Conference Proceedings
DL-CONF-97-001

Recent Advances in Phasing

Proceedings of the CCP4 Study Weekend
January 1997

K S Wilson G Davies A W Ashton and S Bailey

August 1997

COUNCIL FOR THE CENTRAL LABORATORY OF THE RESEARCH COUNCILS

LOW RESOLUTION CRYSTALLOGRAPHIC IMAGES

by A.Urzhumtsev#*, V.Lunin* and A.Podjarny#

#UPR de Biologie Structurale, IGBMC, B.P.163, 67404, Illkirch, c.u. de Strasbourg, France
**IMPB RAN, Puschino, Moscow region, 142292, Russia*

1. Introduction.

The definition of «low resolution» depends on the traditions of a specific laboratory and, first of all, on their typical subjects. In the case of small molecules it can be 3Å. In the case of typical proteins, it is rather about 6-8Å. Another meaning of the term «low resolution» is about 20-25Å, the limit below which X-ray data are quite often not collected.

This paper deals with the analysis of macromolecules, and the resolution below 6-8Å will be referred to as «low resolution» and the one below 20-25Å as «very low resolution» (VLR in what follows). It should be noted that these two limits define the resolution zone where the contribution of the bulk solvent is strong and uncorrelated to that from the macromolecule itself. At higher resolutions the contribution is negligible, and at lower resolutions it is strong but roughly proportional to the one of the macromolecule (Urzhumtsev & Podjarny, 1995).

Measuring the very low resolution X-ray data is technically difficult, and many research groups do not collect them. However, they carry information that can be useful. This paper discuss their importance for improving the molecular images as well as the possibilities of an independent use of these data.

2. Do very low resolution data have any information ?

The basic sources of a noise in macromolecular crystallographic images are *systematic* errors. While in a real case the synthesis is usually worse than expected it is much more difficult to obtain a noisy image in a test calculation. The mean value of *independent* phase errors can reach about 60-70° and the synthesis will still be quite good and close to the ideal image. However, it is easy to destroy an image by introducing *systematic* errors, for example, by error in heavy atoms parameters. Another example is missing of a part of the model, for example the solvent.

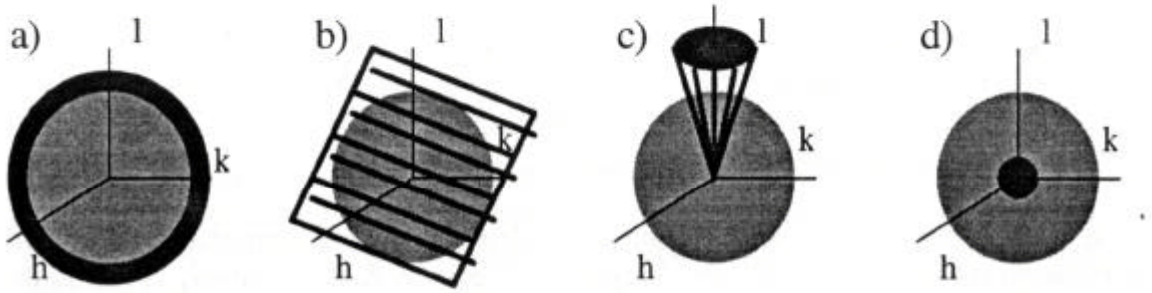

Fig. 1. Schematic presentation of different possibilities of systematic missing of X-ray amplitudes: a) standard resolution cut-off; b) in plane; c) along one axis; d) very low resolution data

The second possibility is *systematically missing* of reflections in the map calculations. Some examples are schematically shown in Fig. 1. Fig 1a corresponds to the *usual* high resolution cut-off. Fig 1b and 1c corresponds to relatively *rare cases* which nevertheless exist. Missing of a plane of reflections causes breaks in the density in the conjugate direction in real space (Lunin, 1991). A systematic absence of reflections along an axis can cause a complete loss of the molecular envelope (Urzhumtsev et al., 1989).

Fig. 1d corresponds to a *usual situation* when VLR data are excluded from the map calculation. They can be either not measured or measured but not phased. Usually they are only a small number of reflections but they are strong and removed systematically.

A phase extension procedure for phasing the VLR data applied by Podjarny et al. (1981) in the case of tRNA demonstrated a drastic improvement of the image. For calculated data (Urzhumtsev, 1991) it was clearly shown that the exclusion of only 1% of the data (29 reflections out from 2500) completely destroy the molecular image at 6Å resolution. In the case, for example, of SIR phase errors, the molecular envelope keeps its position but the electron density peaks are shifted. In the case of missing VLR reflections the effect is inverted: the envelope is lost but the peaks are at their places. This is natural because the exclusion of VLR terms should cause large scale modulations of the density in the unit cell.

The fact that the peaks are at the right place has important consequences. Firstly, when a map is calculated at high resolution, its peaks have a high contrast and such density modulation does not «hide» them completely; this has allowed crystallographers to ignore VLR data for a long period. Secondly, it gives a possibility of automatically determining the molecular envelope from such synthesis.

The knowledge of the envelope can be used to improve the molecular image. The phases of its structure factors can be used as a good approximation to the phase values of VLR reflections. If their amplitudes are available, simple adding them to the Fourier calculation can completely change the map (see Urzhumtsev, 1991, for an example of dractic improvement of the SIR image of the Elongation Factor G). Calculated amplitudes can be used to estimate the quality of the calculated phases and give corresponding weights for the Fourier coefficients through the comparison with the experimental ones.

3. How to use the information from very low resolution data ?

Therefore, VLR data do carry important information, first of all, on the shape of molecule. Such an information can be used in different cases (Podjarny & Urzhumtsev, 1997), for example:

- in density modification procedures for the image improvement;
- in the molecular replacement if the *internal* differences between two molecules are large;
- if diffraction data are not available at higher resolution;
- in the case of very large molecular complexes, like ribosome;

etc.

If VLR amplitudes have been measured, the determination of their phases by isomorphous replacement is difficult while not impossible (Podjarny & Urzhumtsev, 1997). In the case of viruses where practically all VLR reflections are centrosymmetric, a good approximation can be done by calculation of structure factors from a spherical shell. In the general case, a searching procedure based on some *a priori* knowledge of the density can be

applied to find these phases. For this procedure, it is needed to specify: a) the search model (parameters); b) the search space; c) the sampling procedure; d) the available data; e) the criteria of the search.

Sampling of the whole phase space

In the simplest case, the search is either systematic or random with a representative sampling of the whole phase space. In practical terms it means that the space should not be too large, i.e. a small number of reflections can be phased.

An information which can be used to identify the correct solution should be of general type, for example, the knowledge of the correct electron density histogram (Lunin, 1988; 1993). For any generated phase set, a map of a given resolution can be calculated with experimental amplitudes and these phases. A correlation of histograms, target and calculated, can be use as a search criterion.

Table 1 shows the distribution (histogram correlation, C_H, *vs* phase correlation, C_P) of phase sets of a typical search done with calculated data. Most of the generated phase sets have a poor value for both correlations. However, the converse is not true and the phase sets with highest value of the histogram correlation are not necessarily correct. The phase correlation distribution of these phase sets (columns with $C_H \geq 0.9$ in Table 1) shows two groups of phase sets. Some of them have a reasonable phase correlation value, while the others are far from the correct solution. Note also that there are a number of sets with high C_P but low histogram correlation value. The single phase set with the highest C_H value and also the highest C_P value is not statistically significant.

C_H	--->	0.7	-----	-----	-----	-----	0.8	-----	-----	-----	0.9	-----	-----	-----	--->	1.0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	^
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	|
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	|
0	0	0	0	0	0	1	1	4	1	5	1	1	0	0	0	0.950
0	2	2	6	27	32	52	93	84	70	29	7	0	0	0	0	|
78	186	311	486	610	685	727	701	531	284	86	14	0	0	0	0	0.875
2504	3186	3645	3594	3265	2812	2139	1523	846	359	103	6	0	0	0	0	|
8842	8633	7310	5114	4398	2852	1680	872	428	160	42	4	0	0	0	0	|
7495	4763	3298	2235	1220	569	269	110	24	3	0	0	0	0	0	0	^
1465	2385	2734	2876	2746	2123	1502	848	4551	138	44	5	0	0	0	0	|
5157	9999	9999	9999	9999	9999	9732	6301	3384	1290	305	27	0	0	0	0	0.500
2887	7000	9511	9999	9999	9999	7571	4556	2146	689	119	10	0	0	0	0	|
49	303	477	553	508	312	191	71	9	2	0	0	0	0	0	0	C_P

Table 1. Two-dimensional distribution of the generated variants of phase sets for the case of model data at 30Å resolution (29 reflections). The horizontal dimension corresponds to the histogram correlation, C_H, and the vertical one to the weighted phase correlation, C_P. The correct solution should be in the top right corner. Two major clusters are marked by a frame, the variants with highest C_P are indicated by inverted colours.

The behaviour of the C_P values for the sets with highest C_H can be generalised. The search criterion does not select for a single solution, but gives an indication of possible solutions. This solutions are not uniformly distributed with respect to the C_P. Further analysis shows that they appear in clusters (e.g., the two peaks in Table 1 correspond to two clusters); one of these clusters is close to the correct solution.

On the basis of these observations, the following procedure has been suggested to obtain *ab initio* the phases of the VLR reflections (Lunin, Urzhumtsev, Skovoroda, 1990):

a) generation of a large number of phase sets (e.g., one million for 30 phases);
b) calculation of an electron density map for every phase set and calculation of its histogram;
c) selection of the phase sets with highest histogram correlation as admissible ones;
d) after a sufficient number (e.g., one thousand) phase sets are selected, analysis of the distribution of these sets by some clustering procedure;
e) classification of the clusters according to their size in a 'cluster tree'; for every major cluster average the corresponding phase sets in order to obtain the mean phase values and their figures of merits;
f) calculate the corresponding weighted maps and choose, if possible, the correct one.

For the step (d), a proper distance between two phase sets should be defined taking into account different choices of the unit cell origin (Lunin & Lunina, 1996), density flipping and enantiomer.

The procedure was found quite robust in several applications both to the calculated and experimental data. In these cases about 30 reflections were successfully phased which gave images of reasonable quality. The limiting point was the computing time. In order to get finer details, it is necessary to go deeply in the cluster tree to smaller clusters and still have large enough number of phase sets with a high enough value of the criterion.

Another problems is that, unfortunately, while for the middle resolution maps a general method to obtain the corresponding histogram *a priori* has been suggested (Lunin & Skovoroda, 1991), no similar method was found for the very low resolution.

It is important to note that a similar behaviour of the selected phase variants has been observed when the criterion of the histogram closeness was replaced by the criterion of a compact globular envelope.

Simplest parametrisation of the phase space

In order to increase the number of phased reflections for the same level of computing power, the search model should be parametrised. A proper parametrisation should automatically avoid sampling of the «empty» regions of the phase space and the correct phase set should belong to the chosen subspace or, at least, be close enough to it. The number of parameters of every model should be small enough (at least, less that the number of data) in order to make the criterion of choice significant.

The simplest possible way of modelling a molecule is to replace it with a large gaussian sphere, which involves only four parameters (position and radius). Systematic R-factor search with such a model is a known approach to find the centre of the gravity of the molecule. It has been successfully applied in several cases, for example, by Podjarny et al. (1987).

A search with several (N) spheres can be tried but for N>2 it is computationally difficult. In this case, a random sampling can been applied, similarly to the one used for the histogram criterion. A number of test calculations have been carried out using the experimental data of the $tRNA^{Asp}$-RS complex (Giegé et al., 1980; Urzhumtsev et al., 1994).

First, several models of 5-7 large spheres were constructed manually which reproduced the low resolution (30-50Å) image of the complex with a high correlation (0.75-0.80) with the exact one. Then a large number of models, each composed of a small number (2-5) of spheres with randomly distributed centres was generated. Corresponding structure factors were calculated and compared with the correct values, using the amplitude correlation, C_F, as the search criterion. It was found that, similarly to the search with the histogram criterion, the phase sets corresponding to the models with highest C_F are grouped in a small number of clusters, one of which is quite close to the correct phase set. A typical distribution is shown in Table 2 and is schematised in Fig. 2. To check whether this type of the distribution of selected variants is related to the random sapling, two different 2-spheres searches, a random one and a systematic one, have been carried out exactly at the same conditions. The corresponding distribution were very similar.

C_F --->	0.3	0.5	0.7	0.8	1.0	
0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	^
0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	0 0 0 0	|
0 0 0	0 0 0 0	0 0 0 0	0 1 0 0	0 0 0 0	0 0 0 0	0.9
0 0 0	0 0 0 0	0 0 0 0	2 6 2 2	0 0 0 0	0 0 0 0	|
0 0 0	0 0 0 0	0 0 11 17	59 131 214 61	0 0 0 0	0 0 0 0	|
0 0 0	0 0 0 0	0 0 34 116	345 660 754 183	0 0 0 0	0 0 0 0	|
0 0 0	0 0 0 0	1 18 124 516	1118 1884 1464 233	0 0 0 0	0 0 0 0	0.7
0 0 0	0 0 0 0	3 60 344 985	2183 3133 1749 225	0 0 0 0	0 0 0 0	|
0 0 0	0 0 0 0	12 134 601 1645	3209 3898 1943 150	0 0 0 0	0 0 0 0	|
0 0 0	0 0 0 7	23 269 1046 2083	3866 4162 1649 88	0 0 0 0	0 0 0 0	|
0 0 0	0 0 5	77 463 1591 3030	5051 5092 1710 55	0 0 0 0	0 0 0 0	|
0 0 0	0 1	19 105 695 2121	4178 7601 6853 1925	45 0 0 0	0 0 0 0	|
0 0 0	0 4	39 198 883 2738	6433 9999 9090 2494	119 0 0 0	0 0 0 0	0.4
0 0 0	0 8	46 246 1280 4003	9085 9999 9999 2461	116 0 0 0	0 0 0 0	|
0 0 0	1 10	88 411 1721 5320	9999 9999 9999 2514	53 0 0 0	0 0 0 0	|
0 0 0	0 12	125 594 2310 6783	9999 9999 9999 2759	51 0 0 0	0 0 0 0	|
0 0 0	0 17	180 783 2964 8804	9999 9999 9999 3002	156 0 0 0	0 0 0 0	0.2
0 0 0	0 23	245 909 3351 9999	9999 9999 9999 3412	156 0 0 0	0 0 0 0	|
0 0 0	0 30	333 1037 3499 9999	9999 9999 9999 3448	150 0 0 0	0 0 0 0	^
0 0 0	1 44	431 1207 3421 9999	9999 9999 9999 3533	70 0 0 0	0 0 0 0	|
0 0 0	0 56	484 1376 3283 9738	9999 9999 9999 9738	100 0 0 0	0 0 0 0	0.0
0 0 0	3 50	438 1204 3226 9512	9999 9999 9999 6287	167 0 0 0	0 0 0 0	|
0 0 0	5 189	1046 2353 7758 9999	9999 9999 9999 9999	436 0 0 0	0 0 0 0	C_P

Table 2. Two-dimensional distribution of the FAM-generated variants of phase sets for the case of experimental data of the AspRS complex at 50Å resolution (31 reflections). The horizontal line corresponds to the amplitude correlation, C_F, and the vertical one to the weighted phase correlation, C_P. The correct solution should be in the top right corner. The major clusters are marked by a frame, the variant with highest C_P is indicated by inverted colours.

Several important observations should be noted:

1) the best phase set (C_P=0.9, C_F=0.7) does not correspond to the model with the highest amplitude correlation (C_F=0.8);
2) some of the phase sets with high amplitude correlation (C_F=0.8) are close to the correct phase set (C_P=0.7-0.8);
3) a phase set calculated from a model with high amplitude correlation (C_F=0.8) can belong to a cluster quite far (C_P=0.0) from the correct point;
4) averaging of phase sets inside the correct cluster produces a new phase set which is usually better that any individual solution.

A systematic procedure for this search, called FAM (Few Atoms Model), was proposed (Lunin et al., 1995) consisting of the following steps:

a) generation of a large number of simple pseudo-atomic models; every model consists of a the same small (2-10) number of large gaussian spheres; the co-ordinates of the centre of the spheres are distributed randomly in the unit cell;
b) structure factors calculation for every model;
c) comparison of the calculated amplitudes with the experimental ones and selection of the models with the highest CF;
d) merging of the selected phase sets by a clustering procedure;
e) analysis of the cluster tree; averaging of the phase sets inside every major cluster;
f) calculation of corresponding maps and identification, if possible, of the correct one.

This procedure was applied to several calculated and experimental data sets, giving good results. In particular, a 70Å-resolution crystallographic image (about 160 reflections) has been obtained for the 50S ribosomal particle (Volkmann et al., 1990) from *Thermus thermophilus* (T50S; Urzhumtsev et al., 1996). The FAM procedure is in the course of further development.

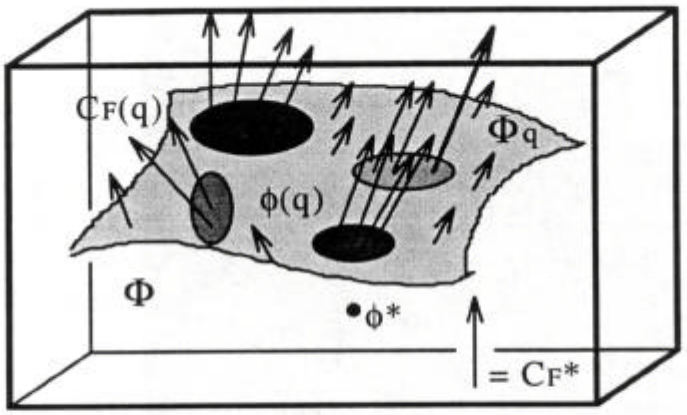

Fig. 2. Schematic presentation of the distribution of the phase sets in the FAM method. Every phase set is presented by a point in the phase space, Φ. Φ_q is a subspace of phase sets, $\phi(q)$, calculated from FAM models. The length of an arrow is proportional to the corresponding amplitude correlation, $C_F(q)$. The variants with $C_F > C_F^*$ are forming few clusters, one of them is close to the correct solution, ϕ^* (thick point).

Further parametrisation of the phase space

In the case were precise information about the three-dimensional molecular structure is available, the search space can be drastically reduced. This leads to the molecular replacement procedure (Rossmann, 1972), which reduces the dimension of this space to six, making possible a quasi-complete search. This procedure has been recently simplified (Navaza, 1994) giving automatically a list of possible positions and orientations of the model. In the case of good data and model, the correct solution corresponds to the maximum amplitude correlation. Alternative (wrong) variants have much lower correlation values, which allows to choose the solution easily. Otherwise, finding the answer is a difficult problem.

Molecular replacement is a standard technique, carried out usually at middle resolution (4-6Å) with an atomic model. At the VLR end the search model becomes a molecular envelope. If the search model is perfect, and the data are very accurate, a similar procedure with some important modifications (Urzhumtsev & Podjarny, 1995) brings the solution with reasonable contrast. In the case of less accurate data and an imperfect model

the contrast is much lower, as it was the case for the T50S particle (Urzhumtsev et al., 1996).

At very low resolution the imperfections of the model envelope can be important. For example, images reconstructed from electron microscopy can be compressed in one direction. When working with such models, molecular replacement puts the envelope either at its correct place (if possible) or into the solvent region but practically never at an intermediate position. This confirms a clustering behaviour of selected variants also for this case.

4. General conclusions

Several different low resolution phasing techniques which explore either the whole phase space or some specific subspace have been analysed. In all cases, the variants with best values of the search criterion are grouped in a small number of clusters which can be easily identified. One of these clusters is usually very close to the correct solution of the phase problem while others can be very far from it. It is important to note that the phase set with the best value of the criterion does not necessarily belong to this correct cluster. This observation explains, in particular, the problems with searches selecting a *single* solution. In general, this typical distribution of phase correlation *vs* search criterion has (by a peculiar coincidence) schematically the shape of the Strasbourg cathedral (Fig. 3; compare, for example, with Table 2). The top corresponds to the best variant which is impossible to identify by the available criteria, the floors correspond to the clusters, and the highest floor is the best cluster.

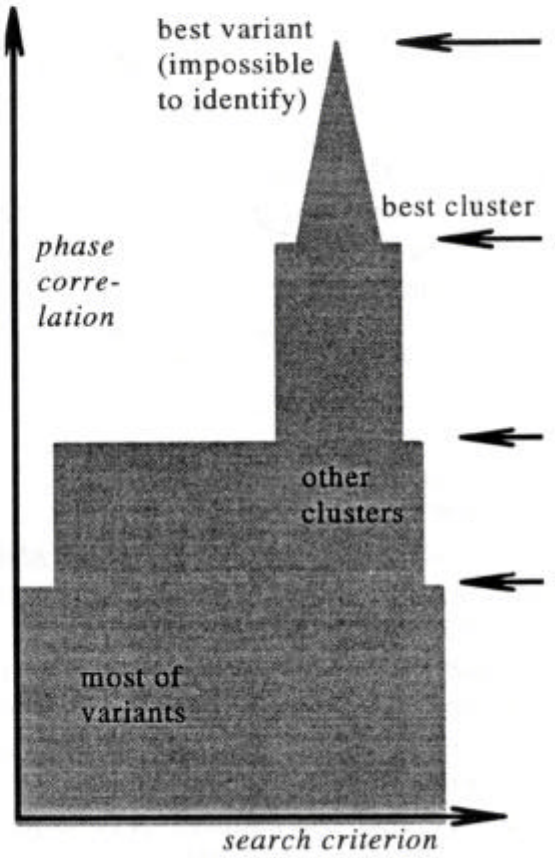

Fig. 3. Schematical profile of the Strasbourg cathedral.

As it was observed, the character of this distribution does not depend on the particular information and criterion used. For example, it can be noted in addition that the LAPS method developed by Volkmann (Volkmann et al., 1995) based on the Bricogne's maximum likelihood criterion found the solution for the T50S case also through a cluster oriented search.

All these observations indicate that at the very low resolution end the available information and search criteria are *weak* in the sense that in general they cannot indicate unambiguously the correct solution; additional information is necessary. At higher resolution, the same information and criteria, e.g., an atomic model and the amplitude correlation, can be *strong* enough to indicate a single solution. The particular low resolution

cases where the information is very accurate and the same criteria can unambiguously identify the right solution remain the exception rather than the rule.

The authors thank D.Moras for his support. This work was supported by the CNRS-RAS collaboration. AGU and ADP were supported by the Centre National de la Recherche Scientifique (CNRS) through the UPR 9004, by the Institut National de la Santé et de la Recherche Médicale, by the Centre Hospitalier Universitaire Régional. VYL was supported by RFBR grant 94-04-12844.

References

Giegé, R., Lorber, B., Ebel, J.-P., Thierry, J.-C. and Moras, D. Comp.Rend.Acad.Sci. (Paris), série D, *291* (1980) 393

Lunin, V.Yu. Acta Cryst., *A44* (1988) 144

Lunin, V.Yu. Dr.Sci.Theses, Institute of Crystallography RAS, Moscow (1991)

Lunin, V.Yu. Acta Cryst., *D49* (1993) 90

Lunin, V.Yu., Urzhumtsev, A.G. and Skovoroda, T.A. Acta Cryst., *A46* (1990) 540

Lunin, V.Yu. and Skovoroda, T.P. (1991) Acta Cryst., *A47* (1991) 45

Lunin, V.Yu., Lunina, N.L., Petrova, T.E., Vernoslova, E.A., Urzhumtsev, A.G. and Podjarny, A.D. Acta Cryst., *D51* (1995) 896

Lunin, V.Yu. and Lunina, N.L. Acta Cryst., *A52* (1990) 365

Navaza J. (1994) Acta Cryst., *A50* (1994) 157

Podjarny, A.D., Schevitz, R.W. and Sigler, P. Acta Cryst., *A37* (1981) 662

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

Podjarny, A.D. and Urzhumtsev, A.G. In *Methods in Enzymology*, (1997) in press

Rossmann, M.G. «The Molecular Replacement Method». Gordon & Breach, New York (1972)

Urzhumtsev, A.G. (1991) Acta Cryst., *A47* (1991) 794

Urzhumtsev, A.G., Lunin, V.Yu. and Luzyanina, T.B. Acta Cryst., *A45* (1989) 34

Urzhumtsev, A.G., Podjarny, A.D. and Navaza, J. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, *30* (1994) 29

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

Urzhumtsev, A.G. and Podjarny, A.D. Acta Cryst., *D51* (1995b) 888

Urzhumtsev, A.G., Vernoslova, E.A. and Podjarny, A.D. Acta Cryst., *D521* (1996) 1092

Volkmann, N., Hottenträger, S., Hansen, H.A.S., Zaytzev-Bashan, A., Sharon, R., Yonath, A. and Wittmann, H.G. (1990) J.Mol.Biol., *216* (1980) 239

Volkmann, N., Schlunzen, F., Urzhumtsev, A.G., Vernoslova, E.A., Podjarny, A.D., Roth, M., Pebay-Peyroula, E., Berkovitch-Yellin, Z., Zaytsev-Bashan, A. and Yonath, A. Joint CCP4 and ESF-EACBM Newsletter on Protein Crystallography, *32* (1995) 23