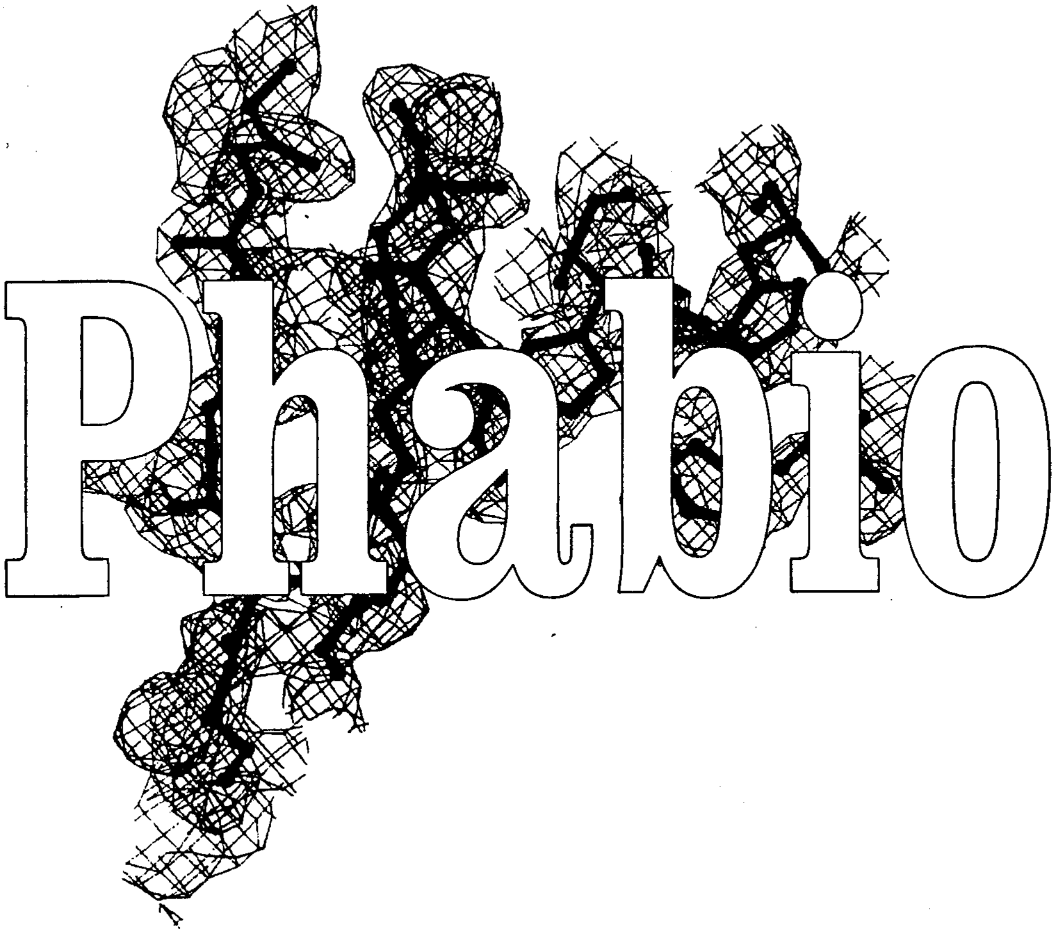
Euroconference on

Phasing
Biological Macromolecules
(Phabio)



Martina Franca – Italy

June 23-27 2001

AB INITIO PHASING STARTING FROM LOW RESOLUTION

By Alexandre Urzhumtsev[1], Alberto Podjarny[2] & Vladimir Y. Lunin[3]

[1]LCM3B, Universite Henri Poincare, Nancy 1, France; sacha@lcm3b.uhp-nancy.fr
[2] IGBMC, Strasbourg, France; podjarny@igbmc.u-strasbg.fr
[3]IMPB RAN, Pushchino, Russia; lunin@impb.psn.ru

Why low resolution first ?

In order to determine a three-dimensional structure of a crystal, all means are good but not all of them always work. In macromolecules, in contrast to small molecular crystals, there is no way for a time being to get a model directly from the diffraction data and it is necessary to pass through the step of Fourier synthesis calculation and its interpretation. In order to do so, a researcher needs to determine the phases corresponding to measured structure factor magnitudes.

Practically all traditional phasing techniques try to find, directly or indirectly, the solution of the phase problem by writing and resolving some kind of equations. Alternatively, the space of phases can be explored and the correct solution is chosen. The two only problems are a huge size of this space and the criterion of the phase selection.

Since the number of phase variants (the size of the search space) grows exponentially with the number of structure factors, it seems to be simpler to start from a small number of reflections. Due to different computational effects, these few reflections is better to be chosen by their resolution than by their magnitudes or other characteristics.

Therefore, from the beginning, low resolution phasing is considered as the first step of the structure determination which should be followed by image improvement. Nevertheless, alternative applications of low resolution images can be suggested, for example :

a) when higher resolution diffraction data are not available, a low-resolution image can be used as a first structural information; a similar quality images can be obtained by electron microscopy (EM) or by small-angle scattering (SAXS); however, these methods need special preparations both in experiment and in data processing and are not always efficient; if all they work, they can be considered as complementary and the results of better quality can be obtained;

b) when a low resolution image can be used in order to place a homologous model in the unit cell like MR method but without need of high quality data and high model homology; such application can be eventually of great importance for Structural Genomic projects.

It can be noted that even for a few tens of lowest-resolution structure factors, an exhaustive search of the phase space is impossible and must be replaced, for example, by random search procedures.

Search strategy and selection criteria

In order to identify the correct phase set among other sets, some its features (or features of the corresponding Fourier synthesis) must be formulated and expressed numerically. Traditionally, it is supposed that there exists a score function such that its global minimum corresponds to the correct solution. In fact, many structure determination techniques, not only for the solution of the phase problem, are based on such assumption.

However, it can happen that the suggested criterion does not have a single minimum or that its global minimum is shifted from the 'best answer' (Fig. 1). Our study of different criteria [6] revealed that such situation is usual for the low resolution *ab initio* phasing in macromolecular crystallography (and in fact in many other crystallographic fields). If no unambiguous criterion is available, the processing of the results can be changed. For example, instead to pick up a single 'optimal' variant, other techniques which treat the results of the search more carefully can bring the correct crystallographic image as discussed below.

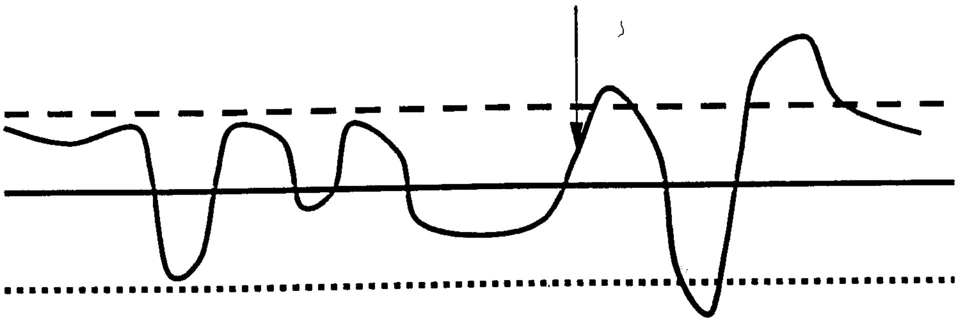


Fig. 1. Schematic presentation of a one-dimensional inappropriate score function. The arrow indicates the correct answer. Too high cut-off level (broken line) contains it in the list of variants but the list is too large. Too low cut-off level (dotted line) gives a single answer which is wrong. An optimal cut-off (solid line) leaves several groups of possible points which form 4 clusters; the correct cluster is the largest one and is somehow in the middle of selected points. Their average is reasonably close to the correct solution.

Search with models

A search with models has two complementary advantages. First, it allows to reduce the search space (instead of all phases, only the phases corresponding to a given class of models can be generated). Second, it allows a classical least-square comparison of experimental magnitudes with those calculated from the model. An old search model which is rediscovered from time to time is a spherical envelope.

It was found that the best fit of corresponding calculated structure factor magnitudes to the experimental data does not necessary happens for the correct envelope position. This is true also for detailed molecular envelopes and for more complicated flexible models consisting of few spheres (e.g., FAM method [2-4]). No search procedure will find the solution with such criterion, it is its intrinsic problem. A more sophisticated criterion of maximum likelihood has the same unfavourable property [8].

An important handicap of searches with models is that no one of them currently includes the contribution of bulk solvent into consideration which is extremely important at low resolution. It can be neglected at very low resolution but creates a barrier in phasing at the resolution of 8-15, sometime at 20 Å [7].

Search directly in the phase space

As an alternative, a search can be done directly with the phases. In this case, instead of comparison of experimental diffraction magnitudes with calculated values, they can be used to calculate Fourier syntheses whose properties are investigated. Some of these properties are formulated locally like an absence of high density at a rotation axis or a size of density fluctuation at any point of the unit cell. However, most often the properties are expressed in some cumulative form. First such example is electron density histograms which prescribe to a given share of the unit cell to have a given density value [1].

Recently, topological features of electron density maps have been used as a selection criterion for direct phasing [5]. These features express directly the demands to a 'good electron density map' as they are formulated by a visual analysis : a known number of connected regions seen at a high density cut-off level, no noisy peaks etc. (This type of score functions reflects that fact that during image analysis one looks for the position and the shape of regions of high density and practically never for the absolute values of electron density).

An extra advantage of such criteria is that they do not use any models and eventually could allow to cross the bulk-solvent barrier in phasing. Unfortunately, these score functions, similarly to model-based criteria described before, do not have the wanted feature neither: to have a unique global minima which is reached at the exact phase set. In this sense, all these score functions are found to be *inappropriate* to the problem. This means that the crystallographic image we are looking for has the property we want but it is not the only one possible with given experimental data and we need more ideas how to recognise it. For all "natural" restraints on electron density or structure factors tried in our works [1-10] the best values of the score function were usually attained for phase sets significantly different from the correct phases while best found phase sets resulted often in intermediate values of the score function.

Use of inappropriate score functions

When a score function does not allow to identify the correct solution unambiguously whether it is useful anyhow ? In the case of such inappropriate score functions, they can have multiple minima with the global minimum does not corresponding to the correct solution or displaced from it (Fig. 1). At the same time, a search with a score function can be seen from another end : it allows to eliminate wrong phase sets. The better the score function, the higher percentage of wrong solutions are eliminated.

Such replacement of the point of view changes the whole search strategy. Instead to search for a single phase set producing the best value of the score function, all phase sets with relatively good values of the score function are selected. If a minimum of the score function is displaced, the exact phase solution can be eliminated but a number of close phase sets will be selected nevertheless. The whole game now is to choose a proper cut-off level with which the score function leaves the phase sets close to the correct solution and eliminate as much as possible irrelevant points.

When such selection is done, a statistical analysis of many selected phases set is needed. An important feature of numerous tested score functions, including those discussed above, is that the minimum corresponding to the correct solution is not necessary the deepest one but is the largest one. In other words, the corresponding crystallographic image is less sensible to perturbations in phases. This means that after phase sets are selected by filtration with an inappropriate score function, the selected ensemble (population) is 'enriched' by phase sets more or less close to the solution.

From our experience, the selected variants are grouped in a small number of clusters and not distributed uniformly in the phase space (these clusters correspond to deepest minima of the score function). The simplest action now is to average all selected phase sets. This averaging will give mean phase values which in the case of a good score function will be somewhere inside the region of the correct minimum, and their figures of merit. The latter are extremely important because they show the quality of such averaging, the dispersion of the phases in the selected variants.

More careful analysis can provide with a few major clusters with the averaging done independently for each of them. In this case, the average values for the correct cluster is of higher quality than in the case of overall averaging but this is compensated by the problem to identify this correct cluster among several available. It is possible to calculate corresponding maps and to check them with an alternative score function (see, for example, [8]) with the hope that the correct solution corresponds a common minimum for both functions while false minima are different.

It can be noted that the averaged solution does not necessary has the same property by which all individual phase sets were selected. On one hand, this enriches the image; for example, in the FAM method, the averaged phases do not correspond to a structure formed by a few pseudo atoms but to some much more complicated image. On the other hand, this can help to identify incorrect answers.

A practical example : LDL structure

In test cases, a single criterion is used in order to study more clear the properties of this criterion. On contrary, in a practical case, in order to find a structure all means are used together including techniques different from crystallography, like electron microscopy, small-angle scattering etc. The case of LDL (Low Density Lipoprotein, alias bad cholesterol) is an example of such complicate macromolecular complex.

The knowledge of the molecular structure of LDL, a large lipoprotein complex, is of great interest for medical investigations. Currently available LDL crystals do not diffract to high resolution and do not allow the application of standard crystallographic techniques. Additional difficulties arise due to a very dense crystal packing and due to the presence of several components with quite different mean densities. Several *ab initio* phasing methods reported previously [2-5] have been successfully applied to find a crystallographic image of LDL at a resolution of 27 Å. The most promising results have been obtained using direct phasing with a connectivity analysis of the electron-density maps. The current image makes it possible to discern a single particle covered by a layer of relatively high density that is asymmetrically distributed on the particle surface. It shows a partition of high and low densities inside the particle and, in particular, strips of varying density in the lipid core [10].

Phase extension to find secondary structure

While the molecular envelope itself can be eventually found by few other techniques (with their own technical difficulties), the crystallographic low resolution phasing has an advantage to be only a first step toward higher resolution image. Next major step would be a phase extension to the resolution when secondary structure elements can be identified and essentially the model can be constructed. It would be logical that for new phases the search is done in the whole space while for previously phased reflections the search is carried out near known values.

Several successful test applications have been done. In one case, a crystallographic image of the protein G was found *ab initio* starting from low resolution [9]. A comparative analysis of this image with the exact maps estimates the efficient resolution of corresponding Fourier synthesis as something between 4 and 5 Å. The map show clearly the α–helix and the β–sheet of the protein.

In the second example, the structure of the ER-1 protein was determined *ab initio* by the connectivity approach. Current maps show unambiguously the secondary structure of this protein. This work is in progress.

Conclusion

Several suggested score functions which can estimate the quality of phase sets or corresponding crystallographic images do not allow unambiguously to identify the correct solution. However, they produce a statistical improvement of the available phase sets, for example, generated randomly. For phase sets reasonably close to the exact values, their percentage is higher in the ensemble of the selected phase sets in comparison with their percentage in a random population. Such 'enrichment' by reasonably good sets allows, as a rule, to obtain a correct crystallographic image through an averaging of the selected phase variants.

A more advanced statistical *ab initio* phasing approach has been developed on the base of this analysis. The search of the solution can be done either directly in the phase space (where every point is a phase set for given structure factors) or in the multidimensional space of parameters (every set of which allows to calculate a corresponding phase set; see for example [2-4]). The procedure consists in several steps :

- generation of a random population of points in the search space and selection of those points which resulted in a reasonable value of the score function;
- cluster analysis of the selected population in order to identify compact groups of the selected variants;
- averaging variants inside every identified cluster.

The procedure provides with a few alternative solutions (usually, a single solution) of the problem considered, which are investigated in the following tests if they are several.

In one of applications, the first crystallographic image of the ribosomal particle 50S has been found *ab initio* [3]. Recently, an application of the developed procedures to the low resolution *ab initio* study of Low Density Lipoprotein complex [10] has been reported. Some latest results show that with the proper choice of the search parameters and the score function, the image resolution at least of 4-5 Å can be reached [9] starting from the low resolution end.

We thank our collaborators N. Lunina, T. Petrova, T. Skovoroda, E. Vernoslova and E. Chabriere for their help at different stages of this research, M. Baumstark, S. Ritter and K.Diederichs for the collaboration on the LDL structure analysis, LORIA for the financial support through CPER. VL was supported by RFBR 00-04-48175 grant

References

[1] Lunin, V., Urzhumtsev, A., Skovoroda, T. (1990) *Acta Cryst.*, **A46**, 540-544

[2] Lunin, V., Lunina, N., Petrova, T., Vernoslova, E., Urzhumtsev, A., Podjarny, A. (1995) *Acta Cryst.*, **D51**, 896-903

[3] Urzhumtsev, A., Vernoslova, E., Podjarny, A. (1996) *Acta Cryst.*, **D52**, 1092-1097

[4] Lunin, V., Lunina, N., Petrova, T., Urzhumtsev, A., Podjarny, A. (1998) *Acta Cryst.*, **D54**, 726-734

[5] Lunin, V., Lunina, N., Urzhumtsev, A. (2000) *Acta Cryst.*, **A56**, 375-382

[6] Lunin, V., Lunina, N., Petrova, T., Skovoroda, T., Urzhumtsev, A., Podjarny, A. (200(*Acta Cryst.,* **D56,** 1223-1232

[7] Urzhumtsev, A., Lunina, N., Skovoroda, T., Podjarny, A., Lunin, V. (2000) *Acta Crysı* **D56,** 1233-1244

[8] Petrova, T., Lunin, V., Podjarny, A. (2000) *Acta Cryst.,* **D56,** 1245-1252

[9] Lunina, N., Lunin, V., Urzhumtsev, A. (2000) *ECM-19 Abstracts, XIXth Europea Cryst.Meeting, 25-31 August 2000, Nancy, France*, 62

[10] Lunin, V., Lunina, N., Ritter, S., Frey, I., Keul, J., Diederichs, K., Podjarny, A Urzhumtsev, A., Baumstark, M. (2001) *Acta Cryst.,* **D57,** 108-121