Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Direct phasing by binary integer programming

Vladimir Y. Lunin, Alexandre Urzhumtsev and Alexander Bockmayr

Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 19 October 2001
Accepted 1 March 2002

Direct phasing by binary integer programming

Vladimir Y. Lunin,[a,b] Alexandre Urzhumtsev[c]* and Alexander Bockmayr[b]

[a]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia, [b]LORIA, UMR 7503, Faculté des Sciences, Université Henri Poincaré, Nancy I, 54506 Vandoeuvre-les-Nancy, France, and [c]LCM3B, UMR 7036 CNRS, Faculté des Sciences, Université Henri Poincaré, Nancy I, 54506 Vandoeuvre-les-Nancy, France. Correspondence e-mail: sacha@lcm3b.uhp-nancy.fr

In the absence of phase information, a variety of electron-density distributions is consistent with the observed magnitudes. This ambiguity may be reduced significantly if the distribution values are restricted to 0 or 1, *i.e.* when the object of search is an envelope rather than a continuous electron-density distribution. The binarizing in both real (the grid-point density values) and reciprocal (the phases) spaces allows the usual structure-factor equations to be replaced by a system of linear inequalities with binary unknowns. A special computer procedure is applied to obtain several sets of values, which satisfy or almost satisfy these inequalities. The averaging of the found phase sets allows the final map to be calculated. The approach was tested with calculated and experimental data for a known protein structure. The size of the grid for the envelope calculation is at the moment the major limitation of the approach. Nevertheless, even for a very small grid, some structure information can be extracted and used as a starting point for further phase improvement or as a way to solve the molecular replacement problem.

1. Introduction

Binary integer programming (BIP below) is an approach for solving a system of linear inequalities in binary unknowns (0 or 1 in what follows). BIP methods have been demonstrated to be extremely useful in a large number of applications, but they have not been applied yet to the solution of the phase problem in crystallography. This paper discusses the ways to reduce the phase problem to a problem of BIP, to overcome methodological and technical difficulties, and to use this powerful method to solve crystallographic problems. The tests described below have been done with protein data while the approach does not depend on the size of the unit cell and the complexity of the studied object.

Crystallographic problems are usually formulated either in terms of real electron-density values or in terms of complex structure factors. These two sets of variables are linked unambiguously by a linear transformation (Fourier transform) if the electron-density values in all points of the unit cell and the full (infinite) set of structure factors are considered. In practice, the search is usually restricted to the density values calculated in the nodes of some grid in the unit cell, and to some subset of structure factors. The usual formulae for grid-function values and calculated structure factors contain some numerical errors, which can be neglected if the grid dimensions are large enough and high-resolution diffraction data are involved. However, the use of grids with a relatively small number of divisions along the unit-cell axes may require special caution.

Quite often, especially when working at low and middle resolution, crystallographers are interested in the position and the shape of the region with density values above a certain level, *i.e.* in a binary function representing this region. At low resolution, this function represents the part of the unit cell occupied by protein molecules, namely a molecular mask or an 'envelope'. If the resolution increases, this binary function may represent elements of the secondary structure or the trace of the polypeptide chain. Replacing the object of search by a binary mask has two important consequences. On the one hand, the restriction of the values to 0 or 1 may enormously reduce the number of possible solutions of the phase problem (see §4 for examples). On the other hand, the equations connecting the search values with the experimental structure factors are no longer strictly valid and require some corrections.

An attempt to restrict the density values to 0 or 1 is not unusual for protein crystallography. This property is equivalent to the condition

$$\rho(\mathbf{r}) = \rho^2(\mathbf{r}) \quad \text{for all } \mathbf{r}, \tag{1}$$

which is in turn equivalent to the equations

$$\mathbf{F}(\mathbf{h}) = (1/V_{\text{cell}}) \sum_{\mathbf{h}'+\mathbf{h}''=\mathbf{h}} \mathbf{F}(\mathbf{h}')\mathbf{F}(\mathbf{h}''). \tag{2}$$

A straightforward consequence of the last equations is the same tangent formula

$$\tan\varphi(\mathbf{h}) = \frac{\sum_{\mathbf{h}'+\mathbf{h}''=\mathbf{h}} F(\mathbf{h}')F(\mathbf{h}'')\sin[\varphi(\mathbf{h}')+\varphi(\mathbf{h}'')]}{\sum_{\mathbf{h}'+\mathbf{h}''=\mathbf{h}} F(\mathbf{h}')F(\mathbf{h}'')\cos[\varphi(\mathbf{h}')+\varphi(\mathbf{h}'')]}, \qquad (3)$$

which follows from the famous Sayre equations (Sayre, 1952) and is extremely widely applied in crystallography. Therefore, the use of the tangent formula may be considered to some extent as an attempt to involve {0 or 1} restriction in phase refinement (Lunin, 1985). A direct binarization of the density in the density-modification procedures has been tried by Cannillo *et al.* (1983). Another way to introduce binary functions is to use wavelet-type approximations of density distributions and to apply {0 or 1} restriction to the wavelet coefficients rather than to the density values (Lunin, 2000).

Originally, the electron-density values are linked to the complex structure factors by linear equations. However, if only the magnitudes of the structure factors are supposed to be known, then the unknown phases and electron-density values are connected with the magnitudes by non-linear equations. These equations may still be considered as linear for centric reflections. Here the phase may take one of only two possible values, and phase uncertainty may be represented by one additional binary variable linearly included in the corresponding equations. For the acentric reflection, the phase can take any value from 0 to 2π. In this case, one may accept the approximation that the phase of the structure factor is restricted to one of four possible values $\pm\pi/4$, $\pm3\pi/4$, and the phase uncertainty may be coded by two additional binary variables, which are linked linearly to the density values.

Being reduced to a BIP problem, the three-dimensional phase problem presents a challenge to BIP methods owing to a large number of unknowns, even when a relatively small grid for density calculation is considered. For this large number of variables, it is often not possible to compute an exact solution of the BIP problem. Local search methods, like those realized in the program *WSATOIP* (Walser, 1997, 1998), are a possible way to overcome this difficulty. This procedure begins the search with some randomly generated start values for the binary unknowns and then tries to improve the solution locally. In general, the optimization does not result in the exact solution but in an 'improved' one, compared to the starting point. Such a local search procedure may be combined with the general approach of low-resolution phasing suggested last decade (see Lunin *et al.*, 2000, for a review). In this approach, a large number of solutions found by a local search starting from random starting points are averaged to get an approximate answer. It must be noted that the phase solutions found by this procedure may correspond to different choices of the origin and enantiomer. Therefore, alignment of phases must be performed before comparing or averaging different solutions (Lunin & Lunina, 1996).

One more feature of crystallographic problems is that usually the electron density possesses some crystallographic symmetry, and not all grid values are independent. In this case, either additional symmetry constraints may be applied or a subset of independent variables may be identified. The second way has been used in our tests.

More details of the suggested approach are discussed below and the results of the first tests are presented.

2. The phase problem and binary integer programming

2.1. Basic equations

We start from the usual formulae that link a real electron-density distribution to its complex structure factors:

$$\rho(\mathbf{x}) = (1/V_{\rm cell}) \sum_{\mathbf{h}\in\mathbf{Z}^3} \mathbf{F}(\mathbf{h}) \exp[-2\pi i(\mathbf{h},\mathbf{x})], \qquad \mathbf{x}\in V, \quad (4)$$

$$\mathbf{F}(\mathbf{h}) = \int_V \rho(\mathbf{x}) \exp[2\pi i(\mathbf{h},\mathbf{x})]\,{\rm d}\mathbf{x}, \qquad \mathbf{h}\in\mathbf{Z}^3. \quad (5)$$

Here, $\mathbf{x} = (x_1, x_2, x_3)^{\rm T}$ and $\mathbf{h} = (h_1, h_2, h_3)^{\rm T}$ represent real- and reciprocal-space vectors, respectively, $V_{\rm cell}$ is the volume of the unit cell V and $\mathbf{Z}^3$ are the vectors with integer coordinates

$$\mathbf{Z}^3 = \{\mathbf{h} = (h_1, h_2, h_3)^{\rm T} : h_1, h_2, h_3 \text{ are integers}\}. \qquad (6)$$

Written in fractional coordinates $\mathbf{x}$, the density distribution $\rho(\mathbf{x})$ has a periodicity with integer periods along all three axes. The structure factors $\mathbf{F}(\mathbf{h}) = F(\mathbf{h})\exp[i\varphi(\mathbf{h})]$ do not reveal any periodicity but fall exponentially when indices increase. If the distribution $\rho(\mathbf{x})$ displays the symmetries of a space group $\Gamma = \{(\mathbf{R}_\nu, \mathbf{t}_\nu)\}_{\nu=1}^{\rm nsym}$:

$$\rho(\mathbf{R}_\nu\mathbf{x} + \mathbf{t}_\nu) = \rho(\mathbf{x}) \quad \text{for all } \mathbf{x} \text{ and } \nu, \qquad (7)$$

then the structure factors reveal the symmetry

$$\mathbf{F}(\mathbf{R}_\nu^{\rm T}\mathbf{h}) = \mathbf{F}(\mathbf{h})\exp[-2\pi i(\mathbf{h},\mathbf{t}_\nu)] \quad \text{for all } \mathbf{h} \text{ and } \nu. \qquad (8)$$

This latter equation implies the extinction conditions:

$$\text{if } \mathbf{R}_\nu^{\rm T}\mathbf{h} = \mathbf{h} \text{ and } (\mathbf{h},\mathbf{t}_\nu)|_{\rm mod1} \neq 0 \text{ then } \mathbf{F}(\mathbf{h}) = 0, \qquad (9)$$

and being coupled with Hermitian symmetry of the structure factors implies the phase restrictions for centric reflections:

$$\text{if } \mathbf{R}_\nu^{\rm T}\mathbf{h} = -\mathbf{h} \text{ then } \varphi(\mathbf{h}) = \psi(\mathbf{h}) \text{ or } \varphi(\mathbf{h}) = \psi(\mathbf{h}) + \pi \qquad (10)$$

with

$$\psi(\mathbf{h}) = \pi(\mathbf{h},\mathbf{t}_\nu). \qquad (11)$$

2.2. Grid functions and grid structure factors

Equations (4)–(5) link unambiguously the structure factors to the electron-density distribution when the full infinite set of structure factors and density values in all points of the unit cell are involved. In practice, the electron-density values are calculated at some grid in the unit cell and the set of structure factors is finite. Let M_1, M_2, M_3 be the number of divisions along the unit-cell axes. Let us suppose also that these numbers are consistent with the symmetry, *i.e.* all symmetry transformations leave the grid invariant. Let $\mathbf{M} = {\rm diag}(M_1, M_2, M_3)$ stand for the diagonal matrix with the diagonal formed by M_1, M_2, M_3, Π is the set of all grid points in the unit cell and $|\mathbf{M}| = M_1M_2M_3$ is the total number of these points:

$$\begin{aligned}\Pi = \{\mathbf{j} = (j_1, j_2, j_3)^{\mathrm{T}} : j_1, j_2, j_3 \text{ are integers;}\\ 0 \le j_1 < M_1; 0 \le j_2 < M_2; 0 \le j_3 < M_3\}.\end{aligned} \tag{12}$$

We introduce the *grid electron-density function* $\{\rho^g(\mathbf{j})\}$ as a set of values of the density distribution at the grid points:

$$\rho^g(\mathbf{j}) = \rho\left(\frac{j_1}{M_1}, \frac{j_2}{M_2}, \frac{j_3}{M_3}\right) = \rho(\mathbf{M}^{-1}\mathbf{j}), \quad \mathbf{j} \in \Pi, \tag{13}$$

and define the *grid structure factors* by the inverse discrete Fourier transform (IDFT):

$$\mathbf{F}^g(\mathbf{h}) = (1/|\mathbf{M}|) \sum_{\mathbf{j}\in\Pi} \rho^g(\mathbf{j}) \exp[2\pi i(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})], \quad \mathbf{h} \in \Pi. \tag{14}$$

The discrete Fourier transform (DFT) may restore the grid density function unambiguously from the grid structure factors:

$$\rho^g(\mathbf{j}) = \sum_{\mathbf{h}\in\Pi} \mathbf{F}^g(\mathbf{h}) \exp[-2\pi i(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})], \quad \mathbf{j} \in \Pi, \tag{15}$$

but the values of the density distribution in the intermediate points cannot be retrieved.

The discrete Fourier transform formulae (14) and (15) are defined primarily for the points of the set Π, but they can be extended to all integer vectors in $\mathbf{Z}^3$ supposing that $\{\rho_{\mathbf{j}}\}$ and $\{\mathbf{F}^g_{\mathbf{h}}\}$ are periodical functions with M_1, M_2, M_3 periods along the axes:

$$\rho^g(\mathbf{j} + \mathbf{M}\mathbf{k}) = \rho^g(\mathbf{j}), \mathbf{F}^g(\mathbf{h} + \mathbf{M}\mathbf{k}) = \mathbf{F}^g(\mathbf{h}), \text{ for every } \mathbf{k} \in \mathbf{Z}^3. \tag{16}$$

The periodicity of the grid function is natural and reflects the periodicity of an electron-density distribution in a crystal. However, the periodicity of the grid structure factors is quite different from the behaviour of the usual structure factors. Other properties that are different for the grid and the usual structure factors are the extinction conditions and the restrictions for centric phases. Owing to the periodicity (16) of the grid structure factors, these conditions have now the form:

$$\text{if } \mathbf{R}_\nu^{\mathrm{T}}\mathbf{h} = \mathbf{h}|_{\mathrm{mod}\mathbf{M}} \text{ and } (\mathbf{h}, \mathbf{t}_\nu)|_{\mathrm{mod}1} \ne 0 \text{ then } \mathbf{F}^g(\mathbf{h}) = 0, \tag{17}$$

$$\begin{aligned}\text{if } \mathbf{R}_\nu^{\mathrm{T}}\mathbf{h} = -\mathbf{h}|_{\mathrm{mod}\mathbf{M}} \text{ then } \varphi^g(\mathbf{h}) = \psi(\mathbf{h}) = \pi(\mathbf{h}, \mathbf{t}_\nu)\\ \text{or } \varphi^g(\mathbf{h}) = \psi(\mathbf{h}) + \pi.\end{aligned} \tag{18}$$

In particular, this means that not only is the $\mathbf{F}^g(0, 0, 0)$ term real, but $\mathbf{F}^g(M_1/2, 0, 0)$, $\mathbf{F}^g(0, M_2/2, 0)$, $\mathbf{F}^g(0, 0, M_3/2)$, $\ldots$, $\mathbf{F}^g(M_1/2, M_2/2, M_3/2)$ are real too. Additionally, the grid structure factors may be absent or reveal properties of centric reflections if one of the indices is equal to half of the corresponding period.

If the grid is fine enough (*i.e.* M_1, M_2, M_3 are large), (5) suggests the common way to calculate structure factors from a model (Sayre, 1951; Ten Eyck, 1977), namely to calculate first the electron density at grid points and then to perform the IDFT (14):

$$\begin{aligned}\mathbf{F}(h_1, h_2, h_3) &\approx \frac{V_{\mathrm{cell}}}{M_1M_2M_3} \sum_{\mathbf{j}\in\Pi} \rho\left(\frac{j_1}{M_1}, \frac{j_2}{M_2}, \frac{j_3}{M_3}\right)\\ &\quad \times \exp\left[2\pi i\left(\frac{h_1j_1}{M_1} + \frac{h_2j_2}{M_2} + \frac{h_3j_3}{M_3}\right)\right]\\ &= \frac{V_{\mathrm{cell}}}{|\mathbf{M}|} \sum \rho(\mathbf{j}) \exp[2\pi i(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\\ &= V_{\mathrm{cell}}\mathbf{F}^g(\mathbf{h}).\end{aligned} \tag{19}$$

Generally speaking, this formula is an approximate one and the error may be significant if the division numbers M_1, M_2, M_3 are small or at least one of the structure-factor indices is large (close to half of the corresponding number of grid points). Sometimes the error may be estimated from the precise formula that connects the grid structure factors with the usual ones (Ten Eyck, 1973):

$$V_{\mathrm{cell}}\mathbf{F}^g(\mathbf{h}) = \mathbf{F}(\mathbf{h}) + \sum_{\substack{\mathbf{k}\in\mathbf{Z}^3\\ \mathbf{k}\ne\mathbf{0}}} \mathbf{F}(\mathbf{h} + \mathbf{M}\mathbf{k}). \tag{20}$$

If a Fourier synthesis of a finite resolution $d_{\min}$ is calculated at the grid whose step is less than $d_{\min}/2$, then all structure factors in the sum on the right-hand side of (20) are supposed to be zero. The formula (19) is therefore exact for finite resolution syntheses calculated at fine enough grids.

2.3. The phase problem as a binary integer programming problem

The main goal of this section is to derive equations or inequalities that allow one to define the grid electron-density values $\{\rho^g(\mathbf{j})\}$ provided the structure-factor magnitudes $\{F(\mathbf{h})\}$ are known. Using formulae (14) and (20), one can write down a system of linear equations defining the values of the grid function $\{\rho^g(\mathbf{j})\}$ in the form

$$\begin{aligned}&\sum_{\mathbf{j}\in\Pi} \cos[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j})\\ &\quad = (|\mathbf{M}|/V_{\mathrm{cell}})F(\mathbf{h}) \cos\varphi(\mathbf{h}) + \operatorname{Re}\mathbf{R}(\mathbf{h}), \quad \mathbf{h} \in \Pi\\ &\sum_{\mathbf{j}\in\Pi} \sin[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j})\\ &\quad = (|\mathbf{M}|/V_{\mathrm{cell}})F(\mathbf{h}) \sin\varphi(\mathbf{h}) + \operatorname{Im}\mathbf{R}(\mathbf{h}), \quad \mathbf{h} \in \Pi,\end{aligned} \tag{21}$$

where

$$\mathbf{R}(\mathbf{h}) = (|\mathbf{M}|/V_{\mathrm{cell}}) \sum_{\substack{\mathbf{k}\in\mathbf{Z}^3\\ \mathbf{k}\ne\mathbf{0}}} \mathbf{F}(\mathbf{h} + \mathbf{M}\mathbf{k}).$$

These equations are linear with respect to the grid density values if both the magnitudes and phases (or the real and imaginary parts) of the structure factors are supposed to be known. However, if not only the density values but also the phases are considered to be unknown, then the equations become non-linear as the phases enter as an argument of trigonometric functions.

The value of $\mathbf{R}(\mathbf{h})$ depends on magnitudes and phases of all structure factors and is generally unknown. Therefore, equations (21) cannot be written in the precise form. The value $R(\mathbf{h})$ may be negligibly small if the grid is fine enough and if the indexes $\mathbf{h}$ are relatively small in comparison with the grid

Table 1
Quality of the approximation of the observed magnitudes by values calculated from binary maps.

The correlation coefficient

$$C_F = \sum_{\mathbf{h}}[F^{\rm bin}(\mathbf{h}) - \langle F^{\rm bin}\rangle][F^{\rm obs}(\mathbf{h}) - \langle F^{\rm obs}\rangle] \bigg/ \left\{ \sum_{\mathbf{h}}[F^{\rm bin}(\mathbf{h}) - \langle F^{\rm bin}\rangle]^2 \sum_{\mathbf{h}}[F^{\rm obs}(\mathbf{h}) - \langle F^{\rm obs}\rangle]^2 \right\}^{1/2}$$

is presented for different resolution zones. The molecular volume defines the number of non-zero grid values. It was adapted for every grid to have the maximal correlation coefficient.

	Resolution range (Å) (number of independent reflections)				
Grid (mol. vol., %)	16–∞ (15)	12–∞ (28)	8–∞ (85)	5–∞ (305)	4–∞ (580)
6*6*6 (50)	0.32	0.39	–	–	–
8*8*8 (35)	0.88	0.92	0.	–	–
10*10*10 (30)	0.68	0.73	0.68	–	–
16*16*16 (20)	0.91	0.79	0.69	0.62	0.03

dimensions. At the same time, it may be significant if one of the indices is close to $M_1/2$, $M_2/2$, $M_3/2$.

In general, this expression can be estimated by the sum of structure-factor magnitudes as

$$|\mathbf{R}(\mathbf{h})| \le \bar{\varepsilon}_1(\mathbf{h}) = (|M|/V_{\rm cell}) \sum_{\substack{\mathbf{k}\in\mathbf{Z}^3\\ \mathbf{k}\neq\mathbf{0}}} F(\mathbf{h}+\mathbf{M}\mathbf{k}). \qquad (22)$$

More sophisticated estimates will be discussed elsewhere. In any case, if an estimate

$$|\mathbf{R}(\mathbf{h})| \le \varepsilon_1(\mathbf{h}) \qquad (23)$$

exists, equations (21) may be replaced by a system of inequalities that restrict the density values in a weaker form but do not require the knowledge of all structure factors:

$$\begin{aligned} -\varepsilon_1(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \cos[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) - (|\mathbf{M}|/V_{\rm cell})F(\mathbf{h}) \\ &\quad \times \cos\varphi(\mathbf{h}) \le \varepsilon_1(\mathbf{h}), \quad \mathbf{h}\in\Pi \\ -\varepsilon_1(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \sin[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) - (|\mathbf{M}|/V_{\rm cell})F(\mathbf{h}) \\ &\quad \times \sin\varphi(\mathbf{h}) \le \varepsilon_1(\mathbf{h}), \quad \mathbf{h}\in\Pi. \end{aligned} \qquad (24)$$

The inequalities (24) contain the phase values $\varphi(\mathbf{h})$ that cannot be determined directly in an X-ray experiment and are the object of our search. The phases enter the inequalities in a non-linear manner. However, if the reflection **h** is centric [*i.e.* it satisfies the condition in (18)], then only two values of the phase, $\psi(\mathbf{h})$ or $\psi(\mathbf{h})+\pi$, with ψ being known, are possible, and (24) may be written as

$$\begin{aligned} -\varepsilon_1(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \cos[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) - \alpha(\mathbf{h})(|\mathbf{M}|/V_{\rm cell})F(\mathbf{h}) \\ &\quad \times \cos\psi(\mathbf{h}) \le \varepsilon_1(\mathbf{h}), \\ -\varepsilon_1(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \sin[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) - \alpha(\mathbf{h})(|\mathbf{M}|/V_{\rm cell})F(\mathbf{h}) \\ &\quad \times \sin\psi(\mathbf{h}) \le \varepsilon_1(\mathbf{h}), \\ &\qquad\qquad \text{for centric } \mathbf{h}. \end{aligned} \qquad (25)$$

Here, the phase ambiguity is represented by a new unknown $\alpha(\mathbf{h})$, which takes one of the two values 1 or -1 and which enters into the inequalities in a linear way. The inequalities (25) become linear $\{\rho^g(\mathbf{j})\}$ and $\{\alpha(\mathbf{h})\}$ provided the structure-factor magnitudes $\{F(\mathbf{h})\}$ are known.

For acentric reflections, one may assume as usual that the phase $\varphi(\mathbf{h})$ can take one of four values: $\pm\pi/4$, $\pm3\pi/4$. Under this hypothesis, the inequalities become

$$\begin{aligned} -\varepsilon_1(\mathbf{h}) - \varepsilon_2(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \cos[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) - \alpha(\mathbf{h})(|\mathbf{M}|/V_{\rm cell}) \\ &\quad \times F(\mathbf{h})2^{-1/2} \le \varepsilon_1(\mathbf{h}) + \varepsilon_2(\mathbf{h}), \\ -\varepsilon_1(\mathbf{h}) - \varepsilon_2(\mathbf{h}) &\le \sum_{\mathbf{j}\in\Pi} \sin[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) - \beta(\mathbf{h})(|\mathbf{M}|/V_{\rm cell}) \\ &\quad \times F(\mathbf{h})2^{-1/2} \le \varepsilon_1(\mathbf{h}) + \varepsilon_2(\mathbf{h}), \\ &\qquad\qquad \text{for acentric } \mathbf{h}, \end{aligned} \qquad (26)$$

where the unknowns $\alpha(\mathbf{h})$ and $\beta(\mathbf{h})$ take one of the two values 1 or -1, and enter the inequalities in a linear way. Here, $\varepsilon_2(\mathbf{h})$ reflects the error introduced by the sampling of the phase value. It can be estimated by

$$\varepsilon_2(\mathbf{h}) \le \bar{\varepsilon}_2(\mathbf{h}) = 2^{-1/2}(|\mathbf{M}|/V_{\rm cell})F(\mathbf{h}). \qquad (27)$$

As a result, we get a system of linear inequalities (26) where the unknowns are the values of the electron density at the grid points $\{\rho^g(\mathbf{j})\}$ and where the additional variables $\alpha(\mathbf{h})$ and $\beta(\mathbf{h})$ represent phase ambiguity. These inequalities are weaker than the initial equations but they reduce the phase problem to linear integer programming, while initially the phase problem is essentially non-linear.

2.4. Binary distributions and integer programming

The inequalities (25)–(26) do not solve the phase ambiguity, as a lot of grid functions corresponding to differently phased magnitudes may satisfy them (see §4.2). The number of possible solutions may be reduced significantly if it is supposed in addition that all unknowns are binary variables. The substitution of the search of the original density distribution by the search of a binary function is not artificial because some basic information obtained from crystallographic Fourier synthesis is the shape of the region of high-density values and not particular values of the electron-density distribution. An exception are ultra-high resolution studies (Lecomte, 1999), where the electron density itself is the subject of investigation. The complexity of this binary function depends on the current resolution. At very low resolution, this binary function (mask function or envelope) presents the overall shape of molecules

Table 2
Map correlation coefficient for phases calculated from binary maps.

The map correlation coefficient

$$C_\varphi = \sum_{\mathbf{h}} F^{\rm obs}(\mathbf{h})^2 \cos[\varphi^{\rm bin}(\mathbf{h}) - \varphi^{\rm exact}(\mathbf{h})] \Big/ \sum_{\mathbf{h}} F^{\rm obs}(\mathbf{h})^2$$

is presented for different resolution zones. The molecular volume defines the number of non-zero grid values. It was adapted for every grid to have the maximal correlation coefficient.

	Resolution range (Å) (number of independent reflections)				
Grid (mol. vol., %)	16–∞ (15)	12–∞ (28)	8–∞ (85)	5–∞ (305)	4–∞ (580)
6*6*6 (50)	0.93	0.74	–	–	–
8*8*8 (35)	0.98	0.94	0.80	–	–
10*10*10 (30)	0.98	0.96	0.90	–	–
16*16*16 (20)	0.99	0.99	0.94	0.87	0.81

and their packing in the unit cell. At a middle resolution, some elements of the secondary structure may appear ('detailed envelope'), *e.g.* helices may by presented by cylindrical regions. At higher resolution, the mask may show the trace of the polypeptide chain and the position of the residues.

The problem appearing immediately with the introduction of binary variables $\{\rho^{\rm bin}(\mathbf{j})\}$ is that the X-ray experiment provides magnitudes $\{F^{\rm obs}(\mathbf{h})\}$ corresponding to a real electron density and not to a binary function approximating it. Nevertheless, tests show (see Tables 1 and 2 and Fig. 1) that the correlation between initial (observed) structure factors and those calculated from binary envelopes may be high enough. So this problem may be overcome by appropriately scaling the observed magnitudes (see Appendix *A*) and by increasing the gaps $\varepsilon(\mathbf{h})$ in (25)–(26). The inequalities may now be written as

$$\begin{aligned} -\varepsilon_{\mathbf{h}} - c^R_{\mathbf{h}} \le \sum_{\mathbf{j}\in\Pi} a^R_{\mathbf{j}} z_{\mathbf{j}} + b^R_{\mathbf{h}} y^R_{\mathbf{h}} \le -c^R_{\mathbf{h}} + \varepsilon_{\mathbf{h}}, \quad \mathbf{h}\in\Pi,\\ -\varepsilon_{\mathbf{h}} - c^I_{\mathbf{h}} \le \sum_{\mathbf{j}\in\Pi} a^I_{\mathbf{j}} z_{\mathbf{j}} + b^I_{\mathbf{h}} y^I_{\mathbf{h}} \le -c^I_{\mathbf{h}} + \varepsilon_{\mathbf{h}}, \quad \mathbf{h}\in\Pi, \end{aligned} \tag{28}$$

where $\{z_{\mathbf{j}}\}_{\mathbf{j}\in\Pi}$, $\{y^R_{\mathbf{h}}, y^I_{\mathbf{h}}\}_{\mathbf{h}\in\Pi}$ are binary variables, which take 0 or 1 values only;

$$y^R_{\mathbf{h}} = [\alpha(\mathbf{h}) + 1]/2, \quad y^I_{\mathbf{h}} = [\beta(\mathbf{h}) + 1/2], \quad (y^R_{\mathbf{h}} = y^I_{\mathbf{h}} \text{ for centric reflections}), \tag{29}$$

$$a^R_{\mathbf{j}} = \cos[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})], \quad a^I_{\mathbf{j}} = \sin[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})], \tag{30}$$

$$b^R_{\mathbf{h}} = -2\kappa F(\mathbf{h})\cos\psi(\mathbf{h}), \quad b^I_{\mathbf{h}} = -2\kappa F(\mathbf{h})\sin\psi(\mathbf{h}), \quad \text{for the centric case}, \tag{31}$$

$$b^R_{\mathbf{h}} = -2\kappa F(\mathbf{h})2^{-1/2}, \quad b^I_{\mathbf{h}} = -2\kappa F(\mathbf{h})2^{-1/2}, \quad \text{for the acentric case}, \tag{32}$$

$$c^R_{\mathbf{h}} = -\kappa F(\mathbf{h})\cos\psi(\mathbf{h}), \quad c^I_{\mathbf{h}} = -\kappa F(\mathbf{h})\sin\psi(\mathbf{h}), \quad \text{for the centric case}, \tag{33}$$

$$c^R_{\mathbf{h}} = -\kappa F(\mathbf{h})2^{-1/2}, \quad c^I_{\mathbf{h}} = -\kappa F(\mathbf{h})2^{-1/2}, \quad \text{for the acentric case}. \tag{34}$$

κ is a scale factor that reduces the observed magnitudes to a 'binary function scale' (see Appendix *A*), and the gap $\varepsilon_{\mathbf{h}}$ reflects three kinds of errors, namely grid sampling errors $\varepsilon_1(\mathbf{h})$, phase sampling errors $\varepsilon_2(\mathbf{h})$ and errors due to replacing the real density distribution by a binary function.

2.5. Symmetry restrictions

If the density distribution possesses a crystallographic symmetry, then the corresponding grid structure factors exhibit the symmetry (17)–(18) and the equations in (21) corresponding to symmetry-related indices are linearly dependent. So the number of equations in (21) and correspondingly the number of inequalities in (28) may be reduced by deleting the dependent ones.

The grid point values $\{\rho^g(\mathbf{j})\}$ are related by symmetry too. A subset of independent values may be selected in this case, and the inequalities may be expressed in these independent unknowns by summation of the coefficients in (28) corresponding to symmetry-related points.

3. Solution of the BIP problem

Linear equations and inequalities in binary variables can be solved by integer programming methods. The general form of an *integer linear programming problem* is

$$\max\{\mathbf{c}^{\rm T}\mathbf{x} | A\mathbf{x} \le \mathbf{b}, \mathbf{x}\in\mathbf{Z}^n\} \tag{35}$$

with a real matrix A of a dimension $m\times n$ and vectors $\mathbf{c}\in\mathbf{R}^n$, $\mathbf{b}\in\mathbf{R}^m$, $\mathbf{c}^{\rm T}\mathbf{x}$ being the scalar product of the vectors $\mathbf{c}$ and $\mathbf{x}$. If the system $A\mathbf{x}\le\mathbf{b}$ includes the constraints $\mathbf{0}\le\mathbf{x}\le\mathbf{1}$, we get a *binary integer linear programming problem (BIP).* A vector $\mathbf{x}^*$ in $\mathbf{Z}^n$ with $A\mathbf{x}^*\le\mathbf{b}$ is called a *feasible solution.* If moreover, $\mathbf{c}^{\rm T}\mathbf{x}^* = \max\{\mathbf{c}^{\rm T}\mathbf{x}|A\mathbf{x}\le\mathbf{b}, \mathbf{x}\in\mathbf{Z}^n\}$, then $\mathbf{x}^*$ is called an optimal solution and $\mathbf{c}^{\rm T}\mathbf{x}^*$ the optimal value. The inequalities (28) form a particular case of general BIP problems as there is no objective function $Q(\mathbf{x}) = \mathbf{c}^{\rm T}\mathbf{x}$ here and the aim is to find all (or at least some) feasible solutions. Additional constraints on the density values may be incorporated by the appropriate choice of an objective function.

Integer linear programming has been studied in mathematics, computer science and operations research for more than 40 years (Bockmayr & Kasper, 1998; Johnson *et al.*, 2000). A huge number of large-scale combinatorial problems can be naturally modelled and solved in this framework. Various exact and heuristic solution procedures have been developed, among them branch-and-bound, branch-and-cut, branch-and-price, and local search. Many commercial and public domain software packages are available in order to compute feasible or optimal solutions.

In our test, we used an approach that combines local search for the solution of BIP problems (Walser, 1997, 1998) with a general strategy of low-resolution phasing developed recently by Lunin *et al.* (2000). First, a set of random initial assignments of values to the binary variables is generated. From every initial assignment, one tries to find a feasible solution of (28) by local flips of the binary variables. This is done by the procedure *WSATOIP* (Walser, 1997, 1998). At each run, the program will try to minimize a *residual*, which is defined on the basis of (28) as

$$R=\sum_{\mathbf{h}}\left[r\left(\sum_{\mathbf{j}}a_{\mathbf{j}}^{R}z_{\mathbf{j}}+b_{\mathbf{h}}^{R}y_{\mathbf{h}}^{R};\,c_{\mathbf{h}}^{R},\,\varepsilon_{\mathbf{h}}\right)+r\left(\sum_{\mathbf{j}}a_{\mathbf{j}}^{I}z_{\mathbf{j}}+b_{\mathbf{h}}^{I}y_{\mathbf{h}}^{I};\,c_{\mathbf{h}}^{I},\,\varepsilon_{\mathbf{h}}\right)\right]. \tag{36}$$

Here

$$r(x;\,q,\,\varepsilon)=\begin{cases}0 & \text{if } -\varepsilon+q\le x\le q+\varepsilon\\ x-(q+\varepsilon) & \text{if } x>q+\varepsilon\\ (q-\varepsilon)-x & \text{if } x<q-\varepsilon\end{cases} \tag{37}$$

so that $r(x;\,q,\,\varepsilon)=0$ if the inequality $-\varepsilon+q\le x\le q+\varepsilon$ is satisfied and $r(x;\,q,\,\varepsilon)$ grows linearly with x otherwise (see Fig. 2). The program stops if the residual has been reduced to 0 (*i.e.* a feasible solution has been found) or if a given maximal number of flips $N_{\rm flip}$ has been reached. So the result of a particular run is not always a feasible solution but a final assignment where the initial residue has been reduced.

For every final assignment, the phases corresponding to the binary function $\{z_{\mathbf{j}}^{\rm fin}\}_{\mathbf{j}\in\Pi}$ are calculated and used together with the observed magnitudes to obtain Fourier syntheses. The calculated syntheses are aligned according to permitted origin and enantiomer choices (Lunin & Lunina, 1996). Then they are averaged to produce a single phase set. In this way, a centroid phase value and an individual figure of merit are defined for every reflection (Lunin *et al.*, 2000).

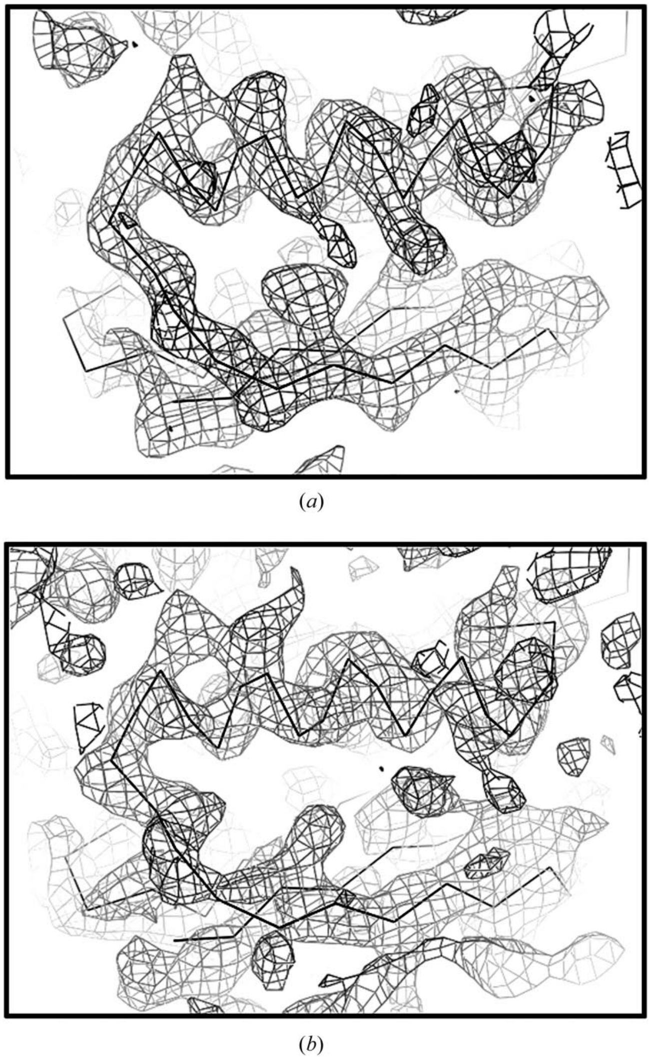


Figure 1
Fragments of 4 Å resolution Fourier syntheses calculated for Protein G with the observed magnitudes and phases: (*a*) calculated from the refined atomic model; (*b*) calculated from the binary approximation on a 16*16*16 grid.

4. Computer tests

The tests were performed with the Protein G data (Derrick & Wigley, 1994). This small protein (61 residues) contains one α-helix and one β-sheet. The protein was crystallized in the space group $P2_12_12_1$ with the unit-cell dimensions 34.9 × 40.3 × 42.2 Å. The complete low-resolution set of experimental diffraction magnitudes was available. The phases calculated from the refined atomic model were considered as the exact ones.

4.1. Binary approximations of Fourier syntheses

The goal of the first series of tests was to check how well small-grid binary functions approximate magnitudes and phases of structure factors. To get a binary approximation for the chosen grid, the Fourier synthesis $\{\rho^{g}(\mathbf{j})\}$ was calculated using the observed magnitudes and the exact phases. The binary approximation values $\{\rho^{\rm bin}(\mathbf{j})\}$ were set to 1 for the given number K of points with highest synthesis values, and to

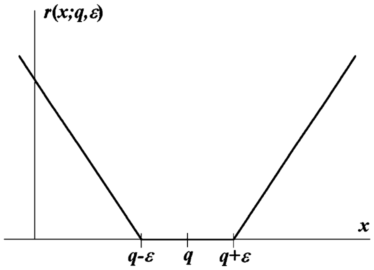


Figure 2
The penalty function used for the solution of the BIP problem.

0 otherwise. The quality of the approximation depends on this parameter K. Special tests were performed to determine optimal K values for different grids. It was found that the optimal ratio of the value K to the full number of grid points (the one which maximizes the correlation) depends on the synthesis resolution and decreases when the resolution increases (Tables 1 and 2). The grid structure factors $\{F^{\rm bin}(\mathbf{h})\exp[i\varphi^{\rm bin}(\mathbf{h})]\}$ were then calculated and their magnitudes and phases were compared to the true ones (Tables 1 and 2; Fig. 1). This test demonstrated that, even at surprisingly small grids, a binary envelope may provide low-resolution phases of a reasonable quality.

4.2. Resolving the phase ambiguity for binary functions

The goal of this test was to study to what extent the condition '0 or 1' allows one to reduce the phase ambiguity. An idealized situation was considered where the exact magnitudes of the real and imaginary parts of the binary structure factors,

$$A(\mathbf{h}) = |F^{\rm bin}(\mathbf{h})\cos\varphi^{\rm bin}(\mathbf{h})|, \quad B(\mathbf{h}) = |F^{\rm bin}(\mathbf{h})\sin\varphi^{\rm bin}(\mathbf{h})|, \tag{38}$$

were supposed to be known. In this case, the grid values satisfy the equations

$$\begin{aligned} \sum_{\mathbf{j}\in\Pi} \cos[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) &= \alpha(\mathbf{h})A(\mathbf{h}), \quad \mathbf{h}\in\Pi, \\ \sum_{\mathbf{j}\in\Pi} \sin[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) &= \beta(\mathbf{h})B(\mathbf{h}), \quad \mathbf{h}\in\Pi, \end{aligned} \tag{39}$$

where the unknowns $\alpha(\mathbf{h})$ and $\beta(\mathbf{h})$ take one of the two values 1 or -1.

The equations (39) have a solution for any particular choice of right-hand values (given by the Fourier transform of these values). So if the grid function is allowed to take any real values, then the known magnitudes $\{A(\mathbf{h}), B(\mathbf{h})\}$ do not define the solution uniquely. Any permutation of signs of $\alpha(\mathbf{h})$ and $\beta(\mathbf{h})$ will result in a solution of (39) possessing the same magnitudes $\{A(\mathbf{h}), B(\mathbf{h})\}$. It may be expected that this is not the case if binary restrictions are added for the unknowns $\{\rho^g(\mathbf{j})\}$:

$$\rho^g(\mathbf{j}) = \{0 \text{ or } 1\}. \tag{40}$$

Now an arbitrary choice of signs $\alpha(\mathbf{h})$ and $\beta(\mathbf{h})$ may result in a solution of (39) that does not satisfy the condition (40). So the binary restrictions may reduce significantly the freedom of the choice of signs and thus may solve the phase problem (or, at least, reduce the phase ambiguity).

4.2.1. 6*6*6 test. In this test, a binary approximation $\{\rho^{\rm bin}(\mathbf{j})\}$ for Fourier synthesis (for Protein G) was constructed at the grid 6*6*6. In this approximation, 50% of the points got the value 1. The corresponding grid structure factors (14) were calculated and the magnitudes (38) were substituted into (39). 100 runs of the program *WSTAOIP* were performed with random starts to solve (39), supposing that the restriction (40) held. The maximal number of flips was set to 50000 (the default value of the program). One run took about 2 min on a Pentium III/500 PC.

In this test, 27 out of 100 runs resulted in a non-zero residual (36), *i.e.* a solution of (39) was not found. The other 73 runs gave in 37 cases a solution equivalent to the true solution $\{\rho^{\rm bin}(\mathbf{j})\}$, while 36 runs resulted in alternative solutions that were equivalent between themselves but not equivalent to the true solution. As usual, we say that the solutions are *equivalent* if they are related by a permitted origin/enantiomer transformation. Furthermore, in our case the solutions $\{\rho^g(\mathbf{j})\}$ and $\{1-\rho^g(\mathbf{j})\}$ must be considered as equivalent as they result in the same magnitudes of structure factors and have the same number of non-zero values (Lunin & Lunina, 1996).

4.2.2. 8*8*8 test. The same tests were performed at the grid 8*8*8 (128 independent grid points). The maximal number of flips was increased to 250000 because the default value was found to be too small to find a solution. Now one run took about 30 min.

20 from 100 runs resulted in a non-zero residual while all the other 80 runs resulted in solutions equivalent to the true one. So when using this grid the binary restriction (40) has eliminated all alternative solutions of (39) corresponding to permutation of the signs of $\alpha(\mathbf{h})$ and $\beta(\mathbf{h})$.

4.2.3. 10*10*10 test. In this test (250 independent grid points), the maximal number of flips was increased to 10000000 because it was not possible to find a solution with a smaller number of flips. Such a large number of trials in the local search procedure required about 70 h of CPU for one run on a Pentium III/500 PC, so that only a small number of runs was performed. In this test, three from five runs resulted in a residual 0 and their results were equivalent to the true solution.

4.3. The use of binary magnitudes

In a more realistic situation, the estimates (38) may be available for centric reflections only, while for acentric reflections only the value $[A(\mathbf{h})^2 + B(\mathbf{h})^2]^{1/2}$ of the magnitude of the complex structure factor may be assumed to be known. The goal of the next test series was to study how such uncertainty affects the solution. It was supposed in these tests that the magnitudes $\{F^{\rm bin}(\mathbf{h})\}$ of the binary structure factors are known exactly, while the magnitudes of their real and imaginary parts were estimated by

$$\tilde{A}(\mathbf{h}) = 2^{-1/2}F^{\rm bin}(\mathbf{h}), \quad \tilde{B}(\mathbf{h}) = 2^{-1/2}F^{\rm bin}(\mathbf{h}). \tag{41}$$

A 'gap' was introduced into equations (39) to take into account the errors caused by this approximation:

$$\begin{aligned} -0.5\tilde{A}(\mathbf{h}) \le \sum_{\mathbf{j}\in\Pi} &\cos[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) - \alpha(\mathbf{h})\tilde{A}(\mathbf{h}) \\ &\le 0.5\tilde{A}(\mathbf{h}), \quad \mathbf{h}\in\Pi \\ -0.5\tilde{B}(\mathbf{h}) \le \sum_{\mathbf{j}\in\Pi} &\sin[2\pi(\mathbf{h}, \mathbf{M}^{-1}\mathbf{j})]\rho^g(\mathbf{j}) - \beta(\mathbf{h})\tilde{B}(\mathbf{h}) \\ &\le 0.5\tilde{B}(\mathbf{h}), \quad \mathbf{h}\in\Pi. \end{aligned} \tag{42}$$

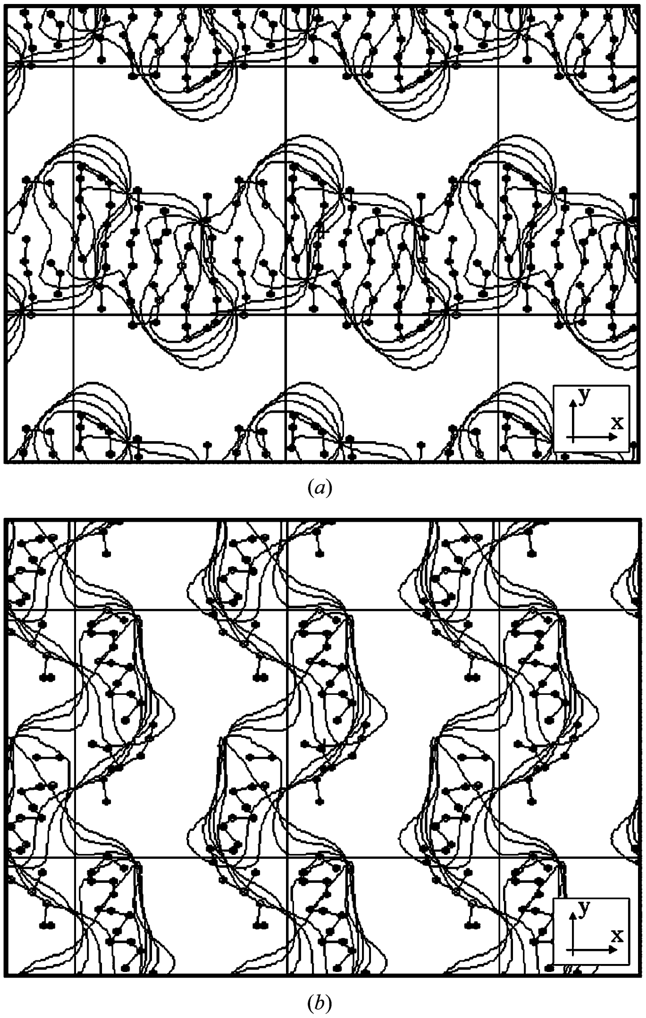
Owing to these approximations, we could not expect any longer that the true solution satisfies (42) and the goal was to make the residual value (36) as small as possible.

4.3.1. 6*6*6 test. All 100 runs of the *WSATOIP* program resulted in non-zero residuals. The analysis of the found solutions showed that 47 of them were equivalent to the exact solution, while 49 were equivalent to the alternative solution found before (§4.2.1).

4.3.2. 8*8*8 test. All 100 runs resulted in a non-zero residual. Cluster analysis of the found solutions revealed a cluster consisting of 19 solutions, all of which were equivalent to the exact solution. Averaging all the 100 solutions gave the phases with a map correlation coefficient with respect to the exact binary phases (Lunin & Woolfson, 1993) equal to 0.95.

4.4. The use of observed magnitudes

When working with real objects, binary magnitudes are not known and must be estimated somehow. In this test, the set of observed magnitudes was used to estimate the binary ones. The scale factor was defined as discussed in Appendix *A*. The grid 8*8*8 was chosen for this test as it allows one to solve BIP problems in a reasonable time using the existing software. On the other hand, the approximation of the binary structure-factor magnitudes using the observed ones is poor at this grid size. This may significantly influence the results. In order to get more reliable results, more powerful BIP methods applicable to larger grids are necessary. The gap in the inequalities (42) was reduced to 25% of the estimated $F^{\rm bin}(\mathbf{h})$ value for acentric structure factors, and to 20% for the centric ones. After 100 runs of *WSATOIP* with random initial assignments, the obtained solutions were aligned and averaged. The found average solution revealed essential features of the 12 Å resolution synthesis and had the map correlation coefficient equal to 0.74 with respect to the exact phases. Fragments of the obtained synthesis overlapped with the atomic model for Protein G is shown in Fig. 3.



(*a*)

(*b*)

Figure 3
Fragments of BIP-phased Fourier synthesis superimposed with C_α atoms of the model for Protein G; several unit cells are shown to illustrate the molecular packing. (*a*) Projection of the slice $z = -2$: 2/40 containing β-sheets; (*b*) projection of the slice $z = 6$: 14/40 containing α-helices. The shown contour isolates 35% of the unit-cell volume (0.4σ cut-off level).

5. Conclusions

The theoretical part of this work shows how the crystallographic phase problem can be reduced to the solution of a system of linear inequalities in binary variables. The practical tests with simulated and experimental protein data illustrate the high potential of this new approach. Crystallographic images found from such phasing can be used for further phase improvement or as an important complementary tool for other techniques like molecular replacement. In order to get images of a higher quality, further work on integer programming methods and their application in crystallography is currently in progress.

APPENDIX *A*
Scaling of observed magnitudes to magnitudes of the binary function

Let $\{B(\mathbf{j})\}_{\mathbf{j}\in\Pi}$ be a binary function defined at a grid Π, $|\mathbf{M}|$ is the number of grid points and K is the number of non-zero values $B(\mathbf{j})$:

$$\sum_{\mathbf{j}\in\Pi} B(\mathbf{j}) = K. \qquad (43)$$

Let $\{\hat{\mathbf{B}}(\mathbf{j})\}_{\mathbf{j}\in\Pi}$ be corresponding structure factors

$$\hat{\mathbf{B}}(\mathbf{h}) = (1/|\mathbf{M}|)\sum_{\mathbf{j}\in\Pi} B(\mathbf{j})\exp[2\pi i(\mathbf{h},\mathbf{M}^{-1}\mathbf{j})], \quad \mathbf{h}\in\Pi, \qquad (44)$$

then

$$\hat{\mathbf{B}}(0) = (1/|\mathbf{M}|)\sum_{\mathbf{j}\in\Pi} B(\mathbf{j}) = K/|\mathbf{M}|. \qquad (45)$$

Owing to the Parseval identity and to the property that $B(\mathbf{j}) = 0$ or 1,

$$|\mathbf{M}|\sum_{\mathbf{h}\in\Pi} \hat{\mathbf{B}}(\mathbf{h})^2 = \sum_{\mathbf{j}\in\Pi} B(\mathbf{j})^2 = \sum_{\mathbf{j}\in\Pi} B(\mathbf{j}) = K \qquad (46)$$

and

$$\sum_{\substack{\mathbf{h}\in\Pi\\ \mathbf{h}\neq\mathbf{0}}} \hat{\mathbf{B}}(\mathbf{h})^2 = (K/|\mathbf{M}|) - (K/|\mathbf{M}|)^2. \qquad (47)$$

If the binary structure factors are supposed to be approximately proportional to some observed values

$$\hat{\mathbf{B}}(\mathbf{h}) \approx \kappa F(\mathbf{h}), \tag{48}$$

then the scale factor κ may be estimated from (47) as

$$\kappa = \left[(K/|\mathbf{M}|) - (K/|\mathbf{M}|)^2 \Big/ \sum_{\substack{\mathbf{h}\in\Pi \\ \mathbf{h}\neq\mathbf{0}}} F(\mathbf{h})^2 \right]^{1/2}. \tag{49}$$

This work was done within the PRST Intellegence Logicielle CPER–Lorraine and was supported by RFBR grants 00-04-48175 and 01-07-90317 and by the University of Nancy. Programs *O* (Jones *et al.*, 1991) and *CAN* (Vernoslova & Lunin, 1993) were used to prepare map illustrations. The authors thank L. Torlay for the computing assistance.

References

Bockmayr, A. & Kasper, T. (1998). *INFORMS J. Comput.* **10**, 287–300.

Cannillo, E., Oberti, R. & Ungaretti, L. (1983). *Acta Cryst.* A**39**, 68–74.

Derrick, J. P. & Wigley, D. B. (1994). *J. Mol. Biol.* **243**, 906–918.

Johnson, E. L., Nemhauser, G. L. & Savelsbergh, M. W. P. (2000). *INFORMS J. Comput.* **12**, 2–23.

Jones, T. A., Zou, J. Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

Lecomte, C. (1999). *Implications of Molecular and Materials Structure for New Technologies, NATO ASI and Euroconference, Series E, Applied Science*, pp. 23–44. Dordrecht: Kluwer Academic Publishers.

Lunin, V. Y. (1985). *Acta Cryst.* A**41**, 551–556.

Lunin, V. Y. (2000). *Acta Cryst.* A**56**, 73–84.

Lunin, V. Y. & Lunina, N. L. (1996). *Acta Cryst.* A**52**, 365–368.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Skovoroda, T. P., Urzhumtsev, A. G. & Podjarny, A. D. (2000). *Acta Cryst.* D**56**, 1223–1232.

Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.

Sayre, D. (1951) *Acta Cryst.* **4**, 362–367.

Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.

Ten Eyck, L. F. (1973). *Acta Cryst.* A**29**, 183–191.

Ten Eyck, L. F. (1977). *Acta Cryst.* A**33**, 486–492.

Vernoslova, E. A. & Lunin, V. Y. (1993). *J. Appl. Cryst.* **26**, 291–294.

Walser, J. P. (1997). *Proceedings of the Fourteenth National Conference on Artificial Intelligence and Ninth Innovative Applications of Artificial Intelligence Conference, AAAI 97*, pp. 269–274. IAAI 97, 27–31 July 1997, Providence, Rhode Island, USA. Cambridge, MA: AAAI Press/The MIT Press.

Walser, J. P. (1998). *Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference, AAAI 98*, pp. 373–379. IAAI 98, 26–30 July 1998, Madison, Wisconsin, USA. Cambridge, MA: AAAI Press/The MIT Press.