Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

On the application of structure-specific bulk-solvent models

Nicholas M. Glykos

Department of Molecular Biology and Genetics, Democritus University of Thrace, University Campus, 68100 Alexandroupolis, Greece

Correspondence e-mail: glykos@mbg.duth.gr

Received 14 May 2011
Accepted 10 June 2011

It is often discussed, mainly in connection with the rather high macromolecular *R* factors, that the treatment of bulk solvent in macromolecular refinement may lack the detail needed for modelling the solvent environment of molecules as complex as proteins and nucleic acids. This line of thought directly leads to the hypothesis that improvements in the modelling of the bulk solvent may substantially improve the agreement between the experimental data and the crystallographic models. Here, part of this hypothesis is being tested through the construction, *via* molecular-dynamics simulations, of a highly detailed, physics-based, structure-specific and crystallographic data-agnostic model of the bulk solvent of a known crystal structure. The water-distribution map obtained from the simulation is converted (after imposing space-group symmetry) to a constant (but scalable) partial structure factor which is then added in a re-refinement of the crystal structure. Compared with the simple Babinet-based correction, a reduction of the totally cross-validated free *R* value by 0.3% is observed. The implications and possible interpretations of these results are discussed.

1. Introduction

The application of a bulk-solvent correction is a standard procedure in all mainstream macromolecular refinement programs. The two most popular correction methods are the exponential scaling model (based on Babinet's principle; Moews & Kretsinger, 1975; Glykos & Kokkinidis, 2000) and the envelope-based methods (Badger, 1997). In their seminal contribution 17 years ago, Jiang & Brünger (1994) examined several models of increasing complexity and found no statistically significant improvement when using bulk-solvent models which included higher order hydration layers about the macromolecular component. Here, I revisit this idea by constructing a highly detailed and protein-structure-specific model for the bulk solvent using molecular-dynamics simulations. The resulting model for the bulk solvent is then converted to a constant (but scalable) partial structure factor, which is then used in a re-refinement of the crystal structure to judge (using total statistical cross-validation) its significance.

2. Methods

For our calculations we used the crystal structure of RM6, a deletion mutant of the repressor of primer protein (Papanikolau *et al.*, 2004; Glykos *et al.*, 2006; Glykos, 2007; PDB entry 1qx8). Our selection is fairly typical of the protein crystal structures contained in the PDB, with data extending to 2 Å resolution and showing a significant amount of diffuse scattering and static disorder (for a typical diffraction image recorded from RM6 crystals, see Glykos, 2007).

The computational procedure we adopted is the following.

(i) A complete model of the crystallographic unit cell (space group *C*2) was constructed using the program *VMD* (Humphrey *et al.*, 1996) and custom scripts. Two views of the resulting model of the crystal structure are shown in Fig. 1. The model included all protein atoms,

doi:10.1107/S090744491102258X

with missing side chains and H atoms added with the program *PSFGEN* from the *NAMD* distribution (Kale *et al.*, 1999). The number of TIP3 water molecules that were added to the system was adjusted in such a way as to maintain a pressure of approximately 1 atm at 298 K under the NVE conditions used for the simulation (noting in connection with this that the crystallographic data were collected at room temperature). The crystallographically determined water molecules were included in their experimentally determined positions. The final system comprised 12 491 atoms, of which 6432 were protein atoms (distributed over a total of eight polypeptide chains), 6051 were water atoms (corresponding to 216 crystallographic waters plus 1801 waters modelling the bulk solvent) and eight were ions needed to neutralize the total charge of the system.

(ii) Four independent molecular-dynamics simulations were performed, amounting to a grand total of 180 ns of simulation time. All four simulations were performed using the program *NAMD* (Kale *et al.*, 1999) in the NVE ensemble with full (PME-based) electrostatics and the CHARMM forcefield with the CMAP correction (MacKerell *et al.*, 1998, 2004). The four simulations differed in the restraints applied to the protein and crystallographic waters, which ranged from light (0.42 kJ mol^{-1} Å^{-2}) restraints applied to C$^{\alpha}$ atoms and crystallographic water O atoms to relatively strong (6.3 kJ mol^{-1} Å^{-2}) restraints applied to the backbone and waters. Two different values of the Langevin friction coefficient were also tested (1 and 10 ps^{-1}). The results from all four simulations were very similar and were internally consistent. For the discussion that follows we used the results from a simulation performed with light (0.42 kJ mol^{-1} Å^{-2}) restraints applied to both C$^{\alpha}$ atoms and crystallographic water O atoms and a value for the Langevin friction of 1 ps^{-1}.

(iii) The cumulative distribution of the O atoms of the noncrystallographic waters was calculated and converted to a *CCP*4 map using the program *CARMA* (Glykos, 2006). Because the simulation is performed in *P*1, the distribution map does not obey exact *C*2 crystallographic symmetry. This allowed us to judge the convergence of the water-distribution map by comparing the mean phase difference and *R* factor (as a function of simulation time) between symmetry-related reflections obtained by Fourier-transforming the distribution map. For all our simulations, and upon convergence, the mean phase difference between symmetry-related structure factors was of the order of 14° (corresponding to a mean figure of merit of 0.97), with an *R* factor of approximately 0.13. In the final step, the crystallographic symmetry was enforced by averaging the symmetry-related structure factors and resetting them in the correct (for *CCP*4) asymmetric unit of reciprocal space. For the centrosymmetric terms, the phases were reset to their closest symmetry-allowed value (0 or π). The result was a complete list of symmetrized partial structure factors (on an arbitrary scale) representing the bulk-solvent distribution.

(iv) The partial structure factors from the previous step were used together with the experimental data and the final model for RM6 (Glykos, 2007) in a re-refinement of the crystal structure using the program *REFMAC* (Murshudov *et al.*, 2011) from the *CCP*4 suite of programs (Winn *et al.*, 2011). To reduce memory effects, 170 cycles of minimization were performed for each of the 20 free *R*-factor sets

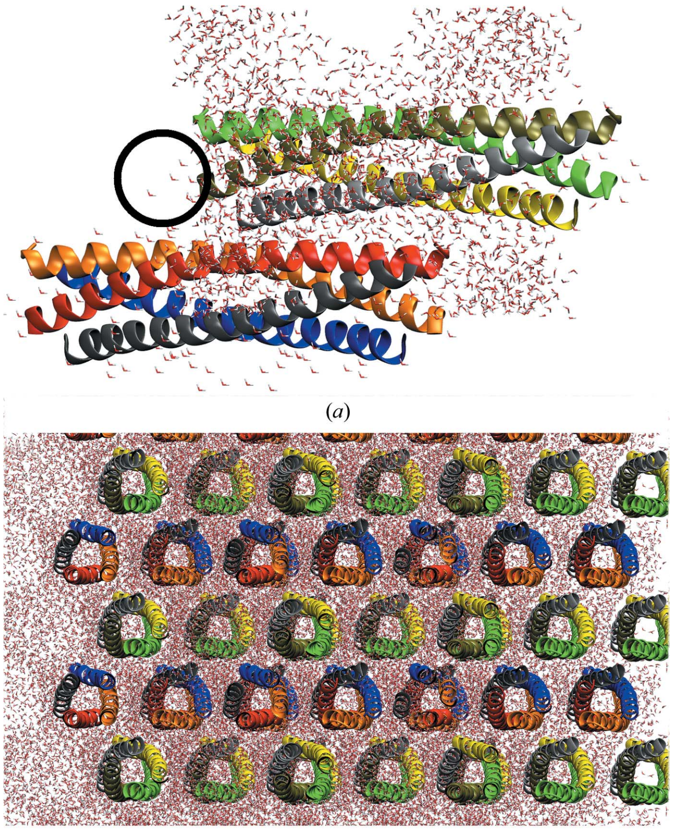

Figure 1
(*a*) shows a view (down orthogonal *z* in the Brookhaven convention) of the unit-cell contents as used for the simulation. The two α-helical bundles (per unit cell) are shown with a cartoon representation to reduce clutter. The circle shows a subset of the crystallographically determined water molecules which all share the same starting orientation. (*b*) shows a view down the bundles' major axes (approximately aligned with the [102] zone axis) covering the equivalent of nine unit cells under the simulation's periodic boundary conditions. The lighter or empty areas of both diagrams [clearly seen at the top of (*a*) and the left of (*b*)] correspond to symmetry-related helical bundles of neighbouring cells that have not been drawn to reduce clutter.

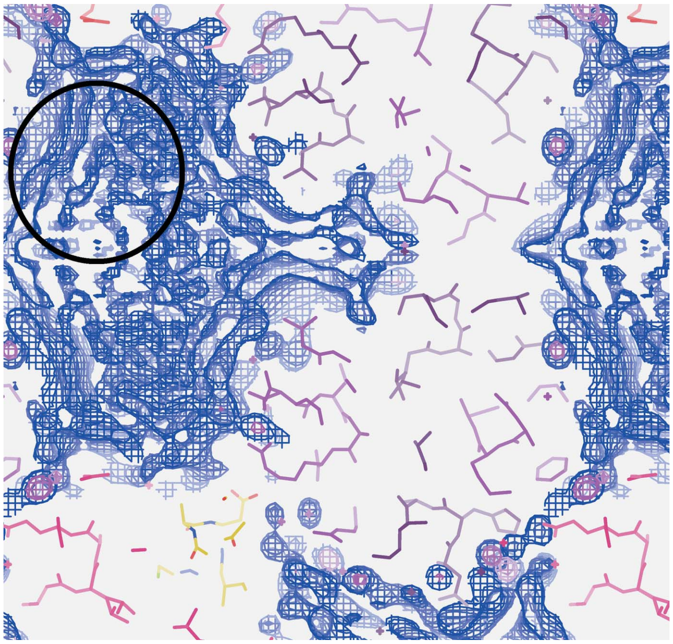

Figure 2
This is a portion of the bulk solvent's final electron-density map obtained after enforcing the *C*2 space-group symmetry on the water-molecule distribution map (see text for details). Two isosurfaces (at 1σ and 2σ above the mean) are drawn for the bulk solvent, with the protein atoms and crystallographic waters indicated as skeletal models and crosses, respectively. The circled area highlights a volume of the bulk solvent where three hydration layers are clearly visible. This figure was prepared with the program *Xfit* (McRee, 1992).

using all data to 2 Å resolution. For the purpose of this paper, we collected and analyzed the statistics for all data to 4.3 Å resolution.

3. Results

Fig. 2 shows an approximately 7 Å thick slab from the final symmetrized water-distribution map obtained from the molecular-dynamics simulation. The map shows the detail and complexity expected from a physics-based simulation, with the first, second and in some areas even the third hydration sphere clearly visible. These hydration spheres are structure-specific in the sense that their presence and density depends on the physical properties of their environment, mainly their neighbouring protein atoms. In addition to the continuous density corresponding to bulk-solvent features, well formed and higher density peaks are also present corresponding to partly ordered (from the simulation's point of view) waters that were not present in the experimentally determined structure. It should be noted here that this map is totally agnostic with respect to the crystallographic data in the sense that the experimental data were in no way used (or referred to) during the simulation.

Incorporation of this map (in the form of a fixed but scalable partial structure factor) in the refinement led to a reduction of both the *R* factor and the free *R* value by 0.4% for all low-resolution data to 4.3 Å (the corresponding improvements in five equally distributed resolution shells were 0.3, 0.3, 0.7, 0.5 and 0.4%, respectively). These numbers were obtained using the same cross-validation test set as that used for the original structure determination. As was noted by Jiang & Brünger (1994) and owing to the relatively small number of reflections entering the calculation of the free *R* value, this is one of the cases where total cross-validation is indeed necessary in order to minimize statistical uncertainty. Repeating the refinement with 20 different non-overlapping test sets, we observed an average reduction of the totally cross-validated free *R* value by 0.3% and of the *R* factor by 0.6% for all data to 4.3 Å resolution.

4. Discussion

We showed that the incorporation of a structure-specific model for the bulk solvent in the macromolecular refinement of the chosen crystal structure only marginally improved the agreement with the experimental data as judged by total statistical cross-validation. Our findings are in good agreement with the results obtained by Jiang & Brünger (1994), who used analytical functions to describe the presence of hydration spheres around the macromolecular surfaces. It would appear at first sight that these results suggest that there is little scope in trying to incorporate complex bulk-solvent models into macromolecular refinement programs. Although this may well be the case, we feel compelled to present a case for the opposing view as well by noting the following.

(i) The marginal improvement that we observed was the result of a purely physics-based data-agnostic simulation, with no adjustable parameters (during refinement) other than an overall scale and temperature factor applied to the bulk-solvent partial structure factor.

(ii) The model for bulk solvent that we used for the simulation (essentially pure water) was not representative of the actual crystallization liquor (which consisted of 900 m*M* NaCl, 45% methanol, 50 m*M* Bis-Tris buffer pH 6.2, 1 m*M* DTT and 1 m*M* EDTA; Papanikolau *et al.*, 2004). Although this very significant difference in the modelling of the solvent environment was unavoidable in order to maintain consistency with the parameterization of the forcefield used for the simulation, it does give additional credit to the physical model responsible for the marginal improvement that we did observe. It should be noted, however, that in the absence of tests with other crystal structures it is difficult to access the importance of the bulk-solvent composition in the calculations described.

(iii) No effort was made to correct for unavoidable artifacts arising from missing side chains and missing (low-occupancy) water molecules from the original structure determination.

(iv) The effects of several important simulation parameters (such as protein-related constraints and the friction coefficient, which determines water viscosity) were not thoroughly examined.

By taking these limitations into account, it is tempting to suggest that there may be some scope for the incorporation of complex physics-based bulk-solvent models in macromolecular refinement programs. Clearly, and for obvious practical reasons, this cannot be based on any form of molecular-dynamics simulations. Rather, it would have to be an analytical function (with possibly several adjustable parameters) that would depend on quantitative physical properties of the macromolecular component such as its electrostatic potential or the distribution of hydrophobic patches on its surface in a procedure rather similar to that described by Lounnas *et al.* (1994). Although it is not feasible to judge *a priori* whether such a model would be statistically useful (in terms of the everyday application of macromolecular refinement programs), little doubt remains in the author's mind that the incorporation of physics-based knowledge in macromolecular refinement can only improve the quality of the deposited crystallographic models.

References

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

Glykos, N. M. (2006). *J. Comput. Chem.* **27**, 1765–1768.

Glykos, N. M. (2007). *Acta Cryst.* D**63**, 705–713.

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 1070–1072.

Glykos, N. M., Papanikolau, Y., Vlassi, M., Kotsifaki, D., Cesareni, G. & Kokkinidis, M. (2006). *Biochemistry*, **45**, 10905–10919.

Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.

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

Kale, L., Skeel, R., Bhandarkar, M., Brunner, R., Gursoy, A., Krawetz, N., Phillips, J., Shinozaki, A., Varadarajan, K. & Schulten, K. (1999). *J. Comput. Phys.* **151**, 283–312.

Lounnas, V., Pettitt, B. M. & Phillips, G. N. (1994). *Biophys. J.* **66**, 601–614.

MacKerell, A. D. *et al.* (1998). *J. Phys. Chem. Ser. B*, **102**, 3586–3616.

Mackerell, A. D., Feig, M. & Brooks, C. L. (2004). *J. Comput. Chem.* **25**, 1400–1415.

McRee, D. E. (1992). *J. Mol. Graph.* **10**, 44–46.

Moews, P. C. & Kretsinger, R. H. (1975). *J. Mol. Biol.* **91**, 201–225.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.

Papanikolau, Y., Kotsifaki, D., Fadouloglou, V. E., Gazi, A. D., Glykos, N. M., Cesareni, G. & Kokkinidis, M. (2004). *Acta Cryst.* D**60**, 1334–1337.

Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.