# JOINT CCP4 AND ESF-EACBM NEWSLETTER ON PROTEIN CRYSTALLOGRAPHY

An informal Newsletter associated with the BBSRC Collaborative Computational Project No. 4 on Protein Crystallography and the ESF Network of the European Association of the Crystallography of Biological Macromolecules.

Number 31 June 1995

## Contents



Editors: Sue Bailey

The Daresbury Laboratory
Daresbury, Warrington WA4 4AD UK

Keith S. Wilson

EMBL c/o DESY
Notkestrasse 85, D-2000 Hamburg 52
Germany

# On the problem of solvent modelling in macromolecular crystals using diffraction data: 1. The low-resolution range.

Alexandre G. Urzhumtsev[1,2] & Alberto D. Podjarny[1]

[1.] *IGBMC. BP 163. 67404 Illkirch. France.*
[2.] *IMPB RAS, Puschino, Moscow region, 142292, Russia*

## 1. Introduction

The properties of biological macromolecules are dependent not only on the molecule itself but also on the solvent distribution around it. The determination of this distribution is often problematic, due to the diverse degrees of disorder of the water molecules, ranging from the buried ones (behaving like solute atoms), to the very disordered ones which have properties close to that of bulk solvent. This disorder is of two types: static (changing from one unit cell to the other) and dynamic (changing in time). In the case of X-ray diffraction, the observed amplitudes are averaged in both time (over the length of data collection) and space (over all the unit cells of the crystal). Therefore, it is not possible to differentiate experimentally these two types of disorder, which together determine the contribution of the water molecules to the structure factors.

This contribution is clearly detached from noise in two resolution ranges:

1) the high one (d< 2Å), where it comes from well ordered water molecules which can be determined by usual crystallographic methods;

2) the low one ( d>8Å) where it comes from disordered water; this contribution is not normally taken into consideration in the crystallographic models, leading to a loss in the quality of the associated images.

The goal of this paper is to describe ways of obtaining the water distribution using X-ray data in the low resolution range, for the case when the molecular envelope is assumed to be known. The middle resolution range (8Å > d > 2Å), where the signal is weak but an atomic model is usually available, is treated in the accompanying paper (Howard et al., 1995).

The solvent region outside the molecular envelope is usually assumed to be flat except for the very ordered water molecules. However, as noted above, there is a strong solvent contribution to the observed amplitudes for d > 8Å, suggesting that this assumption is wrong. To correctly estimate the solvent contribution, special algorithms were developed and applied to the data from pig aldose reductase. This structure, solved to atomic resolution (Rondeau et al., 1992; Tête-Favier et al., 1993), was selected for all the work that follows since accurate structure factors and experimental phases (from MIR + non-crystallographic symmetry averaging) are available in the 25.0 - 2.2Å resolution range. Following Phillips (1980), the distribution of $<|\mathbf{F}_{obs}|>$ and of $<|\mathbf{F}_{cal}|>$ was calculated, which confirmed the observation that the solvent contribution becomes significant at the resolution about 7-8Å. In this particular case, the presence of accurate experimental phases $\phi_{ave}$ in the range 8-25Å enabled the calculation of a difference map with the Fourier coefficients

$$F_{obs} * [\exp(i*\phi_{ave}) - \exp(i*\phi_{cal})]$$

This map confirmed the presence of significant features in the solvent region. In the general case, where no experimental phases are available, the contribution of the solvent region should be estimated from the amplitudes.

## 2. Estimation of solvent structure factors

The contribution from the bulk solvent was estimated as follows. First, the envelope structure factors were calculated by:

a) a molecular envelope was calculated as the set of points inside spheres of approximately 2.5 Å radius surrounding all macromolecular atoms;

b) the envelope was flattened ($\rho=\rho_0$ inside, $\rho=0$ outside);

c) envelope structure factors $\mathbf{F}_{env}$ were calculated from this flat-envelope model.

Then, a first approximation to solvent structure factors was then calculated by

multiplying by a scale function $\lambda(|\mathbf{s}|)$:

$$\mathbf{F}^0_{sol}(\mathbf{s}) = - \mathbf{F}_{env}(\mathbf{s}) * \lambda(|\mathbf{s}|)$$

The values of $\lambda(|\mathbf{s}|)$ were obtained by optimal scaling of calculated and observed amplitudes in each resolution bin (see Figure 1). Clearly, the conventional approximation of $\lambda(|\mathbf{s}|)$ with a single gaussian function would be quite inappropriate.

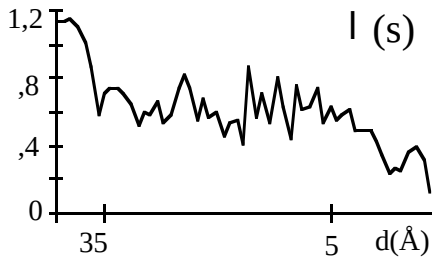


**Figure 1.** Optimal scaling coefficient $\lambda(|\mathbf{s}|)$ for the envelope structure factors as a function of the resolution.

A better approximation, $\mathbf{F}_{sol}(\mathbf{s})$, was then calculated by fixing the amplitude and shifting the phase to close the triangle formed by $\mathbf{F}_{sol}(\mathbf{s})$, $\mathbf{F}_{mod}(\mathbf{s})$ and $|\mathbf{F}|_{obs}(\mathbf{s})$ (see Figure 2). The mean values of this phase correction $\Delta\phi$ varied from 0° for very-low resolution terms (lower than 20 Å) to 70° at the resolution of 5Å.

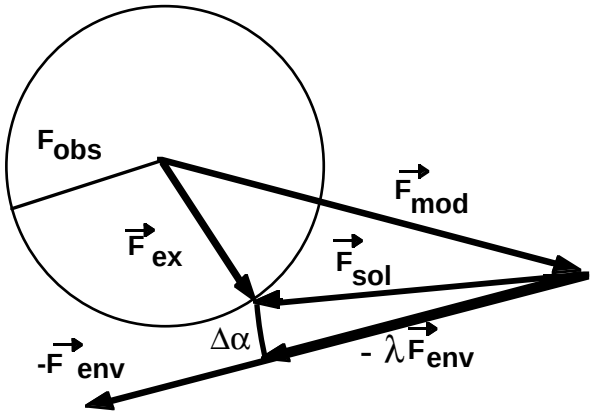


**Figure 2.** The scheme of the modelisation of the solvent structure factors $\mathbf{F}_{sol}(\mathbf{s})$ from the model $\mathbf{F}_{mod}(\mathbf{s})$ and envelope $\mathbf{F}_{env}(\mathbf{s})$ ones and from the experimental amplitudes $|\mathbf{F}|_{obs}(\mathbf{s})$. $\Delta\alpha$ is the phase shift to close the amplitude triangle.

The comparison of the estimated solvent structure factors $\mathbf{F}_{sol}(\mathbf{s})$ with the model ones showed the following:

a) at resolutions higher than 10 Å the amplitude correlation between $\mathbf{F}_{mod}(\mathbf{s})$ and $\mathbf{F}_{sol}(\mathbf{s})$ is small (varying from 0 to 50%) and the phase difference is between 100° and 120° showing that these structure factors are almost independent;

b) at resolutions lower than 10Å both the amplitude correlation and the phase differences start to grow and approach the values of 100% and 180° respectively around 20Å (see Figure 3).

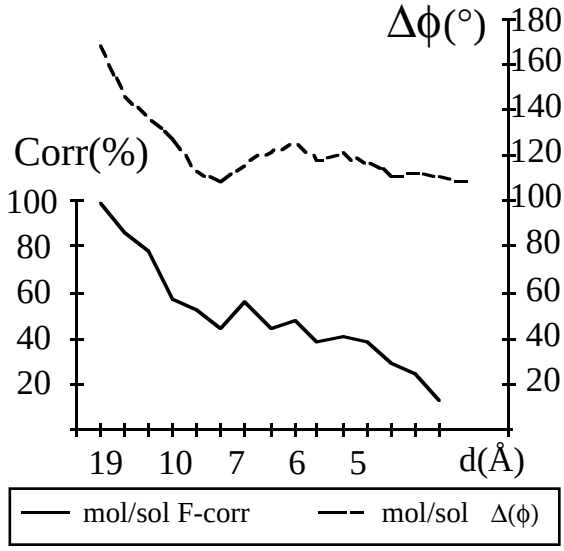


**Figure 3.** Comparison of the model and solvent structure factors at different resolution: amplitude correlation and phase difference.

The last observation means that at resolutions of ≈15Å or lower the solvent structure factors become proportional and collinear with the model ones, and their sum $\mathbf{F}_{obs}(\mathbf{s})$ is also proportional to the model structure factors. Therefore, the low resolution range can be further divided into two approximate zones:

a) ≈8Å < d < ≈15Å. The contribution of the bulk solvent is strong and independent of the atomic model, so specific solvent modelisation is crucial when working at this resolution (note that this does not concern the "experimental" phasing methods like MIR); such modelisations have been attempted, e.g., by Phillips (1980), Cheng & Schoenborn (1988,1990) and Badger & Casper (1991);

b) ≈15Å < d. The solvent structure factors are very strong and collinear to the model ones; therefore the calculated structure factor amplitudes without solvent modelisation can again be directly compared with the observed values.

This last observation implies that both model and solvent structure factors are essentially the transform of flat envelopes. This is important for structure solution methods, e.g. molecular replacement with EM envelopes (Urzhumtsev & Podjarny, 1995) or *ab initio* phasing methods like the Few Atoms Model one (Lunin et al., 1995).

## 3. Envelope-based density modelisation

The previous section shows how to estimate the phase $\phi_{exact}$, which includes the contribution of both the solute and the solvent. The corresponding electron density maps were calculated in the resolution range 6-20Å. The statistical information contained in the density histograms can be used to estimate the image correction when only a binary envelope is known. The resulting density distribution automatically includes both molecular and solvent components, and therefore its "texture" was analysed both inside and outside the envelope. The density histograms $H_d(\rho)$ (d=resolution, ρ=density value; see Lunin, 1988) were expanded to two dimensions by introducing the extra variable r (distance to the border). The resulting ones $H_d(\rho,r)$ measure the number of points with a given value of ρ and r, and are calculated at different resolutions d. The analysis consists in:

a) estimating the expected shape of the histograms;

b) modifying the density to fit the histograms.

To estimate the expected shape of the histograms, electron density maps $\rho_{obs,d}$ were calculated with coefficients ($F_{obs}$, $\phi_{exact}$), where the phases $\phi_{exact}$ are calculated as shown in Fig. 2. The envelope was obtained by keeping all points higher than a given threshold. The mean value of the histograms $H_d(\rho,r)$ (for examples, see Fig. 4), varies monotonically with the distance. The sharpness of the distribution $H_d(\rho,r_0)$ for a fixed distance $r_0$ decreases with increasing d from 20 to 6Å.

Then the possible solutions to (b) were studied. Suppose that a set of molecular envelopes at different resolutions is known, as well as the corresponding $H_d(\rho,r)$ histograms. How precisely the electron density and the corresponding structure factors can be retrieved from this information?

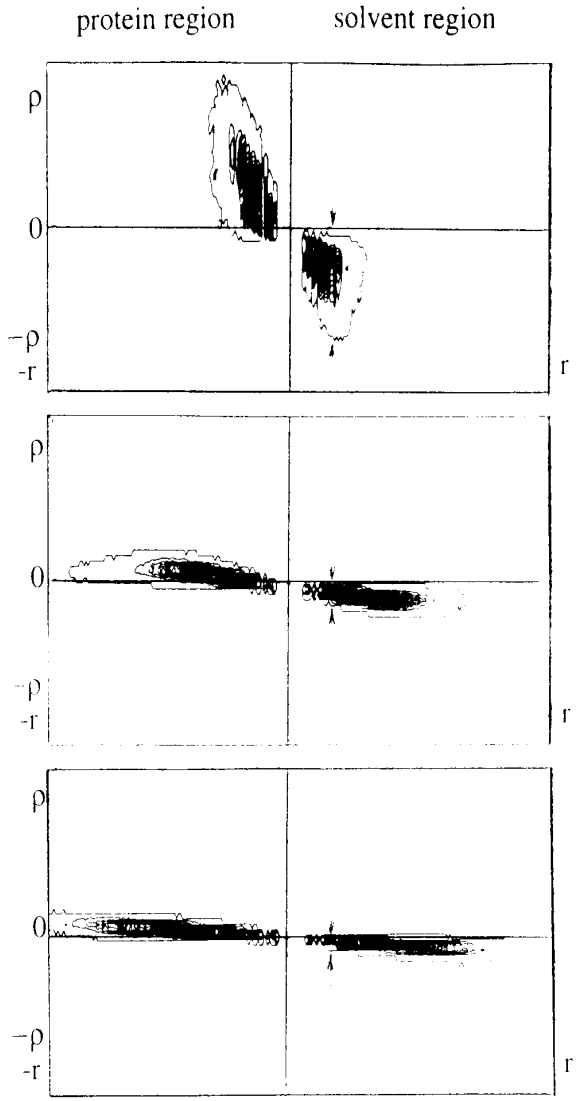


**Figure 4.** 2D-histograms $H_d(\rho,r)$ of density distribution ρ (from down to top) depending on the distance r to the envelope (from left to right) calculated at different resolution d: 6, 15 and 20Å. Density values are calculated in the same scale. The sharpness of the $H_d(\rho,r)$ for a fixed distance is indicated by arrows.

In the simplest case, when one envelope and the corresponding histogram $H_d(\rho,r)$ are known, a grid point which is at the distance $r_0$ from the envelope can be assigned the density value $\langle H_d(\rho,r_0)\rangle$ (averaging over ρ). Tests showed that at very-low resolutions (d>20Å) the replacement of a flat density envelope with such a distance-dependent distribution does not change the quality of the retrieved structure factors, but at middle resolution (e.g., 6Å) improves it (Fig. 5). This effect increases when data at several resolutions are used together

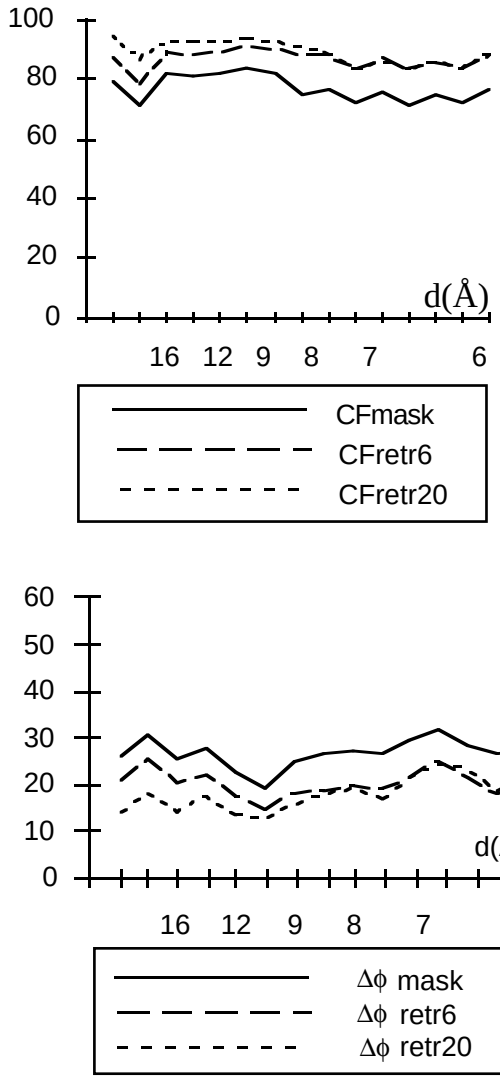


**Figure 5.** The case of the exact envelope. Amplitude correlation (CF,%) and the phase difference (Δϕ,°) in comparison with the exact values. $F_{mask}$ are structure factors calculated from the flat envelope of 6Å resolution; $F_{retr6}$ are ones calculated from the map retrieved using 6Å-resolution 2D-histogram; $F_{retr20}$ are ones calculated from the map retrieved using series of 2D-histogram at resolution from 6 to 20Å simultaneously.

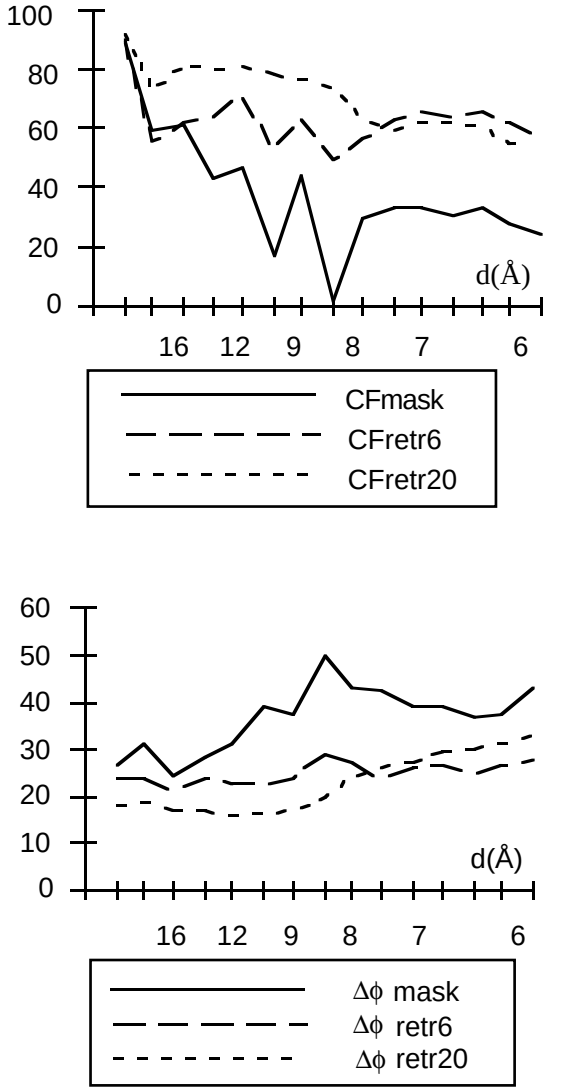


**Figure 6.** The case of approximate envelopes. Amplitude correlation (CF,%) and phase difference (Δϕ,°) in comparison with the exact amplitudes. $F_{mask}$ are structure factors calculated from the flat envelope of 6Å resolution; $F_{retr6}$ are ones calculated from the map retrieved using 6Å-resolution 2D-histogram; $F_{retr20}$ are ones calculated from the map retrieved using series of 2D-histogram at resolution from 6 to 20Å simultaneously.

(e.g., to retrieve 6Å-resolution density, 6Å-, 8Å-, 11Å- and 20Å-resolution envelopes and histograms are used). The use of several envelopes together incorporates the information of resolution-dependent features, specially small pockets and protuberances which are present at 6Å-resolution while they do not exist at 20Å-resolution.

To study a more realistic case, envelopes with errors were used for the retrieval procedure. Again, structure factors were retrieved both from the flat-density syntheses and from density modulated ones. Even when precise values for the comparison of exact and predicted structure factors changed, the behaviour was exactly the same. Note that in this case the improvement of structure factors became even more significant.

To refine the limits where density modulation becomes significant, similar test have been carried out at resolutions of 8 and 11Å. The former one gave results close to the

6Å-resolution case, and the latter one close to the 20Å resolution case.

## 4. Conclusions

In the first part of this paper the behaviour of the solvent contribution to the diffraction amplitudes has been analysed as a function of resolution. It has been proven that for low enough resolutions (d > 15Å) this contribution is essentially proportional and collinear to the model one, and that therefore a binary envelope is a good approximation to the electron density. The resolution range from 8 to 15Å is much more difficult to model, since the contribution of the solvent has significant features, as observed experimentally in a "phase difference" map. A histogram-based technique, which introduces a density modelisation dependent on the distance to the border of the envelope, has been shown to give reasonable results when the signal from several resolution ranges is used simultaneously.

## Acknowledgements:

We thank Dr. D.Moras for his continuous support. We thank Drs. D.Moras, F.Tête-Favier and J.-M.Rondeau for the experimental data of the aldose reductase used in this work The work was supported by the CNRS through the UPR 9004, by the CEE through the project 501168 (proposal 927012), by the Institut National de la Santé et de la Recherche Médicale, the Centre Hospitalier Universitaire Régional.

## References:

Badger,J & Caspar,D.L.D. (1991) *Proc. Natl. Acad. Sci. USA,* **88**, 622-626

Cheng,X & Schoenborn,B.P. (1990) *Acta Cryst.*, **B46,** 195- 208.

Howard,E.I., Grigera,J.R., Grigera,T.S. & Podjarny,A.D. (1995) *this issue*

Lunin,V.Yu. (1988) *Acta Cryst.*, **A44**, 144-150

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

Phillips,S.E.V. (1980) J.Mol.Biol., **142**, 531-554

Rondeau,J.-M., Tête,F., Podjarny,A.D., Reymann,J.M., Barth,P., Biellman,J.F. & Moras,D (1992) *Nature*, **355**, 469-472

Schoenborn,B.P. (1988) *J.Mol.Biol.* **201,** 741-749

Tête-Favier,F., Rondeau,J.M., Podjarny,A.D. & Moras,D (1993) *Acta Cryst.* **D49**, 246-256.

Urzhumtsev,A.G. & Podjarny,A.D. (1995) *Acta Cryst.*, **D51**, in press