Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Ab initio phasing based on topological restraints: automated determination of the space group and the number of molecules in the unit cell

Ludmila Urzhumtseva,[a] Natalia Lunina,[b] Andrei Fokine,[c]‡ Jean-Pierre Samama,[d] Vladimir Y. Lunin[b] and Alexandre Urzhumtsev[c]*

[a]INIST, 2 all. Parc de Brabois, 54505 Vandoeuvre-lès-Nancy, France, [b]Institute of Mathematical Problems of Biology, Russian Academy of Sciences, Pushchino, Moscow Region, 142290, Russia, [c]LCM3B, UMR 7036 CNRS, Faculté des Sciences, Université Henri Poincaré, Nancy 1, BP 239, 54506 Vandoeuvre-lès-Nancy, France, and [d]IGBMC, 1 rue R. Friez, 67404 Illkirch, CU de Strasbourg, France

‡ Current address: Department of Biological Sciences, Purdue University, 915 W. State Street, West Lafayette, IN 47907-2054, USA.

Correspondence e-mail: alexander.ourjoumtsev@lcm3b.uhp-nancy.fr

Received 27 February 2004
Accepted 19 June 2004

The connectivity-based phasing method has been demonstrated to be capable of finding molecular packing and envelopes even for difficult cases of structure determination, as well as of identifying, in favorable cases, secondary-structure elements of protein molecules in the crystal. This method uses a single set of structure factor magnitudes and general topological features of a crystallographic image of the macromolecule under study. This information is expressed through a number of parameters. Most of these parameters are easy to estimate, and the results of phasing are practically independent of these parameters when they are chosen within reasonable limits. By contrast, the correct choice for such parameters as the expected number of connected regions in the unit cell is sometimes ambiguous. To study these dependencies, numerous tests were performed with simulated data, experimental data and mixed data sets, where several reflections missed in the experiment were completed by computed data. This paper demonstrates that the procedure is able to control this choice automatically and helps in difficult cases to identify the correct number of molecules in the asymmetric unit. In addition, the procedure behaves abnormally if the space group is defined incorrectly and therefore may distinguish between the rotation and screw axes even when high-resolution data are not available.

1. Introduction

The possibility of structure factor phasing at early stages of structure determination using a single set of experimental magnitudes is attractive, even when this concerns only several hundred reflections of low resolution. There are several examples, with very different conditions, in which the connectivity-based *ab initio* phasing method (Lunin *et al.*, 2000) was able to find the molecular position and molecular shape (Lunin *et al.*, 2001, 2002) and even revealed some structural details (Lunina *et al.*, 2003). Crystallographers have used the connectivity properties of Fourier syntheses maps for decades for qualitative estimation of the success of phasing. Bhat & Blow (1982), Wilson & Agard (1993) and Baker, Bystroff *et al.* (1993) suggested the use of these properties in phasing as the basis for map modification. In our approach, which is similar to that of Baker, Krukowski & Agard (1993), this information is expressed numerically and becomes a selection criterion.

The phasing procedure consists of several steps. First, a large number of trial phase sets are *generated* randomly. For each of these phase sets, a Fourier synthesis of a specified resolution is calculated and a *binary mask* corresponding to the region of the highest synthesis values is defined in accordance with the specified cut-off level. The *connectivity analysis*

of this mask is aimed at establishing the number of connected components in the mask, to determine if these components are finite or not, to calculate their volumes *etc.* The result of the connectivity analysis is presented in the form of an output list, in which the connected components are ordered by decreasing volume. The trial phase sets are considered as admissible and selected for further treatment if the connectivity-analysis output list satisfies the *selection rules*, formulated in terms of the total number of components, the number of compounds having the same volume, their finiteness *etc.* After a reasonable ensemble of phase sets have been selected (*e.g.* about 100), they are '*aligned*' using the permitted origin/enantiomorph choice (Lunin *et al.*, 1990; Lunin & Lunina, 1996) and *averaged*. For every reflection **h**, 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. Generally, phasing proceeds through several cycles in such a way that the output phase information of a current cycle is used as the input for the next cycle. More details of the method are given by Lunin *et al.* (2000, 2002).

The result of the method depends on a number of parameters. Most of them are easy to estimate, and variation of these parameters within reasonable limits does not significantly affect the results. However, when starting the investigation of a new structure, some important crystallographic information is often absent or ambiguous. Most frequently, this ambiguity concerns the number of protein molecules in the unit cell, and sometimes the identification of the space group. The results presented below with various data sets (Appendix *A*) show to what extent the connectivity-based phasing procedure can help to make an appropriate choice of these parameters. Knowledge of these values may be useful not only for the purpose of direct phasing but also, for example, for the success of molecular replacement.

2. Main steps and parameters of the phasing procedure

2.1. Mask building

Several connectivity-based selection criteria may be applied simultaneously, each of them depending on specific parameters. The most important parameters are those defining the mask construction and the selection rules applied.

The main parameters influencing the mask construction are the resolution, $d_{\rm mask}$, of the Fourier synthesis used to build the mask, the numbers, $N_x \times N_y \times N_z$, defining the grid at which this syntheses is calculated, and the cut-off level ρ*. The mask is defined as the set of grid points such that the corresponding synthesis value exceeds the chosen cut-off. It is convenient to define this value by the corresponding relative volume of the mask that is the ratio of the number of the mask points to the total number of the grid points. Alternatively, we can refer to a specific volume defined as the ratio of the mask volume to the number of amino acid residues in the unit cell. Multiple numerical experiments show that at the first phasing step, when the cut-off level chosen corresponds to the specific volume of $\sim$25 Å^3 per residue, the number of connected components found is usually equal to the number of molecules in the unit cell. This relation can be used to formulate the corresponding selection rule at this stage of the phasing.

The starting resolution, $d_{\rm mask}$, is defined by the number of reflections to be phased at the initial step. This number must not be large, in order to let the initial search in the phase space be exhaustive enough. On the other hand, it cannot be too small, otherwise the Fourier synthesis will not reproduce the details that are required. An experimental estimation of this number is about 15 reflections per expected independent blob, and about 40 reflections can be phased at the first step. After some low-resolution reflections have been phased, more reflections can be added for phasing and the mask resolution can be increased accordingly. The upper resolution limit depends on the goals and on the experimental data available. From our previous experience, we conclude that the method can phase several hundred lowest resolution reflections. For a crystal with a large unit cell, a detailed molecular envelope is obtained (see, for example, Lunin *et al.*, 2001, 2002); when the unit cell is small, the identification of secondary-structure elements is allowed (Lunina *et al.*, 2003).

2.2. Connectivity analysis

The selection rules are formulated in terms of the connectivity-analysis output list. The connectivity analysis is applied to the mask build and is aimed at

(i) defining the number, $N_{\rm tot}$, of connected components from which the mask is composed;

(ii) checking if these components are finite or not;

(iii) calculating the volumes of the components and defining how many components have equal volumes.

In view of the very large number of maps to be analyzed, the algorithms used to identify the connected three-dimensional regions should be fast. Some such algorithms were suggested by Hunt *et al.* (1997) and Lunina *et al.* (2003). In what follows, $V_1, V_2, V_3, \ldots$ (ordered as $V_1 > V_2 > V_3 > \ldots$) and $\mu_1, \mu_2, \mu_3 \ldots$ denote the volumes and the numbers of equal-volume components; in particular, μ_1 is the number of components of the largest volume V_1, μ_2 is the number of second-size components *etc.*; the numbers $\mu_1, \mu_2, \mu_3 \ldots$ are referenced as multipliers. Some examples of the selection rules used are listed below.

Selection rule A. The total number of connected components in the mask must be equal to a given number.

Selection rule B. The multiplier μ_1, corresponding to the largest component, must be equal to a given value. The multipliers μ_2, μ_3, μ_4 *etc.*, corresponding to components next in size, must be equal either to 0 or to another prescribed number.

Selection rule C. The same as the rule *B*, but the ratio of volumes of largest components, corresponding to multipliers μ_1 and μ_2, should be within the given limits.

In particular, selection rule *A* may be used to identify the molecules in the unit cell. Rule *C* helps to identify the blobs linked by non-crystallographic symmetry; the volumes of these

blobs are close to each other but are not necessarily equal. Conversely, occasionally two completely different regions, not linked by any symmetry, can have exactly the same volume; the connectivity analysis used does not distinguish between them.

An important problem of the connectivity analysis is that it is not always clear whether, at a given cut-off level, protein domains appear as separated blobs or, inversely, whether densely packed molecules are seen separately.

2.3. Alignment and averaging

The simplest treatment of the selected variants consists in their alignment and averaging. Two phase sets, $\{\varphi_i(\mathbf{h})\}$, $i = 1, 2$, while being formally very different, may in fact present very close solutions of the phase problem, but referred to different origin/enantiomorph choices. To take this circumstance into account, some kind of alignment must be performed in accordance with the origin/enantiomorph choices permitted (Lunin *et al.*, 1990; Lunin & Lunina, 1996) before a formal phase closeness is calculated. In the present work, the alignment was based on maximization of the map correlation coefficient (Lunin & Woolfson, 1993)

$$C_\varphi = \frac{\int \left[\rho_1(\mathbf{r}) - \langle\rho_1\rangle\right]\left[\rho_2(\mathbf{r}) - \langle\rho_2\rangle\right]\mathrm{d}V_\mathbf{r}}{\left\{\int \left[\rho_1(\mathbf{r}) - \langle\rho_1\rangle\right]^2\mathrm{d}V_\mathbf{r} \int \left[\rho_2(\mathbf{r}) - \langle\rho_2\rangle\right]^2\mathrm{d}V_\mathbf{r}\right\}^{1/2}} = \frac{\sum_\mathbf{h} F_1(\mathbf{h})F_2(\mathbf{h})\cos\left[\varphi_1(\mathbf{h}) - \varphi_2(\mathbf{h})\right]}{\left[\sum_\mathbf{h} F_1(\mathbf{h})^2 \sum_\mathbf{h} F_2(\mathbf{h})^2\right]^{1/2}}$$

For the phase alignment, $\rho_i(\mathbf{r})$ are the Fourier syntheses calculated with the observed magnitudes $\{F_1(\mathbf{h}) = F_2(\mathbf{h}) = F^{\rm obs}(\mathbf{h})\}$ and phases $\{\varphi_i(\mathbf{h})\}$, $i = 1, 2$, to be compared. All permitted origin shifts (and enantiomorph changing, if permitted) are tried for the second set in order to get the highest C_φ value.

2.4. Analysis of results

The results of phasing were compared with the known structural information. As a quantitative measure, the correlation, C_φ, defined above was used. To do so, the first synthesis was calculated with experimental magnitudes weighted by the obtained figures of merit, $\{F_1(\mathbf{h}) = m(\mathbf{h})F^{\rm obs}(\mathbf{h})\}$, and with averaged phases, $\{\varphi_1(\mathbf{h}) = \varphi^{\rm best}(\mathbf{h})\}$, and the second synthesis was calculated with the experimental magnitudes, $\{F_2(\mathbf{h}) = F^{\rm obs}(\mathbf{h})\}$, and with the exact phases $\{\varphi_2(\mathbf{h}) = \varphi^{\rm exact}(\mathbf{h})\}$. The latter were obtained from the atomic model known in these cases.

3. Robustness of the phasing procedure

A large number of tests with various data (described in Appendix *A*) were performed in order to check the relative independence of the results of phasing with respect to variation of the main parameters within reasonable limits, as these parameters can be estimated before the procedure starts.

First, the effect of both the random generator and the number of generated and selected phase sets was studied. Two series of generations were performed for protein G (Appendix *A*1) at a resolution of 16 Å. In the first, the generation process was interrupted after 100 phase variants were selected, and the second series was continued until 200 variants were selected. The results of phase generation vary from one series to another. The averaged values are, however, relatively close to each other. It is important to note that the dispersion of correlation of the averaged phase sets with the exact phases decreases when more random phase sets are generated and selected. Similarly, the average phase values depend on the choice of the starting values of the random numbers generator, but the correlation of the average values with the exact phase values, calculated at 24, 18 or 16 Å resolution, did not vary significantly.

For many crystallographic studies, the grid step is taken to be about one-third of the resolution used. For direct phasing, the grid used can either correspond to the highest expected resolution for the project or be increased with the phasing steps. Tests with the data for RNAse *Sa* (Appendix *A*2) showed that the variation of the grid step between 1 and 4 Å did not much influence the quality of the resulting phases at 12 Å. Interestingly, this observation was true both for the optimal and for the non-optimal choice of other parameters; the phase accuracy was different for different groups of tests, but for the same group the dependence on the grid step was weak. Similar results were obtained for the protein G experimental data.

Finally, a large series of tests were performed with experimental data sets for RNAse *Sa*, protein G and pheromone E*r*-1 (Anderson *et al.*, 1996), for which different phasing protocols were applied. In particular, we varied the separation in resolution zones, the highest resolution used, the number of steps, the number of reflections used at different phasing steps, the cut-off levels, the number of generated random phase sets, the resolution for the alignment *etc*. These tests indicated a relative independence of the results with respect to all these parameters. The discussion of some preferences for the optimal choice of parameters is out of the scope of this article.

A special study was devoted to the contribution of the requested number of connected components in the Fourier maps. First, a correct choice of this parameter seems to be important for the results of phasing. Secondly, the numerical connectivity analysis is still a relatively unusual procedure and currently crystallographers need more experience to work with this value. This study is presented below.

4. Number of connected components

4.1. Unknown number of connected components

In many situations, it is difficult to predict the exact number of connected components at a chosen cut-off level. Even at low resolution, when we would expect to see the number of blobs equal to the number of molecules, this is not always the case. On one hand, for the molecule composed of multiple

domains, the synthesis may show these domains as disconnected blobs. On the other hand, as a result of dense molecular packing in the crystal, the molecules may be seen not as individual blobs but as their aggregates. An example of such a crystal is that of RNAse *Sa*. The unit cell of this crystal (space group $P2_12_12_1$) contains two molecules per asymmetric unit. However, the Fourier synthesis at a resolution of 18 Å (29 reflections used) obtained with the experimental magnitudes and with the phases calculated from the atomic model shows four and not eight blobs when the cut-off level varies in very large limits (corresponding to the relative volume from 0.05 to 0.30). If the model structure-factor magnitudes are used instead of the experimental values, the synthesis shows eight blobs, one for each protein molecule. Therefore, the contribution of the bulk solvent decreases the contrast, masks the border between the molecules and merges the blobs into pairs.

4.2. Variation of the requested number of connected components for RNAse *Sa*

Since RNAse *Sa* has two independent molecules in the asymmetric unit, it is logical to expect the Fourier maps at low resolution to show two independent groups of blobs when starting the phasing procedure. Each group would correspond to molecules linked by crystallographic symmetry of the space group, and the blobs belonging to different groups may be slightly different. The group containing the largest blobs is described by the multiplier four, equal to the number of crystallographic symmetries, as is the group with blobs next in size. There is no reason to expect other blobs that correspond to noise. These three conditions, with the first two multipliers equal to four and the third equal to zero, are denoted in what follows as condition 4–4–0. Additionally, since the blobs belonging to different groups correspond to molecules linked by non-crystallographic symmetry, we can restrain the closeness of their volumes, for example, by the condition $0.7 \leq V_2/V_1 \leq 1$ (see rule *C* in §2.2).

Several alternative independent runs of phasing were tried, assuming (i) the condition described above, (ii) eight components of exactly the same size without noisy drops (condition 8–0) or (iii) four components of the same size without noisy drops (condition 4–0). This latter condition may correspond either to only four molecules in the unit cell or to molecular packing by pairs. In all these runs, the phase sets were generated until 100 phase variants satisfying the selection rule were obtained. When the procedure parameters varied as indicated above, the number of generated phase sets required to select 100 sets was different but remained of similar magnitude for each of the given selection criterion: 15 000–20 000 for rule 4–4–0, 350–400 for rule 4–0 and 100 000–1 000 000 for rule 8–0.

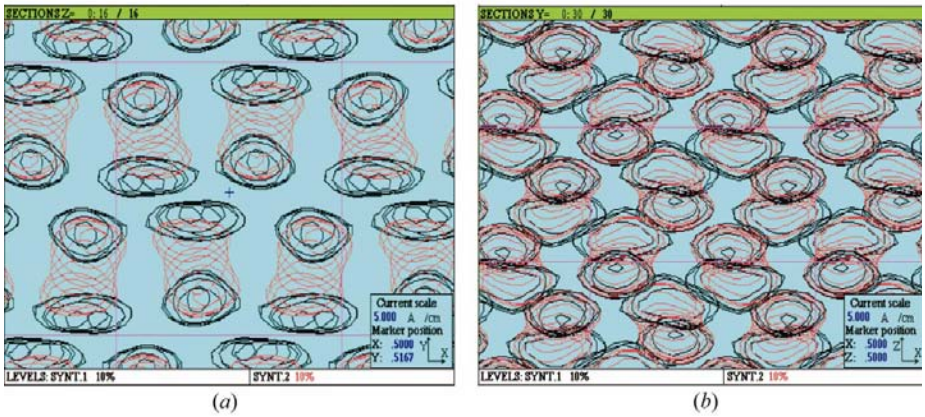
If the number for condition 4–0 is too low we can say that this condition has a selection power that is too weak. Indeed, the correlation, C_φ, of the averaged phase sets for these runs is ~72, 55 and 52% for the reflections at resolutions 24, 18 and 16 Å, respectively, while for the 'correct' rule, 4–4–0, the correlation is as high as 92, 75 and 70%.

Visual analysis confirms the conclusion. As expected, the phases obtained with rule 4–0 produce maps in which the blobs from the two groups are merged (Fig. 1). This merged envelope covers the centers of molecules, which are indeed very densely packed, but does not show their positions unambiguously. The map calculated at 18 Å with the phase set obtained using condition 4–4–0 shows clearly two quartets of blobs (Fig. 1), whose positions coincide well with the known molecular centers.

Rule 8–0 imposes further constraints by forcing the blobs in both quartets to have exactly the same volume. The procedure managed to find such phase sets but their averaging did not improve the result found with the 'soft-restrained' volume ratio.

As a check, the procedure was run with the condition 4–4–4–0 (three groups of blobs, each with multiplicity four, and no other blobs) and the volume ratios $0.7 \leq V_2/V_1 \leq 1$ and $0.7 \leq V_3/V_2 \leq 1$, thus simulating a search for three molecules in the asymmetric unit. With such a condition, no single variant was selected from 10 000 generated and the run was stopped.

From this analysis for RNAse *Sa*, we conclude that the procedure discriminates between the selection rules and leads to the correct composition of the images. To confirm this observation, several calculations were performed on other cases.



(*a*) (*b*)

Figure 1
Ab initio phased syntheses for RNAse *Sa*. Both maps are calculated at a resolution of 18 Å. Red contours show the map obtained by phasing with the selection criterion 4–0, and black contours were obtained by phasing with the criterion 4–4–0. Projections (*a*) *0xy* and (*b*) *0xz* of the whole unit cell are shown. Black contours correspond well to the centers of the molecules (not shown). Both selected regions (red and black contours) occupy 0.10 of the unit-cell volume.

4.3. Variation of the required number of connected domains for AspRS

The asymmetric unit of the AspRS crystal in the cubic form (space group *I*432, 48 crystallographic symmetry

operations; see Appendix *A*3) contains a homodimer of the protein–RNA complex. (Urzhumtsev *et al.*, 1994). In the dimer, the protein molecules associate tightly, while the RNA molecules are separated in space. Direct phasing using a neutron diffraction data set with the concentration of the solvent density masking the protein was successful when the syntheses with two groups of blobs, each one composed of 48 equivalent blobs, were selected (*i.e.* one blob per RNA molecule). Because of the small relative size of the search object, RNA, in the unit cell for the whole complex, a high cut-off level was used. Therefore, a larger number of phase sets, about 200 000, had to be generated in order to select 100 phase sets.

With the same criterion, the search with the second set of diffraction magnitudes, corresponding to RNA being masked, found only a few variants after 500 000 random phase set generations. When one rather than two groups of blobs were searched for (*i.e.* one blob per protein dimer), the procedure found 100 appropriate variants after only 50 000 generations. Corresponding maps clearly show a protein dimer as a single compact object. Some complementary information is given by Fokine *et al.* (2003).

4.4. Variation of the requested number of connected domains for FixJN

In the case of FixJN (Birck *et al.*, 1999; see Appendix *A*4), we can suppose *a priori* that the asymmetric unit may contain two or three monomers per asymmetric unit. Two runs of direct phasing were tried at 25 Å resolution (30 reflections) using structure-factor magnitudes calculated as described in Appendix *A*4. In the first run, the maps were selected if they contained two groups of regions, each with the multiplier four (equal to the number of symmetry operations in space group *C*2), with the ratio of their volumes between 0.5 and 1.0, and with no more 'density drops' (*i.e.* the selection rule 4–4–0 was applied). In the alternative run, the syntheses with three groups of blobs instead of two groups were selected (the selection rule 4–4–4–0; as previously, three number '4's mean the multiplicity four for each of these groups and the '0' denotes the absence of other peaks of density except the three groups defined above). In all cases, the cut-off level was chosen so that the relative volume of the mask was equal to 0.05. This level (which is higher than the usual one of 0.10) was taken because functional species may be protein dimers, whose occurrence would favor merging of the blobs at a low cut-off level. [The map calculated with model-simulated phases at 25 Å resolution shows such blobs (see Fig. 2); at a 0.10 cut-off level, these blobs are indeed merged (data not shown).] In all cases, the phase generation was stopped after selection of 100 variants.

The procedure needed about 14 000 and 113 000 variants to be generated for conditions 4–4–0 and 4–4–4–0, respectively. In contrast to the two previous cases, the search with the correct condition required more phase generations than that for the wrong one. However, in this test, when searching for three blobs, the total number of reflections per blob is only ten, less than the empirical critical number of 15 reflections per blob. It is difficult to reproduce three independent blobs with such a small number of reflections, thus explaining the low percentage of selected variants. (At a higher resolution of 20 Å, where 49 reflections are available, these two searches require roughly the same number of generations in order to select 100 phase sets; however, such a direct search for many phases simultaneously becomes inefficient, giving large phase errors and low figures of merit.)

Nevertheless, both numbers, 14 000 and 113 000, are still within reasonable limits and another criterion is required in order to make the choice. An analysis of the maps calculated with averaged phases shows that both these maps contain essentially three independent blobs, even for the selection with the condition 4–4–0; in this latter case, two of these three well pronounced blobs are 'formally' linked by a very tiny region. The corresponding blobs from these two maps coincide very well with one another and with the known monomer positions in this test case. The correlation of the two maps with the exact map is 0.74 and 0.72, respectively. Therefore, a

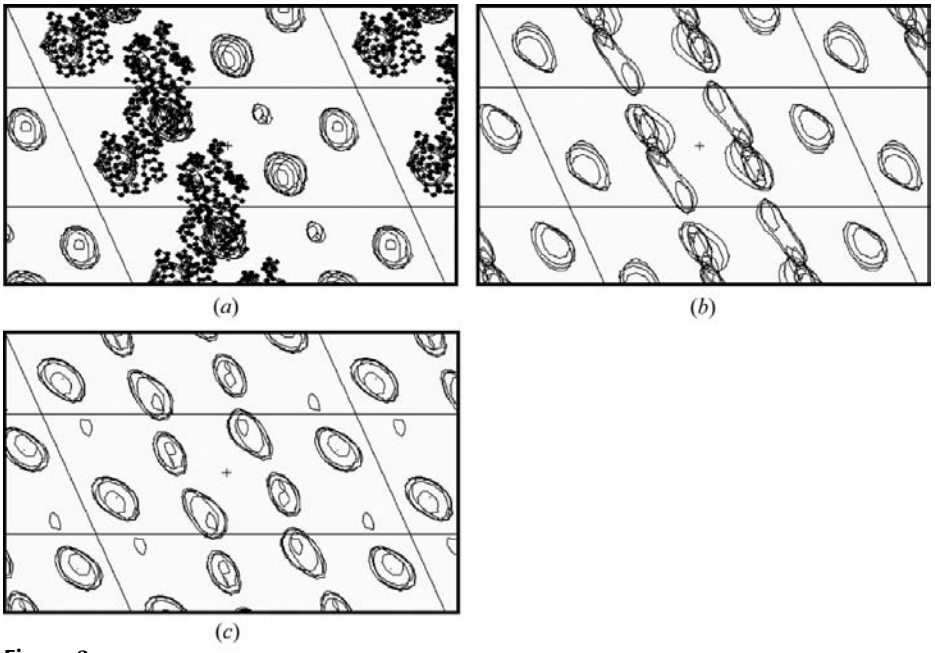


Figure 2
Ab initio phased syntheses for FixJN. All maps are calculated at a resolution of 25 Å using structure factor magnitudes obtained from the model and corrected by bulk solvent contribution. (*a*) was calculated with model-simulated phases. The C_α atoms of one crystallographic copy of the model are shown. (*b*) The result of phasing with condition 4–4–0 (two groups of blobs). (*c*) The same for the selection condition 4–4–4–0 (three groups of blobs). Projection *0xz* of one-half of the unit cell is shown. Selected regions occupy 0.05 of the unit-cell volume.

Table 1
The number of selected variants from 50 000 generated phase sets for different choices of structure factor magnitudes and space groups.

CHFI diffraction data at a resolution of 14 Å are used. F^{obs} denotes the set of experimental magnitudes with the magnitudes missed for four reflections, F_{110}, F_{002}, F_{112} and F_{004}. F^{obs}+ denotes the same set completed by calculated magnitudes for these four reflections. Columns 2, 3 and 4 show the reflections added, removed or with their magnitudes multiplied by the indicated number. The correct space group is $P4_22_12$ (last column, shown in bold). The groups incompatible with the given data set are indicated by the symbol –.

Set	Refl +	Refl −	Refl ×	*P*422	$P42_12$	$P4_122$	$P4_12_12$	$P4_222$	$\mathbf{P4_22_12}$
F^{calc}				1	0	–	–	14	**211**
F^{obs}+				1	1	–	–	7	**244**
F^{obs}				1	0	7	54	3	**56**
F^{obs}+			F_{110} F_{002} × 2.0 F_{112} F_{004} × 2.0	24	8	–	–	39	**575**
F^{obs}+			F_{110} F_{002} × 0.50 F_{112} F_{004} × 0.50	0	0	–	–	2	**98**
F^{obs}+			F_{110} F_{002} × 0.20 F_{112} F_{004} × 0.20	0	0	–	–	1	**71**
F^{obs}+			F_{002} × 0.10	1	0	–	–	10	**120**
F^{obs}+			F_{002} × 0.01	0	0	–	–	11	**121**
F^{obs}+		F_{002}		0	0	13	185	12	**122**
F^{obs}+		F_{002} F_{200}		0	0	15	156	8	**73**
F^{obs}+	$F_{100} = 3$ $F_{300} = 6$			2	–	–	–	11	–
F^{obs}+	$F_{100} = 30$ $F_{300} = 60$			2	–	–	–	9	–
F^{obs}+	$F_{100} = 300$ $F_{300} = 600$			1	–	–	–	13	–
F^{obs}+	$F_{100} = 30$ $F_{300} = 60$	F_{002}		0	–	9	–	14	–

search with condition 4–4–4–0 gives an answer consistent with the selection condition, while a search with 4–4–0 gives a map consistent with 4–4–4–0. Such behavior can be considered as an indication of the presence of three monomers per asymmetric unit.

5. Choice of the space group: CHFI case

Exact knowledge of the space group is crucial for structure resolution. It is the higher resolution reflections that identify systematic extinctions and, as a consequence, distinguish between pure and helical rotation axes. If the crystal does not diffract well, the wrong space group may be chosen.

The dependence of the results of direct phasing on the choice of space group was tested using the data for CHFI (Appendix *A*5).

The direct phasing procedure was applied at 14 Å resolution using different sets of magnitudes and different space groups. The grid was always taken as 28 × 28 × 40, and the cut-off level was always chosen to select the region with a volume equal to 0.1 of the total unit-cell volume. In all these tests, the phase set was selected if it revealed exactly eight connected regions of the same volume (one independent blob thus corresponding to one molecule per asymmetric unit).

When working at a resolution of 14 Å, the presence of reflections 002 and 004, and the absence of 001 and 003, indicate the screw axis 4_2. However, when reflection 002 is not measured, as is the case for the given experimental data set, the data set becomes compatible with the axes 4_1 and 4_3. A fourfold rotation axis can be also considered possible, especially if some weak (noise) signal is accepted as intensity for reflections 001 and 003.

Therefore, the available set of structure-factor magnitudes can be considered in space groups *P*422, $P42_12$ or *P*4222 when several reflections are not measured. It should be remembered that the phasing procedure cannot distinguish between enantiomers, and in what follows space groups $P4_322$ and $P4_32_12$ are treated together with $P4_122$ and $P4_12_12$, respectively.

The first series of random phase generations were performed in all possible groups using the complete set of magnitudes calculated. Table 1 shows the number of selected phase sets from 50 000 generated, indicating (line 1) a clear preference for the correct space group. The maps calculated with the averaged selected phase sets gave a very precise position of the molecule (Fig. 3).

The next calculation (line 2) was performed using the experimental data implemented with calculated and properly scaled structure factor magnitudes for the four missed reflections. This inclusion did not change the behavior of the procedure and illustrated its robustness with respect to typical experimental errors.

Then, a series of similar runs were performed with several reflections removed (thus making other space groups possible) or added (thus removing some groups from possible trials). All averaged images for the correct space group, $P4_22_12$, correspond to the true solution of the phase problem. The weighted correlation of averaged selected phases ranged in general from 0.6 to 0.7. This value is not high for such phasing but may be explained by the fact that no optimal protocol was searched for. The simplest criterion and default conditions were used, and only a single phasing step was performed. In all cases but the one discussed below, the method indicates, through too small a number of selected variants, the wrong space groups, even though these space groups are formally compatible with the data.

It can be observed that replacement of the experimental magnitudes F^{obs} by F^{calc} calculated from the atomic model did not significantly change the efficiency of the procedure used with the complete data set. However, when four missed reflections are not recovered, the selection is slower and the weighted correlation of averaged selected phases is slightly lower. In the intermediate situation, when the magnitudes of these reflections are present but underestimated, the phasing gives results of intermediate quality. When, on the contrary, the magnitudes are overestimated, the procedure selects too

many variants. This result can be explained by the fact that at least two of these reflections are already very strong and already produce a Fourier map with a necessary condition. Their overestimation decreases further the relative contribution of other, weaker, reflections; this contribution to the syntheses calculated becomes negligible, independently of the phases assigned to these reflections.

Addition of several forbidden reflections, assuming that some noise in these positions of the reciprocal space was interpreted as signals, removes the correct space group but does not provide a high rate of selection for any possible space groups. In contrast, deletion of several reflections makes the list of possible space groups larger. A special case was phasing in the absence of reflection 002, rendering space group $P4_12_12$ possible. The selection in this space group is as efficient as the selection for the correct space group (compare the columns $P4_12_12$ and $P4_22_12$ in Table 1). However, when the Fourier syntheses with obtained phases are calculated for both these space groups, their comparison gives a clear preference for the correct space group $P4_22_12$ (Fig. 3). While for space group $P4_22_12$ the map shows isolated compact blobs, the maps in space group $P4_12_12$ show very elongated high-density regions, atypical for proteins, with complicated packing, so that it is difficult to calculate their number, even in this final synthesis.

6. Conclusion

An *ab initio* phasing procedure based on topological properties of Fourier maps has been tested with numerous examples of experimental and semi-experimental data. The tests show a high robustness of the procedure. Variation of basic parameters of the method within reasonable limits does not significantly modify the results. This fact confirms our previous hypothesis that the correct solution of the phase problem differs from other phase combinations in the stability of its features.

Direct phasing in incorrect space groups has a tendency to have too small a share of appropriate phase sets in comparison with phasing using the correct symmetries. Therefore, when the correct space group is ambiguous, a comparison of results of direct phasing in several possible space groups can eliminate wrong candidates.

Similarly, in the case of an unknown number of connected components (in particular, an unknown number of molecules in the asymmetric unit), a different number of connected blobs can be tried. If the percentage of selected variants is too high it is difficult to expect the obtained images to be of high quality. Therefore, the selection criterion is too weak and has to be reinforced. A condition that, on the contrary, gives too small a percentage of selected variants is wrong and can be ruled out.

APPENDIX *A*
Test data

A1. Protein G

The crystals of protein G (Derrick & Wigley, 1994) belong to space group $P2_12_12_1$; the unit-cell parameters are a = 34.9, b = 40.3, c = 42.2 Å. This protein contains 61 residues (about 600 non-H atoms). The diffraction data set contains 15 reflections at a resolution of 16 Å (all magnitudes are measured), 85 reflections at 8 Å and 580 reflections at 4 Å.

A2. RNAse Sa

RNAse *Sa* (Sevcik *et al.*, 1991) is composed of 96 residues and crystallizes in space group $P2_12_12_1$, a = 64.9, b = 78.3, c = 38.8 Å, with two molecules in the asymmetric unit. The data set is complete and contains 29 independent reflections at 18 Å resolution and 39 reflections at 16 Å resolution.

A3. AspRS

The crystals of the cubic form of the $tRNA^{Asp}$–$tRNA^{Asp}$–synthetase complex belong to space group $I432$, have unit-cell dimensions a = 354 Å and contain two molecules of synthetase (478 residues each) and two of tRNA (75 bases each) per asymmetric unit (Ruff *et al.*, 1991; Urzhumtsev *et al.*, 1994).

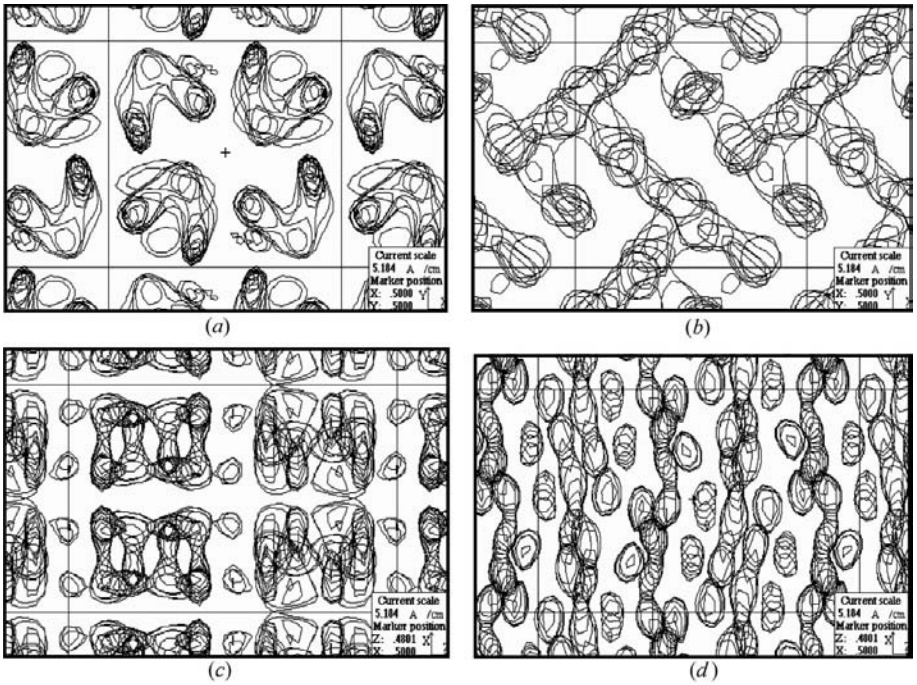


Figure 3
Ab initio phased syntheses for CHFI. Both maps are calculated at a resolution of 14 Å. (*a*) and (*c*) Maps after phasing in space group $P4_22_12$ (correct choice); (*b*) and (*d*) maps after phasing in space group $P4_12_12$. Projection *0xy* (upper figures) shows one-half of the unit cell (four molecules from eight in space group $P4_22_12$), and projection *0xz* (lower figures) shows the whole cell.

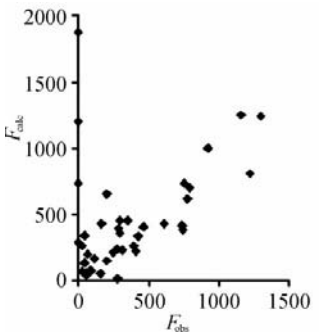


Figure 4
Comparison of experimental and simulated structure-factor magnitudes for CHFI. The four points on the vertical axis represent the unmeasured reflections.

The protein molecules form a compact dimer, and the tRNA molecules are attached to its surface and are separated from one another. There are three low-resolution neutron diffraction data sets, obtained from identical crystals, corresponding to three different contrasts of D_2O/H_2O, showing tRNA, the protein and the whole complex, respectively (Moras *et al.*, 1983). These three data sets are complete and have 37 independent reflections at 45 Å resolution, 49 independent reflections at 40 Å resolution and 110 independent reflections at 30 Å resolution.

A4. FixJN

The FixJN crystals (Birck *et al.*, 1999) belong to space group *C*2 and have unit-cell parameters a = 132.2, b = 91.3, c = 59.1 Å, $\beta = 112^\circ$. The asymmetric unit contains three monomers that correspond to one-and-a-half physiological dimers. The intensities for the reflections at a resolution of less than 20 Å were missed in the experiment. To restore them, the same procedure was applied as described below for CHFI. The resulting magnitudes simulated F^{obs} and were used for direct phasing.

A5. CHFI

CHFI (Behnke *et al.*, 1998) was reported previously (Chen *et al.*, 2000) as a difficult case for molecular replacement. Therefore, direct phasing of its crystals (space group $P4_22_12$, unit-cell parameters $a = b$ = 57.12, c = 80.26 Å) can be considered as an alternative way to obtain molecular packing that can help to position the search model.

The CHFI crystals belong to space group $P4_22_12$ (No. 94). At a resolution of 14 Å and lower, the complete data set consists of 45 reflections, excluding F_{000}. Five of these reflections correspond to the extinctions of the space group: 100, 300, 001, 003 and 005. The experimental magnitudes for four other reflections are not available, those of 110, 002, 112 and 004.

The complete set of structure-factor magnitudes was calculated from the available atomic model (Behnke *et al.*, 1998) by adding the contribution from the flat solvent taken with the parameters k = 0.35 e Å^{-3} and B = 46 Å^2 [see Fokine & Urzhumtsev (2002) for the choice of these parameters]. A comparison of calculated and experimental structure-factor magnitudes is shown in Fig. 4.

The work of NL and VL was supported by grant No. RFBR 03-04-48155. The work of AF was partially supported by the regional administration of Lorraine. AU and VL are partners in the framework of the Pole 'Intelligence Logicielle'. AU and JPS are collaborating in the framework of GdR 2417 of CNRS. The program *CAN* (Vernoslova & Lunin, 1993) was used to prepare maps. The authors thank E. Dodson, D. Moras and D. Teller for making their experimental data available for our tests.

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.
Behnke, C. A., Yee, V. C., Le Trong, I., Pedersen, L. C., Stenkamp, R. E., Kim, S.-S., Reeck, G. R. & Teller, D. C. (1998). *Biochemistry*, **37**, 15277–15288.
Bhat, T. N. & Blow, D. M. (1982). *Acta Cryst.* A**38**, 21–29.
Birck, C., Mourey, L., Gouet, P., Fabry, P., Schumacher, J., Rousseau, P., Kahn, D. & Samama, J.-P. (1999). *Structure*, **7**, 1505–1515.
Chen, W., Kleywegt, G. & Dodson, E. (2000). *Structure*, **8**, 213–220.
Derrick, J. P. & Wigley, D. B. (1994). *J. Mol. Biol.* **243**, 906–918.
Fokine, A., Lunina, N., Lunin, V. Y. & Urzhumtsev, A. (2003). *Acta Cryst.* D**59**, 850–858.
Fokine, A. & Urzhumtsev, A. (2002). *Acta Cryst.* D**58**, 1387–1392.
Hunt, J. F., Vellieux, F. M. D. & Deisenhofer, J. (1997). *Acta Cryst.* D**53**, 434–437.
Lunin, V. Y. & Lunina, N. L. (1996). *Acta Cryst.* A**52**, 365–368.
Lunin, V. Y., 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. Y., Lunina, N. L. & Urzhumtsev, A. (2000). *Acta Cryst.* A**56**, 375–382.
Lunin, V. Y., Urzhumtsev, A. & 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. (2003). *Acta Cryst.* D**59**, 1702–1715.
Moras, D., Lorber, B., Romby, P., Ebel, J. P., Giege, R., Lewit-Bentley, A. & Roth, M. (1983). *J. Biomol. Struct. Dyn.* **1**, 209–223.
Ruff, M., Krishnaswamy, S., Boeglin, M., Poterszman, A., Mitschler, A., Podjarny, A., Rees, B., Thierry, J.-C. & Moras, D. (1991). *Science*, **252**, 1682–1689.
Sevcik, J., Dodson, E. & Dodson, G. (1991). *Acta Cryst.* B**47**, 240–253.
Urzhumtsev, A. G., Podjarny, A. D. & Navaza, J. (1994). *CCP4 Newsl.* **30**, 29–36.
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.