbitrary pair of points $x_1, x_2 \in A$, all the points x_μ in the segment

$$\{x_\mu = (1-\mu)x_2 + \mu x_1, 0 \le \mu \le 1\} \qquad (4)$$

also belong to Y. For subsets of the euclidean plane $\Re^2$ the meaning is intuitive (see Figure 1).

It is easy to show that the subset C_{MOD} is not convex. In fact, given two densities ρ_1, ρ_2 corresponding to the observed moduli $\{F(\mathbf{h})\}$ with the phase sets $\{\phi_1\}, \{\phi_2\}$, the densities on the segment $\rho_\mu = (1-\mu)\rho_2 + \mu\rho_1$ in general will not belong to C_{MOD}, since they will not correspond to the moduli $\{F(\mathbf{h})\}$ unless a very special choice for $\{\phi_1\}, \{\phi_2\}$, is made. As a consequence, the projection on C_{MOD} is not uniquely defined, since a zero-valued $F(\mathbf{h})$ is projected onto the set of points

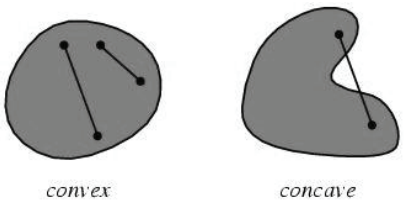


Figure 1. Convex and concave sets in the euclidean plane. Every point lying on the segment drawn between any two points of a convex set belongs to the set itself.

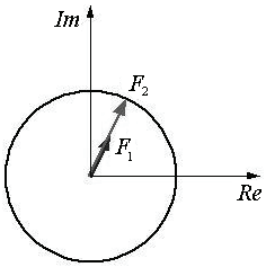


Figure 2. The Fourier modulus projection represented on the Argand plane. The correct modulus subset is a circle of radius $|F_{\mathbf{h}}^{obs}|$; a generic $F_{\mathbf{h}}$ is projected on it by leaving the phase angle unchanged and substituting the modulus with the correct one. A null vector $F_{\mathbf{h}} = 0$ would lie at the same distance from any point of the circle, and the arbitrary choice made in defining Π_{MOD} is to project it with zero phase.

lying on the circle of radius $F^{\text{obs}}(\mathbf{h})$ (Figure 2). In Fourier space the projection of a generic element of $\{F(\mathbf{h})\}$ on C_{MOD} can be written:

$$\tilde{\Pi}_{MOD} : F(\mathbf{h}) \to \begin{cases} F^{obs}(\mathbf{h}) \dfrac{F(\mathbf{h})}{|F(\mathbf{h})|} & \text{if F(h)} \neq 0 \\ F^{obs}(\mathbf{h}) e^{i\Psi_h} & \text{otherwise} \end{cases} \qquad (5)$$

$$\tilde{\Pi}_{MOD} : \rho \to T^{-1}\tilde{\Pi}_{MOD}T(\rho)$$

where the function $\Psi_{\mathbf{h}}$ is an arbitrary one. It is common to select among the many possibilities the projection with $\Psi_{\mathbf{h}} = 0$, which will be called Π_{MOD} in the following.

Another drawback due to non-convexity of the C_{MOD} subset is the presence of traps in a sequence of iterated projections.[1] Traps are fixed points which do not correspond to an intersection between the subsets. When the map is a simple alternation of projections, $\Gamma = \Pi_1\Pi_2$, and the constraints are non-convex, traps can represent a serious problem. If the trajectory of the representative point gets to a trap, in each successive iteration the density will oscillate between $\rho_1 \in C_1$ and $\rho_2 \in C_2$, each being the projection of the other, *i.e.* $\Pi_1(\rho_2) = \rho_1$ and $\Pi_2(\rho_1) = \rho_2$ (Figure 3). This can be

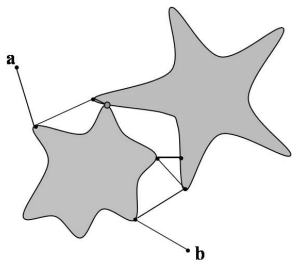


Figure 3. Two trajectories constructed by alternated projections on non-convex subsets. The succession of points starting from **a** converges to the intersection, while the one beginning in **b** ends in a trap. This means that the representative point is projected back and forth between two points lying at a local minimum of distance between the sets.

viewed as a consequence of the two subsets attaining a local minimum of distance; if their boundaries are continuous, the surface of the subset C_1 in ρ_1 and that of subset C_2 in ρ_2 will be parallel. In cases of nearly parallel surfaces the evolution is not completely blocked but becomes very slow; in that case we say the algorithm has entered a tunnel. These undesirable phenomena are known as stagnation.

AN OVERVIEW OF SOME EXISTING PHASING ALGORITHMS

Traps and tunnels potentially occur in phasing when using the Gerchberg-Saxton (*GS*) algorithm,[2] which constitutes the first non-crystallographic phase retrieval algorithm ever proposed. The real space constraint allowing image reconstruction is represented by the knowledge of the *object support S*, defined as the region in which the density is expected to be non zero. The *GS* map is simply the repeated projection on support and moduli subsets:

$$\Gamma_{GS} = \Pi_S \Pi_{MOD} \tag{6}$$

The projection onto the correct support subset is obtained by simply setting to zero the density values outside the region *S*:

$$\Pi_S : \rho_x \rightarrow \begin{cases} \rho_x & \text{if } x \in S \\ 0 & \text{if } x \notin S \end{cases} \tag{7}$$

The success of the reconstruction obviously relies on some knowledge about the object size and shape. An upper bound for the support can be inferred from its autocorrelation function, directly computable from Fourier moduli. In general, for an *N*-point sampling, a necessary (but not sufficient) condition for solution uniqueness is that the sum of the dimensions of the two subspaces must not exceed the dimension of the search space, that is, $\dim C_S + \dim C_{MOD} \leq N$, otherwise the intersection cannot be empty. For this to be true, since the subset defined by known moduli has a dimension of *N*, we must have $\dim C_S < N/2$; in other words, the problem is well posed only when the object is smaller than half of the image.

The progress of the iterations can be followed by monitoring the summed distance error *J*, which corresponds to the sum of the distances between the current density and its projections on the two subsets:

$$J(\rho) = \left\| \Pi_{MOD}(\rho) - \rho \right\| + \left\| \Pi_S(\rho) - \rho \right\| \tag{8}$$

Since this quantity can vanish only at the intersection of the subsets, a trap is characterized by the fact that *J* stabilizes on a non-zero value. A powerful alternative to the *GS* map was introduced by Fienup algorithms,[3] the most effective being the so-called *Hybrid Input-Output* (*HiO*):

$$HiO : \rho_x \rightarrow \begin{cases} \Pi_{MOD}(\rho_x) & \text{if } x \in S \\ \rho_x - \beta \Pi_{MOD}(\rho_x) & \text{if } x \notin S \end{cases} \tag{9}$$

Density within the support is modified by imposing the observed moduli, like in the *GS* algorithm; the difference lies in the outside region, where the density is no more set to zero but rather to its previous value diminished by the feedback term $\beta \Pi_{MOD}(\rho_x)$, which increases with the difference between the projected density outside the support and its expected value of zero. When the intersection has been found, the resulting density $\hat{\rho}$ is consistent with the observed moduli and is also zero outside the support, so that no further evolution is observed:

$$\Pi_{MOD}(\hat{\rho}) = \hat{\rho}_x = 0 \qquad \forall x \in S \tag{10}$$

Compared to *GS*, the *HiO* algorithm does not suffer from traps, and the convergence is faster. In terms of projections, the *HiO* map can be written as

$$\Gamma_{MOD} = \Pi_S \Pi_{MOD} + (1 - \Pi_S)(1 - \beta \Pi_{MOD}) \tag{11}$$

Recently a general form of map has been proposed,[4] the *difference map* (*DM*), which avoids stagnation and can be applied to any kind of non-convex constraints. The *HiO* algorithm turns out to be a particular case of *DM* in which the support constraint is used and a given choice of the parameters is made. The *DM* operator is defined by

$$\Gamma_{DM} = 1 + \beta \Delta \tag{12}$$

$$\Delta = \Pi_1 f_2 - \Pi_2 f_1 \tag{13}$$

The operator Γ_{DM} adds to the density a quantity Δ proportional to the difference of two composed maps. Each of these two maps results from the successive application of a map f_i and a projection Π_j on one of the two constraint subsets.

A fixed point $\hat{\rho}$ of the difference map is characterized by $\Delta = 0$, so that

$$\Pi_1 f_2(\hat{\rho}) = \Pi_2 f_1(\hat{\rho}) = \rho_{1 \cap 2} \tag{14}$$

where the element $\rho_{1 \cap 2}$, lying at the intersection between the subsets C_1 and C_2, represents the solution to the phase problem. It should be pointed out that here the solution does not coincide with the fixed point $\hat{\rho}$. Since in a fixed point Δ must vanish, its norm

$$\varepsilon_i = \left\| \Delta(\rho_i) \right\| \tag{15}$$

can be used to follow the progress of the iterations.

While the global behavior of the algorithm does not depend on the nature of the f_i, a careful choice of them is necessary to allow convergence. Setting for instance $f_1 = f_2 = 1$ (the identity map) does not give attractive fixed points. A possible choice is to construct f_i in a way that its operation on ρ produces a point on the line joining ρ to $\Pi_i(\rho)$:

$$f_i(\rho) = (1+\gamma_i)\Pi_i(\rho) - \gamma_i\rho \tag{16}$$

The optimal parameter values are $\gamma_1 = -\beta^{-1}$, $\gamma_2 = \beta^{-1}$, as found by taking into account the local behavior in the proximity of a fixed point. It can be shown that the difference map can escape traps; these cannot behave like fixed points because they do not allow the quantity Δ to vanish.

THE BINARY APPROXIMATION

A possibility for restraining the number of solutions is to approximate the electron density in the unit cell to a binary function. This approximation is motivated by the physical reality of separated solvent and protein regions. The densities of the two zones differ in average value and in variance, both quantities being greater in the protein region. The solvent density can be assumed to be flat to a good approximation, while in the protein region the density can deviate much from its average value.[5] Numerical tests show that approximating an image with a binary one leads, in Fourier space, to essentially correct phases, while the moduli are more seriously affected. In terms of constraint subsets, the binary densities subset is not expected to intersect the moduli subset, so that an approximate solution would lie between the closest points of the two sets. Moreover, the (euclidean) distance between these two elements of the two sets should be appreciable. However, the two-value approximation can be justified to some extent if the resolution is low (> 4 Å). A search for a binary mask has been successful in reconstructing the density at a resolution of about 12 Å.[6] In that case, a Binary Integer Programming (BIP) approach was used, where the main drawback is that the computing time grows exponentially with the complexity of the problem (*i.e.* with the number of grid points chosen to sample the electron density). In this perspective a more efficient search method, as an iterative one, could perhaps help in extending the resolution limit (at least in the range where the binary approximation is justified). A two-valued function can be scaled to a binary one (having only 0 and 1 as possible values), by shifting and scaling its values. To operate this scaling in Fourier space one needs to know the expected fraction of ones in the unit cell, that is, the volume defined by the molecular envelope that is to be searched for.

TWO BINARY ALGORITHMS

The subset of binary densities $C_{01} = \{\rho(\mathbf{x}) \in \{0,1\} \forall \mathbf{x}\}$ is formed by disjoint points (the corners of an hypercube) and so it is not convex. The projection of ρ on C_{01} is the element $\tilde{\rho}_{01} \in C_{01}$ which minimizes the distance

$$\left\| \rho - \rho_{01} \right\| = \sum_k [\rho(\mathbf{x}_k) - \rho_{01}(\mathbf{x}_k)]^2 \tag{17}$$

and this means that the quantities $|\rho(\mathbf{x}_k) - \rho_{01}(x_k)|$ must be minimum for every pixel k. This leads to the simple expression for the binary projector:

$$\Pi_{01} : \rho(\mathbf{x}) \rightarrow \begin{cases} 0 & : \rho(\mathbf{x}) < 1/2 \\ \{0,1\} & : \rho(\mathbf{x}) = 1/2 \\ 1 & : \rho(\mathbf{x}) > 1/2 \end{cases} \tag{18}$$

This projector is not single-valued and some arbitrary choice has to be made about the treatment of densities with value ½ , since they can be indifferently set to 0 or 1.

Here both subsets are non-convex, so that alternate projections will fail. In fact, iteration of a map $\Pi_{01}\Pi_{MOD}$ rapidly gets to a trap, because many different $\rho(\mathbf{x})$ possess the same projection. Once $\Pi_{MOD}(\rho_{n+1})$ becomes too close to $\Pi_{MOD}(\rho_n)$ the evolution stops, since Π_{01} projects both of them on the same point of C_{01}.

To find a solution to the binary phase problem it is thus necessary to avoid that any iteration ρ_n exactly belongs to the subset C_{01}. For this reason in the present work a heuristic algorithm inspired to the *HiO* map, and in particular to the feedback concept, was conceived. It is based on a map Γ_B, consisting in the alternate application of the two operations $\Xi_{MOD}^{(\gamma)}$ and $\Xi_{01}^{(\beta,\delta)}$, each one flipping the density or the moduli about their 'expected values':

$$\Gamma_B = \Xi_{MOD}^{(\gamma)} \Xi_{01}^{(\beta,\delta)}, \quad \beta, \gamma, \delta > 0 \tag{19}$$

$$\Xi_{01} : \rho \rightarrow \begin{cases} -\beta\rho & : \rho < \delta \\ \rho & : \delta \le \rho \le 1-\delta \\ 1+\beta(1-\delta) & : \rho > 1-\delta \end{cases} \tag{20}$$

$$\begin{gathered} \Xi_{MOD} = T^{-1} \tilde{\Xi}_{MOD} T \\ \tilde{\Xi}_{MOD} : |F_{\mathbf{h}}| e^{i\varphi_h} \rightarrow [|F_h^0| + \gamma(|F_h^0| - |F_h|)] e^{i\varphi_h} \end{gathered} \tag{21}$$

(The symbols T and T^{-1} stand for direct and inverse Fourier transform, respectively). The $\Xi_{01}^{(\beta,\delta)}$ operator leaves unchanged the density values falling into the interval $[\delta, 1-\delta]$, while the remaining are flipped about the nearest expected value (0 or 1) (Figure 4); the extent by which each pixel value is flipped is proportional to the parameter β. A similar operation is carried out in reciprocal space on the values of the moduli by the

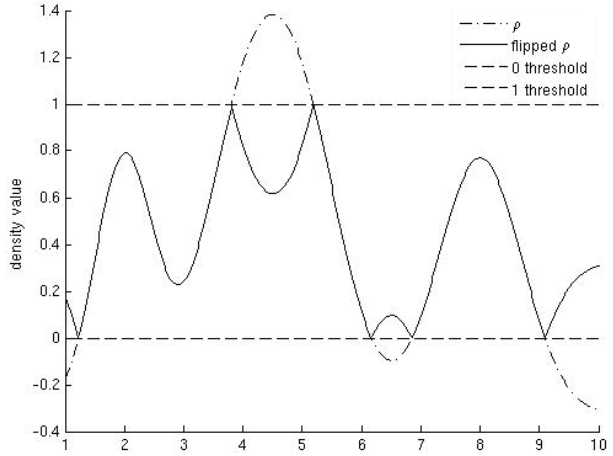


Figure 4. The flipping operation in real space. The values for a one-dimensional density are reported as a function of a spatial coordinate. Values greatest than 1 or smaller than 0 are inverted with respect to their nearest binary value.

operator $\Xi_{MOD}^{(\gamma)}$; in that case, the expected value of each Fourier modulus $|F_{\mathbf{h}}|$ is simply the known quantity $|F_{\mathbf{h}}^{0}|$, and every modulus is flipped by a quantity proportional to γ. It must be noted that both flipping operations, in real and reciprocal space, are needed for the iterations to converge. Moreover, the previous knowledge of the zero-frequency term F_0 (which usually is experimentally unmeasurable) is also necessary, and a separated flipping parameter γ_0 was introduced for it. It must be observed that, in terms of elementary projections, the map Γ_B results to be a rather complex one. The real space operation can be written as

$$\Xi_{01} = \begin{cases} (1+\beta)\Pi_{01} - \beta & : \rho \in Z_\delta \\ 1 & : \rho \notin Z_\delta \end{cases} \qquad (22)$$
$$Z_\delta = \{\mathbf{x} \,|\, \rho(\mathbf{x}) \le \delta \vee \rho(\mathbf{x}) \ge 1-\delta\}$$

where the domain Z_δ is defined as the set of points with a density falling outside the range $[\delta, 1-\delta]$. The flipping in Fourier space, in turn, can be expressed as:

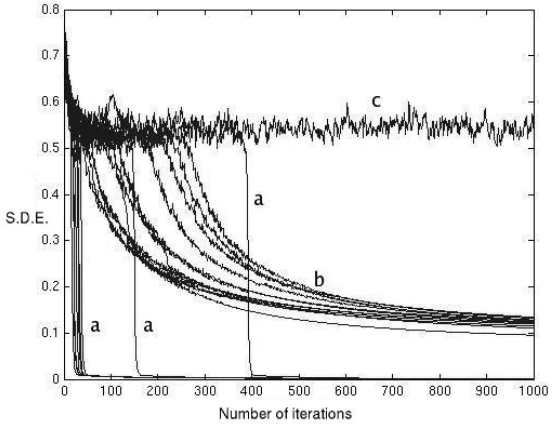
$$\tilde{\Xi}_{MOD} = (1+\gamma)\,\Pi_{MOD} - \gamma$$
$$\Xi_{MOD} = T^{-1}\,\tilde{\Xi}_{MOD}\,T = (1+\gamma)\,\Pi_{MOD} - \gamma \qquad (23)$$

The action of the operators Ξ_{01} and Ξ_{MOD} is to move the density on a point which lies on the segment joining the starting density with the projected one (in the case of Ξ_{01} this is only an approximate picture).[(a)]

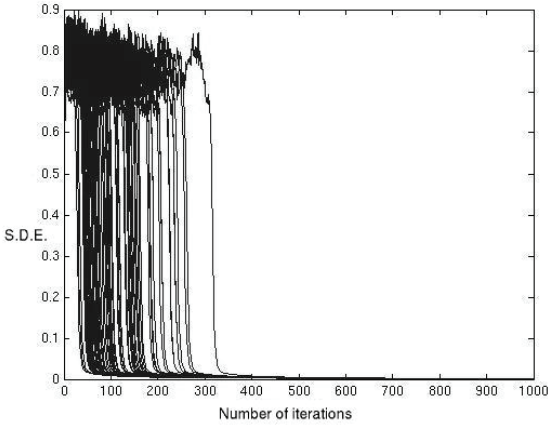
The progress of the iterations can be followed by means of a type of a summed distance error (SDE):

$$\mathrm{SDE} = N^{-1}\left[\begin{array}{l} \sum_{k=1}^{N} |\,\Pi_{MOD}[\rho(\mathbf{x}_k)] - \rho(\mathbf{x}_k)\,| + \\ \sum_{k=1}^{N} |\,\Pi_{01}[\rho(\mathbf{x}_k)] - \rho(\mathbf{x}_k)\,| \end{array}\right] \qquad (24)$$
$$k = \text{pixels}$$

The algorithm was implemented in Fortran 90 for the two-dimensional case, using the static libraries GFT[7] for FFT computation. Its behaviour has been studied for different values of β, γ, γ_0, δ, in order to identify the set of parameters giving the quickest convergence. Some test results are reported with a 2D trial density (20×20 pixels). In Figure 5a the SDE plots are shown for 20 independent runs of the algorithm (each relates to a



(a)



(b)

Figure 5. (a) SDE versus iteration number for a non-optimal parameter setting (β = 0.5, δ = 0.2, γ_0 = 1.2, γ = 1.3). 20 plots, corresponding to different runs, are displayed. Three behaviors **a**,**b**,**c** can be observed, as discussed in the main text; (b) SDE for optimized parameters (β = 0.5, δ = 0.2, γ_0 = 1.2, γ = 1.6). The plots for 100 different runs are displayed. Of the three behaviors shown in Figure 8, **a** (quick convergence to the true solution) has become the preferred one.

[(a)] These expressions show an interesting similarity with the *difference map* algorithm discussed below.

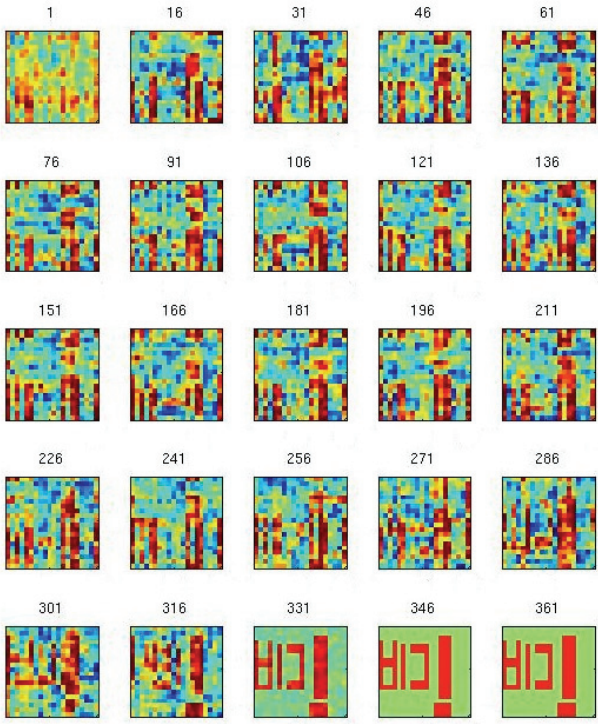


Figure 6. Snapshots of the density during its evolution, taken every 15 iterations. The abrupt change (Figure 6, case (a)) in the figure of merit (SDE) occurs near cycle 320, when the density suddenly begins to converge to the correct (binary) one. According to the color scale used here, negative values are represented in blue, and positive ones in red. Zero valued pixels are green.

different starting set of random phases). In each run, 1000 iterations were performed. Three cases can be identified:

(a) convergence to the true solution. It occurs suddenly, once the algorithm enters the basin of attraction of the solution after a chaotic trajectory. Very low values of SDE are attained (≈ 0.01). The density evolution during a converging run is shown in Figure 6.

(b) stagnation. At some moment the figure of merit begin to decrease, but slowly sets to a non-zero value (≈ 0.1) because some kind of trap has been entered.

(c) the trajectory extends over the performed 1000 iterations without entering any basin of attraction.

The dependence of the behaviour on the different parameters can be rationalized as:

• δ affects mostly the speed of convergence, which increases with δ until it rapidly goes to zero above $\delta \approx 4$, probably because the basins of attraction of the fixed points become very small.

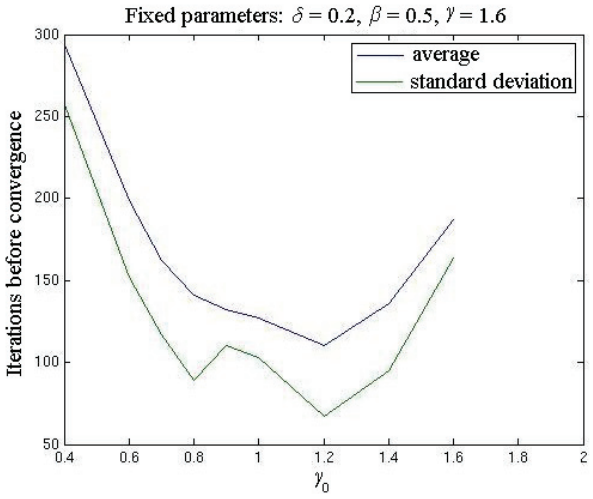


Figure 7. Optimization plot for the parameter γ_0 for fixed values of the other three parameters. A similar trend is observed for the general parameter γ.

• β and γ, since they determine the flipping magnitude, influence the ability of the algorithm to 'jump over' local minima (traps). Setting these parameters to small values leads to stagnation, while, at the other extreme, too high values prevent convergence. Since these two quantities play a similar role, they cannot be optimized independently; in fact, for each β value there exists a given range of γ in which convergence is possible (Figure 7).

The situation after choosing the optimal parameters can be seen in Figure 5b. Traps are avoided, and at the same time the basin of attraction of the true solution has been enlarged, so that the two unwanted situations (**b**) and (**c**) of Figure 5a are both much less probable. The number of iterations before convergence (I_C) probably depends on the ratio between the volume of attraction basins and the total volume of the search space; the I_C distribution (shown in Figure 8) is an exponential one, as expected for a memory-less process.

The algorithm does not need any knowledge about the support, but only about the fraction κ_1 of non-zero pixels in the solution (which relates to the zero frequency term through $\kappa_1 = F_0/N$, where N is the number of pixels); the object can appear anywhere in the cell and obviously the two possible enantiomorph choices are equally probable. Since the origin cannot be fixed *a priori*, such a kind of algorithm will always work with a $P1$ cell, independently from crystallographic symmetry, which cannot be taken into account. Symmetry can only emerge by itself and for this reason it could be used to test the correctness of the solution. For other phase retrieval algorithms without support it has been shown

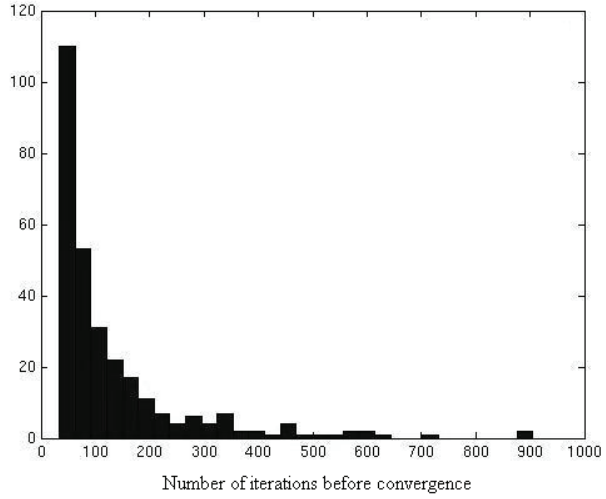


Figure 8. A histogram showing the distribution of number of iterations needed for convergence (a sort of trajectory length) for the binary flipping algorithm. The distribution has an approximately exponential decay, suggesting a memory-less process.

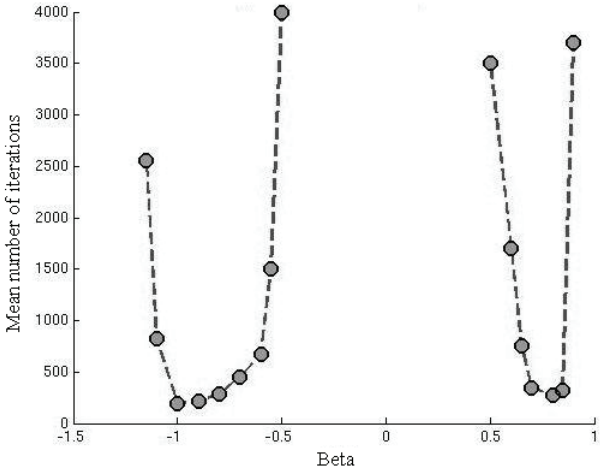


Figure 9. Optimization plot for the binary difference map. The average number of iterations needed for convergence is shown as function of the single parameter β. Two optimal ranges are found, the first (centered on $\beta = -1$, the global minimum) being larger and deeper.

that any attempt to fix the origin results in a reduced convergence speed, probably because the solution space collapses to a single point.

An alternative algorithm can be derived as a special case of the difference map $D = 1 + \beta\Delta$ with

$$\Delta = \Pi_{01}\left[(1+\beta^{-1})\Pi_{MOD} - \beta^{-1}\right] - \Pi_{MOD}\left[(1-\beta^{-1})\Pi_{01} + \beta^{-1}\right] \quad (25)$$

where the binary projector Π_{01} has been defined according to one of the two possible choices in Eq. (18). An advantage over the binary flipping algorithm is that the zero-frequency term F_0 can be unknown, as it will be found automatically by the algorithm itself; moreover, there is one single parameter to be optimized.

Various experiments have been conducted with different trial densities to determine the influence of β on the speed of convergence and to compare the behavior of the two algorithms. Two different optimal ranges of β have been found, one centered about –1 and the other about 0.8 (Figure 9). This is in agreement with the literature,[4] where the optimum values for the β parameter are found to be close to ±1. The comparison between binary flipping and difference map shows that their effectiveness varies greatly with the nature of the object to be reconstructed, but the dependence differs from one algorithm to the other. The two methods are, to some degree, complementary; putting aside very simple cases, often one of the two appears to perform well in those situations where the other exhibits a very slow convergence.

BINARY APPROXIMATIONS AND REAL CASES

Once established that a method existed to solve the binary problem, a more realistic case was considered, consisting in pseudo-molecular data in two dimensions. The moduli were obtained by Fourier transforming the density of benzene molecules projected onto the molecular plane. The cell was a square of 10 Å edge in which one, two or four benzene molecules had been placed. Data were used up to a resolution of 2 Å.

A binary approximation to the real density can be constructed by scaling the density and then setting a threshold z. The points with values higher than z are given the new value of 1 and the others of 0.

$$\rho_{01}(\mathbf{x}) = \begin{cases} 1 & : \rho(\mathbf{x}) \geq z \\ 0 & : \rho(\mathbf{x}) < z \end{cases} \qquad F(\rho_{01}(\mathbf{x})) = F_{01}(\mathbf{h}) \quad (26)$$

The structure factors corresponding to the binary density, $F_{01}(\mathbf{h})$, can be assumed proportional to the true ones, as in Ref. 6:

$$F_{01}(\mathbf{h}) \approx kF(\mathbf{h}) \quad (27)$$

where the constant k can be calculated from the knowledge of the fraction κ_1 of non-zero pixels in ρ_{01}:

$$k = \left[\frac{\kappa_1 - \kappa_1^2}{\sum_{\mathbf{h}\neq 0} |F(\mathbf{h})|^2}\right], \quad \kappa_1 = \frac{\left[\sum_i \rho_{01}(\mathbf{x}_i)\right]}{N} \quad (28)$$

The data from molecular structures were scaled in this way and then given as input to the binary flipping algorithm. No convergence was observed, for none of

the β, γ, δ parameter sets that had worked better for the ideal binary cases. This can be explained assuming that there is no intersection between the two constraint subsets, that is, no binary density exists that could reproduce the non-binary moduli. In fact, binarization of a density not only will affect the moduli in the chosen resolution sphere (in 2D, a circle), but it will create non-zero frequency components outside the sphere (where the original moduli had been set to be zero). To allow the two subsets to intersect in some point, out-of-sphere moduli should be allowed to deviate to some extent from their expected value of zero; it is not clear, however, if any physically meaningful solution could be found in this way.

A SIMPLIFIED SAYRE EQUATION FOR BINARY IMAGES

Another possibility for phasing diffraction data from a binary object can be derived outside the iterative methods context, taking inspiration from the Sayre equation.[8] While this relationship has been derived to exploit the atomicity property, it can be shown that it holds, in a simplified form, for binary densities too. In fact, the Sayre equation presupposes that density and squared density are related by convolution with a spread function *g*:

$$\rho = g \cdot \rho^2 \tag{29}$$

This is true for a density made of identical, well resolved, spherical peaks (equal atom structure); nevertheless, it is also consistent with a binary function, in which case g reduces to a constant. Assuming the density can take only the values 0 or a, we have

$$a\rho = \rho^2 \tag{30}$$

which in reciprocal space is equivalent to:

$$F_h = (aV)^{-1} \sum_{\mathbf{k}} \mathbf{F_k} \mathbf{F_{h-k}} \tag{31}$$

where V is the unit cell volume (in the 3D case). This convolution relationship would allow the solution search to be carried out entirely in reciprocal space, borrowing a variety of existing algorithms from the field of direct methods. Moreover, a binary approximation to a non-binary object can be found by minimizing the deviation between the two sides of the equation, while iterative algorithms fail in this task. In fact, from the lack of intersection between the constraint subsets follows that only a global minimum of the distance between the subsets can be searched. But this minimum is not qualitatively different from those non-meaningful local minima (traps) that a good algorithm is expected to avoid.

MODIFICATIONS OF THE CHARGE FLIPPING ALGORITHM

A possible criticism to the application of the binary flipping approach to non-binary density is that, while the lowest density region (corresponding to solvent in protein structures and to vacuum in small molecule structures) can be effectively assumed to be sharply distributed around zero, the object (molecular) density has a broader distribution. The behavior of the algorithm becomes more interesting after suppression of the flipping about the upper value of 1, letting β tend to 1 and γ to 0, and giving F_0 the freedom to vary during the iterations: the density of a single benzene ring in the cell could be slowly reconstructed. With these modifications, the algorithm reduces to the known method of *charge flipping*,[9] which alternates moduli projection to a change in sign of low-valued density:

$$\Gamma_{CF} = \Pi_{MOD} \Xi_0^{\delta} \tag{32}$$

$$\Xi_0^{\delta} : \rho \rightarrow \begin{cases} \rho & : \rho \geq \delta \\ -\rho & : \rho < \delta \end{cases} \quad (\delta > 0) \tag{33}$$

In term of projections, the flip operator can be written

$$\Xi_0^{\delta} = 2\Pi_{S(\delta)} - 1 \tag{34}$$

where $\Pi_{S(\delta)}$ stands for support projection. The important thing is that the support $S(\delta)$ is a dynamic one, being updated at each iteration by selecting the points with $\rho \geq \delta$. The *CF* algorithm has been proposed in crystallography for reconstructing atomic (< 1.2 Å) resolution structures, but it has been shown to be also applicable to the phase retrieval of non-periodic objects that lack atomicity. In both cases, however, the uniqueness of solution is guaranteed by the presence of extended regions of density with near-zero values and by (not strict) positivity. For non-atomic objects the algorithm tends more to stagnation, so that it has been used in conjunction with the *HiO* map: *CF* provides support evolution, while *HiO* drives to convergence because it is insensitive to traps.

The 2D benzene ring at 2 Å resolution does not display atomicity, but the presence of a vast majority of pixels with small absolute values of density still causes the solution to be unique. Because of the lack of atomicity sudden convergence is never observed; what happens is instead a slow, gradual approach to the solution. This good behaviour is compromised in going from one molecule to two and four molecules per cell, because the ratio of null pixels to the total number of pixels decreases. With two molecules, although the null pixels still occupy more than half of the cell, the algorithm fails in reconstructing the rings, whose density is rather flat, and shows a preference for 'peaky' solutions with higher

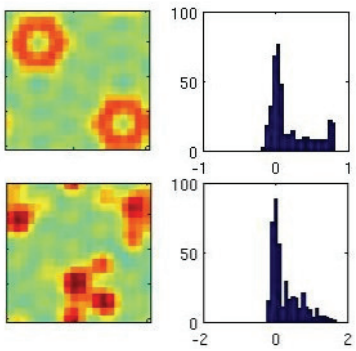


Figure 10. Test of the *CF* algorithm (δ = 0.2) on the 2D projection of two benzene molecules (2 Å resolution). Upper plots: density and its histogram for the true map. Lower plots: same for the reconstructed map.

density variance (Figure 10). The only way to find a solution with the required characteristics is to introduce new restraints; for example, an upper limit to density values can be used to force density flatness.

A choice that has been proven to be effective is to set a proportionality constant α between average density (calculated with the values above the flipping threshold δ) and the maximum allowed density s; at each *CF* cycle, the density values are modified by inversion about the expected maximum (*plateau*) value.

$$\rho \to s - \eta(\rho - s),\ s = \alpha \langle \rho \rangle_{\rho > \delta} \tag{35}$$

The value of s is calculated at each cycle. With this additional restraint, correct solutions could be found for the cases of 2 and 4 molecules per cell (Figure 11). The best values for the parameters were $\alpha \approx 1.3$, $\eta \approx 2$; the first one depends on the expected maximum value for the density, and can be varied only in a very narrow range if wrong solutions are to be avoided. A 3D case was then considered, to test if the modified *CF* algorithm with upper bound restraint could phase bigger structures. Synthetic trial data were calculated with the software SHELX[10] from the PDB coordinates of one molecule of *Fatty Acid Binding Protein* (FABP, PDB code 2HMB[11]). This protein comprises 131 aminoacids, organized in a β structure which defines an internal cavity. The reflections were computed from a single molecule positioned in a *P*1 cell (for simplicity, $a = b = c$, $\alpha = \beta = \gamma = 90°$ were chosen). Since zero density zones (which can be identified here with the solvent regions) define the degree of determinacy of the problem, different tests have been carried out varying the length of the cell edge, *i.e.*, the unit cell volume. The effect of data resolution was also investigated, across the range 20–2.5 Å.

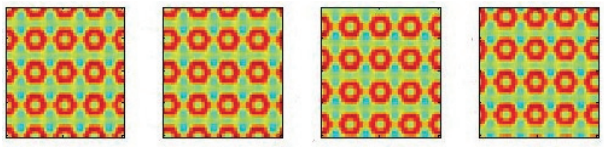


Figure 11. Densities reconstructed in four different runs of the upper-bounded *CF* algorithm. Four molecules per cell are present. Four unit cells are shown for clarity. Note the different origin positions, which depend on the (random) starting point.

It has been found that setting an upper bound for the density has no or little effect on converging to the correct solution, which could be retrieved in a small percentage of runs only when the solvent content is very high (at least 85 % of the unit cell volume, far too high to be found in any real crystal). This probably means that, below a given fraction of null pixels, the correct solution ceases to be a strong attractor for the *CF* algorithm, and this happens well before the problem becomes underdetermined. In fact it was noted that, even starting from the correct phases, there is a tendency to escape from the correct solution; the rate of this process increases with the flipping threshold δ. Another modification of the *CF* algorithm was tested in the perspective of phase extension applied to protein diffraction data. It consists in imposing on the electron density a topological restraint motivated by very general features of protein structures. A key process consists in dividing the image in the connected components, *i.e.*, separated features appearing in density when the isosurface for a given cutoff value is constructed. For a given threshold κ a mask Ω is defined as

$$\Omega_k = \{\mathbf{x} : \rho(\mathbf{x}) \geq \kappa\} \tag{36}$$

the set of points Ω_k can be decomposed in a number M of *connected components* $\omega_k^{(i)}$, each with a given volume $v_k^{(i)}$. A subset of points $\omega \subset \Omega$ is said to be a connected component when every pair of points $\{\mathbf{x}_1, \mathbf{x}_2\} \in \omega$ can be joined by a curve entirely contained in ω.

While connected component analysis identifies volume segments, without saying nothing about their shape, we can define a useful quantity for estimating the linear length of density pieces. This topological property, named *connectivity*, is computed by tracing the *skeleton*, that is, the set of lines joining all neighboring points above a given threshold.[12] With density defined on a grid, the procedure is to select grid points having density greater than 1.4 standard deviations above the mean, connecting by edges the points which are nearest neighbors. Two grid points belong to the same graph if they are connected by a continuous set of edges.

By means of the skeleton we define the connectivity as:

$$\text{Connectivity} = \frac{\text{number of points in the longest graph}}{\text{total number of points in graphs}} \quad (37)$$ [(b)]

This quantity is obviously a function of the threshold and the phase set. If the threshold is appropriately chosen, the global maximum for connectivity should coincide with the true phases, for which the electron density shows a single continuous polypeptide chain. Connectivity values relative to random phase sets are smaller than 0.1 while for correct phases a value above 0.9 is expected. It has been shown that the addition of an increasing phase error to the correct phase set always decreases the connectivity in a gradual way.

The connectivity restraint could be exploited into an iterative algorithm by selectively eliminating the densities that belong to the shortest graphs. Although it is impossible to know if those small segments would result to be correctly placed in the final density, one surely knows that correct density should not show small, isolated blobs. The idea is to force the density to evolve by growth of the longest fragments rather than by fusion of many small segments. An encouraging observation is that connectivity only depends on strong reflections and it is preserved even if a consistent fraction of moduli are given completely random phases (up to 80 % of the weakest ones – test carried out at 4 Å resolution).

An implementation was tried in this work using a weaker topological constraint, based on the segment volume rather than on graph length. The volume constraint is expected to be weaker than connectivity (as defined by Baker *et al.*) because it involves no restriction on the shape of the density; there is no reason to think that a general relationship between the volume of a connected component and its skeleton length should exist. However, for a densities in a neighborhood of the solution (so that the phase error is acceptable and connectivity is not too low) some kind of local relationship should arise, since the longest elements will also be the largest ones. For that reason, one expects that the requirement for the density to display a minimum number of volume elements $\omega_k^{(i)}$ could be used to improve or extend a set of known phases. Thus, a modified *CF* algorithm was devised, introducing supplementary real space operations:

- a binary mask is created to distinguish between points above and below a fixed threshold.
- a segmentation algorithm is used to identify the connected components into the density;
- a sorted list of segments is created on the basis of their volume (number of voxels);
- segments with volume below a certain minimum value $v_{\min}$ are set to zero in the density map.

The segmentation method used here was essentially the 'burning grass' algorithm described by Lunina *et al.*,[13] which consists in the following steps (Figure 12):

- *Initialization*: the points above the threshold are given a value 1, the others 0. No found components are present.
- *Search for a new component*: the nodes of the grid are scanned until a node with value '1' is found. The number of found components is increased by one. A 'current front' is defined as a set consisting of this node only. The new found component is marked with a consecutive number *m*. If no more '1' nodes are present the algorithm stops.
- *Isolation of a connected component*: the 'future front' is defined as the set of the nodes with value 1 that are neighbouring to one of the nodes of the 'current front'.
- *Propagation of the front*: the nodes of the current front are marked as belonging to the *m*-th component. The 'future front' becomes the 'current front' and

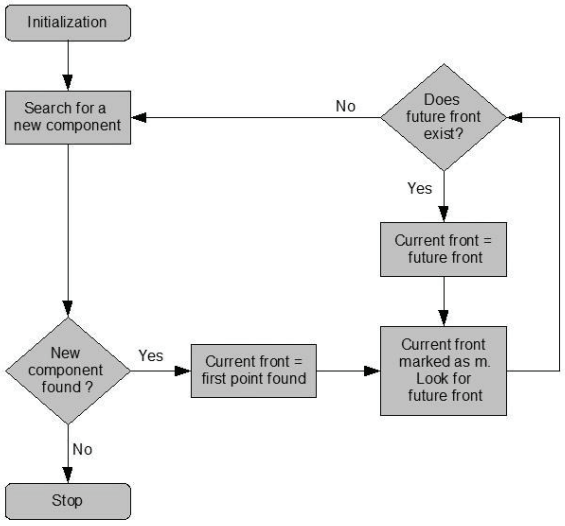


Figure 12. Flow chart for the 'burning grass' segmentation algorithm. Each *m*-th time a new initial point is found, the propagation loop is entered. The loop defines a 'future front' as the list of those points which are nearest-neighbors to points of the 'current front'; these latter are then marked as belonging to the *m*-th segment and the procedure is repeated until no more nearest neighbors are found and all the *m*-th connected components have been isolated.

[(b)] An alternative definition for connectivity is the total number of graphs.

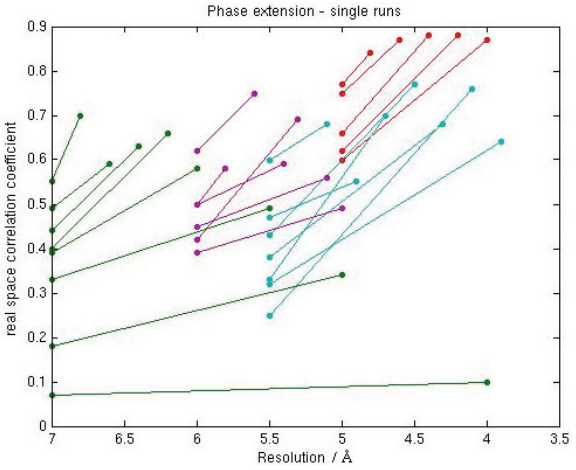


Figure 13. Starting and final correlation coefficient for some runs of the connectivity-restrained *CF* performed on ideal data from the protein FABP with an exact starting set. The correlation coefficient was calculated according to the relationship:

$$C_{\phi} = \frac{\sum_{\mathbf{h}} F^{obs}(\mathbf{h})^2 \cos[\phi_1(\mathbf{h}) - \phi_2(\mathbf{h})]}{\sum_{\mathbf{h}} F^{obs}(\mathbf{h})^2}$$

The map correlation coefficient has been reported as a function of starting and final resolution. Several hundreds of cycles were carried out, but in many cases the density ceased to evolve after only 50–100 iterations.

the algorithm goes back to the preceding point. This loop is repeated until the 'future front' is empty, then the search for a new component is performed.

This last variant of the *CF* algorithm has shown some phase extension power in a series of error-free tests conducted with a starting set of exact phases (Figure 13), that were extended to cover a larger sphere of reflections. Phases not belonging to the starting set were initially given random values, while known phases were kept constant in each run. It must be noted, however, that the algorithm is not able to improve a set of error-affected phases if these are given the freedom to vary from one cycle to the other. In fact, a divergent behavior was always observed in that case, probably because of underdeterminacy. For this reason, a phase combination step should be introduced; the best way to carry out the phase extension would probably follow the density modification scheme.

Acknowledgement. We are grateful to Anke Seydel for carefully reading the manuscript.

REFERENCES

1. H. Stark and Y. Yang, *Vector Space Projections*, John Wiley & Sons, New York , 1998.
2. R. W. Gerchberg and W. O. Saxton, *Optik* **35** (1972) 237–246.
3. J. R. Fienup, *Opt. Lett.* **3** (1978) 27–29.
4. V. Elser, *J. Opt. Soc. Am. A* **20** (2003) 40–55.
5. A. Goldstein and K. Y. J Zhang, *Acta Crystallogr., Sect D* **54** (1998) 1230–1244.
6. V. Y. Lunin, A. Urzhumtsev, and A. Bockmayr, *Acta Crystallogr., Sect. A* **58** (2002) 283–291.
7. J. Chergui, *GFT*, *Generic Fourier Transform*, copyright CNRS/ IDRIS, France, 2002, http://www.idris.fr/data/publications/GFT/
8. V. Y. Lunin, *Acta Crystallogr., Sect. A* **41** (1985) 551–556.
9. G. Oszlányi and A. Sütő, *Acta Crystallogr., Sect. A* **60** (2004) 134–41.
10. G. M. Sheldrick, *SHELX-96, Programs for X-Ray Crystallography*, University of Göttingen, 1996.
11. G. Zanotti, G. Scapin, P. Spadon, J. H. Veerkamp, and J. C. Sacchettini, *J. Biol. Chem.* **267** (1992) 18541–18550.
12. D. Baker, A. E. Krukowski, and D. A. Agard, *Acta Crystallogr., Sect. D* **49** (1993) 186–192.
13. N. Lunina, V. Lunin, and A. Urzhumtsev, *Acta Crystallogr., Sect. D* **59** (2003) 1702–1715.

SAŽETAK

Iterativne metode za rješavanje faznog problema u proteinskoj kristalografiji

Anton Thumiger[a] i Giuseppe Zanotti[a,b]

[a]*Department of Biological Chemistry, University of Padua, Viale G. Colombo 3, 35131 Padua, Italy*
[b]*Venetian Institute of Molecular Medicine (VIMM), Via Orus 2, 35129 Padua, Italy*

Problem faze predstavlja najveći izazov prilikom određivanja strukture rentgenskom difrakcijom. To naročito dolazi do izražaja kad su objekti proučavanja makromolekulski kristali koji slabo difraktiraju i sadrže mnogo atoma. Zbog toga uobičajene direktne metode, koje su pogodne za kristale malih i srednje velikih molekula, uglavnom ne mogu rješiti strukturu proteina čiji kristali ne difraktiraju do atomske rezolucije. U ovom je radu dan pregled nekih iterativnih metoda određivanja faza koje se koriste u optici. Prikazani su naši vlastiti rezultati pokušaja korištenja tih metoda i u makromolekulskoj kristalografiji. Binarno ograničenje elektronske gustoće ugrađeno je u novi itera-

tivni algoritam kao i u diferentnu mapu elektronske gustoće s namjerom određivanja kristalografskih faza. Drugi postojeći algoritam, "charge flipping", modificiran je da bi se ispitalo određivanje faza temeljeno na međusobnoj povezanosti atoma. Metoda binarnih gustoća nije polučila rezultate u realnim slučajevima no pokazalo se da kriterij konektivnosti ima određenog potencijala kao metoda za proširenje faze.