Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Reintroducing electrostatics into protein X-ray structure refinement: bulk solvent treated as a dielectric continuum

Luc Moulinier,[a] David A. Case[b] and Thomas Simonson[a,c]*

[a]Département de Biologie et Génomique Structurales, Institut de Génétique et Biologie Moléculaire et Cellulaire (CNRS), 1 Rue Laurent Fries, 67404 Illkirch-Strasbourg, France, [b]Department of Molecular Biology, Scripps Research Institute, La Jolla, CA, USA, and [c]Laboratoire de Biochimie (UMR7654 du CNRS), Département de Biologie, Ecole Polytechnique, 91128 Palaiseau, France

Correspondence e-mail:
thomas.simonson@polytechnique.fr

Received 9 May 2003
Accepted 15 August 2003

This article is dedicated to the memory of Alec Hodel.

Structural refinement of proteins involves the minimization of a target function that combines X-ray data with a set of restraints enforcing stereochemistry and packing. Electrostatic interactions are not ordinarily included in the target function, partly because they cannot be calculated reliably without a description of dielectric screening by solvent in the crystal. With the recent development of accurate implicit solvent models to describe this screening, the question arises as to whether a more detailed target function including electrostatic and solvation terms can yield more accurate structures or somewhat different structures of equivalent accuracy. The Generalized Born (GB) model is one such model that describes the solvent as a dielectric continuum, taking into account its heterogeneous distribution within the crystal. It is used here for X-ray refinements of three protein structures with experimental diffraction data to 2.4, 2.9 and 3.2 Å, respectively. In each case, a higher resolution structure is available for comparison. The new target function includes stereochemical restraints, van der Waals, Coulomb and solvation interactions, along with the usual X-ray pseudo-energy term, which employs the likelihood estimator of Pannu and Read. Multiple simulated-annealing refinements were performed in torsion-angle space with a conventional target function and the new GB target function, yielding ensembles of refined structures. The new target function yields structures of similar accuracy, as measured by the free *R* factor, map/model correlations and deviations from the high-resolution structures. About 10% of side-chain conformations differ between the two sets of refinements, in the sense that the two ensembles of conformations do not completely overlap. Over 75% of the differences correspond to surface side chains. For one of the proteins, the GB set has a greater dispersion, indicating that for this case the conventional target function overestimates the true precision. As GB parameterization continues to improve, we expect that this approach will become increasingly useful.

1. Introduction

In recent years, methods to refine protein models against crystallographic data have improved in several respects. Improved likelihood estimators have been constructed as target functions (Bricogne, 1993, 1997; Pannu & Read, 1996; Adams *et al.*, 1997; Murshudov *et al.*, 1997), cross-validation is routinely used to avoid overfitting (Brünger, 1992*a*; Kleywegt & Brünger, 1996); torsion-angle dynamics limit the number of model parameters to be fitted (Rice & Brünger, 1994) and increased computer power makes it possible to perform multiple simulated-annealing refinements (Rice & Brünger, 1998); for a review, see Brunger & Adams (2002). The 'chemical' portion of the target function has also evolved. Early energy functions were borrowed directly from the molecular-mechanics and molecular-dynamics community (Brünger *et al.*, 1987; Brooks *et al.*, 1983; Gros *et al.*, 1990). They included 6–12 van der Waals interactions, non-aliphatic H atoms and, in some cases, Coulombic electrostatic interactions between protein atoms. It quickly became apparent that protein–protein electrostatic interactions could not be correctly modelled in the absence of solvent and were better left out. In that case, H atoms can be left out as well, since they typically do not carry van der Waals interaction terms and their X-ray scattering is normally neglected. The attractive dispersion part of the van der Waals term can also be left out (Hendrickson, 1985) to avoid artificial over-packing of the protein in the absence of explicit solvent. With these assumptions, the chemical part of the target function is only used to maintain correct stereochemistry and to avoid atomic overlap. It makes no attempt to provide (possibly biased) information on the detailed molecular interactions and the resulting structural features. However, these chemical target functions are not as completely unbiased as they may appear. Indeed, for charged (and possibly for polar) side chains, to score conformations solely by stereochemistry and sterics is to over-weight conformations where the side chain packs against the protein and under-weight conformations where it extends into solvent.

In the meantime, the energy functions of the molecular-mechanics and molecular-dynamics community have continued to improve in several ways, some of which may already be relevant to structure refinement (Schiffer & Hermans, 2003). In particular, an efficient and reasonably accurate treatment of electrostatic interactions with solvent has become possible through various implicit solvent models (Roux & Simonson, 1999). One of the most successful is the 'Generalized Born' (GB) model (Still *et al.*, 1990; Hawkins *et al.*, 1995; Schaefer & Karplus, 1996; Qiu *et al.*, 1997; Bashford & Case, 2000; Simonson, 2001). It describes the solvent around the biomolecule as a dielectric continuum. However, the numerical complexities of the inhomogeneous solute/solvent dielectric system are effectively swept away and replaced by approximate efficient analytical formulas. The model allows the computation of the electrostatic interactions between a macromolecule and its surrounding solvent without explicitly including individual solvent molecules in the calculation. The accuracy of this model is surprisingly good and continues to improve as variations and better parameters are introduced (Schaefer *et al.*, 1998; Ghosh *et al.*, 1998; Dominy & Brooks, 1999; David *et al.*, 2000; Onufriev *et al.*, 2000; Calimet *et al.*, 2001; Lee *et al.*, 2002). It provides an accuracy for structures and thermodynamics that can already approach that of explicit solvent simulations in favourable cases (Lee *et al.*, 2002). This raises the question whether a more detailed chemical target function, once again including van der Waals and electrostatic interactions but now also including an accurate implicit solvent model, may lead to improved X-ray structures.

A target function including electrostatics and a GB implicit solvent was used very recently for the refinement of a protein structure against NMR data (Xia *et al.*, 2002). In the X-ray context, it would be expected to have several effects. In regions at the protein surface where the electron-density map is poorly defined, backbone and side-chain positions would presumably be more accurately modelled with such a target function than with merely the 'null' hypothesis of good stereochemistry and sterics. In some positions where electrostatic interactions are particularly strong, we expect that alternate conformations will be obtained with the electrostatic model, without necessarily reducing the level of agreement with the diffraction data. Finally, multiple refinements with electrostatics and implicit solvent could lead to a greater or lesser dispersion between models, *i.e.* to a different picture of disorder.

To test these hypotheses, the GB model was implemented for systems with crystal symmetry[1] in the *CNS*, *X-PLOR* and *NIH-XPLOR* programs (Brünger *et al.*, 1998; Brünger, 1992*b*; Schweiters *et al.*, 2003). We have used the model to refine three protein structures taken from the PDB, with experimental diffraction data at medium to poor resolution: aspartyl-tRNA synthetase, with experimental data to 2.4 Å resolution (Schmitt *et al.*, 1998), an MHC-I molecule, with data to 3.2 Å resolution (Menssen *et al.*, 1999), and formylase, with data to 2.9 Å resolution (Schmitt *et al.*, 1996). Multiple refinements were performed with and without electrostatics and GB solvent. Differences between the two ensembles of structures are analyzed. In addition, for each protein tested, a structure of the same protein has been solved at higher resolution (1.9, 2.0 and 2.0 Å, respectively) which can be used as a benchmark structure (with certain limitations, discussed below). Simulated-annealing (SA) refinements were performed using torsion-angle dynamics and a maximum-likelihood crystallographic target function (Pannu & Read, 1996). The GB refinements give comparable or very slightly improved agreement with the experimental data, as measured by the free *R* factor, map correlations and deviations from the higher resolution reference structures. Despite this, they exhibit alternate positions for some side chains and, in one case, somewhat greater structural variations between SA runs.

This paper is organized as follows. In §2, we recall the basics of the GB model and derive the relevant equations for systems with crystal symmetry. §3 describes the systems studied and the computational methods. §4 describes the results. The last section is a discussion.

2. Theory

2.1. GB forces in the absence of symmetry

The electrostatic interaction between two charges *i* and *j* includes both a direct Coulomb term and a contribution from the solvent, polarized by the solute charges. Treating the

[1] The GB code was written by one of us (TS), with contributions from François Wagner (IGBMC) and David Case. It is available in the *NIH-XPLOR* program distributed by M. Clore (National Institutes of Health, Bethesda, MD, USA) or on request from TS (thomas.simonson@polytechnique.fr).

solvent as a linear homogeneous dielectric medium, the total electrostatic energy has the form

$$E^{\mathrm{elec}} = \frac{1}{2}\sum_{i\neq j}\frac{q_i q_j}{r_{ij}} + \frac{1}{2}\sum_{ij} g_{ij}, \tag{1}$$

where the sums are over all pairs of protein charges. The term g_{ij} in the second sum represents the interaction between a protein charge q_i and the solvent polarization induced by another charge, q_j. In the generalized Born model (Still *et al.*, 1990), this term is approximated by

$$g_{ij} = g(\mathbf{r}_i, \mathbf{r}_j) = \frac{\tau q_i q_j}{[r_{ij}^2 + b_i b_j \exp(-r_{ij}^2/4b_i b_j)]^{1/2}}, \tag{2}$$

where $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$, $\tau = (1/\varepsilon_w) - 1$, ε_w is the solvent dielectric constant (80 at room temperature) and b_i and b_j are the effective 'solvation radii' of the charges i and j. The interaction term g_{ij} depends explicitly on the atomic positions $\mathbf{r}_i$, $\mathbf{r}_j$ and implicitly on all the other atomic positions through the solvation radii. Indeed, the solvation radius b_i is determined by the 'self' energy E_i^{self} of charge i,

$$E_i^{\mathrm{self}} = \frac{1}{2} g_{ii} = \frac{\tau q_i^2}{2b_i}. \tag{3}$$

E_i^{self} is the interaction energy between q_i and the polarization it creates in the solvent. In practice, b_i is roughly equal to the shortest distance between q_i and the protein surface. In the GB model, it is approximated by a simple analytical function of the positions of all the solute atoms (including those that have a zero partial charge): $b_i = b_i(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$. Different GB variants use different functional forms; see below. In most variants, including those considered here, the self-energy takes the form of a pairwise sum over atoms,

$$E_i^{\mathrm{self}} = \sum_j E_{ij}^{\mathrm{self}}(\mathbf{r}_i, \mathbf{r}_j). \tag{4}$$

The force on atom n includes contributions from both the Coulombic and the solvation terms. Taking the gradient of g_{ij} with respect to the position of a solute atom n and using the chain rule for differentiation, we have

$$\nabla_n g_{ij} = \frac{\partial g_{ij}}{\partial r_{ij}}\nabla_n r_{ij} + \frac{\partial g_{ij}}{\partial b_i}\nabla_n b_i + \frac{\partial g_{ij}}{\partial b_j}\nabla_n b_j. \tag{5}$$

Taking into account the relation between the b_i and the self-energy terms E_{ij}^{self}, the total solvation force can be arranged to read

$$\begin{aligned}\nabla_n \frac{1}{2}\sum_{i,j} g_{ij} = \sum_{i\neq n}\bigg(&\frac{\partial g_{in}}{\partial r_{in}} + dE_n^{\mathrm{int},b}\frac{\partial b_n}{\partial \Delta E_n^{\mathrm{self}}}\frac{\partial E_{ni}^{\mathrm{self}}}{\partial r_{in}}\\ &+ dE_i^{\mathrm{int},b}\frac{\partial b_i}{\partial \Delta E_i^{\mathrm{self}}}\frac{\partial E_{in}^{\mathrm{self}}}{\partial r_{in}}\bigg)\frac{\mathbf{r}_n - \mathbf{r}_i}{r_{in}},\end{aligned} \tag{6}$$

with

$$dE_i^{\mathrm{int},b} = \sum_j \lambda_{ij}\frac{\partial g_{ij}}{\partial b_i}$$
$$\begin{cases}\lambda_{ij} = 1 & \text{if } i \neq j\\ \lambda_{ij} = 1/2 & \text{if } i = j.\end{cases} \tag{7}$$

2.2. Including crystal symmetry

The system is now assumed to have n_G symmetry elements, which are isometries of the form

$$S : \mathbf{r} \rightarrow \mathbf{R}\,\mathbf{r} + \boldsymbol{\rho}. \tag{8}$$

$\mathbf{R}$ is a rotation or an inversion with respect to a plane or a point and $\boldsymbol{\rho}$ is a translation vector. The total solvation energy now involves a sum over symmetry images; the solvation energy E per asymetric unit is

$$E = \frac{1}{2n_G}\sum_{iS}\sum_{jS'} g(S\mathbf{r}_i, S'\mathbf{r}_j) = \frac{1}{2}\sum_{ijS} g(\mathbf{r}_i, S\mathbf{r}_j), \tag{9}$$

where n_G is the order of the symmetry group (which is infinite for an infinite crystal). In practice, the infinite summation over all crystal translations can be truncated with a minimum image convention (Allen *et al.*, 1991), since the total electrostatic interaction energy (Coulomb plus solvation) is rather short-ranged, in contrast to the Coulomb energy alone.

To obtain the solvation forces, we use the relations

$$\begin{aligned}\nabla_n g(\mathbf{r}_n, S\mathbf{r}_j) &= g'(\mathbf{r}_n, S\mathbf{r}_j)\frac{\mathbf{r}_n - S\mathbf{r}_j}{|\mathbf{r}_n - S\mathbf{r}_j|},\\ \nabla_n g(\mathbf{r}_i, S\mathbf{r}_n) &= R^{-1}g'(\mathbf{r}_i, S\mathbf{r}_n)\frac{\mathbf{r}_i - S\mathbf{r}_n}{|\mathbf{r}_i - S\mathbf{r}_n|},\\ \nabla_n g(\mathbf{r}_n, S\mathbf{r}_n) &= 2g'(\mathbf{r}_n, S\mathbf{r}_n)\frac{\mathbf{r}_n - S\mathbf{r}_n}{|\mathbf{r}_n - S\mathbf{r}_n|}.\end{aligned}$$

Here, $g'(\mathbf{r}_i, \mathbf{r}_j)$ represents differentiation of $g_{ij} = g(\mathbf{r}_i, \mathbf{r}_j)$ considered as a function of the scalar variable $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$. The gradient of the solvation energy takes the form

$$\begin{aligned}\nabla_i E = &\sum_{i\leq j,S}\lambda_{ij} g'(r_{iJ})\frac{\mathbf{r}_i - \mathbf{r}_J}{r_{iJ}}\\ &+ \sum_{j\leq i,S} R^{-1}\lambda_{ij} g'(r_{Ij})\frac{\mathbf{r}_I - \mathbf{r}_j}{r_{Ij}}.\end{aligned} \tag{10}$$

The indices I, J correspond to the images of the particles i, j under S.

The energy and forces can be accumulated by summing over the interacting pairs (i, j) where $i \leq j$ (Verlet, 1967). While processing the (i, j) term, we do two things: (i) we accumulate the contribution of j to the force on i ('direct' contribution) and (ii) we calculate and set aside $G_{ij} = \lambda_{ij} R^{-1} g'(r_{iJ})[(\mathbf{r}_i - \mathbf{r}_J)/r_{iJ}]$, which represents the contribution of i to the force on j ('scatter' contribution). In the vectorized code of *CNS* or *X-PLOR*, once the loop over all j is finished, the G_{ij} are 'scattered' (Allen *et al.*, 1991) or added to the appropriate atomic forces, F_j.

Table 1
Structure-determination conditions for the proteins studied.

For each protein, the low-resolution data is used for refinement; the high-resolution structure is used for comparison.

Protein	AspRS		MHC-I		Formylase	
Resolution (Å)	2.4	1.95	3.2	2.0	2.9	2.0
PDB code	—	—	1a9b	1a1n	—	1fmt
Temperature (K)	120	120	100	100	193	193
Space group	$P2_12_12$		$P2_12_12_1$	$P2_12_12_1$	$P3_221$	
Unit-cell parameters (Å)	a = 124.1, b = 125.1, c = 87.3		a = 45.7, b = 116.4, c = 169.3	a = 50.0, b = 80.5, c = 105.9	a = 152.6, b = 152.6, c = 82.8	
Solvent content (%)	59		46	49	72	

2.3. GB/ACE self-energy term

The self-energy and the associated forces depend on the GB variant. By partitioning the solute into atomic volumes (following Lee and Richards; for example, Lee & Richards, 1971), one can express the self-energy $E_i^{\rm self}$ as a sum over all the solute atoms (Schaefer & Karplus, 1996; Hawkins *et al.*, 1995),

$$E_i^{\rm self} = \frac{\tau q_i^2}{2R_i} + \sum_{k \neq i} E_{ik}^{\rm self}, \qquad (11)$$

where R_i is a constant atomic radius to be determined (close to the van der Waals radius) and $E_{ik}^{\rm self}$ is related to the integral of the electrostatic energy over the volume of atom k. Notice that the charges of the other atoms, q_k, do not appear here. The effect of these atoms is merely to exclude solvent from the vicinity of atom i (Schaefer & Froemmel, 1990).

The volume integral $E_{ik}^{\rm self}$ is approximated in two steps. The first step is to approximate the electric field by the 'Coulombic field' of charge i (Schaefer & Froemmel, 1990; Calef & Wolynes, 1983; Sklenar *et al.*, 1990). This is simply the unscreened field that would exist if q_i were in a vacuum; it radiates uniformly in all directions and falls off as $1/r^2$ with distance; the corresponding energy density is $1/r^4$. The next step is to calculate the integral of $1/r^4$ over the volume of atom k. The different GB variants do this in different ways. In GB/ACE, Schaefer and Karplus assume the density of each solute atom is a Gaussian centred at the atom's position. The integral $E_{ik}^{\rm self}$ then has a tractable form, which can be approximated by interpolating between a Gaussian form at short ranges and a $1/r^4$ form at long range, leading to the Ansatz (Schaefer & Karplus, 1996)

$$E_{ik}^{\rm self} = \frac{1}{\omega_{ik}} \exp(-r_{ik}^2/\sigma_{ik}^2) + \frac{V_k}{8\pi} \left(\frac{r_{ik}^3}{r_{ik}^4 + \mu_{ik}^4} \right)^4. \qquad (12)$$

Here, ω_{ik} and μ_{ik} are simple functions of the atomic volume V_k, the atomic radii R_i, R_k [= $(3V_k/4\pi)^{1/3}$] and an adjustable 'smoothing' parameter α which determines the width of the atomic Gaussian distributions. The atomic charges are taken directly from the existing force field. The adjustable parameters of the model are then the volumes V_k and the smoothing parameter α. Ionic strength is not included, although methods to do so have been proposed (Onufriev *et al.*, 2000; Srinivasan *et al.*, 1999). Volumes V_k can be either calculated using Voronoi polyhedra [using an external program (Lee & Richards, 1971) and reading them into *CNS* or *X-PLOR*] or assigned values from existing libraries (Schaefer & Karplus, 1996; Onufriev *et al.*, 2000; Schaefer *et al.*, 2001). Note that the V_k are considered to be constants independent of the solute conformation. This is essential to obtain tractable expressions for the GB forces (above). Although the gradients of the atomic volumes have recently been derived (Edelsbrunner & Koehl, 2003), including them would complicate the formalism considerably and is probably not justified for the present application, where volume fluctuations are small.

With these approximations, $E_i^{\rm self}$ can sometimes become positive, so that the (necessarily positive) solvation radius can no longer be defined by (3). Therefore, we use a definition proposed by Schaefer *et al.* (1998),

$$b_i = \begin{cases} (\tau q_i^2/2E_i^{\rm self}) & \text{if } E_i^{\rm self} \leq E_{\rm min} = (\tau q_i^2/2b_{\rm max}) \\ b_{\rm max}[2 - (E_i^{\rm self}/E_{\rm min})] & \text{if } E_i^{\rm self} \geq E_{\rm min} \end{cases}. \qquad (13)$$

Here, $b_{\rm max}$ is an upper limit for the solvation radius, which can be set to the largest linear dimension of the solute, for example. This definition leads to continuous energies and forces.

A different model for the self-energy, proposed by Hawkins *et al.* (1995), has also been implemented in *CNS* and *X-PLOR*;[2] see the code documentation (available from TS) for details and Tsui & Case (2001) for a review of applications of this model to free protein and nucleic acid simulations.

3. Methods

3.1. The systems

Our first test system was the 'low-resolution' (2.4 Å) structure of aspartyl-tRNA synthetase from *Pyrococcus abysii* (AspRS; Schmitt *et al.*, 1998). A 'high-resolution' structure (1.9 Å) of the same protein was available for comparison (Schmitt *et al.*, 1998). The two structures were crystallized in the same space group with very similar unit-cell parameters (Table 1). AspRS is a functional dimer; the 'high-resolution' structure contains an aspartate ligand in just one of the two monomers. The other monomer has an empty active site, as in the lower resolution structure. Multiple simulated-annealing refinements were performed against the low-resolution data. The ligand-free monomer from the high-resolution structure is used as a reference to judge the quality of the refined structures. Solvent content, data-collection temperature and other experimental parameters are given in Table 1.

[2] Using code by David Case.

The second test system was an MHC-I protein complexed with a nine-residue peptide solved at 3.2 Å resolution (PDB code 1a9b; Menssen *et al.*, 1999). Solvent content was 46% of the unit-cell volume. For comparison, we used a structure of the same protein complexed with a slightly different eight-residue peptide solved at 2 Å resolution (PDB code 1a1n; Smith *et al.*, 1996). The latter 'high-resolution' structure has a unit cell that is approximately doubled along one direction (Table 1). It was solved at a slightly higher pH (6.5 compared with 5.6), with somewhat different crystallization conditions. Data for both structures were collected at 100 K. Experimental intensities were not available in the PDB for the high-resolution structure.

The third system was formylase solved at 2.0 Å resolution (PDB code 1fmt; Schmitt *et al.*, 1996). A separate 2.9 Å resolution data set (E. Schmitt & Y. Méchulam, personal communication) was used for the present refinements. Solvent content was 72% by volume.

3.2. Starting structures

For both AspRS and MHC-I, complete sets of simulated-annealing runs were performed with two different starting structures. For AspS, they were (i) the final refined 'low-resolution' structure and (ii) a model taken from an intermediate point during the original structure refinement (E. Schmitt, personal communication). All the results reported below correspond to the second case, which was considered to be more representative of the situation that would arise in the determination of a new structure. The C^{α} r.m.s. deviation between this structure and the high-resolution 1.9 Å structure is 0.6 Å.

For MHC-I, the two starting structures were (i) the previously refined low-resolution structure (1a9b; Menssen *et al.*, 1999) and (ii) the high-resolution structure, 1a1n, with the ligand side chains removed (Smith *et al.*, 1996). Only results for the second structure are reported below. The C^{α} r.m.s. deviation between this structure and the structure 1a9b, originally refined against the 2.9 Å data, is 1.2 Å.

For formylase, the PDB structure 1fmt was refined at 2.0 Å resolution (Schmitt *et al.*, 1996). An earlier data set collected at 2.9 Å resolution was also available (E. Schmitt and Y. Méchulam, personal communication). The 2.0 Å structure was used as a model in a molecular-replacement search against the 2.9 Å data; the resulting structure was used for the refinements.

3.3. Target function: chemical terms

The target function for the 'chemical' terms includes a set of stereochemical restraints, van der Waals parameters, atomic partial charges (for the runs with electrostatics) and a parameterization of the Generalized Born model (when used). Here, the stereochemical parameters were the usual Engh amd Huber set (Engh & Huber, 1991). A 6–12 van der Waals potential was used; van der Waals parameters and atomic charges were taken from the CHARMM19 force field (Brooks *et al.*, 1983). Finally, the GB/ACE solvent model was used (Schaefer & Karplus, 1996; Schaefer *et al.*, 1998), with parameters optimized earlier for protein simulations in conjunction with the CHARMM19 force field (Calimet *et al.*, 2001).

3.4. Simulated-annealing refinements

For each test system, series of 12–24 simulated-annealing refinements were run with and without the electrostatic energy terms (intra-protein plus solvation). All other conditions were the same for the two sets of runs: starting structure, annealing schedule, run length and test set of reflections used for cross-validation.

Torsion-angle molecular dynamics were used (Rice & Brünger, 1994). Results from a complete set of AspRS runs with Cartesian dynamics were similar and are not reported. The annealing temperature was 5000 K; MD segments were performed every 25 K for 0.1 ps for a total run length of 20 ps. This is eight times longer than the 'standard' run length implemented in the default *CNS* task files (Brünger *et al.*, 1998), allowing increased exploration of conformational space. The same conditions were employed in the runs with and without electrostatics. However, because the simulations without electrostatics are about eight times faster, the corresponding series included more runs: 24 runs compared with 12 with electrostatics for AspRS; 24 runs compared with 15 with electrostatics for MHC-I; 24 compared with 12 for formylase.

The crystallographic target function was the maximum-likelihood function of Pannu & Read (1996). H atoms were included in the 'chemical' portion of the target function (above); however, they did not contribute to the calculated structure factors, *i.e.* their X-ray scattering factor was zero (as usual). The scattering contribution of bulk solvent was included with the method of Jiang & Brünger (1994), which assigns a uniform adjustable density and *B* factor to the solvent volume.

The relative weights of the crystallographic and chemical portions of the target function were determined by standard methods. Initial weights were chosen to balance the average crystallographic and chemical forces. These were then refined manually to minimize the free *R* factor from short SA runs. The initial and refined weights differed by a factor of only two and gave very similar results.

4. Results

The quality of the refined models was measured by the free *R* factor (Brünger, 1992*a*), by correlations between electron-density maps, by r.m.s. deviations of atomic positions and torsion angles relative to the 'high-resolution' structures and by visual inspection. The refinements with and without electrostatics are referred to as the GB and NE ('no electrostatics') sets, respectively. The 'high-resolution' structures are referred to as the HR structures. We first describe results for the AspRS system; results for MHC-I and formylase are described subsequently.

Table 2
Summary of refinement results.

	AspRS		MHC-I		Formylase	
Method	NE†	GB‡	NE	GB	NE	GB
R (%)	29.0	29.2	26.1	26.7	23.2	23.6
$R_{\rm free}$ (%)	33.8	33.7	35.9	35.2	28.1	28.3
⟨Deviation from HR⟩ (Å)	0.6	0.6	0.9	0.9	0.4	0.4
$\langle\Delta\varphi\rangle$ (°)	7	7	12	11	5	5
$\langle\Delta\psi\rangle$ (°)	7	7	12	12	5	5
$\langle\Delta\chi_1\rangle$ (°)	19	18	20	21	6	7
$\langle\Delta\chi_2\rangle$ (°)	26	27	27	27	11	14
R.m.s. fluctuations (Å)	0.2	0.3	0.6	0.6	0.2	0.2
$\sigma(\varphi)$ (°)	2	2	9	8	2	2
$\sigma(\psi)$ (°)	2	2	10	8	2	2
$\sigma(\chi_1)$ (°)	7	9	19	18	6	6
$\sigma(\chi_2)$ (°)	18	24	38	37	16	18

† Standard 'non-electrostatic' target function. ‡ Generalized Born target function, including solvation and electrostatics.

4.1. AspRS results

At this resolution level (2.4 Å), the agreement with experiment is very similar with and without electrostatics, as summarized in Table 2. Without electrostatics, $R_{\rm free}$ for the top 12 SA models ranges from 33.5 to 34.0%. With electrostatics, the range is from 33.3 to 34.2%. The deviations of atomic positions relative to the high-resolution (HR) structure were averaged over the top 12 models and over backbone and C^{β} atoms. With and without electrostatics, the mean deviation is 0.6 Å. Torsion angles were treated in the same way. For the backbone (φ, ψ) angles, the mean deviations were (7, 7°) with or without electrostatics. For the side-chain angles (χ_1, χ_2), the deviations were (18–19, 26–27°) with or without electrostatics. Notice that the magnitude of the deviations from the HR structure is not unusual when comparing a completely refined structure (HR) and structures refined at a lower resolution without any explicit waters or *B*-factor optimization (GB and NE).

An electron-density map was calculated from the 2.4 Å experimental structure-factor amplitudes, with phases calculated from the 2.4 Å refined structure. A real-space correlation coefficient was calculated for each protein side chain. The average correlation was 91.8% with electrostatics and 91.9% without. Overall, at this resolution, refinement with electrostatic interactions, *including solvation*, does not affect the level of agreement with experiment. It does lead to local structural differences and to a somewhat different picture of structural disorder, as described next.

To characterize local structural differences between the GB and NE models, we identified amino acids where the ensembles of SA models from the two methods did not coincide. To count an amino acid as 'different', at least two models in either ensemble (out of 12 or 24) had to differ from the other ensemble. Side chains with very large *B* factors in the final refined structure (greater then 70 Å^2) were not considered. Out of 984 residues in two monomers, there are 82 (9%) with such conformational differences between the NE and GB ensembles. 21 of them are buried. Only half are polar side chains; the other half are hydrophobic (including the 21 buried side chains). This is a significant result which may appear surprising to some. It makes physical sense, however, since in the continuum dielectric model hydrophobic residues exclude high-dielectric solvent and therefore modify the dielectric environment of the polar residues.

Several positions with differing GB and NE ensembles are illustrated in Fig. 1. Additional examples, including three-dimensional views, will be shown below for MHC-I. For Thr49 in AspRS, for example, the NE refinements gave 24 models with χ_1 values around 70°, compared with about −70° in the HR structure. The GB refinements gave a mixture of models with either χ_1 value. Although most of the GB models agree with the NE refinements, three out of 12 agree with the HR structure. For Val55, as well as Glu69, the situation is reversed: all the NE models agree with the high-resolution structure and some of the GB models differ. For Glu75, all the GB models agree with the HR structure for χ_1, χ_2, whereas the NE ensemble includes a mixture of conformations.

Thus, while the level of agreement is similar with the two methods, the structural ensembles are noticeably different. The dispersion within the GB ensemble is also greater. For example, the r.m.s. coordinate dispersion, averaged over non-H atoms is 0.24 Å with NE and 0.28 Å with GB. The average torsional fluctuations for φ, ψ, χ_1 and χ_2 are 2, 2, 7 and 18° with NE compared with 2, 2, 9 and 24° with GB. The dispersion in side-chain positions is thus significantly greater with GB, despite the smaller number of runs. Overall, when the two ensembles are taken together the structural diversity is much greater than with NE alone. Since the GB structures reproduce the experimental data equally well, all the models must be considered plausible, so that the apparent precision of the structure is noticeably lower than would have been assumed from conventional NE refinements alone.

In a 'real' structure determination, it is the improvement in map quality after a round of simulating annealing that will guide the next stage of model building and adjustment. To illustrate the effect of the GB and NE refinements on map quality, we compared peaks in the resulting maps to the positions of water molecules in the fully refined structure (Schmitt *et al.*, 1998). The NE refinement led to 64 peaks that could be interpreted as water peaks and were each within 1 Å of a water position in the fully refined model. With GB, there were 90 such peaks, an improvement of 40%.

4.2. MHC-I results

MHC-I was refined against the 3.2 Å experimental data set, starting from the HR structure (PDB code 1a1n; Smith *et al.*, 1996). When SA runs were started from 1a9b, which had already been refined against the 3.2 Å data by the original authors (Menssen *et al.*, 1999), the final $R_{\rm free}$ values were somewhat lower, but the relative behaviour of NE and GB (not shown) was very similar to that described below.

Results are qualitatively similar to AspRS (Table 2). However, the differences beween NE and GB are more pronounced and there is a small but noticeable $R_{\rm free}$ improvement with GB, from 35.9% (NE) to 35.2% (GB), a

0.7% improvement (averaged over the five best structures with each method; for the ten best structures the improvement is 0.9%). Map correlations are 83% (GB) and 82% (NE), respectively. (The HR structure factors were not available in the PDB, so a map calculated from the HR structure was used.) Agreement for torsion angles is almost the same with the two methods: r.m.s. deviations are 11–12° for φ, ψ, 20–21° for χ_1 and 27° for χ_2. The r.m.s. deviation from the HR structure is 0.9 Å in both cases (averaged over backbone and C^β atoms).

Local structural differences were characterized as before by identifying amino acids where the GB and NE ensembles of structures did not coincide. Out of 750 residues, 84 (11%) have different ensembles with GB and NE. All but 12 of the differences correspond to surface residues. 34 correspond to hydrophobic side chains (including the 12 buried side chains). Examples are shown in Fig. 2. Side-chain positional fluctuations are 0.6 Å with both methods. Torsional fluctuations agree within 1–2° (Table 2).

Residue accessibilities to solvent were also calculated. The GB and NE results are similar: the mean difference per side chain is 2.0 Å^2, the same as the standard deviation for a given side chain within the GB or NE ensembles. Thus, the differences in accessibility are of the same magnitude as the differences between individual SA runs.

Figs. 2 and 3 illustrate typical differences between the GB and NE refinements. The distribution of torsion angles is shown in Fig. 2 for selected residues whose GB and NE ensembles are not identical. Many of the torsion angles in the figure are highly disordered because the corresponding residues are intrinsically flexible and the electron density is imprecise at this resolution level. Fig. 3 shows three-dimensional views of the same residues, along with the experimental 3.2 Å electron-density map. For Asn80 (Fig. 3*a*), the HR structure (green) and the structure refined at 3.2 Å (purple) have two different orientations of the terminal group, with the O atom pointing to the left and right, respectively. Both orientations lie within the electron-density lobe. The HR orientation is clearly incorrect, while the 3.2 Å structure makes hydrogen bonds with the ligand backbone on the left and the Arg79 side chain on the right. The GB structures all have the correct orientation, while the NE ensemble (not shown) is split between the two orientations (see Fig. 2). The HR Arg79 points to the back, out of density. In Fig. 3(*b*), we show the amino acids Glu128 and Arg111. The HR positions (green) are out of density, with Arg111 actually reaching into the density

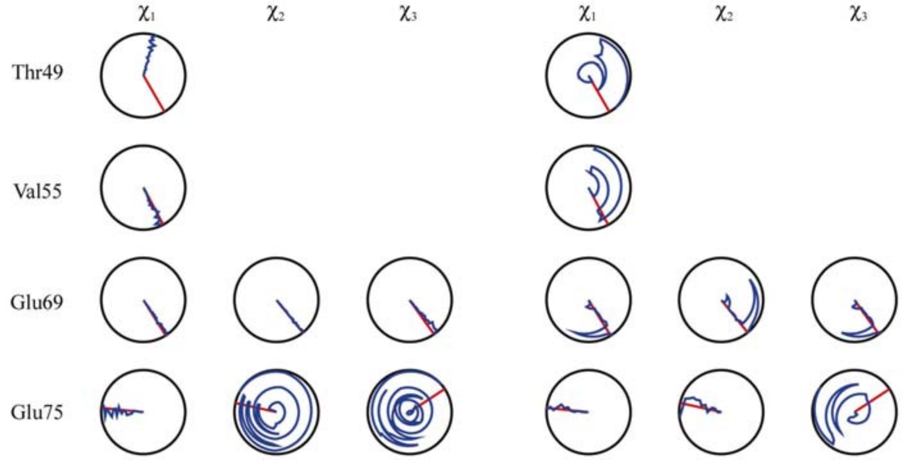

Figure 1
Dials plot of selected AspRS torsion angles in the NE (left) and GB (right) ensembles. The angular origin is on the horizontal axis to the right of each dial; the positive direction is counterclockwise. The structures are ordered by decreasing $R_{\rm free}$ from the centre to the outer edge. The straight dotted line corresponds to the angle in the HR structure. For Glu69, for example, all the NE structures (left-hand dials) agree with HR; a few of the GB structures disagree (right-hand dials).

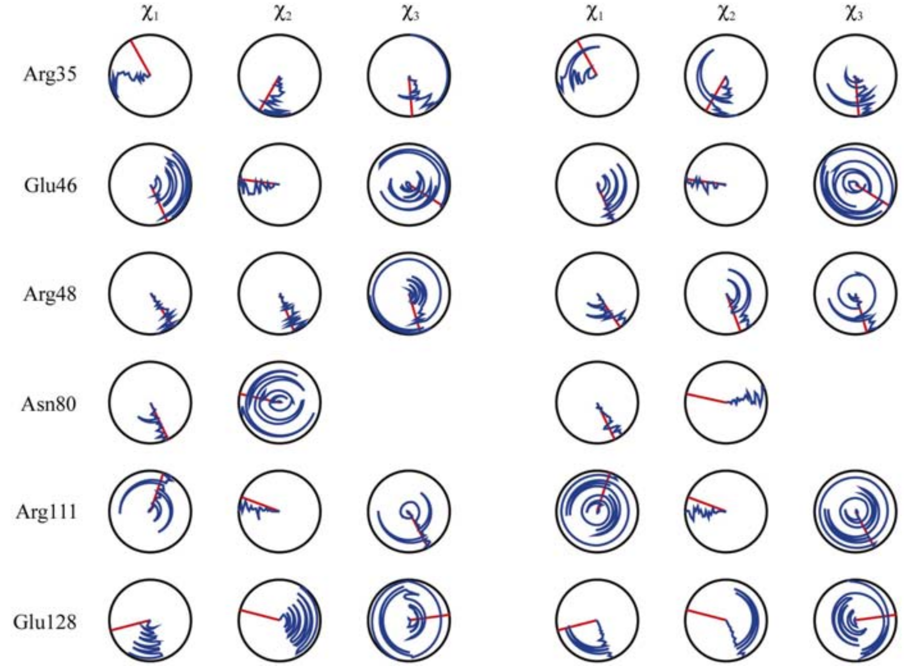

Figure 2
Dials plot of selected MHC torsion angles in the NE (left) and GB (right) ensembles. The straight dotted lines correspond to the HR structure. For Asn80, for example, all the GB structures disagree with the HR χ_2 (which is probably incorrect; see text), while the NE ensemble has a mixture of structures. Several of the torsion angles are highly disordered, reflecting both the flexibility of this region and the moderate resolution of the electron-density map.

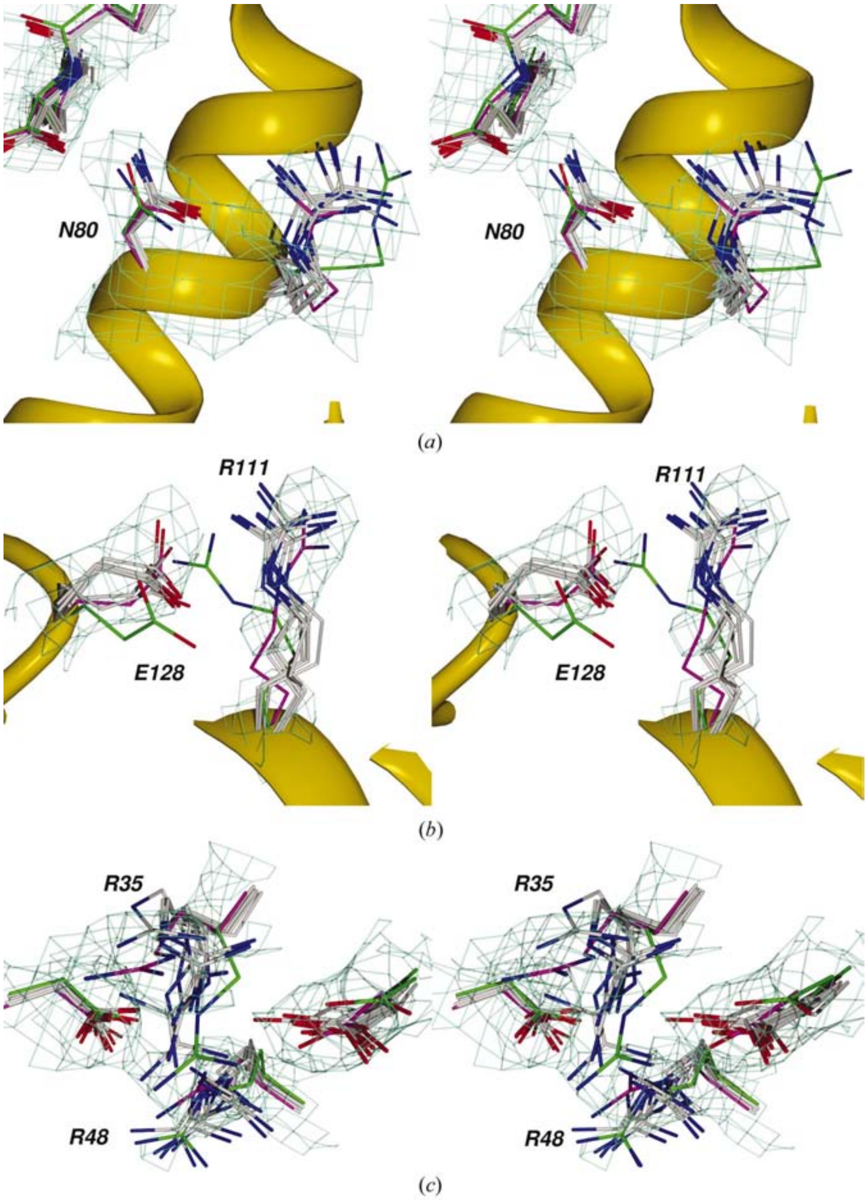

Figure 3
Stereoviews of selected MHC side chains, comparing the different refined structures. In each part, the HR structure is green, GB structures are grey and the 3.2 Å structure 1a9b is purple; the electron-density map is contoured at one standard deviation. (*a*) Asn80 interacting with Arg79 on the right and the ligand on the left. (*b*) Glu128 interacting with Arg111. (*c*) Cluster of four charged residues: Arg35*A*, Glu46*A*, Arg48*A* and Asp54*B* (*A* and *B* are chain identifiers).

corresponding to Glu128. In the 3.2 Å refined structure 1a9b (purple), the side chains are in density but are too far apart to make hydrogen bonds. In the GB structures, they have shifted slightly and reoriented so as to make two hydrogen bonds and form a salt bridge. This 'sideways' hydrogen-bonding salt bridge is very common for Arg-Glu pairs. Fig. 3(*c*) shows a cluster of four charged residues: Arg35*A*, Glu46*A*, Arg48*A* and Asp54*B* (*A* and *B* are chain identifiers). The arginines in between the carboxylates occupy two very different positions in the HR (green) and 3.2 Å (1a9b; purple) structures. In HR (green), Arg35 reaches down into the Arg48 density, while Arg48 points down and away. In 1a9b (purple), Arg35 points to the upper left of the figure. The GB structures (grey) represent a variety of intermediate conformations, some of which make hydrogen bonds to one or both carboxylates on either side.

4.3. Formylase results

For formylase, the mean $R_{\rm free}$ is very slightly higher with GB: 28.3% compared with 28.1% for NE (averaged over the 12 best models; the best GB model has $R_{\rm free}$ = 28.2 compared with 27.9% for the best NE model). The backbone and C^{β} structure are very similar in the two methods: the r.m.s. deviations from HR are 0.4 Å with both methods and the deviations of φ, ψ and χ_1 are also almost identical (5–7° in all cases). The mean χ_2 deviation is 14° with GB compared with 11° with NE. The two methods lead to a similar description of structural disorder; *e.g.* the side-chain torsion angles χ_1, χ_2 have standard deviations of 6 and 16° with NE compared with 6 and 18° with GB. 40 side chains out of 608 have somewhat different conformations with the two methods, as defined above (non-identical GB and NE ensembles). One pair of adjacent residues has different backbone conformations.

5. Conclusions

This work represents the first protein crystal structure refinement with electrostatic and solvation forces. We suggested at the outset that while this approach might not improve the agreement with the X-ray data, it would lead to alternate structures, not sampled with the usual target function, yet providing similar agreement with the data. For three proteins with 2.4–3.2 Å resolution data we see that this is indeed the case. Compared with the usual NE target function, the GB $R_{\rm free}$ values were equivalent (AspRS), slightly better (MHC-I) or only marginally worse (formylase). Agreement with the high-resolution structures is very similar with the two methods. For AspRS, the GB refinement led to greater map improvement, as measured by the appearance of density peaks corresponding to known water molecules (*i.e.* molecules seen in the high-resolution structure). The GB solvation forces are critical for such good performance: X-ray refinement with electrostatics but no solvation is known to give poor results (Brünger *et al.*, 1987).

Despite the global agreement between GB and NE, the conformations sampled with GB are different in many local regions, primarily at the surface. In all three proteins, about 7–11% of the side chains had overlapping but non-identical ensembles of conformations with GB and NE. For AspRS, the dispersion or disorder within the GB ensemble was somewhat greater than within the NE ensemble. For all three proteins, taking the GB and NE ensembles together, the overall precision is significantly lower than would be assumed from NE alone. In addition to the data reported above, additional sets of runs were performed with a different MD algorithm (Cartesian dynamics), a different annealing schedule, a different weighting of the X-ray term and different starting structures, showing that the conclusions are robust.

The Generalized Born model is a sophisticated and physically sound representation of the electrostatic interactions between protein and solvent. The GB variant used here represents an improvement over several earlier GB implementations (Calimet *et al.*, 2001; Simonson, 2001). For four small proteins, simulations without any X-ray restraints led to structures in good agreement with experiment (mean backbone deviation of less than 2 Å; T. Simonson, unpublished data). Simpler solvation treatments, such as a distance-dependent dielectric constant or a surface-area model, gave much poorer agreement (Schaefer *et al.*, 1999; Calimet *et al.*, 2001). Meanwhile, GB models continue to improve. Corrections to the Coulomb field approximation have been proposed (Lee *et al.*, 2002; Lévy, 2002), as well as additional energy terms describing hydrophobic contributions (Schaefer *et al.*, 1998; Wagner & Simonson, 1999). The most recent variants can provide mean backbone devations as small as 1 or 1.5 Å (in the absence of any X-ray restraints; Lee *et al.*, 2002). They have even been used to fold two proteins *ab initio*, starting from extended conformations, without any restraints or bias towards the experimental structure (Simmerling *et al.*, 2002; A. Onufriev, personal communication). The GB refinements are rather expensive, almost an order of magnitude slower than NE, and a complete set of GB refinements takes several days on a small cluster of PCs. However, with computer speed continuing to increase rapidly, it will soon be possible to perform the same calculations routinely on a desktop machine.

More work is obviously needed to test additional proteins in different resolution ranges and to explore in much more detail the effect of modern solvation treatments on map and model quality. We have shown that current GB models can already give equivalent agreement compared with traditional NE refinements. The next generation of GB models may be expected to give a superior description of both the mean structures and structural disorder. For current structural genomics efforts, as well as for high-throughput ligand-screening efforts, it is very important to develop the best possible 'default' protocol lending itself to automation. Generalized Born target functions should be increasingly useful in this context.

We acknowledge financial support from the Centre National de la Recherche Scientifique and the National Science Foundation (CNRS/NSF travel grant INT-0128783 to DAC and TS) and from the French Ministère de la Recherche (ACI Bioinformatique to TS), as well as helpful discussions with Emmanuelle Schmitt and Michael Schaefer. We thank E. Schmitt and Y. Méchulam for communicating unpublished data.

References

Adams, P., Pannu, N., Read, R. & Brünger, A. (1997). *Proc. Natl Acad. Sci. USA*, **94**, 5018–5023.

Allen, B. K., Allen, M. & Tildesley, D. (1991). *Computer Simulations of Liquids*. Oxford: Clarendon Press.

Bashford, D. & Case, D. (2000). *Annu. Rev. Phys. Chem.* **51**, 129–152.

Bricogne, G. (1993). *Acta Cryst.* D**49**, 37–60.

Bricogne, G. (1997). *Methods Enzymol.* **276**, 361–423.

Brooks, B., Bruccoleri, R., Olafson, B., States, D., Swaminathan, S. & Karplus, M. (1983). *J. Comput. Chem.* **4**, 187–217.

Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–475.

Brünger, A. T. (1992*b*). *X-PLOR v. 3.1. A System for X-ray Crystallography and NMR*. New Haven: Yale University Press.

Brunger, A. & Adams, P. (2002). *Acc. Chem. Res.* **35**, 404–412.

Brünger, 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., Simonson, T. & Warren, G. (1998). *Acta Cryst.* D**54**, 905–921.

Brünger, A., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Calef, D. & Wolynes, P. (1983). *J. Chem. Phys.* **87**, 3387.

Calimet, N., Schaefer, M. & Simonson, T. (2001). *Proteins*, **45**, 144–158.

David, L., Luo, R. & Gilson, M. (2000). *J. Comput. Chem.* **21**, 295–309.

Dominy, B. & Brooks, C. III (1999). *J. Phys. Chem. B*, **103**, 3765–3773.

Edelsbrunner, H. & Koehl, P. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 2203–2208.

Engh, R. & Huber, R. (1991). *Acta Cryst.* A**47**, 392–400.

Ghosh, A., Rapp, C. & Friesner, R. A. (1998). *J. Phys. Chem. B*, **102**, 10983–10990.

Gros, P., van Gunsteren, W. & Hol, W. (1990). *Science*, **249**, 1149–1152.

Hawkins, G., Cramer, C. & Truhlar, D. (1995). *Chem. Phys. Lett.* **246**, 122–129.

Hendrickson, W. (1985). *Methods Enzymol.* **115**, 252–270.

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

Kleywegt, G. & Brünger, A. (1996). *Structure*, **4**, 897–904.

Lee, B. & Richards, F. (1971). *J. Mol. Biol.* **55**, 379–400.

Lee, M., Salsbury, F. Jr & Brooks, C. III (2002). *J. Chem. Phys.* **116**, 10606–10614.

Lévy, N. (2002). PhD thesis, University of Evry, France.

Menssen, R., Orth, P., Ziegler, A. & Saenger, W. (1999). *J. Mol. Biol.* **285**, 645–653.

Murshudov, G., Vagin, A. & Dodson, E. (1997). *Acta Cryst.* D**53**, 240–255.

Onufriev, A., Bashford, D. & Case, D. (2000). *J. Phys. Chem. B*, **104**, 3712–3720.

Pannu, N. & Read, R. (1996). *Acta Cryst.* A**52**, 659–668.

Qiu, D., Shenkin, P., Hollinger, F. & Still, W. (1997). *J. Phys. Chem. A*, **101**, 3005–3014.

Rice, L. & Brünger, A. (1998). *J. Appl. Cryst.*, **31**, 798–805.

Rice, L. & Brünger, A. T. (1994). *Proteins*, **19**, 277–290.

Roux, B. & Simonson, T. (1999). *Biophys. Chem.* **78**, 1–20.

Schaefer, M., Bartels, C. & Karplus, M. (1998). *J. Mol. Biol.* **284**, 835–847.

Schaefer, M., Bartels, C. & Karplus, M. (1999). *Theor. Chem. Acc.* **101**, 194–204.

Schaefer, M., Bartels, C., Leclerc, F. & Karplus, M. (2001). *J. Comput. Chem.* **22**, 1857–1879.

Schaefer, M. & Froemmel, C. (1990). *J. Mol. Biol.* **216**, 1045–1066.

Schaefer, M. & Karplus, M. (1996). *J. Phys. Chem.* **100**, 1578–1599.

Schiffer, C. & Hermans, J. (2003). *Methods Enzymol.* In the press.

Schmitt, E., Blanquet, S. & Méchulam, Y. (1996). *EMBO J.* **15**, 4749–4758.

Schmitt, E., Moulinier, L., Fujiwara, S., Imanaka, T., Thierry, J. & Moras, D. (1998). *EMBO J.* **17**, 5227–5237.

Schweiters, C., Kuszewski, J., Tjandra, N. & Clore, G. (2003). *J. Biomol. NMR*, **160**, 65–73.

Simmerling, C., Strockbine, B. & Roitberg, A. (2002). *J. Am. Chem. Soc.* **124**, 11258–11259.

Simonson, T. (2001). *Curr. Opin. Struct. Biol.* **11**, 243–252.

Sklenar, H., Eisenhaber, F., Poncin, M. & Lavery, R. (1990). *Theoretical Biochemistry and Molecular Biophysics*, edited by D. Beveridge & R. Lavery, pp. 317–335. New York: Adenine Press.

Smith, K., Reid, S., Stuart, D., McMichael, A., Jones, E. & Bell, J. (1996). *Immunity*, **4**, 215–228.

Srinivasan, J., Trevatan, M., Beroza, P. & Case, D. (1999). *Theor. Chem. Acc.* **101**, 426–434.

Still, W. C., Tempczyk, A., Hawley, R. & Hendrickson, T. (1990). *J. Am. Chem. Soc.* **112**, 6127–6129.

Tsui, V. & Case, D. (2001). *Biopolymers*, **56**, 275–291.

Verlet, L. (1967). *Phys. Rev.* **159**, 98–103.

Wagner, F. & Simonson, T. (1999). *J. Comput. Chem.* **20**, 322–335.

Xia, B., Tsui, V., Case, D., Dyson, H. & Wright, P. (2002). *J. Biomol. NMR*, **22**, 317–331.