Acta Cryst. (1999). A**55**, 916–925

Seminvariant density decomposition and connectivity analysis and their application to very low resolution macromolecular phasing

V. Y. Lunin,[a,b] N. L. Lunina[a] and A. G. Urzhumtsev[b]*

[a]*Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region 142292, Russia, and* [b]*LCM*^{3}*B, UPRESA 7036 CNRS, Faculté des Sciences, Université Henri Poincaré, Nancy 1, 54506 Vandoeuvre-lès-Nancy, France. E-mail: sacha@lcm3b.u-nancy.fr*

(*Received 26 October 1998; accepted 6 April 1999*)

Abstract

A low-resolution Fourier synthesis is thought to show a molecule as a compact region of a high electron density. As a consequence, the number of such regions, chosen at a proper cut-off level, should be equal to the number of molecules in the unit cell. This hypothesis may be used as a basis for selection criteria in multisolution *ab initio* phasing procedures. However, when working with a small number of reflections, this hypothesis may break down. The suggested Fourier-synthesis decomposition explains some reasons for failure and provides a connectivity-based procedure for the determination of macromolecular position in the crystal unit cell and the phasing of several low-resolution reflections. The simplest decomposition consists in separating the reflections into two sets according to whether their phases do or do not depend on a permitted origin shift. It is shown that the partial Fourier syntheses corresponding to these subsets are simply a half-sum and a half-difference of the initial electron-density distribution with its shifted copy. Therefore, they display the true images overlapped with the shifted ones (or with shifted and additionally flipped copies for the latter synthesis). The paper generalizes the decomposition for the case of a finite subgroup of the group of permitted origin shifts and reveals the role of one-phase seminvariants.

1. Introduction

The development of *ab initio* phasing methods applicable at very low resolution (VLR) is stimulated by an increasing interest by crystallographers in large macromolecules and their complexes. Standard approaches such as isomorphous and molecular replacement, multiwavelength anomalous diffraction (MAD) or direct methods while demonstrating success on the way to such structures (Luger *et al.*, 1997; Ban *et al.*, 1998; Miller & Weeks, 1998; Sheldrick, 1998; Woolfson, 1998) are as yet far from routine tools. The goal of VLR phasing methods is to find phases for a relatively small number of reflections and these methods do not pretend to give directly an image interpretable in terms of individual atoms or even residues. The information that can be extracted from low-resolution Fourier syntheses is confined mostly to the position of the object in the unit cell and its outlines. These data may be used as a starting point for the following phase-extension procedures, which bring new structural details and are very useful when using information from alternative methods like electron microscopy.

VLR phasing methods differ between themselves by the search procedure and by the complementary information (or hypotheses) used to recognize suitable phase sets. For one type of procedure, the search is carried out in a parameterized phase space where all phase sets are calculated from some kind of simple macromolecular model. Basically, they are models composed from a set of spheres (or pseudoatoms or globs) and contain one (Podjarny *et al.*, 1987; Harris, 1995; Andersson & Hovmöller, 1996), a small number (Podjarny *et al.*, 1987; Lunin *et al.*, 1995, 1998*a*) or a large number (Subbiah, 1991, 1993) of them. Such approximations are reasonable at VLR, and the positions of the spheres can sometimes be determined from the best correspondence of calculated structure-factor magnitudes to the experimental ones. Pseudoatom approximations are useful in electron crystallography as well (Dorset, 1997; Dorset & Jap, 1998).

Alternatively, all phase sets (or a representative ensemble of phase sets) can be checked in order to choose the best one according to some criterion. Such a representative ensemble can be chosen randomly (Lunin *et al.*, 1990) or as some 'regular grid' in the configuration space of all phase sets, *e.g.* constructed using error-correcting codes (Woolfson, 1954; Gilmore *et al.*, 1999). The key problem in such approaches is the choice of a selection criterion that allows identification of the true phase set among all the considered ones (in classical direct methods, these criteria are called figures of merit, FOMs). Traditional FOMs (Gilmore, 1998) are not always applicable at low resolution when the assumptions forming the basis of these criteria fail.

Some new criteria, *e.g.* ones using the specificity of electron-density histograms (Lunin *et al.*, 1990; Lunin, 1993) or likelihood-based criteria (Bricogne & Gilmore, 1990; Lunin *et al.*, 1998*a*) were found to be more appropriate when working at VLR.

It was found that different approaches of both types being applied to VLR data have a general feature, the presence of multiple minima (or maxima, in relevant cases) of the search criteria. These minima regions are similar in size and depth, and the correct solution does not necessarily belong to the vicinity of the deepest minimum. In order to overcome this difficulty, a special technique was suggested (Lunin *et al.*, 1990, 1995, 1998*a*), which is based on the cluster analysis of all reasonably good candidates for solution rather than a few best ones. This technique reduces the phase ambiguity to a choice among a small number of alternative phase sets. In order to make this last choice, two additional criteria were successfully used in the course of VLR *ab initio* phasing of a ribosomal 50S particle from *Thermus thermophilus* (Urzhumtsev *et al.*, 1996; Lunin *et al.*, 1998*b*), namely generalized likelihood (Lunin *et al.*, 1998*a*) and a topological criterion based on a connectivity analysis of the density distribution.

The connectivity properties were used for many years to estimate qualitatively electron-density maps. Baker *et al.* (1993) formalized this in a quantitative criterion for middle- or high-resolution maps. This criterion is based on the observation that a well phased Fourier synthesis reveals extended continuous regions of high electron density corresponding to polypeptide and side chains and that, on the contrary, the presence of a large number of small isolated 'drops' indicates ill defined phases. A minimal principle was formulated as follows: the larger are the connected components in the synthesis and the less is their number, the better are the phases. A modification of this principle extended for skeletonized maps was recently tested by Mishnev (1998). At VLR, this minimal principle can hardly be applied in the usual form because a poorly phased synthesis shows merged molecular images rather than multiple drops. As is discussed in §2, the principle could be replaced by one saying that the region of high density of a VLR synthesis must contain as many roughly equal connected regions as the number of molecules in the unit cell and that these regions must be as large as possible while being separated. Nevertheless, even in such a modified form, the connectivity criterion may not be applicable if a too small number of reflections is taken which is necessary nowadays to perform an exhaustive phase search. These changes in connectivity characteristics at VLR may be explained to some extent in the frame of a Fourier-syntheses decomposition as discussed in §§3 and 4. For every permitted origin shift, the reflections are separated into two sets according to the property of their phase to be independent of this shift or not. The partial Fourier syntheses calculated for these two sets of reflections are shown to be simply a half-sum and a half-difference of the initial electron-density distribution with its shifted copy. The former, origin-independent, partial synthesis necessarily shows a superposition of the correct molecular image with its shifted copy. This means that, when the reflections of the first type dominate in the synthesis, the number of macromolecular images in the synthesis tends to be greater than the number of macromolecules, the images merge and the topological criterion formulated as above is useless. The latter, origin-variable, component represents some kind of difference synthesis with a correct number of sharpened individual macromolecular images but possibly deformed; this happens if the centre of the shifted and flipped image is close to one of the true object images. It should be stressed that this synthesis may have desired connectivity characteristics even when the full syntheses does not reveal them. Therefore, the procedures of a connectivity-based search applied to the origin-variable component rather than to a full synthesis may provide one with the true object position and reasonable values for origin-variable phases.

A cumulative effect of several permitted origin shifts is considered in §4. In this case, the special role is played by seminvariant reflections which do not change phase values under all permitted origin shifts and are of special interest in classical direct methods (Giacovazzo, 1980).

The main ideas of the Fourier-synthesis decomposition are illustrated in §5 by the example of an artificial simplified object. The seminvariant density decomposition analysis was developed in the course of the phasing of a ribosomal 50S particle (with the use of experimental data obtained in the laboratories of A. Yonath) which will be discussed elsewhere. The test object used in this paper displays the particles packing in the unit cell and their features important for the analysed problem.

2. Connectivity analysis

2.1. *Connectivity-analysis procedure*

A mask of the molecular region Ω is usually defined as a region in the unit cell which contains the points with highest Fourier-synthesis values above some cut-off level κ. This level depends on the magnitude scale and on the resolution of the synthesis.

When operating with a high-resolution synthesis, κ is usually chosen high enough and the mask reproduces, with more or less detail, the trace of the polypeptide and side chains. At such a resolution, a mask of the molecular region consisting of a large number of small components usually indicates a noisy poorly phased synthesis, contrary to the case with a few large connected components. Thus, the number M of the connected components in Ω may be used as a figure of merit of the phase set (Baker *et al.*, 1993; Mishnev, 1998)

and the decrease of M (probably up to some limit) may be considered as the indication of the progress in phasing.

For the low-resolution syntheses, the chains are not distinguishable and the mask represents the shape and the position of the macromolecule rather than fine molecular details. This molecular region mask Ω may correspond to a lower κ value and occupy a larger share of the unit cell than in the case of a higher-resolution 'chain' mask. When analysing VLR syntheses, one expects that for a reasonable cut-off level κ the connected components at this synthesis correspond to the masks of symmetry-related copies of the molecule and therefore the number of these components $M(\kappa)$ should be equal to the number $N_{\rm mol}$ of molecules in the unit cell. Obviously, this is not the case when the cut-off level κ is chosen too low and the molecular images merge together. It can also be not true if κ is taken too high and Ω can contain several local maxima of density per molecule rather than the connected molecular envelope.

The topological analysis of an electron-density distribution $\rho(\mathbf{r})$ may consist in the calculation of the size and the number of connected components in the regions Ω corresponding to different cut-off levels. It is convenient to introduce a parameter p which varies with an equal step in the interval $0 < p < 100\%$ and to define the levels κ_p and corresponding Ω_p regions,

$$\Omega_p = \{\mathbf{r} : \rho(\mathbf{r}) \geq \kappa_p\}, \tag{1}$$

in a such way that every Ω_p region has the relative volume equal to this p value:

$$\frac{\text{volume}\,(\Omega_p)}{\text{unit-cell volume}} \times 100\% = p. \tag{2}$$

An example of the application of the connectivity analysis is shown in Table 2 (Appendix A). Its use in a phase-determination process is discussed in §5.

2.2. *Connectivity-based VLR phasing*

At VLR, the electron-density distribution, in general, is higher at the centre of the molecular image and lower at the molecular border. Therefore, the principal effect of phase errors is losing the envelope and merging the molecular images in a Fourier synthesis while a higher density at the centre of the molecules is kept until the phase error reaches some limit. The higher the phase error, the smaller is the size of the isolated regions, corresponding to $N_{\rm mol}$ macromolecules in the unit cell. For any synthesis, we can define the value of the parameter $p^{\rm max}(N_{\rm mol})$ as the highest relative volume p such that the Ω_p region consists of exactly $N_{\rm mol}$ similar connected components. If two syntheses are calculated with the same amplitudes and with different phases, one can expect that the worst of them has a smaller $p^{\rm max}(N_{\rm mol})$, which thus can be used as a criterion to judge alternative VLR phase sets in an *ab initio* structure determination. Indeed, it was extremely useful for the phasing of the 50S ribosomal particle from *Thermus thermophilus* (Lunin *et al.*, 1998*b*).

When the number of reflections is relatively small (*e.g.* in the search for the phases of a few strongest VLR reflections), it is possible to carry out an exhaustive search and to check all possible phase combinations. (Obviously, some reasonable sampling should be applied to the phases of noncentrosymmetric reflections.) The maximization of $p^{\rm max}(N_{\rm mol})$ could be used as a selection criterion in this search. Unfortunately, a synthesis calculated with a small number of reflections, even when their magnitude and phase values are exact, can lose desired connectivity properties (for example, see Fig. 1) and the suggested procedure may fail. Some possible reasons behind this as well as a way to overcome the problem are discussed below.

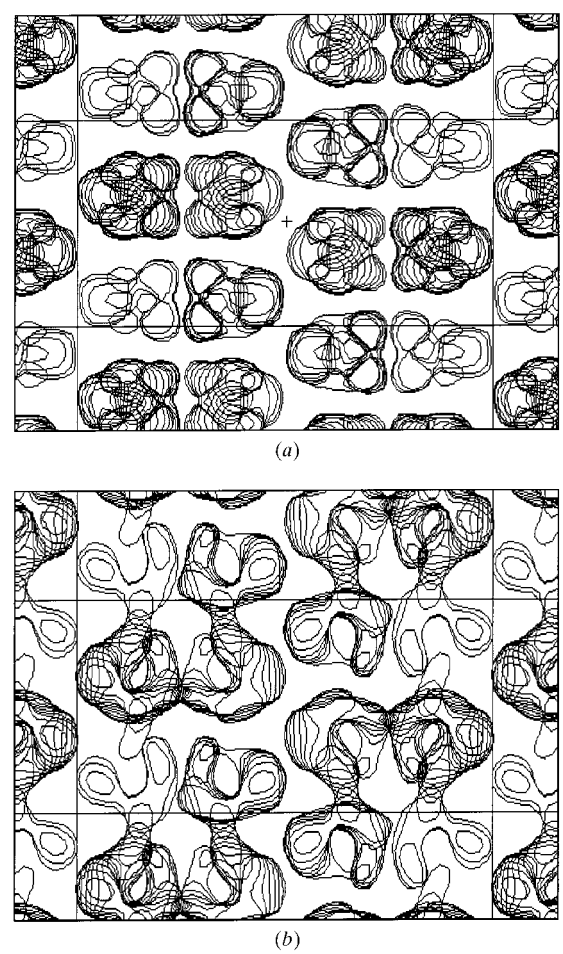

Fig. 1. Fourier syntheses for the test object: (*a*) synthesis calculated with all 52 reflections within the resolution zone $d > 60$ Å. (*b*) The 11 strongest VLR reflections (S_0 set) are used. X sections are shown.

3. One-shift density decomposition

3.1. *Density decomposition*

Let S be a set of reflections used to calculate a studied Fourier synthesis $\rho(\mathbf{r})$ in some coordinate system:

$$\rho(\mathbf{r}) = (1/V) \sum_{\mathbf{h}\in S} F_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}}) \exp[2\pi i(\mathbf{h}, \mathbf{r})]. \quad (3)$$

While structure-factor amplitudes are invariant for a shift of the origin of this coordinate system, it is not generally the case for the phases. The origin shift by the vector $\mathbf{u}$ results in the shift $2\pi(\mathbf{h}, \mathbf{u})$ of the phase $\varphi_{\mathbf{h}}$. According to this property of changing phase values, for every particular shift $\mathbf{u}$ the set S can be split into two parts:

$$S = S_{\rm oi} \cup S_{\rm ov}. \quad (4)$$

Here, $S_{\rm oi}$ stands for reflections that are invariant under the origin shift, while the set $S_{\rm ov}$ consists of reflections whose phases vary with the shift $\mathbf{u}$. It follows from the definition that

$$(\mathbf{h}, \mathbf{r}) = 0|_{\rm mod\,1} \quad \text{for } \mathbf{h} \in S_{\rm oi} \quad (5)$$

and

$$(\mathbf{h}, \mathbf{r}) \neq 0|_{\rm mod\,1} \quad \text{for } \mathbf{h} \in S_{\rm ov}. \quad (6)$$

The decomposition (4)–(6) induces two complementary partial Fourier syntheses:

$$\rho_{\rm oi}(\mathbf{r}) = (1/V) \sum_{\mathbf{h}\in S_{\rm oi}} F_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}}) \exp[2\pi i(\mathbf{h}, \mathbf{r})] \quad (7)$$

$$\rho_{\rm ov}(\mathbf{r}) = (1/V) \sum_{\mathbf{h}\in S_{\rm ov}} F_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}}) \exp[2\pi i(\mathbf{h}, \mathbf{r})], \quad (8)$$

which forms two parts of the $\rho(\mathbf{r})$ distribution:

$$\rho(\mathbf{r}) = \rho_{\rm oi}(\mathbf{r}) + \rho_{\rm ov}(\mathbf{r}). \quad (9)$$

Although the decomposition (7)–(9) may be performed for an arbitrary origin shift, it has some important features when $\mathbf{u}$ is a permitted origin shift, *i.e.* one that does not change the symmetry properties of $\rho(\mathbf{r})$ (see §4). The most important feature in this case is that the partial syntheses $\rho_{\rm oi}(\mathbf{r})$ and $\rho_{\rm ov}(\mathbf{r})$ are simply the half-sum and half-difference of the full synthesis $\rho(\mathbf{r})$ and its shifted copy $\rho(\mathbf{r} - \mathbf{u})$:

$$\rho_{\rm oi} = \tfrac{1}{2}[\rho(\mathbf{r}) + \rho(\mathbf{r} - \mathbf{u})] \quad (10)$$

$$\rho_{\rm ov} = \tfrac{1}{2}[\rho(\mathbf{r}) - \rho(\mathbf{r} - \mathbf{u})]. \quad (11)$$

This follows from the theorem 1 proven in Appendix *B*.

The former synthesis (10) gives an overlapped picture of two shifted copies of the true image [by the true one we mean the image presented in the full synthesis (3)]. This is governed by origin-invariant (o.i.) reflections from the $S_{\rm oi}$ set. If such reflections dominate in a small set of low-resolution reflections, then, owing to extra molecular copies, it may result in merged globs at the electron-density map rather than in separated molecular images. The latter synthesis (11) gives the true image surrounded (and, possibly, distorted) by its flipped and shifted copies (Fig. 2). This is governed by the origin-variable (o.v.) reflections from the $S_{\rm ov}$ set. It does not induce the merging of images but may provoke image distortions if one of the symmetry-related flipped images is in the vicinity of the true one.

It must be noted that the decomposition (4)–(6) of the set of reflections into two parts is not a unique way to separate the structure factors. The advantage of this decomposition is that the partial Fourier syntheses (7)–(8) have a simple interpretation (10)–(11) in terms of the whole electron density $\rho(\mathbf{r})$. Other decompositions of this type are considered in §4.

3.2. *Connectivity-based determination of VLR origin-variable phases*

The origin-variable partial Fourier synthesis does not contain reflections that cause the overlapping of the image with $\mathbf{u}$-shifted phantoms. So this partial synthesis may reveal sharpened and isolated compact regions separated by the equal but negative ones. Therefore, the origin-variable part $\rho_{\rm ov}(\mathbf{r})$ of the electron density may be more suitable for connectivity-based *ab initio* phasing than the $\rho(\mathbf{r})$ while two important features of this phasing must be emphasized. Firstly, the o.v. partial synthesis (8) by no means represents the true image of the object. It represents some artificially constructed image, positioned similar to the true one and possessing the same o.v. phases. As a consequence, only the position of the object in the unit cell and a subset of origin-variable phases could be found in such a search. Secondly, the properties of $\rho_{\rm ov}(\mathbf{r})$ may depend strongly on the choice of the permitted origin shift $\mathbf{u}$. If the position of the $\mathbf{u}$-shifted negative image is close to the position of one of the symmetry-related positive images, then the difference image may be completely destroyed and the phasing procedure may fail. Usually it is not known in advance how the shifted copies are arranged with respect to the initial ones. Nevertheless, different permitted origin shifts can be tried in the hope that at least one of them leads to success.

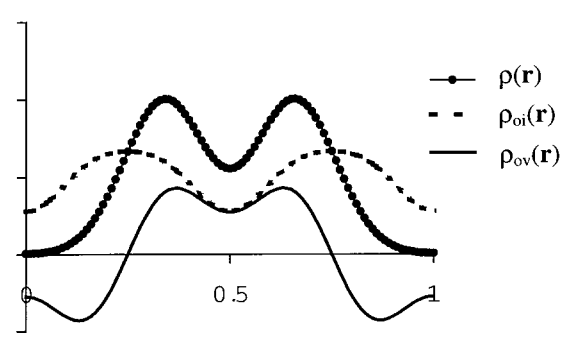

Fig. 2. One-dimensional example of density decomposition corresponding to an origin shift of 1/2.

4. Multiple-shifts density decomposition

In this section, a more general case of density decomposition is studied. All considered distributions are supposed to be periodical ones so that all equalities are considered implicitly by modulo integer.

4.1. *Permitted origin shifts.*

Let $\Gamma = \{(\mathbf{R}_\nu, \mathbf{t}_\nu)\}_{\nu=0}^{n-1}$ be a crystallographic space group, $\Gamma^t = \{(\mathbf{I}, \mathbf{t}_\mu)\}_{\mu=0}^{m-1}$ its translation subgroup and $(\mathbf{R}_0, \mathbf{t}_0) = (\mathbf{I}, \mathbf{0})$ the identical transformation. (For the space group with primitive unit cell, the subgroup Γ^t consists of the identical transformation only.) Two real-space vectors $\mathbf{u}$ and $\mathbf{v}$ are equal by modulo Γ^t ($\mathbf{u} = \mathbf{v}|_{\mathrm{mod}\,\Gamma^t}$), if they differ by a translation from Γ^t:

$$\mathbf{u} = \mathbf{v}|_{\mathrm{mod}\,\Gamma^t} \text{ if and only if } \mathbf{u} - \mathbf{v} = \mathbf{t}_\mu \text{ for some } 0 \le \mu < m - 1. \quad (12)$$

Permitted (or allowed) origins may be defined in terms of structure-factor properties as 'the points which, taken as origins, retain the same functional form of the structure factor' (Giacovazzo, 1974). The permitted origins are related to each other by the permitted translations or shifts. Since originally the symmetry of a space group is defined in the real space, an equivalent and more formal introduction of this concept may be given (Lunin & Lunina, 1996) by the following two definitions:

Definition 1. A function $\rho(\mathbf{r})$ possesses the Γ symmetry if

$$\rho(\mathbf{R}_\nu \mathbf{r} + \mathbf{t}_\nu) = \rho(\mathbf{r}) \text{ for all } \mathbf{r} \in \mathbf{R}^3 \text{ and all } 0 \le \nu < n - 1. \quad (13)$$

Definition 2. A vector $\mathbf{u}$ is an origin shift permitted for the group Γ if it retains the Γ symmetry, *i.e.* any function $\rho(\mathbf{r})$ possesses the Γ symmetry if and only if $\rho(\mathbf{r} - \mathbf{u})$ does.

The necessary and sufficient condition for $\mathbf{u}$ to be a permitted origin shift may be formulated as the following (Giacovazzo, 1980; Lunin & Lunina, 1996):

Lemma 1. $\mathbf{u}$ is a permitted origin shift if and only if

$$\mathbf{R}_\nu \mathbf{u} = \mathbf{u}|_{\mathrm{mod}\,\Gamma^t} \quad \text{for all } 0 \le \nu < n - 1. \quad (14)$$

The permitted origin shifts form a group Γ^{pos}. This group may be an infinite (*e.g.* for triclinic or monoclinic space groups) or a finite one (for example, for orthorhombic groups).

4.2. *Shift-dependent reflections*

A shift of the origin by a vector $\mathbf{u}$ has no influence on the structure-factor magnitudes but changes the phase of a reflection with the index $\mathbf{h}$ by $2\pi(\mathbf{h}, \mathbf{u})$. Thus, the phase remains unchanged if $(\mathbf{h}, \mathbf{u}) = 0|_{\mathrm{mod}\,1}$.

Definition 3. Reflection with the index $\mathbf{h}$ and the corresponding phase $\varphi_{\mathbf{h}}$ are $\mathbf{u}$-*invariant* if $(\mathbf{h}, \mathbf{u}) = 0|_{\mathrm{mod}\,1}$ and is $\mathbf{u}$-*variable* otherwise. If Γ_0 is a subgroup of a group of permitted origin shifts, a reflection is Γ_0-*invariant* if it is invariant with respect to every shift from the Γ_0 group.

It follows from this definition that the phases of Γ_0-invariant reflections do not change their values for any origin shift from Γ_0. In particular, the phases of the reflections that are invariant with respect to *all* permitted origin shifts are one-phase structure seminvariants.

4.3. *Density decomposition*

Let $\Gamma_0 = \{(\mathbf{I}, \mathbf{u}\}_{\kappa=0}^{K-1}$ be a finite subgroup of the group of permitted origin shifts and $\mathbf{u}_0 = \mathbf{0}$. Any function $\rho(\mathbf{r})$ can be represented formally as

$$\rho(\mathbf{r}) = \rho_{\mathrm{oi}}(\mathbf{r}) + \rho_{\mathrm{ov}}(\mathbf{r}), \quad (15)$$

where

$$\rho_{\mathrm{oi}}(\mathbf{r}) = (1/K)\rho(\mathbf{r}) + (1/K) \sum_{\kappa=1}^{K-1} \rho(\mathbf{r} - \mathbf{u}_\kappa) \quad (16)$$

$$\rho_{\mathrm{ov}}(\mathbf{r}) = [(K-1)/K]\rho(\mathbf{r}) - (1/K) \sum_{\kappa=1}^{K-1} \rho(\mathbf{r} - \mathbf{u}_\kappa). \quad (17)$$

In this decomposition, the origin-invariant part $\rho_{\mathrm{oi}}(\mathbf{r})$ is the sum of K equally weighted shifted copies while the origin-variable part $\rho_{\mathrm{ov}}(\mathbf{r})$ is formed by the original image distorted by shifted and flipped images taken with a relatively low weight. The ratio of the weights of these components increases linearly with the size of the group Γ_0. Fig. 3 illustrates the density decomposition in the one-dimensional case.

It is important to note that, as follows from the definition, the functions $\rho_{\mathrm{oi}}(\mathbf{r})$ and $\rho_{\mathrm{ov}}(\mathbf{r})$ have the same symmetry as $\rho(\mathbf{r})$ has and that

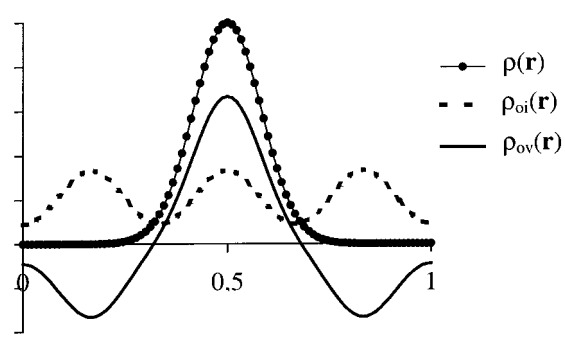

Fig. 3. One-dimensional example of density decomposition corresponding to the {0, 1/3, 2/3} group of origin shifts.

$$\rho_{\rm oi}(\mathbf{r}-\mathbf{u}_\kappa)=\rho_{\rm oi}(\mathbf{r}) \text{ for all } \mathbf{u}_\kappa\in\Gamma_0. \quad (18)$$

4.4. *Fourier-series decomposition*

The density decomposition (15)–(17) in the real space is coupled with the Fourier-series decomposition as is established by the following theorem 2 (for the proof see Appendix *C*).

Theorem 2. Let $\Gamma_0=\{(\mathbf{I},\mathbf{u}_\kappa)\}_{\kappa=0}^{K-1}$ be a finite subgroup of the group of permitted origin shifts ($\mathbf{u}_0=\mathbf{0}$) and the reciprocal space be split into Γ_0-invariant and Γ_0-variable parts:

$$S_{\rm oi}(\Gamma_0)=\{\mathbf{h}:(\mathbf{h},\mathbf{u}_\kappa)=0 \text{ for all } \kappa=0,\ldots,K-1\} \quad (19)$$

$$S_{\rm ov}(\Gamma_0)=\{\mathbf{h}:(\mathbf{h},\mathbf{u}_\kappa)\neq 0 \text{ for some } \kappa=0,\ldots,K-1\}. \quad (20)$$

If

$$\rho(\mathbf{r})=(1/V)\sum_{\mathbf{h}}F_{\mathbf{h}}\exp(i\varphi_{\mathbf{h}})\exp[2\pi i(\mathbf{h},\mathbf{r})] \quad (21)$$

and the functions $\rho_{\rm oi}(\mathbf{r})$ and $\rho_{\rm ov}(\mathbf{r})$ are defined by (16) and (17), then

$$\rho_{\rm oi}(\mathbf{r})=(1/V)\sum_{\mathbf{h}\in S_{\rm oi}(\Gamma_0)}F_{\mathbf{h}}\exp(i\varphi_{\mathbf{h}})\exp[2\pi i(\mathbf{h},\mathbf{r})] \quad (22)$$

and

$$\rho_{\rm ov}(\mathbf{r})=(1/V)\sum_{\mathbf{h}\in S_{\rm ov}(\Gamma_0)}F_{\mathbf{h}}\exp(i\varphi_{\mathbf{h}})\exp[2\pi i(\mathbf{h},\mathbf{r})]. \quad (23)$$

This theorem shows that $\rho_{\rm oi}(\mathbf{r})$ is simply a part of the $\rho(\mathbf{r})$ Fourier series calculated over all Γ_0-invariant reflections and $\rho_{\rm ov}(\mathbf{r})$ is the corresponding sum calculated over the complementary subset. In particular, if Γ_0 is the group of all permitted origin shifts, the Fourier series for $\rho_{\rm oi}(\mathbf{r})$ is formed by the reflections possessing seminvariant phases. This is why formulae (22)–(23) are called the *seminvariant density decomposition*. All considerations of §3 are applicable to the decomposition (22)–(23).

5. Test example of the seminvariant density decomposition

The following simplified model example is chosen to illustrate transparently the ideas of the decomposition discussed above. The test calculations were performed with model data simulating a low-resolution analysis of a large macromolecular complex (Appendix *A*1). It can be mentioned that the seminvariant density decomposition study was originated from the structural analysis of the H50S ribosomal particle (data provided by A. Yonath), these results will be discussed elsewhere.

The exact 60 Å resolution synthesis (52 independent reflections) shows the right number of molecular images in the unit cell (Fig. 1*a*). Unfortunately, this number of reflections is too high to perform an exhaustive search using the connectivity as the selection criterion. Such a search is feasible for a smaller number of reflections, *e.g.* for 11 strongest reflections at this resolution. An attempt to find the correct crystallographic image with such a data set failed which was not surprising as the syntheses calculated with these 11 structure factors (even when their values are exact) does not show the right number of molecules (Fig. 1*b*). The seminvariant decomposition analysis explains this behaviour. The phases of most of these strongest reflections are seminvariants which makes the phantom images quite strong in this case.

In order to get an object possessing more favourable topological properties, the origin-variable part $\rho_{\rm ov}(\mathbf{r})$ of the Fourier synthesis was studied separately for three independent permitted origin shifts (Appendix *A*2). Table 2 represents the results of the connectivity analysis for them and shows that the shift $\mathbf{t}_1=(1/2,0,0)$ only provides the desired connectivity characteristics compatible with those of the full 60 Å-resolution synthesis. An attempt to find the phases of nine origin-variable reflections corresponding to $\mathbf{t}_1$ using the exhaustive connectivity-based search (with $\pi/4$, $3\pi/4$, $5\pi/4$, $7\pi/4$ sampling for noncentrosymmetric phases) allowed definition of phases possessing an 84% phase-correlation coefficient with respect to the exact phases. Sections of the corresponding exact and *ab initio* phased syntheses are shown in Fig. 4.

It should be emphasized that not every permitted origin shift allows the desired connectivity characteristics of $\rho_{\rm ov}(\mathbf{r})$ to be obtained by means of the decomposition (7)–(8). This is illustrated by Figs. 5 and 6. The presence of shifted and flipped images at $\rho_{\rm ov}(\mathbf{r})$ distorts the original one and the greater is the overlapping the stronger is the distortion. It is not known in advance how large the overlapping is for a particular origin shift. However, it is possible to try to phase origin-variable sets corresponding to different shifts and to select the one resulting in the best connectivity characteristics of the found map. Naturally, when the phases are found for some subset of reflections, it is possible to fix them and to repeat the phasing procedure for an extended phase set.

6. Conclusions

The decomposition analysis explains the reasons for the loss of the desired connectivity and suggests a way to recover necessary features. The connectivity-based phase search might be performed with the subsets composed of the origin-variable structure factors and not with the full data set. The number of such subsets depends on the number of possible choices of the origin. If for one of the permitted origins the initial and shifted images are not strongly superimposed, the origin-variable part of corresponding synthesis shows the *contrasted* molecular position of a *single macromolecular object* in the crystal and the connectivity

principle may work now. In the test case discussed in §5, this information was sufficient to solve the problem for one of the shift-independent phase sets and not for the set of strongest reflections. Even if the shape of the molecule is perturbed, this information can be important to identify the correct phase subset and to get preliminary packing formation. Moreover, seminvariant decomposition can also predict a way in which the image is perturbed by overlapping the original and shifted copies.

APPENDIX *A*

*A*1. *Test object*

The main goal of the suggested method is the phasing of very large macromolecular complexes. Therefore, a test object was constructed to simulate a large ribosomal particle. The envelope was formed by five spheres centred at the vertexes of a pyramid with a square base. The radii of spheres was chosen as 55 Å, the base side as 60 Å and the pyramid height as 60 Å (Fig. 7). This envelope was placed without overlapping in the unit cell with the parameters $210 \times 300 \times 573$ Å and the symmetry space group $C222_1$ simulating the real experimental data. Eight symmetry-related envelopes per unit cell occupied about 52% of the unit-cell volume. The envelopes were filled with dummy atoms. The magnitudes and phases of the structure factors calculated from this model were used as the observed values of structure-factor magnitudes and the exact phases.

It is worth noting that at very low resolution the structure factors calculated from the solvent content of the unit cell are nearly proportional to structure factors corresponding to the protein part of the unit cell (Urzhumtsev & Podjarny, 1995). So the absence of the solvent part in the calculated structure factors does not produce principal differences and results in the scale factor only, which is not important when studying topological properties.

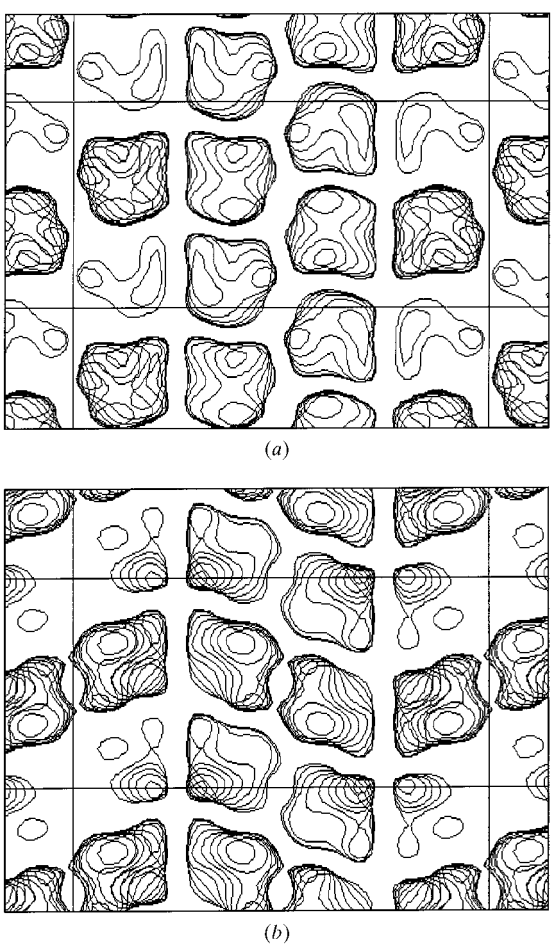

Fig. 4. Origin-variable part $\rho_{ov}(\mathbf{r})$ corresponding to the origin shift $(1/2, 0, 0)$. Nine VLR reflections (S_1 set) are used, X sections are shown. (*a*) Exact phases; (*b*) phases found by the connectivity-based search.

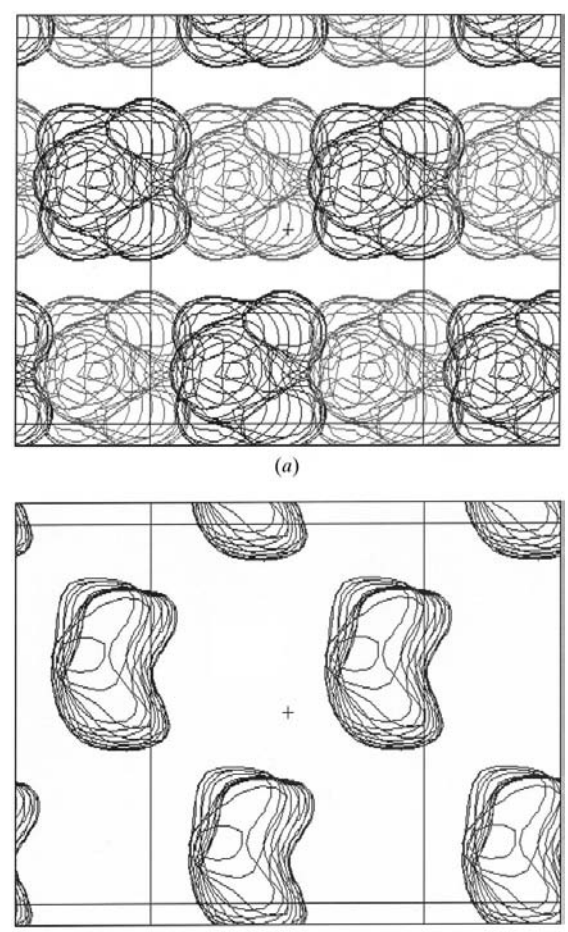

Fig. 5. Fourier syntheses for: (*a*) $\rho(\mathbf{r})$ (black) and $\rho(\mathbf{r} - \mathbf{u})$ (grey); (*b*) $\rho_{ov}(\mathbf{r})$. Nine reflections of S_1 data set are used, $\mathbf{u} = (0, 0, 1/2)$, Z sections are shown.

A2. Permitted origin shifts

The space group $C222_1$ allows eight origin shifts. However, only three of them, namely

$$\mathbf{t}_1 = (1/2, 0, 0), \quad \mathbf{t}_2 = (0, 0, 1/2), \quad \mathbf{t}_3 = (1/2, 0, 1/2), \tag{24}$$

are essential because the rest are the zero shift and the shifts connected with the shifts (24) by crystallographic symmetry, *i.e.* by the translation $(1/2, 1/2, 0)$.

A3. Connectivity analysis for the sets of shift-dependent reflections

Five different VLR sets were studied which included:
(i) all 52 reflections of the 60 Å resolution zone;
(ii) the 11 strongest VLR reflections (S_0);
(iii) the 9 strongest $\mathbf{t}_1$-dependent reflections (S_1);
(iv) the 10 strongest $\mathbf{t}_2$-dependent reflections (S_2);
(v) the 11 strongest $\mathbf{t}_3$-dependent reflections (S_3).

The lists of these reflections are given in Table 1. The results of the connectivity analysis presented in Table 2 show that the use of a small number of VLR reflections (both origin variable and origin invariant, set S_0) deteriorate the connectivity characteristics. At the same time, the $\mathbf{t}_1$-based origin-variable component $\rho_{\rm ov}(\mathbf{r})$ calculated with only nine reflections (set S_1) reveals the connectivity characteristics similar to the ones of 60 Å resolution synthesis calculated with all 52 VLR reflections. This table shows also that $\mathbf{t}_1$ is the only permitted origin shift which is favourable for the connectivity-based search for phases of the $\rho_{\rm ov}(\mathbf{r})$ component.

Table 1. *Very low resolution sets of reflections (see Appendix A3)*

S_0	S_1	S_2	S_3
111	111	111	112
112	112	113	110
110	110	201	201
113	113	203	203
201	133	133	116
203	116	025	025
204	130	021	130
133	114	023	114
116	131	131	021
025		115	023
130			132

APPENDIX *B*
Theorem 1

Let subsets $S_{\rm oi}$ and $S_{\rm ov}$ of structure factors and partial Fourier syntheses $\rho_{\rm oi}(\mathbf{r})$ and $\rho_{\rm ov}(\mathbf{r})$ are defined for a function $\rho(\mathbf{r})$ in accordance with the formulae (5)–(8). If $\mathbf{u}$ is a permitted origin shift and $2\mathbf{u} = \mathbf{0}|_{\rm mod\,1}$, then $\rho_{\rm oi}(\mathbf{r})$ is the half-sum while $\rho_{\rm ov}(\mathbf{r})$ is the half-difference of $\rho(\mathbf{r})$ and its shifted copy $\rho(\mathbf{r} - \mathbf{u})$.

Proof of Theorem 1. Let $\{F_{\mathbf{h}}^{\rm oi} \exp(i\varphi_{\mathbf{h}}^{oi})\}$ be the structure factors corresponding to the sum

$$S(\mathbf{r}) = \tfrac{1}{2}[\rho(\mathbf{r}) + \rho(\mathbf{r} - \mathbf{u})]. \tag{25}$$

The structure factors corresponding to $\rho(\mathbf{r} - \mathbf{u})$ are $\{F_{\mathbf{h}} \exp i[\varphi_{\mathbf{h}} + 2\pi(\mathbf{h}, \mathbf{u})]\}$, thus

$$\{F_{\mathbf{h}}^{\rm oi} \exp(i\varphi_{\mathbf{h}}^{\rm oi})\} = \tfrac{1}{2} F_{\mathbf{h}}\{1 + \exp[2\pi i(\mathbf{h}, \mathbf{u})]\} \exp(i\varphi_{\mathbf{h}}). \tag{26}$$

Owing to the condition $2\mathbf{u} = \mathbf{0}|_{\rm mod\,1}$, the scalar product $(\mathbf{h}, \mathbf{u})$ is integer if $(\mathbf{h}, \mathbf{u}) = 0|_{\rm mod\,1}$ and half-integer otherwise. Therefore,

$$\{F_{\mathbf{h}}^{\rm oi} \exp(i\varphi_{\mathbf{h}}^{\rm oi})\} = \begin{cases} F_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}}) & \text{if } (\mathbf{h}, \mathbf{u}) = 0|_{\rm mod\,1} \\ 0 & \text{otherwise.} \end{cases} \tag{27}$$

As a result, the Fourier series for the $S(\mathbf{r})$ coincides with the partial Fourier synthesis (7) and consequently $S(\mathbf{r})$ coincides with $\rho_{\rm oi}(\mathbf{r})$.

Let $\{F_{\mathbf{h}}^{\rm ov} \exp(i\varphi_{\mathbf{h}}^{\rm ov})\}$ be the structure factors corresponding to the difference

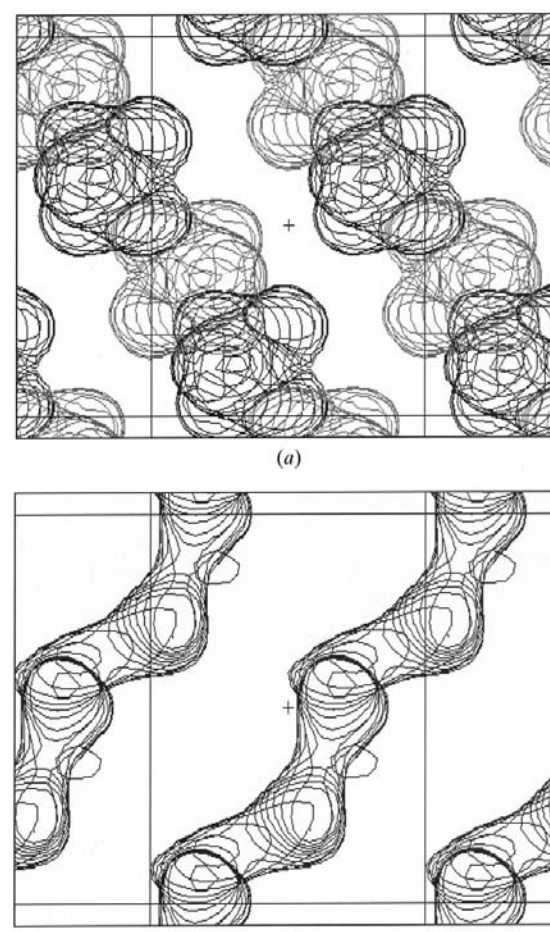

Fig. 6. Fourier syntheses for: (*a*) $\rho(\mathbf{r})$ (black) and $\rho(\mathbf{r} - \mathbf{u})$ (grey); (*b*) $\rho_{\rm ov}(\mathbf{r})$. Ten reflections of S_2 data set are used, $\mathbf{u} = (0, 0, 1/2)$, Z sections are shown.

Table 2. *Connectivity analysis for different VLR phase sets*

Exact phase values are used for the calculations. The number of centrosymmetric and noncentrosymmetric reflections is given in the column headers.

	Connected components and their size (in grid points)				
Ω_p relative volume (%)	26+26 reflections of the 60 Å zone	S_0: 6+5 strongest VLR reflections	S_1: 2+7 t_1-variable reflections	S_2: 5+5 t_2-variable reflections	S_3: 7+4 t_3-variable reflections
5	8*148+8*80+8*43+8*35	8*260+8*46	8*306	8*306	8*306
10	8*422+8*190	4*1224	8*612	4*1224	4*1224
15	8*918	2*3672	8*918	4*1836	4*1836
20	8*1224	2*4896	8*1224	4*2448	4*2448
25	8*1530	2*6120	8*1530	2*6120	1*12240
30	4*3872	1*14688	1*14688	1*14688	1*14688

$$D(\mathbf{r}) = \tfrac{1}{2}[\rho(\mathbf{r}) - \rho(\mathbf{r}-\mathbf{u})]. \tag{28}$$

It follows from $\rho(\mathbf{r}) = D(\mathbf{r}) + S(\mathbf{r})$ that

$$\{F_{\mathbf{h}}^{\mathrm{ov}} \exp(i\varphi_{\mathbf{h}}^{\mathrm{ov}})\} = F_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}}) - F_{\mathbf{h}}^{\mathrm{oi}} \exp(i\varphi_{\mathbf{h}}^{\mathrm{oi}}) = \begin{cases} F_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}}) & \text{if } (\mathbf{h}, \mathbf{u}) \neq 0|_{\mathrm{mod}\,1} \\ 0 & \text{otherwise,} \end{cases} \tag{29}$$

so that the Fourier series for the difference (28) coincides with partial Fourier synthesis (8) and thus $D(\mathbf{r}) = \rho_{\mathrm{ov}}(\mathbf{r})$.

APPENDIX *C*
Proof of Theorem 2 (§4.4)

It follows from (16) that the structure factors $F_{\mathbf{h}}^{\mathrm{oi}} \exp(i\varphi_{\mathbf{h}}^{\mathrm{oi}})$ of the function $\rho_{\mathrm{oi}}(\mathbf{r})$ are

$$F_{\mathbf{h}}^{\mathrm{oi}} \exp(i\varphi_{\mathbf{h}}^{\mathrm{oi}}) = \tau_{\mathbf{h}} F_{\mathbf{h}} \exp(i\varphi_{\mathbf{h}}), \tag{30}$$

where

$$\tau_{\mathbf{h}} = (1/K) \sum_{\kappa=0}^{K-1} \exp[2\pi i(\mathbf{h}, \mathbf{u}_\kappa)]. \tag{31}$$

It is obvious that

$$\tau_{\mathbf{h}} = 1 \quad \text{for } \mathbf{h} \in S_{\mathrm{oi}}(\Gamma_0). \tag{32}$$

Let us show that $\tau_{\mathbf{h}} = 0$ otherwise.

Γ_0 being a finite Abelian group, it may be represented as a direct sum of its primary cyclic subgroups. In other words, every $\mathbf{u}_\kappa$ may be uniquely represented in the form

$$\mathbf{u}_\kappa = m_1\mathbf{u}_1^0 + m_2\mathbf{u}_2^0 + \ldots + m_n\mathbf{u}_n^0, \tag{33}$$

where $\mathbf{u}_1^0, \ldots, \mathbf{u}_n^0$ are some shifts from $\{\mathbf{u}_\kappa\}_{\kappa=0}^{K-1}$, numbers $0 \le m_j \le P_j - 1 \quad (j = 1, \ldots, n)$ are integers and $P_1, \ldots, P_n$ are prime numbers such that $P_1P_2 \ldots P_n = K$ and

$$P_j\mathbf{u}_j^0 = \mathbf{0}|_{\mathrm{mod}\,\mathbf{1}}. \tag{34}$$

Therefore,

$$\tau_{\mathbf{h}} = (1/K) \sum_{m_1=0}^{P_1-1} (z_1)^{m_1} \sum_{m_2=0}^{P_2-1} (z_2)^{m_2} \ldots \sum_{m_n=0}^{P_n-1} (z_n)^{m_n} \tag{35}$$

with

$$z_j = \exp[2\pi i(\mathbf{h}, \mathbf{u}_j^0)]. \tag{36}$$

If for some j the value of $(\mathbf{h}, \mathbf{u}_j^0) \neq 0$ then, from (34),

$$\sum_{m_j=0}^{P_j-1} (z_j)^{m_j} = 1 - (z_j)^{P_j}/(1 - z_j) = \{1 - \exp[2\pi i(\mathbf{h}, P_j\mathbf{u}_j^0)]\}/(1 - z_j) = 0 \tag{37}$$

and the corresponding $\tau_{\mathbf{h}} = 0$. As a consequence, $\tau_{\mathbf{h}} \neq 0$ if and only if $(\mathbf{h}, \mathbf{u}_j^0) = 0$ for all j, which gives $\tau_{\mathbf{h}} = 1$. As

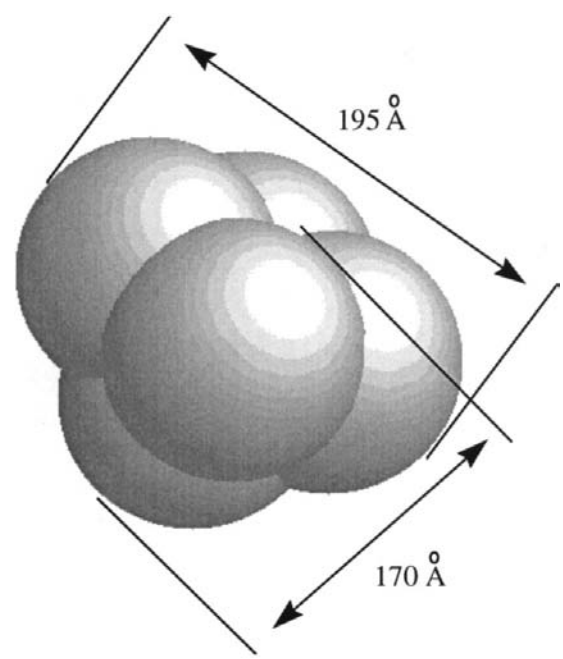

Fig. 7. The envelope of the test object.

follows from (33), in this case $(\mathbf{h}, \mathbf{u}_\kappa) = 0$ for all $\kappa = 0, \ldots, K-1$, *i.e.* $\mathbf{h} \in S_{\rm oi}(\Gamma_0)$.

The authors thank Professor C. Lecomte and the Université Henri Poincaré, Nancy 1, for the invitation to VL for a work visit. The decomposition analysis was initiated by the low-resolution phasing of the ribosomal particles, and the authors thank Professor A. Yonath and Dr I. Agmon for access to the experimental ribosome data and Dr A. Podjarny for intensive and continuous collaboration. This work was supported by grant RFBR 97-04-48319.

References

Andersson, K. M. & Hovmöller, S. (1996). *Acta Cryst.* D**52**, 1174–1180.

Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 186–192.

Ban, N., Freeborn, B., Nissen, P., Penszek, P., Grassucci, R. A., Sweet, R., Frank, J., Moore, P. B. & Steitz, T. A. (1998). *Cell*, **93**, 1105–1115.

Bricogne, G. & Gilmore, C. J. (1990). *Acta Cryst.* A**46**, 284–297.

Dorset, D. L. (1997). *Acta Cryst.* A**53**, 445–455.

Dorset, D. L. & Jap, B. K. (1998). *Acta Cryst.* D**54**, 615–621.

Giacovazzo, C. (1974). *Acta Cryst.* A**30**, 390–395.

Giacovazzo, C. (1980). *Direct Methods in Crystallography*. New York: Academic Press.

Gilmore, C. J. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 159–167. Dordrecht: Kluwer Academic Publishers.

Gilmore, C., Dong, W. & Bricogne, G. (1999). *Acta Cryst.* A**55**, 70–83.

Harris, G. W. (1995). *Acta Cryst.* D**51**, 695–702.

Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. (1997). *Nature (London)*, **389**, 251–260.

Lunin, V. Y. (1993). *Acta Cryst.* D**49**, 90–99.

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

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Urzhumtsev, A. G. & Podjarny, A. D. (1998*a*). *Acta Cryst.* D**54**, 726–734.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Urzhumtsev, A. G. & Podjarny, A. D. (1998*b*). 18th European Crystallographic Meeting, Praha, 1998, Abstracts (A), pp. 131–132.

Lunin, V. Y., Lunina, N. L., Petrova, T. E., Vernoslova, E. A., Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Acta Cryst.* D**51**, 896–903.

Lunin, V. Y., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.*, A**46**, 540–544.

Miller, R. & Weeks, C. M. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 389–400. Dordrecht: Kluwer Academic Publishers.

Mishnev, A. F. (1998). Personal communication.

Podjarny, A. D., Rees, B., Thierry, J.-C., Cavarelli, J., Jesior, J. C., Roth, M., Lewitt-Bentley, A., Kahn, R., Lorber, B., Ebel, J.-P., Giegé, R. & Moras, D. (1987). *J. Biomol. Struct. Dynam.* **5**, 187–198.

Sheldrick, G. M. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer Academic Publishers.

Subbiah, S. (1991). *Science*, **252**, 128–133.

Subbiah, S. (1993). *Acta Cryst.* D**49**, 108–119.

Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography*, **32**, 12–16.

Urzhumtsev, A. G., Vernoslova, E. A. & Podjarny, A. D. (1996). *Acta Cryst.* D**52**, 1092–1097.

Woolfson, M. M. (1954). *Acta Cryst.* **7**, 65–67.

Woolfson, M. M. (1998). Personal communication