Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Connectivity-based *ab initio* phasing: from low resolution to a secondary structure

Natalia Lunina,[a] Vladimir Y. Lunin[a] and Alexandre Urzhumtsev[b]*

[a]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region 142290, Russia, and [b]LCM3B, UMR 7036 CNRS, Faculté des Sciences, Université Henry Poincaré, Nancy I, 54506 Vandoeuvre-lés-Nancy, France

Correspondence e-mail: sacha@lcm3b.uhp-nancy.fr

Received 5 May 2003
Accepted 8 July 2003

The connectivity-based phasing method currently allows *ab initio* determination of phases for several hundred reflections. In the case of large macromolecular crystals, these reflections correspond to a very low resolution and the structural information deduced essentially consists of the molecular packing and an approximate molecular envelope. However, when the unit cell is relatively small, such a phasing procedure can produce phases such that secondary-structure elements can be identified in the corresponding maps. In the case of the pheromone E*r*-1, all three α-helices present are seen in the *ab initio* phased maps. In the case of protein G, not only the α-helix but also some individual β-strands are distinguishable.

1. Introduction

Direct phasing of a single set of experimental magnitudes at low resolution can provide useful information for structure determination when conventional phasing techniques are not applicable. The incorporation of the connectivity analysis of Fourier syntheses (Lunin *et al.*, 2000) into the framework of a general low-resolution *ab initio* phasing method (Lunin *et al.*, 1990, 2002) makes this method one of the most efficient in this resolution range. Several examples of successful application of this method to the determination of molecular positions have been published previously (Lunina *et al.*, 2000). A series of test calculations (all performed with experimental structure-factor magnitudes) was completed by the successful application of the method to the determination of the structures of LDL (Lunin *et al.*, 2001) and lectin SML-2 (Müller & Lunina, in preparation). It is worthy of note that from the very beginning of protein crystallography the connectivity properties of Fourier syntheses maps have been used for qualitative estimation of the success of phasing. Bhat & Blow (1982) and essentially Wilson & Agard (1993) and Baker, Bystroff *et al.* (1993) used these properties as the basis for map modification in phasing. In our approach, similar to that of Baker, Krukowski *et al.* (1993), this information is expressed numerically and becomes a selection criterion.

The goal of this work was to study the possibility of using this method to extend phase information to medium resolution and to examine the kinds of structural details that can be obtained using such an approach. We do not discuss all the details of the protocols applied and the methods by which these protocols were obtained; the choice of protocols and definition of optimal parameters could be the topic of an independent paper. Our basic aim here is to demonstrate that with some reasonable protocols *ab initio* phasing of several hundred of the lowest resolution reflections for small proteins can lead to visualization of secondary-structure elements and not just a rough envelope.

The observed structure-factor magnitudes were used in all the tests discussed below. The comparison of phases found *ab initio* was performed with those calculated directly from atomic model. We did not introduce corrections for bulk solvent (Urzhumtsev & Podjarny, 1995) in order to avoid discussion of the choice of the bulk-solvent model and the estimation of corresponding parameters, which is a separate issue (we note that some of these models modify the magnitudes only and not the phases). Secondly, the Fourier maps calculated with the model phases show the macromolecules well, sometimes even better than the maps calculated with the corrected phases, and for our goal of visualization of secondary-structure elements these phases are quite suitable for following the progress of phasing.

2. Phasing method

2.1. Phasing procedure

The connectivity-based phasing method (Lunin *et al.*, 2000) is founded on the observation that the topological properties of high-density regions of Fourier syntheses differ between properly phased syntheses and those calculated with random or erroneous phases.

Let $\rho(\mathbf{r})$ be a Fourier syntheses calculated on a grid in the unit cell, ρ^* be some cutoff level and $\Omega(\rho^*) = \{\mathbf{r}: \rho(\mathbf{r}) \geq \rho^*\}$ be the high-values region mask, which is defined in this paper as the region composed of all real-space points such that the Fourier synthesis value at the point exceeds the specified cutoff level. If the cutoff level ρ^* is chosen appropriately, the mask corresponding to the correctly phased low-resolution macromolecular synthesis is usually composed of a small number of isolated components. At very low resolution, the number of these components is equal to the number of molecules in the unit cell, while at a higher resolution there may be several components per molecule. On the other hand, the masks for randomly phased syntheses are likely to contain infinite merged regions (for low-resolution syntheses) or/and a large number of small 'drops' (for medium-resolution syntheses).

It is important to keep in mind that in the following the region $\Omega(\rho^*)$ used for connectivity study does not cover the whole molecule but corresponds to its central part or to its mostly dense molecular domains.

Such differences in the features of the high-density regions may serve as a basis for a phasing procedure that follows the general scheme suggested previously by Lunin *et al.* (1990) and consists of the following steps.

(i) For the available independent structure-factor magnitudes, a large number of random phase sets are generated. The phases are generated uniformly at the beginning or in accordance with known phase distributions if this information is already available (*e.g.* from the previous steps of *ab initio* phasing). The phase distribution is supposed to be continuous for acentric reflections and is limited to two allowed values for centric reflections.

(ii) For each generated phase set, the Fourier synthesis is calculated using the experimental structure-factor magnitudes and the generated phases.

(iii) The connectivity analysis is performed for the mask of the Fourier synthesis corresponding to the chosen cutoff level ρ^*. This analysis implies the identification of connected components in the mask and the calculation of the number of such components per unit cell and their individual volumes. The connectivity-analysis algorithm used is discussed in Appendix *A*.

(iv) If the results of the connectivity analysis satisfy the imposed conditions (selection rules), the generated phase set is selected for further analysis. In the simplest mode, the phase set is selected if the number of components is equal to the expected number of molecules in the unit cell. Further selection rules are discussed below.

(v) After a reasonable number of phase sets (*e.g.* about 100) have been selected, they are 'aligned' using the permitted origin/enantiomorph choices (Lunin *et al.*, 1990; Lunin & Lunina, 1996) and averaged. For each reflection $\mathbf{h}$, this averaging produces the 'best' phase $\varphi^{\rm best}(\mathbf{h})$ and its individual figure of merit $m(\mathbf{h})$ which reflects the spread of the phase values corresponding to this reflection in different selected phase sets. It is worthy of note that the result of averaging may depend on the alignment used.

The details of the method are given in Lunin *et al.* (2000) and are not discussed here. At the same time, it is useful to restate some terms used below in the description of the tests. The most important managing parameters of the method are the cutoff level ρ^* used when preparing Fourier synthesis masks and the set of selection rules.

2.2. Fourier synthesis mask

The first important parameter of the method is the cutoff level used to define the high-values region. In our tests, it is defined mostly through the relative volume of the corresponding mask. For a given Fourier synthesis calculated on a grid, each cutoff level ρ^* separates all points of the unit cell into two sets: those with a synthesis value higher than ρ^* and those lower. If N is the total number of grid points in the unit cell and N^* is the number of those with a high density, *i.e.* belonging to the mask $\Omega(\rho^*) = \{\mathbf{r}: \rho(\mathbf{r}) \geq \rho^*\}$, the relative volume is the ratio $\alpha = N^*/N$. The higher ρ^*, the smaller α and *vice versa*. Sometimes it is convenient to use a more invariant way of specifying the cutoff level, *e.g.* to specify it by the ratio of the mask volume to the number of residues, where the volume and the number of residues correspond to the content of a unit cell.

It was found empirically (Lunin *et al.*, 2000) that in many cases a specific volume near to 25 Å^3 per residue is a suitable choice for the beginning of the work, which roughly corresponds to $\alpha = 0.1$ for 'usual' proteins. This volume is very small and confirms the remark made in the previous section that the region $\Omega(\rho^*)$ should not be confused with the molecular region and that the parameter is not directly linked to the molecular volume (while the choice of ρ^* may be influenced

by this latter). Also, this value of 25 Å^3 is only indicative and both smaller and larger values can be used in practice depending on particular problems.

2.3. Connectivity analysis

For the chosen parameter α (or corresponding ρ^*), the region $\Omega(\rho^*)$ may be composed of several connected components whose number and shape vary with α. The method of composition of the mask from the connected components is the object of the connectivity analysis. Some maps show high-density regions continuously crossing the whole space. This is rather unusual for correctly phased maps and may be used to eliminate wrong phase sets. If the space group contains $N_{\rm sym} > 1$ symmetry operations, each connected component should be presented, generally speaking, by $N_{\rm sym}$ copies, but this is not always the case when the symmetry-linked regions are merged together, for example. Naturally, the regions linked by crystallographic symmetry have exactly the same volume. On the other hand, the map can contain several connected regions not linked by symmetry that accidentally have exactly the same volume. The selection criteria discussed in this paper characterize each domain by its volume only and therefore cannot distinguish between these regions. The finite component volume is estimated in the tables below by the number of grid nodes covered by the component. The number of regions with exactly the same volume is referred to below as 'the multiplier corresponding to the considered volume'. The list of component volumes and their multipliers (arranged in the decreasing order) produces the output of the connectivity analysis.

Strictly speaking, the features listed characterize some mask and not the phase set. Nevertheless, in this paper the only masks considered are those corresponding to Fourier syntheses calculated with the observed magnitudes and trial phases. Therefore, we will link the connectivity-analysis output list and the connectivity properties to the trial phase set as well.

The full connectivity analysis implies the study of a variety of masks corresponding to different α. Nevertheless, in this paper we restrict the analysis to one previously chosen level only.

2.4. Selection rules

The selection rules in the test discussed below were based on the results of the connectivity analysis of a mask of Fourier synthesis calculated with trial phases and observed magnitudes. For the chosen cutoff level, the number of mask components is calculated in the whole unit cell, taking its periodicity into account. In the tests discussed in the current paper, the next three selection rules were used in various combinations.

Selection rule *A*. The mask must consist of a specified number $N_{\rm comp}$ of connected components of equal volume.

Selection rule *B*. The total number of connected components is less than or equal to some value $N_{\rm comp}$ that is specified in advance; additionally, the multiplier corresponding to the largest component must be equal to a specified value $\mu_{\rm largest}$.

Selection rule *C*. This is similar to rule *B*, but a permitted multiplier $\mu_{\rm second}$ corresponding to the component second in size is also specified.

A selection criterion, in particular one based on the rules formulated above, may depend on some inner parameters such as the cutoff level in the mask building or the resolution of the Fourier synthesis. Obviously, several criteria with different parameter values, as well as criteria of different kinds, may be used together.

2.5. Map alignment

To evaluate the 'formal' closeness of two phase sets $\{\varphi_1(\mathbf{h})\}$ and $\{\varphi_1(\mathbf{h})\}$, the map correlation coefficient, which is equivalent to the phase correlation weighted by magnitudes, was used (Lunin & Woolfson, 1993),

$$C_\varphi = \frac{\int[\rho_1(\mathbf{r}) - \langle\rho_1\rangle][\rho_2(\mathbf{r}) - \langle\rho_2\rangle]\ {\rm d}V_{\mathbf{r}}}{\left\{\int[\rho_1(\mathbf{r}) - \langle\rho_1\rangle]^2\ {\rm d}V_{\mathbf{r}} \int[\rho_2(\mathbf{r}) - \langle\rho_2\rangle]^2\ {\rm d}V_{\mathbf{r}}\right\}^{1/2}} = \frac{\sum\limits_{\mathbf{h}} F^{\rm obs}(\mathbf{h})^2 \cos[\varphi_1(\mathbf{h}) - \varphi_2(\mathbf{h})]}{\sum\limits_{\mathbf{h}} F^{\rm obs}(\mathbf{h})^2}. \quad (1)$$

Here, $\rho_i(\mathbf{r})$ is the Fourier synthesis calculated with the observed magnitudes $\{F^{\rm obs}(\mathbf{h})\}$ and phases $\{\varphi_i(\mathbf{h})\}$, $i = 1, 2$. This measure was found to be more appropriate than the mean phase difference when comparing low-resolution maps, where the strongest reflections are more important than the weaker ones. Two phase sets that are formally very different may in fact present very close solutions of the phase problem, but linked to different origin/enantiomorph choices. To take this into account, some kind of alignment must be performed in accordance with the origin/enantiomorph choices permitted before calculation of the formal phase closeness is performed (Lunin *et al.*, 1990; Lunin & Lunina, 1996). In the present work, the alignment was based on maximization of the map correlation coefficient.

The correlation coefficient (1) for two phase sets depends on the resolution $d_{\rm align}$ at which these sets are compared; variation of this resolution may change the optimal phase alignment. In order to simplify the presentation of the material, in the current article we do not discuss the choice of this essential but technical parameter of the alignment procedure. It may be considered in the following that this resolution is equal to the highest resolution used in the selection rules in each phasing step.

2.6. Phase refinement

In the first run of the phasing of a particular reflection $\mathbf{h}$, the corresponding phase value $\varphi(\mathbf{h})$ is generated as a random variable distributed uniformly over the $[0, 2\pi]$ interval. In subsequent phasing cycles this phase is generated in accordance with the Von Mises distribution (circular normal distribution; Evans *et al.*, 2000),

$$P(\varphi) \simeq \exp\{t(\mathbf{h}) \cos[\varphi - \varphi^{\rm best}(\mathbf{h})]\}. \quad (2)$$

Here, $\varphi^{\rm best}(\mathbf{h})$ is the phase obtained as a result of the averaging in the previous cycle of phasing and the parameter $t(\mathbf{h})$ in the distribution is adjusted so that $\langle\cos(\varphi - \varphi^{\rm best})\rangle = m(\mathbf{h})$, where m is the available estimate of the figure of merit for the given phase.

For centric reflections, the choice is performed for two possible values that are taken with equal probability at the first step and with the estimated probability for phase extension.

3. *Ab initio* phasing of pheromone E*r*-1

3.1. Crystals of E*r*-1

The structure of the pheromone E*r*-1 has been reported by Anderson *et al.* (1996), who considered this crystal to be 'a challenging case for protein crystal structure determination'. The crystals of E*r*-1 belong to space group *C*2, with unit-cell parameters a = 53.91, b = 23.08, c = 23.11 Å, β = 110.4°. A complete data set at a resolution of 1 Å is available for this crystal. Very dense packing of the macromolecule (Matthews coefficient $V_{\rm M}$ = 1.53 Å^3 Da^{-1}) made its structure determination impossible using some conventional methods and difficult using others (Anderson *et al.*, 1996). This packing suggests E*r*-1 to be a very interesting example for our connectivity-based phasing method, which is expected to work more easily if the macromolecules are well separated in the unit cell. Another feature of this protein is that it is formed of three α-helices. Owing to the relatively small size of the protein, the number of independent reflections is small even at the resolution of 4–5 Å (Table 1) at which the helices can be recognized if the quality of the Fourier syntheses is reasonable.

3.2. General strategy of the search

Phasing of the diffraction data was performed in several steps. Firstly, phases were found for a few tens of reflections of lowest resolution. This phase set was then extended in several iterations, with several tens of reflections added each time. This choice of the number of reflections allowed a relatively complete search of the phase space in each step. The space group *C*2 has four symmetry transformations, so that the first (low-resolution) selection rule may be formulated as 'the unit-cell mask falls into four connected components of the equal volumes' (one component per the molecule), which is the selection rule *A* listed above with the multiplier $N_{\rm comp}$ = 4 (see §2.4). When the resolution increases, these components can decompose into several smaller ones. The way in which this decomposition happens is unknown and it is difficult to predict the exact composition of such an image. However, the number of connected components should not be large (maps with many drops are considered to be noisy maps). It can also be supposed that the number of major connected regions is equal to the number of molecules, *i.e.* to four. In all cases, infinite components were forbidden (condition not shown in Table 2).

An important parameter of the connectivity analysis is the cutoff level at which the Fourier synthesis masks are built and analyzed. In the current case of a densely packed protein, it can be expected that the components for individual molecules would merge into each other if a cutoff corresponding to the recommended specific volume of 25 Å^3 per residue (the relative volume occupied by the mask is 0.17) is chosen. Instead, the relative volume of the mask was taken as 0.13 (roughly 19 Å^3 per residue).

In each step, random phase sets were generated until 100 of them were selected. To simplify the analysis and comparison of the results, all maps were calculated on the same grid with grid size 40 × 20 × 20, which is sufficient to work at a resolution d of approximately 4 Å, assuming as appropriate a grid step of d/3.

Table 1
Number of reflections in resolution zones for pheromone E*r*-1.

The last row shows the number of additional reflections in each resolution zone compared with the previous one.

Resolution (Å)	11.0–∞	7.5–∞	5.8–∞	5.0–∞	4.5–∞	4.1–∞	3.9–∞
No. reflections	13	39	82	132	178	230	264
+ No. reflections		26	43	50	46	52	34

Table 2
Scheme of connectivity-based phasing of the pheromone E*r*-1.

Here, $d_{\rm mask}$ is the resolution of the synthesis subjected to connectivity analysis and $N_{\rm comp}$, $\mu_{\rm largest}$ and $\mu_{\rm second}$ are the total number of connected components and the multiplicity of the largest and the second components, respectively. *A*, *B* and *C* are the selection rules described in §2.4.

Cycle	$d_{\rm mask}$ (Å)	Selection rule	$N_{\rm comp}$	$\mu_{\rm largest}$	$\mu_{\rm second}$	No. of generated phase sets
1	11.0–∞	*A*	4	4	—	72818
	17.5–∞	*C*	≤24	4	4 or 0	
2	11.0–∞	*A*	4	4	—	17876
	7.5–∞	*C*	≤24	4	4 or 0	
	5.8–∞	*B*	≤36	4	Any	
3	11.0–∞	*A*	4	4	—	81127
	7.5–∞	*C*	≤24	4	4 or 0	
	5.8–∞	*B*	≤32	4	Any	
	15.0–∞	*B*	≤36	4	Any	
4	11.0–∞	*A*	4	4	—	53100
	7.5–∞	*C*	≤24	4	4 or 0	
	5.8–∞	*B*	≤32	4	Any	
	4.5–∞	*B*	≤60	4	Any	
5	11.0–∞	*A*	4	4	—	28354
	7.5–∞	*C*	≤24	4	4 or 0	
	5.8–∞	*B*	≤32	4	Any	
	4.1–∞	*B*	≤100	4	Any	
6	11.0–∞	*A*	4	4	—	46733
	7.5–∞	*C*	≤24	4	4 or 0	
	5.8–∞	*B*	≤32	4	Any	
	3.9–∞	*B*	≤120	4	Any	

3.3. Results of phasing

Details of the application of the general strategy described above are given in Table 2. For each generated phase set, the Fourier map masks were analyzed at several resolutions

Table 3
Comparison of the phases obtained *ab initio* with the phases calculated from the atomic model for pheromone E*r*-1.

The map correlation coefficient (1) multiplied by 100 is shown for different resolution shells and different cycles of the phasing.

Resolution (Å)	3.9–4.1	4.1–4.5	4.5–5.0	5.0–5.8	5.8–7.5	7.5–11.0	11.0–∞	7.5–∞	5.8–∞	5.0–∞	4.5–∞	4.1–∞	3.9–∞
No. of reflections	34	52	46	50	43	26	13	39	82	132	178	230	264
Map correlation													
Step 1	41	43	39	41	38	66	69	66	50	40	33	29	29
Step 2	29	48	33	44	39	70	76	70	54	41	32	28	26
Step 3	23	41	37	37	48	69	82	70	49	41	32	25	23
Step 4	19	49	45	39	49	68	80	68	51	40	32	26	23
Step 5	29	69	55	49	42	66	67	64	50	43	30	27	26
Step 6	17	63	58	40	42	74	69	72	51	40	31	28	26

simultaneously and different rules (*A*–*C*) were used in different combinations. As seen in Table 2, the allowed number of connected components increased with the resolution. The first phasing step required the largest number of phase sets to be checked. When the phases for the first 39 reflections were evaluated (step 1), the phasing of the next 43 reflections (step 2) required many fewer phase sets to select the required 100 phase sets for averaging. This probably means that the connectivity properties of the maps at 5.8 Å resolution are essentially defined by the reflections of lower resolution. This is different from step 3, where the number of analyzed phase sets again increased.

In order to estimate *a posteriori* the quality of the generated phases in this test case where the atomic model is already known, the weighted phase correlation for the aligned phase sets was used. The map correlation was calculated for two phase sets, one obtained *ab initio* and the second one calculated from the available atomic model (Anderson *et al.*, 1996). This correlation is shown in Table 3 as a function of the resolution zone.

The correlation shown for the maps both at a given resolution and in different resolution shells indicates that the conditions imposed at a higher resolution were important in order to increase the quality of the corresponding phases (see, for example, columns '4.1–4.5' and '4.5–5.0'). At the same time, when higher resolution reflections are included in the selection, the quality of the lower resolution reflections (column '11.0–∞') first increased somewhat and then fell. This has been observed previously and shows the need to improve the procedure further and the ways to achieve this. It follows from the first column ('3.9–4.1') that the applied procedure was not sufficient to obtain good phases for the reflections of resolution about 4 Å. The reason may be an insufficient number of iterations, the use of inadequate criteria or the resolution being too high for the given procedure; this analysis is outside the scope of the current study.

In any case, the main criterion of the map quality is the structural details that can be found in the corresponding maps. Figs. 1 and 2 show the weighted maps calculated with the experimental magnitudes and the phases and figures of merit (FOMs) obtained *ab initio*.

Individual molecules, represented by the peaks at their centres, are seen only at very low resolution, for example 11 Å (Fig. 1*a*). These maps show four clear peaks which are isolated even at quite a low cutoff level corresponding to 0.35 of the unit-cell volume. At 7.5 Å, the map already shows eight or 12 isolated regions, depending on the cutoff level chosen (Fig. 1*b*) and their composition into individual molecules is impossible.

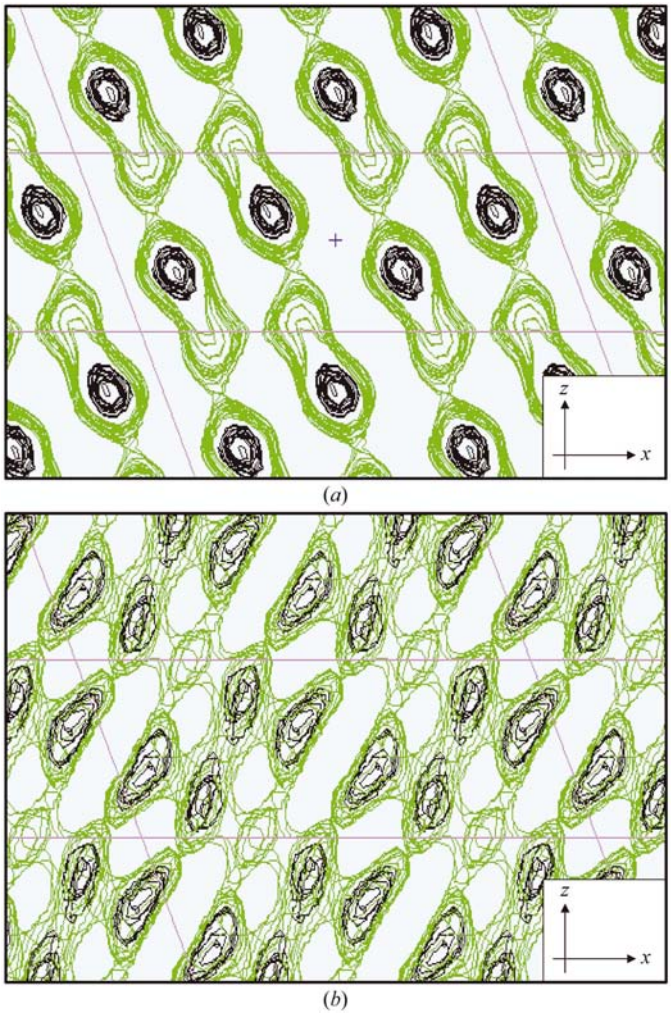

Figure 1
Projections of weighted Fourier synthesis maps (the whole unit cell) for the pheromone E*r*-1 calculated with the phases found *ab initio*. (*a*) Map resolution is 11.0 Å; the contours correspond to the relative mask volumes 0.04 (black) and 0.35 (green); (*b*) map resolution is 7.5 Å; the contours correspond to the relative mask volumes 0.04 (black) and 0.20 (green).

Table 4
Number of reflections in resolution zones for protein G.

The last line shows the number of additional reflections in each resolution zone compared with the previous one.

Resolution (Å)	16.0–∞	12.0–∞	10.0–∞	9.0–∞	8.0–∞	7.0–∞	6.5–∞	6.0–∞	5.0–∞	4.5–∞	4.0–∞	3.5–∞	3.2–∞	3.0–∞
No. reflections	15	28	46	61	85	117	150	181	305	411	580	847	1101	1323
+ No. reflections		13	18	15	24	32	33	31	93	106	169	267	254	222

If all reflections to a resolution of 7 Å are included in the map calculation, the map obtained (Fig. 2*a*) clearly shows three connected elongated regions corresponding to all three α-helices of the protein. The size of these regions correlates well with the length of the corresponding helices and they are well positioned; however, the direction of two of them is slightly different from the direction of the corresponding helices. Interestingly, while the helices are clearly seen in the *ab initio* obtained maps, the dense packing does not allow the separation of molecules. Similar results in the connectivity-based phasing of E*r*-1 were obtained independently by Chabrière *et al.* (2001).

When a map is calculated at a formal resolution of 4 Å, the effective resolution of the map is lower owing to weighting by FOMs (we do not formalize this notion of the effective resolution here). Nevertheless, the addition of reflections provides us with some further detail (Fig. 2*b*). In particular, the density at the left part of this image clearly follows several turns of the helix. While this demonstrates the potential of the method, the noise present in this map and the less clear identification of the other two helices show the current limitations.

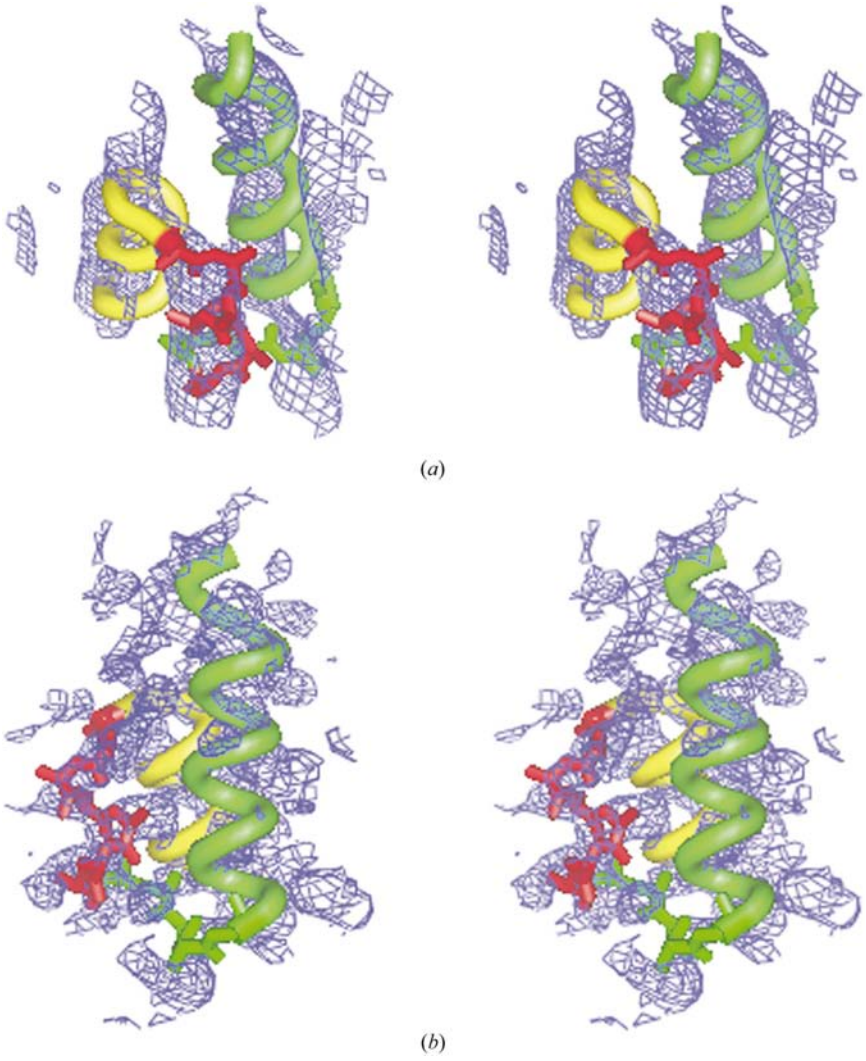

Figure 2
Ab initio phased weighted Fourier synthesis maps for pheromone E*r*-1 superposed with its atomic model. (*a*) Resolution 7.0 Å; note the continuous density for each of the three helices; (*b*) resolution 4.0 Å; the image is rotated in comparison with (*a*) in order to optimally present the density corresponding to the helix shown in red.

4. *Ab initio* phasing for protein G

4.1. Crystals of protein G

The crystals of protein G belong to space group $P2_12_12_1$, with unit-cell parameters a = 34.9, b = 40.3, c = 42.2 Å. The asymmetric part of the unit cell contains a single molecule. The structure of this protein has been solved by Derrick & Wigley (1994) at a resolution of 1.1 Å, for which a complete data set is available. This protein contains 61 residues (about 600 non-H atoms) and its secondary structure is composed of one α-helix and one β-sheet formed by four strands. The number of independent reflections for different resolution shells is shown in Table 4. For this crystal, the specific volume of 25 Å^3 per residue corresponds to the relative volume of a Fourier synthesis mask near to α = 0.1. Masks with this relative volume were used in the connectivity analysis.

4.2. Phasing strategy

Analysis of the number of reflections allows the establishment of a general strategy for phasing. The procedure can be started from the phasing of 15 reflections at a resolution of 16 Å where individual molecules can be searched for. Since there are four molecules in the unit cell, the number of connected components expected in the correct phase synthesis is four. The same condition of four connected compo-

Table 5
Protocols for different start trials of the connectivity-based phasing for the protein G.

Here, d_{mask} is the resolution of the synthesis subjected to connectivity analysis and N_{comp} and $\mu_{largest}$ are the total number of connected components and the multiplicity of the largest component, respectively. *A* and *C* are the selection rules described in §2.4. The last three columns present the result of the comparison of the averaged phases with the exact phases. The total number of generated phase sets (the first value) and the number of phase sets kept after application of one selection criterion after another are given in the column 'No. sets'.

						Comparison with the exact phases		
Protocol	d_{mask} (Å)	Selection rule	N_{comp}	$\mu_{largest}$	No. sets	Resol.	$\langle m(\mathbf{h})\rangle$	Corr.
*P*1.1	16.0	*A*	4	4	216	16.0	0.38	0.75
					100	12.0	0.24	0.50
						10.0	0.18	0.36
						9.0	0.16	0.30
						8.0	0.14	0.26
*P*2.1	16.0	*A*	4	4	2247	16.0	0.21	0.72
	12.0	*A*	4	4	1027	12.0	0.25	0.65
					100	10.0	0.18	0.40
						9.0	0.15	0.36
						8.0	0.13	0.31
*P*3.1	16.0	*A*	4	4	10661	16.0	0.22	0.77
	12.0	*A*	4	4	5038	12.0	0.27	0.69
	10.0	*C*	≤8	4	496	10.0	0.20	0.51
					100	9.0	0.17	0.46
						8.0	0.14	0.41
*P*3.2	16.0	*A*	4	4	2020	16.0	0.23	0.74
	8.0	C	≤12	4	1210	12.0	0.24	0.67
					100	10.0	0.24	0.56
						9.0	0.20	0.50
						8.0	0.17	0.43

nents can be also tried at 12 Å since the number of reflections is still small. At higher resolution this is no longer true and each of symmetry-linked molecules can already be seen as two or three compounds. Therefore, another selection criterion can be tried and the permitted number of regions can be defined as eight or 12, with the first multiplier equal to four.

At a resolution of 8 Å the data set consists of 85 reflections, which is too many for the initial exhaustive search. Therefore, if we want to arrive at such a resolution, we need to perform this in several iterations. It is simpler to choose the grid once for all phasing steps, where the last step will be performed at a resolution of about 8 Å. The traditional rule 'grid step = resolution/3' can be applied, giving a grid size of $18 \times 20 \times 24$. Such a conservation of the grid for all steps is not a necessary requirement and is performed only for convenience. Otherwise, the grid can be increased with phase extension.

Several phasing protocols were subsequently applied as discussed below. Their results are summarized in Table 5. Two of the protocols (*P*1.1 and *P*2.1) consist of one phasing cycle and the third protocol consists of two phasing cycles (*P*3.1 and *P*3.2).

4.3. Initial phasing

4.3.1. Protocol *P*1. Following the analysis discussed above, the connectivity analysis was performed at 16 Å resolution. The generated phase set was selected if the corresponding mask contained exactly four connected components. Infinite components were always forbidden (condition not shown in Table 5). The procedure had already selected 100 variants from the first 216 phase sets generated, allowing us to believe that the mean phase set obtained with such parameters would not be much better than a random set. Nevertheless, these selected sets were aligned at the resolution of 16 Å and averaged. This averaging resulted in a mean value of the FOM of near 0.38 and a relatively high correlation with the exact phases (0.75 at 16 Å resolution; Table 5).

It is worthy of note that when generating random phase sets the phases of reflections at any resolution may be generated. Similarly, when *a posteriori* comparing the found estimates of phase values with the exact phases, the map correlation coefficient may be calculated for different resolution zones. Nevertheless, if a particular reflection was not involved in calculation of the analyzed syntheses (*e.g.* it is of too high a resolution), then it will be presented by arbitrary values of the corresponding phase in the selected phase sets, as this phase value was not restricted. This is usually reflected by a near-zero value of the corresponding figure of merit obtained in the averaging.

The maps calculated at a resolution of 16 Å were superimposed with the corresponding exact maps. Two different projections perpendicular to the crystallographic axes 0*Z* and 0*X* show (Fig. 3*a*) that the molecular centre, indicated by the major peak in the map calculated with the model phases, was determined relatively well, but the molecular envelope could yet be determined from this *ab initio* phased map. In the following, this phase set is referred to as *P*1.1. It can be remarked that even the map with the model phases does not show the molecular envelope (there is practically no density for the α-helix as it appears in Figs. 3*b*–3*d*); this lack of information is a consequence of too low a resolution rather than of imprecise phase values.

4.3.2. Initial phasing with a stronger selection rule: protocol *P*2. In order to increase the quality of the obtained phases, a more strict selection was applied in the initial phasing cycle by simultaneously checking two conditions instead of the single one used in protocol *P*1. It was now requested that four connected components appear not only in a 16 Å resolution-based mask but also in a 12 Å resolution-based mask. Other parameters were taken as previously.

The procedure now selected 100 variants after 2247 phase-set generations. Similarly to *P*1, the first condition reduced the number of phase sets by close to twofold (1027 sets satisfied the first selection rule), but only 100 of these 1027 phase sets also satisfied the second rule. The new averaged phases obtained were generally better than the previous ones. While the phase correlation in the 16 Å resolution zone slightly decreased, from 0.75 to 0.72, it significantly increased for higher resolution reflections (Table 5). Therefore, this second phase set, referred to in the following as *P*2.1, should be a better starting point for further phase extension.

The improved quality of this map in comparison with the results of previous phasing can be identified by visual analysis. The corresponding map shows density corresponding to the

α-helix and not only to the centre of the molecule (Fig. 3*b*). This map is formally calculated at a resolution of 8 Å, although its effective resolution is much lower because of weighting by FOMs.

When the model and the exact phases are unknown, several phasing strategies can be tested and the quality of several obtained phase sets can be compared by analyzing the connectivity properties of the corresponding maps. In order to illustrate this, two maps at a resolution of 12 Å were calculated with the phases *P*1.1 and *P*2.1, respectively. For these maps, the number and the volume of connected components were analyzed for different cutoff levels. The map with phases *P*1.1 showed four connected components up to a cutoff corresponding to a relative volume of 0.10. At low cutoff levels the components started to merge and there were only two connected components when the selected relative volume was varied from 0.10 to 0.28.

For the map calculated with the phases *P*2.1, this switching from four to two isolated regions happens at a significantly lower level, at only 0.18. This better separation of regions, which for an unknown structure are believed to correspond to individual molecules, lets us choose of the second set as the better one in a practical situation when the exact phases are unknown.

4.3.3. Further reinforcement of the starting selection: protocol *P*3. An attempt to improve phases further was made using one extra condition. This new condition was defined similarly to the two used previously, but was applied to a higher resolution (10 Å in this test) mask. However, at such a resolution the maps do not necessarily have four connected components; they can already be split into smaller parts. Therefore, this extra condition was formulated as the selection of maps that at 10 Å resolution show not more than eight connected regions (in other words, each molecule can be represented by one or two regions).

More phase sets, 10 661 in total, were generated before 100 of them were selected. It may be noted that increasing the number of criteria correspondingly increases the CPU time. The corresponding averaged phases were called phase set *P*3.1. This resulting map at a formal resolution of 8 Å is shown in Fig. 3(*c*), in which an α-helix can be distinguished already.

4.3.4. Phase refinement and extension: protocol *P*3.2. The resolution of the synthesis should be increased if more structural details are to be searched for. It is not possible to perform the search at a high resolution from the beginning because the number of reflections to be phased becomes too large and the search becomes either impossibly long or too incomplete. Therefore, the already obtained phase information should be used as a starting point for such a phase extension. For the reflections in a higher resolution shell which were excluded from the previous phasing, the phases are searched uniformly as before, while for the previously phased reflections new phase values are generated using the Von Mises distribution (see §2.6). Figures of merit calculated in the initial step are used to define the parameter of this distribution for each reflection (Lunin *et al.*, 1990). The higher the FOM, the sharper the probability distribution around the already available phase is and the lower is the chance of shifting it far away from this value.

Table 6
Analysis of connectivity of the 8 Å resolution maps calculated with the phases obtained by the averaging of the variants in the *A*3 and *A*4 clusters (see §4.4).

N_{tot} is the total number of connected components in the Fourier synthesis mask.

Relative mask volume	Phase set *A*3 N_{tot}	Comp. 1	Comp. 2	Comp. 3	*A*4 N_{tot}	Comp. 1	Comp. 2	Comp. 3
0.014	8	4 × 171	4 × 28	0	8	4 × 196	4 × 2	
0.028	12	4 × 283	4 × 105	4 × 8	12	4 × 354	4 × 42	4 × 6
0.042	12	4 × 377	4 × 165	4 × 58	12	4 × 485	4 × 99	4 × 18
0.056	12	4 × 561	4 × 225	4 × 14	12	4 × 606	4 × 152	4 × 37
0.070	8	4 × 680	4 × 319	0	12	4 × 724	4 × 219	4 × 58
0.083	4	4 × 1204	0		12	4 × 843	4 × 276	4 × 82
0.097	4	4 × 1398	0		12	4 × 958	4 × 343	4 × 99
0.100	4	4 × 1441	0		12	4 × 979	4 × 358	4 × 105
0.125	2	2 × 3598	0		10	4 × 1170	2 × 960	4 × 155
0.130	2	2 × 3738	0		10	4 × 1203	2 × 1008	4 × 169
0.153	2	2 × 4410	0		4	2 × 3234	2 × 1190	0
0.167	2	2 × 4798	0		4	2 × 3502	2 × 1306	0
0.181	2	2 × 5196	0		1	1 × 10412	0	
0.195	2	2 × 5606	0		1	1 × 11212	0	
0.209	2	2 × 6006	0		1	1 × 12036	0	
0.222	2	2 × 6410	0		1	1 × 12816	0	
0.236	2	2 × 6802	0		1	1 × 13616	0	
0.250	2	2 × 7210	0		1	1 × 14420	0	
0.264	2	2 × 7610	0		1	1 × 15228	0	
0.278	2	2 × 8014	0		1	1 × 16048	0	

For such a phase extension starting from phase set *P*3.1, we applied two selection criteria simultaneously: a previous condition for the presence of four components at 16 Å and a new condition at 8 Å where we required a map to have not more than 12 connected components and to have the first multiplier equal to four. The FOMs of the resulting phase set *P*3.2 are similar to those of *P*3.1; however, the correlation with the exact phases, especially in the higher resolution zones, continues to increase (Table 5). The map calculated at a resolution of 8 Å (Fig. 3*d*) shows the shape of the α-helix quite unambiguously. Extra density also appears corresponding to other parts of the model (not shown).

4.4. Phase extension using a clustering procedure

The averaging of the variants together is an example of a simple procedure for the treatment of the selected variants. A more intelligent procedure consists of separating the selected sets into groups of close phase sets (these groups are called clusters) and averaging inside every isolated cluster separately (Lunin *et al.*, 1990, 1995, 2000; see these papers and also Lunin *et al.*, 2002 for further technical details). The clustering procedure uses the matrix of one-to-one distances between the selected variants as the input and does not require other phase information. The closer the phase sets are, the earlier they are merged. A graphical presentation of this merging procedure is a cluster tree (Fig. 4). Each point on the bottom line corresponds to a single phase set. Merging of two phases sets into a cluster is indicated by connection of the corre-

Table 7
Correlation of the exact phases with the phases obtained by the averaging in clusters *A*1–*A*4.

The map correlation coefficient (1) multiplied by 100 is shown.

	Resolution (Å)											
Cluster	16.0–∞	12.0–∞	10.0–∞	9.0–∞	8.0–∞	7.0–∞	12.0–16.0	10.0–12.0	9.0–10.0	8.0–9.0	7.0–8.0	6.5–7.0
*A*1	78	70	50	45	40	34	73	63	65	12	46	24
*A*2	76	70	49	44	39	33	72	64	59	12	53	26
*A*3	76	77	57	49	43	36	78	55	45	14	49	31
*A*4	75	37	29	21	26	21	41	48	63	38	48	23

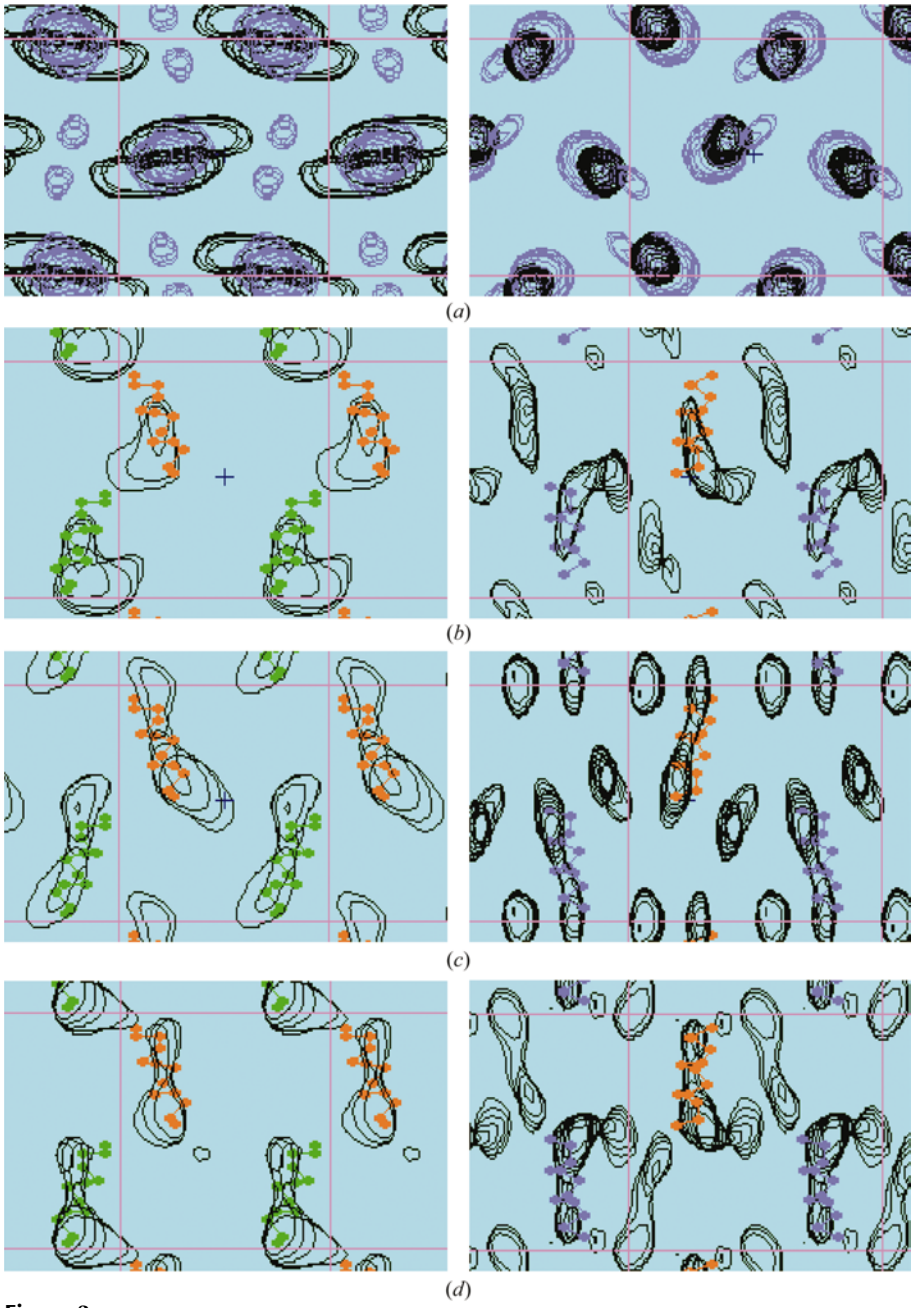

Figure 3
Fourier synthesis maps for protein G calculated with the phases found *ab initio* (see §4.3). Two projections, perpendicular to the axes 0*Z* and 0*X*, are shown on the left and right, respectively. The *ab initio* phased maps (black contours) are superimposed with the map calculated at 16 Å with the model phases (*a*; blue contours) or with the α-helix of the atomic model (*b*–*d*). (*a*) Phase set *P*1.1, resolution 16 Å; (*b*) phase set *P*2.1, resolution 8 Å; (*c*) phase set *P*3.1, resolution 8 Å; (*d*) phase set *P*3.2, resolution 8 Å.

sponding lines; the lower the point of the connection, the closer the phase sets. Such an approach may result in several alternative averaged phase sets that must be tested to select the best one or used together in the framework of a multi-solution strategy. A serious obstacle in this approach is that it is difficult to automate and human intervention is currently required to make a choice.

The cluster analysis needs a larger number of selected phase variants because the whole variety of selected sets will be divided into clusters and main clusters should contain a reasonable number of phase sets. To do so, the phase-generation protocol *P*3 was repeated with exactly the same conditions but searching for a larger number of selected phase sets, namely 150. These variants (set *P*4.1) were the subject of a cluster-analysis procedure.

Fig. 4 shows the cluster tree calculated for set *P*4.1. The whole ensemble of phase variants can be taken together and considered to be a single large cluster *A*1. In this case, the whole procedure will be exactly the same as previously. However, at its right the cluster tree shows a group of phase sets (20 variants) that are separated from the others. Our experience shows that such a small isolated group usually contains noise variants and can be excluded from further analysis, thus already improving the result. The remaining 130 variants can be considered as an alternative cluster *A*2 and can be averaged. However, the tree can be studied at the next level. These 130 variants of the cluster *A*2 are, in turn, separated into two clusters: the left one, *A*3, composed of 71 variants and the right one, *A*4, composed of 59 variants. The averaging can be performed sepa-

rately for the phase variants included in these clusters, thus giving two phase sets.

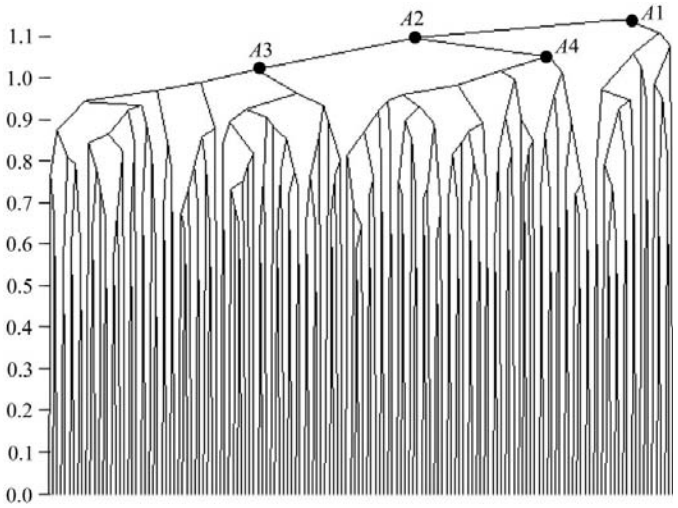

Figure 4
Cluster tree calculated for the phase sets *P*4.1 (see §4.4). *A*3 and *A*4 are the two major clusters discussed in the text.

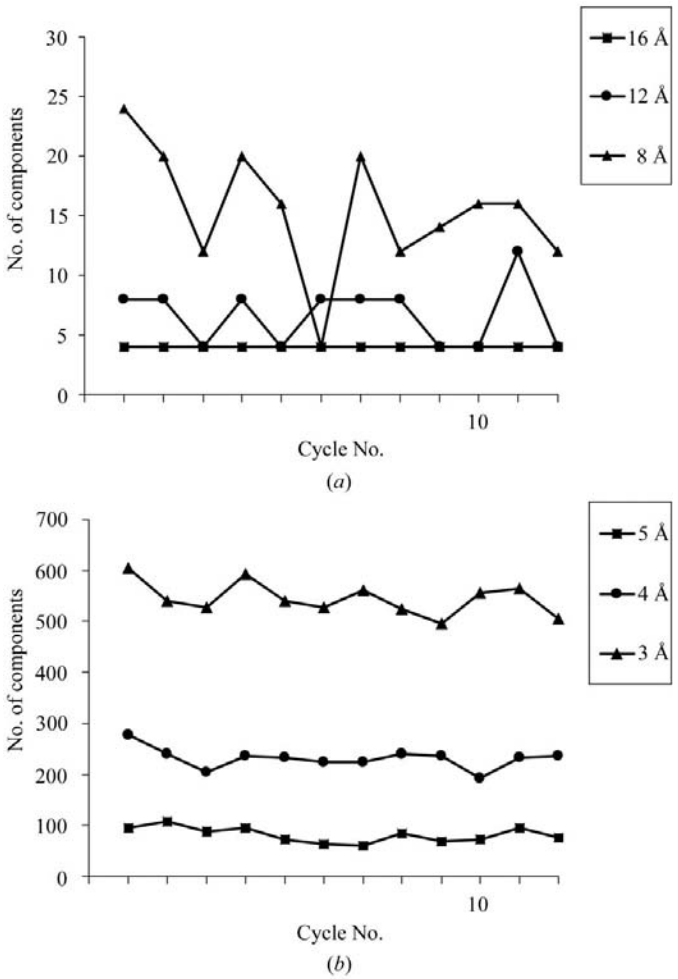

Figure 5
The number of connected components in the *ab initio* phased Fourier synthesis masks for protein G as a function of the phasing cycle number. The relative volume of the masks is 0.1. Different curves correspond to syntheses of different resolutions.

Going to lower levels of the cluster tree does not make much sense because the size of these clusters becomes too small and the phase averaging would be statistically meaningless. However, if a larger number of variants are generated and selected, such a procedure could be possible. On the other hand, a growing number of clusters complicates the problem of selection of a single variant as the solution of the phase problem, as discussed below.

In the current case, the four suggested clusters *A*1–*A*4 were analyzed as follows. The clusters *A*1 and *A*2 were close to each other and the right group of 20 variant of *A*1 was considered to be noise; therefore, cluster *A*2 was considered to be preferable to cluster *A*1. The clusters *A*3 and *A*4 were quite different, being complementary components of the cluster *A*2. After they were merged into *A*2, the quality of the best cluster (we do not know yet which one) could be decreased and the best strategy would be to calculate two independent phase sets, one for *A*3 and one for *A*4, and to try to identify the better one. An *a priori* choice was cluster *A*3 because it was slightly larger than *A*4 and more compact; its top point is lower that the top point for the cluster *A*4. This choice can be performed (or confirmed) by a more objective numerical method. The connectivity properties of the maps calculated at 8 Å with the averaged phases from *A*3 and from *A*4 were studied. Table 6 shows that the phases *A*3 give a well connected map with a

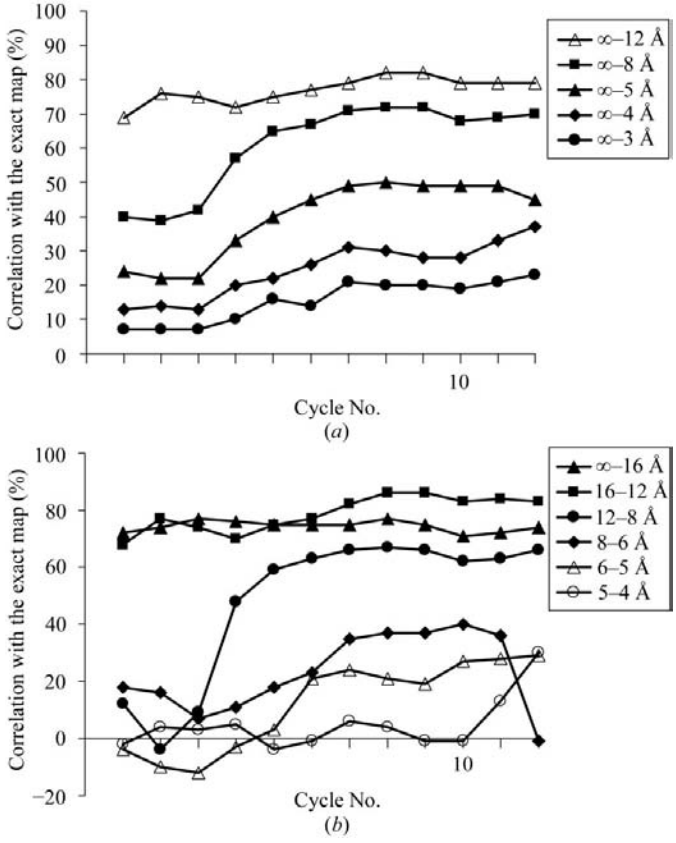

Figure 6
Variation of the phase quality in the course of *ab initio* phasing for protein G. The comparison with the phases calculated from the refined atomic model is performed. The map correlation coefficient (1) for phase sets at different resolution is shown as a function of the iteration number.

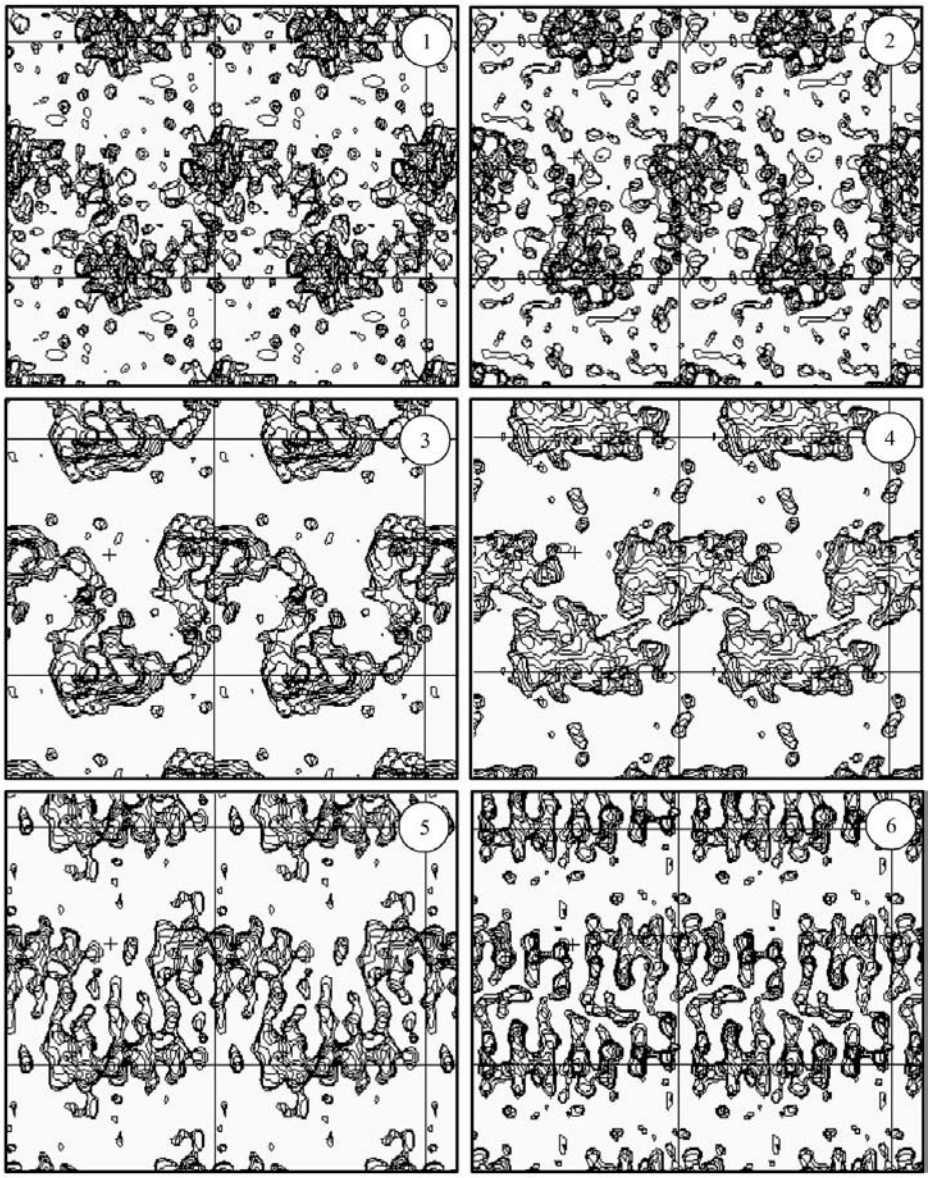

Figure 7
Variation of the map quality in the course of *ab initio* phasing for protein G. Weighted 3 Å resolution Fourier syntheses maps are shown in projection perpendicular to the z axis of at different stages of phasing. A slab $-6/72 \leq z \leq 6/72$ in the unit cell is shown.

cutoff corresponding to a relative volume of up to 0.1, while the map calculated with the phases *A*4 for the same level is much more disconnected, thus confirming the choice of *A*3 for the phase estimation.

For this test case with the known exact phase values, the correlation of an obtained phase set with these exact phases can be calculated. This comparison was performed for phases corresponding to all four clusters *A*1–*A*4. It shows (Table 7) that the cluster *A*3 is indeed the best one and that phase information is available to a resolution of 8–9 Å. Table 7 shows also that phases *A*1 and *A*2 have practically the same quality and that phases *A*4 are clearly the worst of the four sets. The phase correlation for *A*3 is practically the same as for the phases *P*3.2 obtained after two steps of the phasing procedure, but this set was obtained after a single step of phase generation. This makes us believe that application of the phase extension described above to the phases *A*3, the best available from the clustering analysis, can improve the phases further.

4.5. Phase extension to a higher resolution

In this favourable case of a relatively small protein and an experimental data set of very high quality, including completeness, it was interesting to continue the phasing to higher resolution. The phases *P*4.1, cluster *A*3, were taken as the results of the first phasing step and a procedure similar to that described above was applied to iteratively increase the resolution of the phase set and to improve the quality of the already available phases. As might be expected, more selection conditions were used when working at a higher resolution. In all cases, we limited the total number of connected regions and this number increased with the resolution. In most cases, the choice was performed empirically from the rule that each new condition roughly reduced the variants by the same percentage as any previous one.

During the entire procedure, the phases were generated for all reflections at a resolution of 3 Å and lower, even when these reflections did not contribute to any of the verified conditions. Naturally, the averaging of phases excluded from all filtrations gave lower FOMs and higher phase errors, which could approach 90°. However, this allowed a formal calculation of the Fourier maps at all resolutions up to 3 Å at any step of phase extension. Fig. 5 illustrates the analysis of the connectivity properties of these maps. All connected regions were determined with a cutoff level selecting 0.1 of the volume. For all phase sets, the synthesis of the largest resolution used (16 Å) always has four connected domains, one per molecule. For the syntheses of an intermediate resolution, 8 or 12 Å, this number oscillated and each molecule was represented by one to five regions, depending on the iteration. The maps at a higher resolution had a tendency for the number of regions to decrease with the iterations, even when these reflections were excluded from the calculations. This can be considered as some kind of a self-verification, the analysis of 'free reflec-

tions'. This kind of analysis does not require knowledge of the exact phases and can be performed for any phasing project.

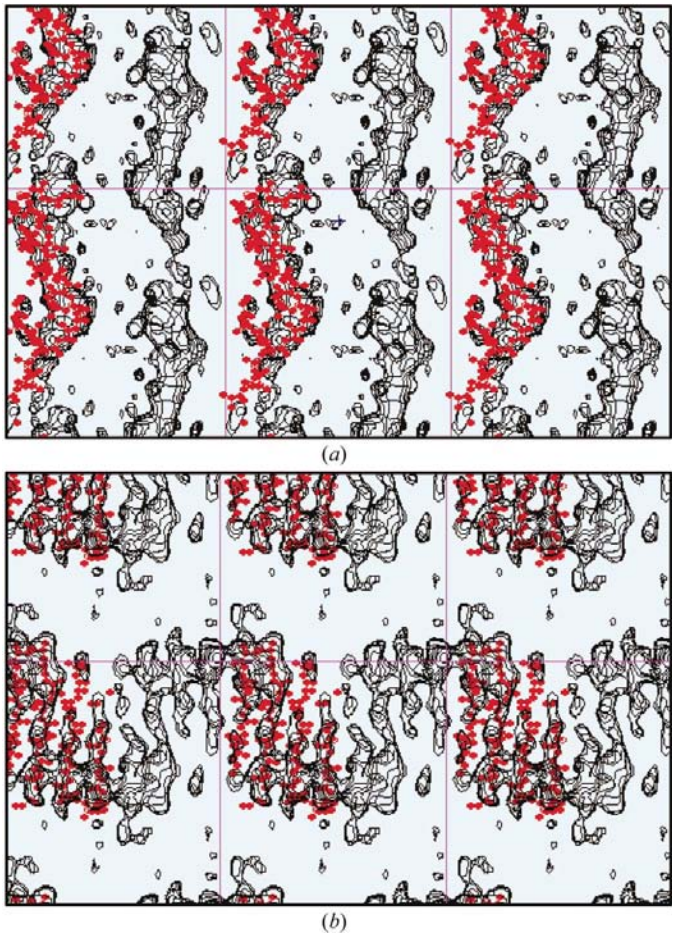

Figure 8
Superposition of an *ab initio* phased map with the known atomic model of protein G. The slabs shown correspond to the part of the model containing (*a*) α-helix (slab $11/72 \leq z \leq 25/72$) and (*b*) β-sheet (slab $30/72 \leq z \leq 42/72$). The atom positions are marked for one of the symmetry-related molecule copies only.

In our test case with the known exact phases their correlation with the phases obtained after each phasing step can be calculated, which is obviously not possible in the practical situation. Fig. 6 shows that the evolution of the phase quality is not monotonous. The phases of the lowest resolution reflections remain practically the same during the entire process. The phases of the reflections in the resolution shell 8–12 Å were drastically improved after the first few iterations. For some reason, the quality of the phases of the reflections with resolution 6–8 Å fell in the last step. However, it could be remarked that the quality of the phases for reflections with resolution higher than 8 Å was generally quite low. This low quality was reflected in very low FOMs, which practically excluded these reflections from the synthesis calculation even when they were formally present. Nevertheless, all these reflections gradually started to play a role in the map quality. Fig. 7 illustrates the evolution of maps formally calculated at 3 Å resolution at different stages of this phasing. While no specific conditions on the shape of the density were applied, the evolution of the maps shows a slow appearance of elongated regions corresponding to individual β-strands.

Fig. 8 shows the obtained map superposed with the available atomic model. This map does not simply show a rough molecular envelope as is usually expected for maps obtained by phasing starting from the low-resolution end, but shows the α-helix and individual β-strands.

Another presentation of the map obtained is given in Fig. 9. As previously for E*r*-1, even some features of the main chain can be observed, as is the case for the α-helix and for the β-strand, at the depth of the image, but these features are irregular and the goal of further developments is to let the phasing method be available to reproduce them systematically. To calculate this map, all reflections to a resolution of 3 Å were used, being weighted with the corresponding FOMs. In order to estimate the effective resolution of the final map, the map was compared with maps calculated with the exact phases. Such a comparison (Fig. 10) shows that the effective resolution can be estimated at somewhere between 4 and 5 Å.

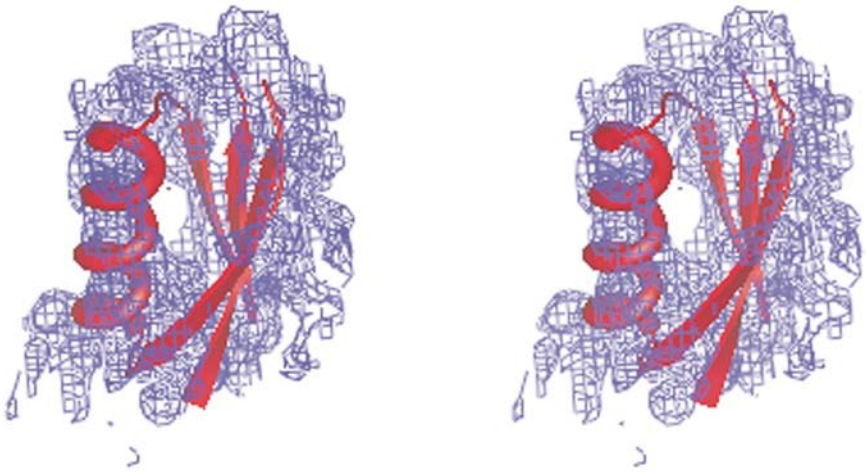

Figure 9
An *ab initio* phased Fourier synthesis map calculated at a resolution of 4 Å superposed with a protein G model.

5. Conclusions

Usually, phasing at low resolution is considered to be a way to obtain the molecular envelope and the molecular packing in the crystal. One of the most powerful approaches in this resolution range (for a review, see Lunin *et al.*, 2002) is connectivity-based phasing (Lunin *et al.*, 2000).

This phasing procedure is relatively straightforward; however, at some stages the procedure may (but does not necessarily) require human intervention. In the initial step, a few tens of reflections are phased; this phase information is then extended to a larger set of structure factors.

This phasing method has previously been tested with several known structures (Lunin *et al.*, 2000) and then successfully applied to the experimental data of LDL, allowing its structure determination at a resolution of about 27 Å (Lunin *et al.*, 2001). In the present work, an attempt was made to extend the resolution zone. The results obtained prove that in a favourable situation direct phasing and, in particular, the connectivity-based approach allow the determination of maps of medium resolution (about 4–5 Å). The quality of these maps may be high enough to allow secondary-structure elements such as α-helices and individual β-strands to be identified.

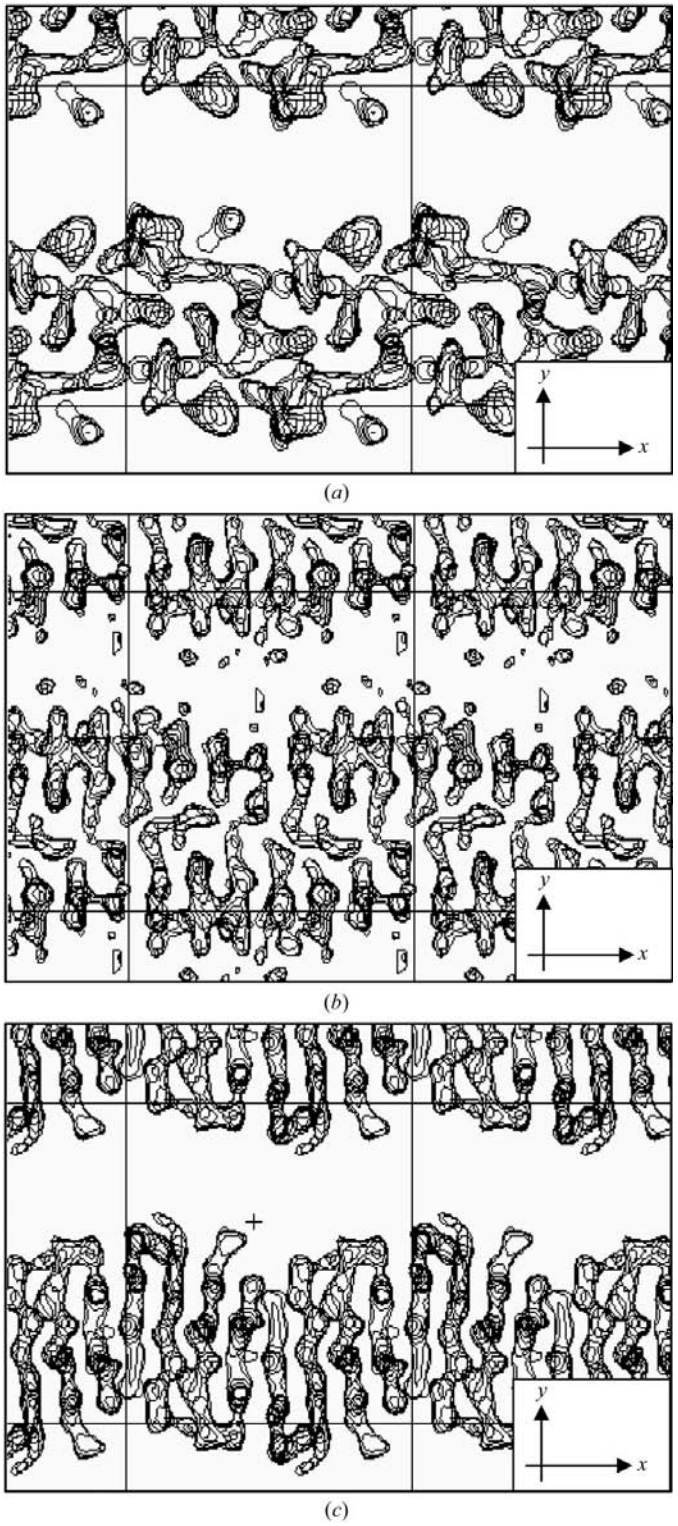

Figure 10
The comparison of the *ab initio* phased synthesis (*b*) with the exactly phased 5 Å resolution (*a*) and 4 Å resolution (*c*) syntheses allows the estimation of the effective resolution of the *ab initio* phased synthesis as 4.5 Å. A slab $-6/72 \leq z \leq 6/72$ in the unit cell is shown.

APPENDIX *A*
Connectivity-analysis algorithm

Let a grid with $N_x \times N_y \times N_z$ divisions be introduced into the unit cell of a crystal and Ω be some selected subset of the nodes of this grid. The goal of connectivity analysis is to find the number of connected components in Ω and to define their size.

The concept of the connected component is based on the definition of neighbouring nodes. For a given node, the set of its neighbours may be defined differently and the results of the connectivity analysis may depend on this definition. In this paper, the simplest definition was explored. Let $\mathbf{r} = (i_1, i_2, i_3)$ be the indices of the grid nodes which are supposed to vary in the limits $1 \leq i_1 \leq N_x$, $1 \leq i_2 \leq N_y$, $1 \leq i_3 \leq N_z$. For an inner node (i_1, i_2, i_3) in the unit cell, the following six nodes will be considered as *neighbours*,

$$(i_1 - 1, i_2, i_3), (i_1 + 1, i_2, i_3), (i_1, i_2 - 1, i_3), (i_1, i_2 + 1, i_3), (i_1, i_2, i_3 - 1), (i_1, i_2, i_3 + 1). \quad (3)$$

For the nodes at the border of the unit cell, the definition must be corrected using the periodicity of the crystal. Following the definition chosen above, the grid node with indices (1, 1, 1) has neighbours $(N_x, 1, 1)$, (2, 1, 1), $(1, N_y, 1)$, (1, 2, 1), $(1, 1, N_z)$ and (1, 1, 2). If $1 < i_1 < N_x$ and $1 < i_3 < N_z$, then the neighbours of the node (i_1, N_y, i_3) are $(i_1 - 1, N_y, i_3)$, $(i_1 + 1, N_y, i_3)$, $(i_1, N_y - 1, i_3)$, $(i_1, 1, i_3)$, $(i_1, N_y, i_3 - 1)$, $(i_1, N_y, i_3 + 1)$ *etc*.

A sequence of grid nodes $\mathbf{r}^1, \mathbf{r}^2, \ldots, \mathbf{r}^n$ is called a *chain* if every two consecutive nodes are neighbours.

A subset D of the set Ω is called a *connected component* if for every two of its nodes **a** and **b** it is possible to find a chain $\mathbf{r}^1, \mathbf{r}^2, \ldots, \mathbf{r}^n$ of the nodes of D such that $\mathbf{r}^1 = \mathbf{a}$ and $\mathbf{r}^n = \mathbf{b}$.

An algorithm based on analogies with the process of burning of dry grass was used in our calculations. Its principal steps may be described as follows.

Step 1. Initialization. The nodes of the set Ω are marked as 'fuel'. All others nodes of the grid are marked as 'empty'. The number of found components is initially equal to zero.

Step 2. Search for a new component. The nodes of the grid are scanned until a 'fuel' node is found. A 'current front' is defined as a set consisting of this node only. The number of found components is increased by one. A symbol m is chosen to mark the new component found; for example, this can be a consecutive number of the connected component.

If no 'fuel' nodes were found at this step then the work is finished.

Step 3. Isolation of a connected component. The next two steps are repeated several times until the termination condition of Step 3.1 is fulfilled.

Step 3.1. The 'future front' is defined as the set of the nodes that are neighbouring to one of the nodes of the 'current front' and are 'fuel' nodes.

If the 'future front' is found to be empty, then the connected component is isolated and the algorithm returns to the Step 2 to search for the next connected component.

Step 3.2. Propagation of the front. The nodes of the current front are considered to be burned away and became 'empty'. They are marked by the symbol m chosen for the current component. The 'future front' is renamed the 'current front' and Step 3.1 of the procedure is repeated.

When no more 'fuel' nodes found at Step 2, the algorithm terminates the work. The result of the procedure is the number of the found connected components; the nodes of every component are marked by a symbol specific for this component. The number of nodes of every component may easily be calculated.

It is easy to see that the suggested algorithm does not depend on the particular definition of the set of neighbouring nodes and can be easily adapted to any of them. For example, the set of neighbouring nodes (1) may be extended by including 12 'diagonal' nodes

$$(i_1-1, i_2-1, i_3), (i_1+1, i_2-1, i_3), (i_1-1, i_2+1, i_3), \ldots, (i_1, i_2+1, i_3+1) \quad (4)$$

or, additionally to (4), by eight further nodes

$$(i_1-1, i_2-1, i_3-1), (i_1+1, i_2-1, i_3-1), \ldots, (i_1+1, i_2+1, i_3+1) \quad (5)$$

etc.

The authors thank Dr A. Podjarny for fruitful discussions and for help with this work. The work of NL and VL was supported by grants from RFBR 00-04-48175 and 03-04-48155. AU is a member of GdR 2417 CNRS. The programs *CAN* (Vernoslova & Lunin, 1993) and *PyMOL* (DeLano, 2002) were used to show maps.

References

Anderson, D. H., Weiss, M. S. & Eisenberg, D. (1996). *Acta Cryst.* D**52**, 469–480.
Baker, D., Bystroff, C., Fletterick, J. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 429–439.
Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 186–192.
Bhat, T. N. & Blow, D. M. (1982). *Acta Cryst.* A**38**, 21–29.
Chabrière, E., Lunina, N., Lunin, V. Y. & Urzhumtsev, A. (2001). *ECM-20 Abstracts, XXth Eur. Crystallogr. Meet.*, p. 70.
DeLano, W. L. (2002). *The PyMOL Molecular Graphics System*, http://www.pymol.org.
Derrick, J. P. & Wigley, D. B. (1994). *J. Mol. Biol.* **243**, 906–918.
Evans, M., Hastings, N. & Peacock, B. (2000). *Statistical Distributions*, 3rd ed., pp. 189–191. New York: Wiley.
Lunin, V. Y. & Lunina, N. L. (1996). *Acta Cryst.* A**52**, 365–368.
Lunin, V. Yu., Lunina, N. L., Petrova, T. E., Vernoslova, E. A., Urzhumtsev, A. G. & Podjarny, A. D. (1995). *Acta Cryst.* D**51**, 896–903.
Lunin, V. Yu., Lunina, N., Podjarny, A., Bockmayr, A. & Urzhumtsev, A. (2002). *Z. Kristallogr.* **35**, 668–685.
Lunin, V. Y., Lunina, N. L., Ritter, S., Frey, I., Berg, A., Diederichs, K., Podjarny, A. D., Urzhumtsev, A. & Baumstark, M. W. (2001). *Acta Cryst.* D**57**, 108–121.
Lunin, V. Yu., Lunina, N. L. & Urzhumtsev, A. G. (2000). *Acta Cryst.* A**56**, 375–382.
Lunin, V. Yu., Urzhumtsev, A. G. & Skovoroda, T. P. (1990). *Acta Cryst.* A**46**, 540–544.
Lunin, V. Y. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 530–533.
Lunina, N. L., Lunin, V. Y. & Urzhumtsev, A. (2000). *ECM-19 Abstracts, XIXth Eur. Crystallogr. Meet.*, p. 62. Abstract s7.m6.o4.
Urzhumtsev, A. & Podjarny, A. D. (1995). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 12–16.
Vernoslova, E. A. & Lunin, V. Y. (1993). *J. Appl. Cryst.* **26**, 291–294.
Wilson, C. & Agard, D. A. (1993). *Acta Cryst.* A**49**, 97–104.