electronic reprint

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the distribution of the bulk-solvent correction parameters

Nicholas M. Glykos and Michael Kokkinidis

Copyright © International Union of Crystallography

Author(s) of this paper may load this reprint on their own web site provided that this cover page is retained. Republication of this article or its storage in electronic databases or the like is not permitted without prior permission in writing from the IUCr.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the distribution of the bulk-solvent correction parameters

Nicholas M. Glykos[a] and Michael Kokkinidis[a,b]*

[a]IMBB, FORTH, PO Box 1527, 71110 Heraklion, Crete, Greece, and [b]Department of Biology, University of Crete, PO Box 2208, 71409 Heraklion, Crete, Greece

Correspondence e-mail: kokkinid@imbb.forth.gr

Received 21 January 2000
Accepted 3 May 2000

The distribution of the bulk-solvent correction parameters (B_{sol}, k_{sol}) (as determined with an exponential scaling algorithm based on Babinet's principle) for 219 crystal structures deposited in the Protein Data Bank is presented. The distribution shows that (i) the range of values observed is far wider than the usually cited parameter range, (ii) the observed k_{sol} values do not correlate with their assumed physical meaning and (iii) the two parameters are not independent and a reasonable agreement with the experimental data can be obtained through the application of a simple exponential function. These findings are interpreted in terms of the inability of the currently used algorithms to uncouple the values of the two parameters during macromolecular refinement.

1. Introduction

The application of a bulk-solvent correction during macromolecular refinement has become a standard procedure in recent years. Several solvent models of increasing complexity have been described in the literature (see reviews by Jiang & Brünger, 1994; Badger, 1997) and some of these models are now widely available through their implementation in several popular refinement programs [for example, *X-PLOR* (Brünger, 1992), *REFMAC* (Murshudov *et al.*, 1996), *BUSTER-TNT* (Bricogne & Irwin, 1996), *TNT* (Tronrud, 1997), *SHELXL* (Sheldrick & Schneider, 1997) and *CNS* (Brunger *et al.*, 1998)].

The simplest bulk-solvent correction algorithm (hereafter referred to as the exponential scaling model) is based on the assumption of a uniform electron-density distribution for both the macromolecular and solvent components. This leads – through the application of Babinet's principle – to a simple expression for the corrected structure-factor amplitude,

$$F = F_P\left(1.0 - k_{sol}\exp\{-B_{sol}[\sin(\theta)/\lambda]^2\}\right),$$

where F is the corrected structure-factor amplitude, F_P is the amplitude of the protein component alone, $\sin(\theta)/\lambda$ is the reciprocal resolution, k_{sol} is the ratio of the mean electron densities of the solvent and macromolecule, and B_{sol} is a measure of the diffuseness (or sharpness) of the boundary between the two components (Moews & Kretsinger, 1975; Tronrud, 1997). This physical interpretation of the meaning of k_{sol} and B_{sol} is only valid when the initial assumption is satisfied; that is, when the electron-density distribution for both the macromolecular and solvent components is uniform. Because this can only be true at very low resolution, it is common practice not to fix their values but instead to allow k_{sol} and B_{sol} to enter the refinement as two independent (adjustable) parameters whose values determine the contribution from the bulk solvent (see Tronrud, 1997 for a discussion of the refinement procedure). Although this is one of the least accurate of the solvent models presently available (see Jiang & Brünger, 1994; Kostrewa, 1997 for a compar-

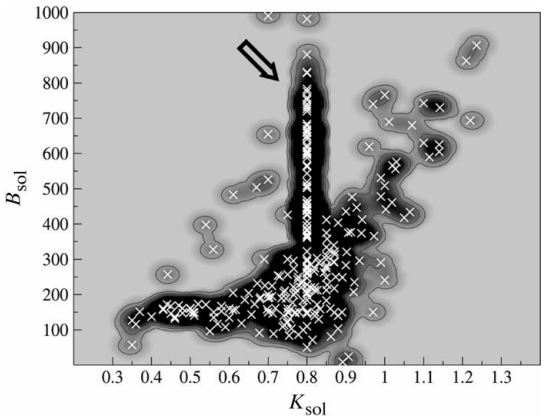

Figure 1
Distribution of the bulk-solvent correction parameters for 301 structures deposited in the PDB. Each cross represents the parameters of one structure. The underlying contour greyscale representation was calculated as described in §2. All diagrams were prepared using the program *KUPLOT*, http://www.uni-wuerzburg.de/mineralogie/crystal/discus/kuplot.html.

© 2000 International Union of Crystallography
Printed in Denmark – all rights reserved

electronic reprint

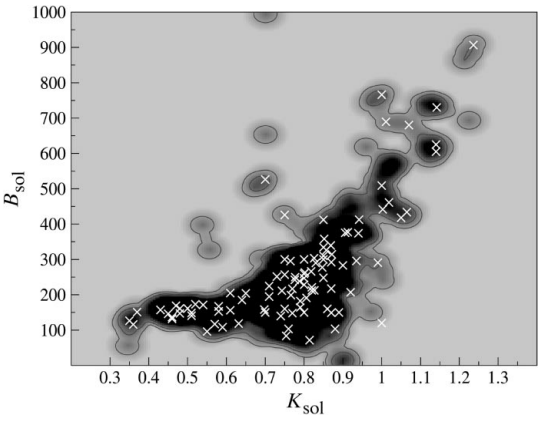

Figure 2
Distributions of the bulk-solvent correction parameters for 103 high-resolution structures (shown by crosses) and of the general sample of 219 structures (contour greyscale representation). See text for details.

ison), its computational efficacy has made it one of the most popular algorithms (used, for example, in the *TNT*, *REFMAC* and – in a modified form – *SHELXL* programs).

We determined the distribution of the exponential scaling model parameters (k_{sol}, B_{sol}) for crystal structures deposited in the Protein Data Bank (PDB; Bernstein *et al.*, 1977; http://www.rcsb.org/pdb/). Our hope was the distribution would show a tight clustering somewhere near the values usually cited in the literature [0.75–0.95 for k_{sol} and 150–350 Å^2 for B_{sol}, as given for example by Tronrud (1997), Badger (1997) or Kostrewa (1997)]. This would allow us to incorporate a rough (but computationally efficient) bulk-solvent correction algorithm into a 6*n*-dimensional molecular-replacement program that we have recently developed.[1]

2. Methods

The PDB was accessed through its mirror site at the European Bioinformatics Institute (http://pdb-browsers.ebi.ac.uk/) and was interrogated with the software provided (*3DB Browser*; http://pdb-browsers.ebi.ac.uk/pdb-bin/pdbmain). From a total of 8823 crystallographically determined structures, 765 contained any of the search strings ('BSOL' or 'KSOL') and of these 327 had a NULL entry for either of the two parameters. Of the remaining 438 structures, 134 were refined using a different solvent model and were eliminated from the subsequent analysis and 304 structures were refined with the exponential scaling model. The parameters of three structures were eliminated from the sample because the values provided in the PDB header were outside the limits set by one of the computer programs used for this analysis (see below). These entries were 9ice (B_{sol} = 8429.017), 2msp (k_{sol} = 0.050) and 8icl (B_{sol} = 1080.078).

To aid data analysis and presentation, the sample points were transformed and interpolated on a fine square grid using a modified nearest-neighbour gridding algorithm implemented with a locally written computer program. The grid covered a range of values equal to 0.2–1.4 in k_{sol} and 0–1000 Å^2 in B_{sol}. All other calculations were performed with the programs *Xmgr* (http://plasma-gate.weizmann.ac.il/Xmgr/) and *KUPLOT* (http://www.uni-wuerzburg.de/mineralogie/crystal/discus/kuplot.html).

3. Results

Fig. 1 shows the distribution of the bulk-solvent correction parameters (k_{sol}, B_{sol}) for all 301 structures that contained this information in their PDB file headers. Probably the most notable feature of this distribution is the dense line of points at k_{sol} = 0.800 (indicated by an arrow in Fig. 1). All these points correspond to structures refined with a fixed (constrained) value for the k_{sol} parameter while B_{sol} was allowed to vary freely. The great majority of these structures correspond to crystallographic determinations of closely related macromolecules (entries 1zqj, 7icl, 7icp, 8ica, 8ice, 8icg, 8ici, 8icj, 8ick, 8icn, ...) and because the parameters obtained from these structures are no longer independent – and thus do not belong to the sought (k_{sol}, B_{sol}) distribution – these data points were excluded from further consideration. The resulting distribution (after the elimination of constrained pairs) consisted of 219 data points and is shown as a contour greyscale representation in Fig. 2. Overlaid on this map is shown the distribution of all structures determined at a resolution of 2 Å or better.[2]

Two conclusions can immediately been drawn from this figure. The first is that the range of values attained by the two parameters is far wider than the usually cited interval (0.75–0.95 for k_{sol} and 150–350 Å^2 for B_{sol}; Tronrud, 1997; Badger, 1997; Kostrewa, 1997), with some of the k_{sol} values being clearly outside the limits posed by the physical interpretation of the parameter (given as the ratio of the mean electron densities of the solvent and macromolecule).

The second observation is that the two parameters are not uncorrelated; they appear to be related by a simple exponential function of the form $B_{sol} = B_0 + a\exp(bk_{sol})$. To quantify this observation, we calculated the linear correlation coefficient (C) between the observed data and the data calculated from the best (in the least-squares sense) set of parameters (B_0, a, b) for both the general sample and its high-resolution subset. For the 219 structures from the general sample, we obtained a value of C_{all} = 0.750, which increased to C_{high} = 0.824 for the high-resolution subset (with B_0 = 99.1 Å^2, a = 5.79 and b = 4.03).[3]

The unrealistically wide range of values attained by k_{sol} prompted us to examine the crystallization conditions (when given) for

[1] This we would achieve by keeping the values of k_{sol} and B_{sol} constant and equal to the centroid of the observed distribution, thus eliminating the need for repetitive cycles of parameter refinement. It is worth noting here that the absence of a bulk-solvent correction from molecular-replacement calculations represents a serious problem, especially for algorithms that attempt to simultaneously determine the rotational and translational parameters of the search model(s) through a direct optimization of a suitably chosen statistic such as the linear correlation coefficient or the R factor (as described by Chang & Lewis, 1997; Kissinger *et al.*, 1999; Glykos & Kokkinidis, 2000). This is because the absence of a bulk-solvent correction not only introduces a systematic error for all data to about 5 Å resolution, but also makes necessary a low-resolution cutoff (usually in the 15–10 Å range). This cutoff introduces series-termination errors which further complicate the target-function landscape, making the identification of the global minimum more difficult.

[2] It would appear at first sight that the high-resolution limit is irrelevant with respect to the accuracy of the bulk-solvent correction parameters. Although this is in principle correct, we believe that because high-resolution structures can fully define the contribution from ordered solvent and because disordered regions are less frequent, they result in a better definition of the bulk-solvent volume and thus to a higher accuracy of the corresponding parameters.

[3] It should be noted here that in the absence of error estimates for the values of k_{sol} and B_{sol} – which would be difficult both to define and to calculate – it is impossible to calculate a goodness-of-fit statistic. Consequently, in the absence of other information it is impossible to rule out a multitude of other models which also correlate well with the observed distribution, such as a two-line model (giving C_{all} = 0.767 and C_{high} = 0.821), a quadratic model (giving C_{all} = 0.751 and C_{high} = 0.818), a cubic equation (giving C_{all} = 0.755 and C_{high} = 0.825) *etc*. Although most of these models could possibly be excluded on the ground of making unreasonable extrapolating predictions about the curve, it is still true that the data alone is consistent with a whole family of related equations.

electronic reprint

structures lying at both the low and high end of the curve. Not unexpectedly, this examination failed to show a strong correlation between k_{sol} and the electron density of the solvent of crystallization. This was especially true for the low k_{sol} range, where proteins crystallized, for example, from $\sim$1.2 *M* phosphate buffer exhibited k_{sol} values around 0.40, whereas a protein crystallized from 1 *M* sodium malate gave a k_{sol} value of 1.07.

4. Conclusions

In summary, we have shown that the bulk-solvent parameter values obtained from the practical application of the exponential scaling model in macromolecular refinement have no consistent physical meaning and are not independent. We believe that the key to interpreting these findings rests with an understanding of why these two parameters appear to be related by a simple exponential function. In the following paragraphs, we will discuss some of the possible interpretations of this dependence.

The first and probably most elegant interpretation is that the dependence of B_{sol} on k_{sol} may be a manifestation of an underlying physical phenomenon: higher values of k_{sol} correspond to higher electron density for the solvent, which – at least in the case of macromolecular crystallography – can safely be correlated with an increase of the solvent's ionic strength.[4] This in turn leads to the formation of an ionic atmosphere which interacts strongly with the macromolecule, resulting in a reduction of the sharpness of the boundary between the two components and to a corresponding increase of B_{sol}. This line of argument introduces nothing new in the interpretation of k_{sol} and B_{sol}; its testable prediction is that the values of k_{sol} should be correlated with the ionic strength of the solvent. As discussed in the previous section, this prediction is not supported by the data.

[4] This statement rests on the assumption that high molecular-weight polyethylene glycols, which are very common precipitants, do not enter the crystals.

Another interpretation is that the observed distribution may be a consequence of a variation in the data-collection or refinement procedures between individual structures. Examination of a random sample of structures with respect to their low-resolution limit, their data-collection statistics (completeness and R_{symm}) and their final refinement statistics (R and R_{free}), revealed no systematic trend. Furthermore, as Fig. 2 shows, the distribution of parameters for the subset of high-resolution structures (determined at a resolution of 2 Å or better), closely follows the pattern of the general sample. Although some of the data points in the sample may indeed be artifacts arising from less-than-ideal data collection and refinement practices, it would appear rather unlikely that so many structures would suffer from such problems.

A third possible interpretation is that the observed dependence may arise not from the solvent structure but from the macromolecular component: if the surface of a macromolecule contains many high-mobility elements (long side chains, disordered loops *etc*.), then the boundary between the two phases becomes less sharp, leading to an increase in B_{sol}. On the other hand, these high-mobility elements contribute to the observed structure factors at low resolution but do not do so at high resolution; this, combined with the increased value of B_{sol}, leads to abnormally high values for k_{sol}. Again, it is difficult to reconcile this model with the observation that the high-resolution structures – which on the average are better defined – also show the same behaviour (see Fig. 2).

The most convincing argument (to us at least) is that the observed distribution is simply the consequence of trying to make the best of a bad model: although the assumptions of the exponential scaling model can only be satisfied at very low resolution, the common practice is to refine the parameters against all data. Additionally, because k_{sol} and B_{sol} are correlated with the overall scale and temperature factor applied to the calculated amplitudes during refinement, the normal procedure is to refine all four scaling parameters simultaneously (as described by Tronrud, 1997 and implemented in *TNT*, *REFMAC* and, in a modified form, in *SHELXL*). The result is that (at least formally) k_{sol} and B_{sol} are not, and never were, independent. They are part of a larger distribution containing all four scaling parameters; the observed exponential relation is the consequence of their dependence on the other parameters.

We believe that with the computing power presently available, there seems to be no reason for not using the more accurate (envelope-based) algorithms.

References

Badger, J. (1997). *Methods Enzymol.* **277**, 344–352.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bricogne, G. & Irwin, J. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. J. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 85–92. Warrington: Daresbury Laboratory.

Brünger, A. T. (1992). *X-PLOR Version* 3.1. *A System for X-ray Crystallography and NMR.* Connecticut, USA: Yale University Press.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.

Jiang, J.-S. & Brünger, A. T. (1994). *J. Mol. Biol.* **243**, 100–115.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Kostrewa, D. (1997). *CCP4 Newslett.* **34**, 9–22.

Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–228.

Murshudov, G. N., Dodson, E. J. & Vagin, A. A. (1996). *Proceedings of the CCP4 Study Weekend. Macromolecular Refinement*, edited by E. J. Dodson, M. Moore, A. Ralph & S. Bailey, pp. 93–104. Warrington: Daresbury Laboratory.

Sheldrick, G. M. & Schneider, T. R. (1997). *Methods Enzymol.* **277**, 319–343.

Tronrud, D. E. (1997). *Methods Enzymol.* **277**, 306–319.

electronic reprint