*Molecular Simulation*, 1993, Vol. 10(2-6), pp. 97-120
Reprints available directly from the publisher
Photocopying permitted by license only

# AN EFFECTIVE SOLVATION TERM BASED ON ATOMIC OCCUPANCIES FOR USE IN PROTEIN SIMULATIONS

PIETER F.W. STOUTEN[1,4], CORNELIUS FRÖMMEL[2], HARUKI NAKAMURA[3] and CHRIS SANDER[1]

*[1]European Molecular Biology Laboratory, Meyerhofstrasse 1, W-6900 Heidelberg, Germany, [2]Institut für Physiologische und Biologische Chemie, Humboldt-Universität, Sektion Chemie, Hessische Strasse 3-4, O-1040 Berlin, Germany, [3]Protein Engineering Research Institute, 6-2-3 Furuedai, Suita, Osaka, 565 Japan, [4]Present Address: Du Pont Merck Pharmaceutical Company, P.O. Box 80353, Wilmington, DE 19880-0353, U.S.A.*

*(Received January 1993, accepted January 1993)*

Several approaches to the treatment of solvent effects based on continuum models are reviewed and a new method based on occupied atomic volumes (occupancies) is proposed and tested. The new method describes protein-water interactions in terms of atomic solvation parameters, which represent the solvation free energy per unit of volume. These parameters were determined for six different atoms types, using experimental free energies of solvation. The method was implemented in the GROMOS and PRESTO molecular simulation program suites. Simulations with the solvation term require 20-50% more CPU time than the corresponding vacuum simulations and are approximately 20 times faster than explicit water simulations. The method and parameters were tested by carrying out 200 ps simulations of BPTI in water, *in vacuo*, and with the solvation term. The performance of the solvation term was assessed by comparing the structures and energies from the solvation simulations with the equivalent quantities derived from several BPTI crystal structures and from the explicit water and vacuum simulations. The model structures were evaluated in terms of exposed total surface, buried and exposed polar surfaces, secondary structure preservation, number of hydrogen bonds, energy contributions, and positional deviations from BPTI crystal structures. Vacuum simulations produced unrealistic structures with respect to all criteria applied. The structures resulting from the simulations with explicit water were closer to the 5PTI crystal structure, although part of the secondary structure dissolved. The simulations with the effective solvation term produce structures that are normal according to all evaluations and in most respects are remarkably similar to the 5PTI crystal structure despite considerable positional fluctuations during the simulations. The segments where the model and crystal structures differ are known to be flexible and the observed difference may be physically realistic. The effective solvation term based on occupancies is not only very efficient in terms of computer time but also results in meaningful structural properties for BPTI. It may therefore be generally useful in molecular dynamics of macromolecules.

KEY WORDS: effective solvation, atomic occupancies, protein simulations, molecular dynamics, stochastic dynamics

## 1 REVIEW OF APPROACHES TO THE TREATMENT OF SOLVENT EFFECTS

Most essential chemical and biological processes that involve proteins, such as chemical reactions and formation of structure, are governed by interactions with solvent molecules, especially water. Analysis of these processes by means of computer simulations can reveal the physical nature of the underlying mechanisms. Simulations can also be used as a powerful predictive tool in the rational design process of novel proteins with desired properties. Accurate analysis and design require realistic computer simulations that treat the protein-solvent interactions adequately.

Protein-solvent interactions can be divided into two classes: bulk and specific effects [1]. The hydrophobic effect and dielectric shielding are bulk effects, which involve the configurational rearrangement of large numbers of solvent molecules. The hydrophobic effect (whose long-range nature has been observed directly in an experiment by Israelachvili & Pashley [2]) plays a major role in the formation and stability of the compact structures of globular proteins in aqueous solution [3]. It causes most non-polar groups to come together during the folding process to form a hydrophobic core in the interior of the protein [4]. The electrostatic interaction constitutes another major influence on the formation of protein structure [5–7]. Since the rearrangement of electric dipoles of water molecules is rapid (on the protein folding time scale) and the electric force has a long-range nature, the reaction field due to the rearrangement of water molecules is quite significant. Describing these bulk effects adequately is difficult because they involve essentially multibody (non-pairwise) interactions. The specific part of protein-solvent interactions is the formation of hydrogen bonds between polar and charged groups of the protein and water molecules [8]. Treating these interactions in computer simulations is in principle simpler than dealing with bulk effects, but the local structure of water molecules and exposed protein side-chains, which change dynamically, must be known. For these reasons it is difficult to describe protein-solvent interactions adequately by a single simple model.

Protein-solvent interactions can be simulated and analysed using two distinctly different models. These models are based either on continuum or on molecular theories. Continuum models are well suited to characterize and reproduce the bulk properties of protein-solvent interactions. The hydrophobic effect has been described by assuming a linear relationship between hydration free energies and either solvent accessible surface areas [9–14] or first hydration shell volumes [15, 16]. Using empirically determined parameters, this approach has successfully reproduced hydration free energies of small organic molecules [11] and proteins [17] alike. Continuum theory has also been applied to the description of electrostatic interaction between protein and solvent. Precise electrostatic potentials around protein molecules have been calculated by numerically solving the Poisson-Boltzmann equation [18–25]. These continuum approaches have two drawbacks, however. First, they do not adequately describe specific interactions between the protein and nearby water molecules, and second, they are based on static protein structures.

Molecular models and inter-atomic forces are the basis of the popular Molecular Dynamics (MD) techniques. Molecular dynamics simulations of classical particles have become a powerful tool to analyse protein-solvent interaction at the atomic level [26]. The first ever computer simulation of a protein was carried out *in vacuo* with complete neglect of solvent except for some tightly bound "bio-water" molecules [27]. Recent molecular dynamics studies have shown that taking into account massive amounts of solvent is crucial for a proper description of the behaviour of proteins [26, 28, 29]. And, because of the relatively slow rotational relaxation of water molecules, long simulations are necessary to just reproduce bulk properties of the water, such as its dielectric constant [30]. Incorporating solvent effects correctly in molecular dynamics simulations, therefore, requires vast amounts of CPU time, most of which is spent on evaluating water-water interactions. In order to reduce CPU time electrostatic forces are often artificially truncated at short distances, but this gives rise to large errors [31]. Also, polarization effects, which are very important especially at the protein-water interface, are generally ignored. So, although molecular theory in principle can describe both the hydrophobic interaction and dielectric shielding, the present practical limitations prevent adequate simulation of bulk properties.

In order to circumvent these problems a "third" approach has been developed. It uses a hybrid of the continuum and molecular models in that the protein is described on the molecular level while a continuum description is employed to characterize the bulk properties of the protein-solvent interactions. In attempts to solve the problem of dielectric shielding, several approaches have combined the continuum model with molecular theory [32–37]. However, as King & Warshel have pointed out [33], positive feedback can occur between the total dipole moment of the space treated by the molecular theory and the reaction field force from the continuum space. This mechanism is inherently due to the system Hamiltonian and without careful treatment of the reaction field unrealistic results might occur in the course of a molecular dynamics simulation. The hydrophobic effect has also been combined with molecular models. Here, atomic hydration free energies are calculated from accessible surface areas at every step of the molecular dynamics simulation. These hydration free energies are added to the solute's conventional intra-molecular energies. This method is quite expensive, however, as it requires calculating complicated derivatives of the accessible surface area [14]. Recently, Schiffer *et al.* [38] reported a procedure based on this approach which demands about 8 times as much CPU time as *in vacuo* simulations. Another hybrid approach is followed in Stochastic Dynamics (SD) techniques [39]. Here, solvent effects are mimicked by introducing an effective friction term, which damps the velocity of exposed atoms, and a stochastic term, which represents random collisions of the solute with solvent molecules. This approach describes the behaviour of peptides in an apolar solvent such as $CCl_4$ satisfactorily since here the hydrophobic effect, dielectric shielding and specific solute-solvent interactions are of minor importance, but it does not work well with water as solvent [40].

We propose here a new variant of this "third" approach, which is an extension of the solvent contact model [16] and can easily be combined with stochastic dynamics techniques. We intend to provide a method to describe protein-solvent interactions, which is simple, very cheap, accurate enough for most practical purposes, and which reproduces structural features satisfactorily. Our approach is based on the assumed linear relationship between the solvation free energy of an atom and the percentage of the volume around that atom which is "empty" (i.e., available for water to occupy). This direct link between occupancies (occupied volumes) and solvation free energies effectively bypasses surface area calculations, and hence does not require complicated functions to be evaluated. The method can be applied for many purposes such as fast and realistic exploration of conformational space (especially where surface residues are concerned) using SD techniques, improvement of any (X-ray, *de novo*, mutant) model structure using MD, SD and energy minimization (EM), and assessment of the quality of protein models by means of static energy calculations. Our initial goal is to use the method routinely to generate plausible protein models in modelling by homology projects. Therefore, we are primarily concerned with reproducing protein structural features. Our second concern is performance. The CPU requirements of MD simulations *in vacuo* are generally tolerable, but considerable increases will not be acceptable in situations where simulations are used as a tool to improve modelled protein structures.

## 2 METHOD AND IMPLEMENTATION OF THE SOLVATION TERM

We make several assumptions:

(1) the free energy of solvation of a given protein can be approximated by a sum of individual atomic contributions,
(2) the individual solvation energy of an atom varies linearly with the degree of exposure of that atom,
(3) the atomic degree of exposure can be adequately expressed as the percentage of the volume around an atom which is not occupied,
(4) the occupied volume around an atom (occupancy) can be approximated by summing tabulated fragmental volumes of all other atoms in the protein using a suitable envelope function.

In molecular physics potential energies are basically sums over pair interactions and it is, therefore, reasonable to assume that the enthalpy of solvation can be regarded as a sum of atomic contributions. The solvent contact model [16] gives a plausible justification for the linear dependence of the solvation enthalpy on the occupancy. The entropic contributions to the free energy of solvation of a protein mainly stem from the entropy loss of the water and these effects are considerable [1, 2]. Any effective solvation term should therefore seek to produce free energies of solvation rather than enthalpies. The entropy loss of water is caused by the formation of rigid water structures around groups of atoms at the protein surface and we assume that these group contributions are additive.

One can define occupancies in various ways. Since we want to incorporate the solvation term in molecular dynamics programs, we need a differentiable, short-range function and we chose a Gaussian. The occupancy of a given atom $i$ is calculated from

$$\mathrm{Occ}(i) = \sum_{j \neq i} \mathrm{Vol}(j) \cdot \exp(-r_{ij}^2/2\sigma^2) \tag{1}$$

where Vol($j$) is the fragmental volume of atom $j$ [41], $r_{ij}$ the distance between the centres of atoms $i$ and $j$, and $\exp(-r_{ij}^2/2\sigma^2)$ the envelope function. The summation extends over all non-hydrogen atoms $j \neq i$. Since we deal with heavy atoms only we use one overall $\sigma$-value. We have taken $\sigma = 3.5$ Å because that distance corresponds roughly to the minimum of the van der Waals potential for two heavy atoms. Using a square-well envelope function or using different $\sigma$-values we obtained essentially the same results (data not shown). The atomic free energy of solvation $\Delta g_{\mathrm{solv}}(i)$ of a given atom $i$ is defined as

$$\Delta g_{\mathrm{solv}}(i) = \mathrm{SolPar}(i) \cdot (\mathrm{Occ}_{\mathrm{max}}(i) - \mathrm{Occ}(i)) \tag{2}$$

where SolPar($i$) and $\mathrm{Occ}_{\mathrm{max}}(i)$ are the atomic solvation parameter and the maximum occupancy of atom $i$. SolPar($i$) is assumed to have a constant value which only depends on the atom type and represents the solvation free energy per unit of volume. The actual occupancy Occ($i$) of atom $i$ is the only variable that depends on the environment of this atom. Note the analogy between our occupancy approach and the accessible surface area approach [14]. Substituting into (2) the occupancy as defined in (1) we obtain

$$\Delta g_{\text{solv}}(i) = \text{SolPar}(i)\cdot \left(\text{Occ}_{\max}(i) - \sum_{j \neq i} \left[\text{Vol}(j)\cdot \exp(-r_{ij}^2/2\sigma^2)\right]\right) \tag{3}$$

The term SolPar($i$)·Occ$_{\max}$($i$) is a constant, which only depends on the atom type and represents the free energy of solvation of atom $i$ at zero occupancy. In deriving the corresponding forces exerted on atom $i$ by all other atoms only the summation in (3) contributes. The total atomic solvation-dependent force $f_{\text{solv},w}(i)$ exerted on atom $i$ in Cartesian direction $w(w = x, y$ or $z)$ is obtained by differentiation of (3):

$$f_{\text{solv},w}(i) = -\text{SolPar}(i) \sum_{j \neq i} \text{Vol}(j)\cdot \frac{\partial \exp(-(r_{ij,x}^2 + r_{ij,y}^2 + r_{ij,z}^2)/2\sigma^2)}{\partial r_{ij,w}} \tag{4}$$

where the subscript $w$ in $f_{\text{solv},w}(i)$ and $r_{ij,w}$ stands for any Cartesian direction and the subscripts $x$, $y$ and $z$ for the three orthogonal directions. Carrying out the differentiation, we obtain

$$f_{\text{solv},w}(i) = \text{SolPar}(i) \sum_{j \neq i} \text{Vol}(j)\cdot \exp(-r_{ij}^2/2\sigma^2) \cdot \frac{r_{ij,w}}{\sigma^2} \tag{5}$$

This result implies that the solvation term based on excluded volume described here reduces to a sum of two-body terms only. These can be conveniently evaluated at the same time as the van der Waals and Coulomb terms. The effect of solvation can be easily visualized as a force that draws hydrophilic atoms out into the solvent (i.e., in the directions of least occupancy) and pushes hydrophobic ones towards the interior. The advantage of this description over accessible surface area approaches is obvious: no construction of any surface is necessary and no complicated derivatives have to be evaluated. How well this simple approach works is being described below.

The solvation (free) energy term has been implemented in two simulation packages. It has been added to the GROMOS [42] energy minimization and stochastic dynamics programs and to the PRESTO [43] energy minimization/molecular dynamics program. GROMOS does not allow the solvation term to be easily incorporated in the bonded and non-bonded routines. In order to ensure correctness the solvation term has been implemented as a separate subroutine in the main minimization and dynamics engines. Unavoidably, this results in longer execution times. In PRESTO, which is highly modular, the solvation term was added to the 1-2, 1-3, 1-4 and 1-5 interaction routines in a straightforward fashion. The most time-consuming part in the evaluation of the solvation term is the exponential envelope function (see (3)). In order to reduce the CPU time a table with function values of the exponential is constructed when execution starts and all further solvation force and energy calculations merely require lookups in this table. Use of the solvation term increases the required CPU time (compared to the equivalent vacuum simulations) by 20% in case of PRESTO and 50% in case of GROMOS.

## 3 PARAMETER FITTING

In order to use (3) and (5) to calculate atomic solvation free energies and forces several parameters have to be defined or determined. Exactly which parameters should be determined and how they are defined depends on the data used in the fitting procedure. Here we attempt to capture protein solvation effects by relying on experimental free energies of solvation, using carefully chosen model compounds. Isolated amino acids don't constitute good model compounds for amino acid residues in proteins since their charged amino and carboxyl groups make them very different. Indeed, Wolfenden *et al.* [44] could not determine their presence at all in the vapour phase above aqueous solutions. The alternatives are to use amino acids that are blocked at both ends, or many small blocked di- or tri-peptides or side-chain analogues. Wesson and Eisenberg [14] used data on side-chain analogs to derive atomic solvation parameters based on accessible surface areas. The disadvantage of their approach is that no direct information on backbone atoms can be obtained. This problem was bypassed by using only five different atom types (see Table 2) and assuming that a single solvation parameter applies to backbone and side-chain atoms of the same type alike.

We adopted a similar approach and chose a limited set of atom types. The basis of the parameter fitting procedure is that the free energy of solvation of a given residue or side-chain analogue is assumed to be a sum over atomic free energies of solvation. Using (2), this solvation free energy can be expressed in atomic contributions as follows.

$$\Delta G_{\text{solv}} = \sum_{i}^{\text{atoms}} (\text{Occ}_{\text{max}}(i) - \text{Occ}(i)) \cdot \text{SolPar}(i) \qquad (6)$$

where the summation extends over all non-hydrogen atoms in the residue or side-chain analogue. According to (6) free energies of solvation depend on the difference between actual and maximum occupancies, and on the atomic solvation parameters. The occupancies in turn depend on fragmental volumes (see (1) and Table 1), which were taken from [41]. These volumes had been calculated by numerical integration of the van der Waals envelope and they allow accurate estimation of the molecular volume as the sum of appropriate fragmental volumes [41]. For some atoms no data are provided in [41]. In these cases we estimated fragmental volumes on the basis of related atoms. Atomic solvation parameters are based on side-chain analogue data, so the related occupancy data must also refer to these analogues. Assuming that the conformational behaviour of side-chains in proteins and in the side-chain analogs are similar, actual side-chain analogue occupancies can be determined by taking side-chains from protein structures and simply averaging the calculated atomic occupancies over identical isolated side-chains. An atom in a protein which is completely surrounded by other protein atoms does not interact directly with water and in that sense is analogous to an atom *in vacuo*. We use maximum occupancies derived from protein structures in the fitting procedure for side-chain analogues. These maximum occupancies were obtained by using residues with less than 8% relative accessibility in a linear regression with relative residue accessibilities and atomic occupancies as variables. Extrapolation to a relative accessibility

**Table 1** Residue library with atomic volumes, occupancies and solvation free energies. Vol is the atomic fragmental volume [41], $\langle \mathrm{Occ} \rangle$ the average atomic occupancy based on isolated side-chain data for a set of 67 selected protein structures, and $\mathrm{Occ}_{\max}$ the maximum atomic occupancy derived from data on the entire proteins in the same set of structures. $\Delta g$ is the atomic free energy of solvation calculated from the average and maximum occupancies (this Table) and the fitted atomic solvation parameters (Table 2, 6-parameter set). Vol, $\langle \mathrm{Occ} \rangle$ and $\mathrm{Occ}_{\max}$ are in Å$^3$, $\Delta g$ is in kcal mol$^{-1}$.

| | *Vol* | $\langle Occ \rangle$ | $Occ_{max}$ | $\Delta g$ |
|---|---|---|---|---|
| **Ala** | | | | |
| CB | 16.15 | 0.00 | 344.22 | 1.368 |
| **Arg** | | | | |
| CB | 12.77 | 36.89 | 364.88 | 1.303 |
| CG | 12.77 | 43.30 | 360.36 | 1.260 |
| CD | 12.77 | 46.76 | 354.13 | 1.221 |
| NE | 9.00 | 50.62 | 352.99 | −5.262 |
| CZ | 6.95 | 48.74 | 349.52 | 1.195 |
| NH1 | 9.00 | 41.23 | 339.26 | −5.186 |
| NH2 | 9.00 | 38.73 | 344.52 | −11.988 |
| **Asn** | | | | |
| CB | 12.77 | 25.88 | 356.41 | 1.313 |
| CG | 9.82 | 31.58 | 358.25 | 1.298 |
| OD1 | 8.17 | 30.09 | 358.78 | −5.720 |
| ND2 | 13.25 | 25.88 | 348.87 | −5.621 |
| **Asp** | | | | |
| CB | 12.77 | 21.93 | 353.03 | 1.316 |
| CG | 9.82 | 26.85 | 353.74 | 1.299 |
| OD1 | 8.17 | 26.01 | 350.53 | −5.647 |
| OD2 | 8.17 | 26.06 | 353.12 | −6.701 |
| **Cys** | | | | |
| CB | 12.77 | 17.50 | 350.93 | 1.325 |
| SG | 19.93 | 11.14 | 340.27 | −2.113 |
| **Css** | | | | |
| CB | 12.77 | 14.43 | 347.77 | 1.324 |
| SG | 16.39 | 11.10 | 341.98 | −2.124 |
| **Glu** | | | | |
| CB | 12.77 | 29.79 | 364.31 | 1.329 |
| CG | 12.77 | 33.56 | 359.17 | 1.294 |
| CD | 9.82 | 36.76 | 356.41 | 1.270 |
| OE1 | 8.17 | 34.57 | 353.12 | −5.543 |
| OE2 | 8.17 | 34.22 | 357.07 | −6.615 |
| **Phe** | | | | |
| CB | 12.77 | 40.36 | 353.17 | 1.243 |
| CG | 7.26 | 56.50 | 355.97 | 0.169 |
| CD1 | 10.80 | 51.41 | 349.09 | 0.168 |
| CD2 | 10.80 | 51.37 | 349.35 | 0.168 |
| CE1 | 10.80 | 49.13 | 344.30 | 0.166 |
| CE2 | 10.80 | 49.13 | 344.17 | 0.166 |
| CZ | 10.80 | 48.30 | 342.42 | 0.166 |
| **Pro** | | | | |
| CB | 12.77 | 29.30 | 352.73 | 1.285 |
| CG | 12.77 | 30.79 | 349.44 | 1.266 |
| CD | 12.77 | 30.31 | 351.98 | 1.278 |
| **Gln** | | | | |
| CB | 12.77 | 33.03 | 358.74 | 1.294 |
| CG | 12.77 | 37.42 | 350.62 | 1.244 |
| CD | 9.82 | 41.50 | 348.03 | 1.218 |
| OE1 | 8.17 | 38.56 | 348.12 | −5.387 |
| NE2 | 13.25 | 33.95 | 336.58 | −5.266 |
| **His** | | | | |
| CB | 12.77 | 33.60 | 363.08 | 1.309 |
| CG | 7.26 | 46.63 | 367.51 | 0.728 |
| ND1 | 9.25 | 42.99 | 363.65 | −6.279 |
| CD2 | 10.80 | 41.45 | 360.01 | 0.723 |
| CE1 | 10.80 | 39.44 | 358.25 | 0.723 |
| NE2 | 9.25 | 41.01 | 358.08 | −5.517 |
| **Ile** | | | | |
| CB | 9.40 | 38.51 | 352.99 | 1.250 |
| CG1 | 12.77 | 35.53 | 346.63 | 1.236 |
| CG2 | 16.15 | 28.91 | 340.66 | 1.239 |
| CD1 | 16.15 | 29.35 | 337.72 | 1.225 |
| **Leu** | | | | |
| CB | 12.77 | 25.00 | 349.74 | 1.290 |
| CG | 9.40 | 40.93 | 347.81 | 1.219 |
| CD1 | 16.15 | 22.42 | 336.01 | 1.246 |
| CD2 | 16.15 | 22.42 | 336.15 | 1.247 |
| **Lys** | | | | |
| CB | 12.77 | 34.52 | 360.10 | 1.294 |
| CG | 12.77 | 40.97 | 352.29 | 1.237 |
| CD | 12.77 | 43.34 | 348.95 | 1.214 |
| CE | 12.77 | 40.88 | 346.80 | 1.216 |
| NZ | 13.25 | 34.61 | 335.53 | −11.797 |
| **Met** | | | | |
| CB | 12.77 | 33.51 | 355.18 | 1.278 |
| CG | 12.77 | 37.72 | 352.42 | 1.250 |
| SD | 16.39 | 34.70 | 342.55 | −1.976 |
| CE | 16.15 | 31.45 | 341.10 | 1.230 |
| **Trp** | | | | |
| CB | 12.77 | 47.64 | 358.34 | 1.235 |
| CG | 7.26 | 66.63 | 364.79 | 0.168 |
| CD1 | 10.80 | 57.95 | 358.65 | 0.169 |
| CD2 | 6.80 | 73.65 | 363.91 | 0.163 |
| NE1 | 9.00 | 61.59 | 358.82 | −5.172 |
| CE2 | 6.80 | 71.50 | 361.15 | 0.163 |
| CE3 | 10.80 | 63.34 | 356.67 | 0.165 |
| CZ2 | 10.80 | 60.01 | 351.32 | 0.164 |
| CZ3 | 10.80 | 57.42 | 351.67 | 0.166 |
| CH2 | 10.80 | 56.28 | 348.82 | 0.165 |

**Table 1** (*cont.*)

| | *Vol* | $\langle Occ\rangle$ | $Occ_{max}$ | $\Delta g$ | | *Vol* | $\langle Occ\rangle$ | $Occ_{max}$ | $\Delta g$ |
|---|---|---|---|---|---|---|---|---|---|
| **Ser** | | | | | **Tyr** | | | | |
| CB | 12.77 | 10.09 | 355.67 | 1.373 | CB | 12.77 | 41.54 | 354.13 | 1.242 |
| OG | 11.04 | 11.84 | 353.82 | −5.951 | CG | 7.26 | 59.22 | 360.14 | 0.169 |
| | | | | | CD1 | 10.80 | 54.92 | 355.58 | 0.169 |
| **Thr** | | | | | CD2 | 10.80 | 55.01 | 352.07 | 0.167 |
| CB | 9.40 | 24.87 | 359.13 | 1.328 | CE1 | 10.80 | 54.44 | 351.28 | 0.167 |
| OG1 | 11.04 | 21.28 | 356.37 | −5.831 | CE2 | 10.80 | 54.48 | 347.51 | 0.165 |
| CG2 | 16.15 | 17.15 | 347.95 | 1.314 | CZ | 7.26 | 58.39 | 351.41 | 0.165 |
| | | | | | OH | 10.94 | 43.34 | 340.97 | −5.179 |
| | | | | | **Val** | | | | |
| | | | | | CB | 9.40 | 29.39 | 351.32 | 1.279 |
| | | | | | CG1 | 16.15 | 21.06 | 340.66 | 1.270 |
| | | | | | CG2 | 16.15 | 21.06 | 340.62 | 1.270 |

of zero gave the maximum occupancies. We used extrapolated values rather than the observed maxima in order to reduce the effect of coordinate errors in the crystal structures. The actual and maximum occupancies were derived from 67 well-determined protein structures ($R < 20\%$ and resolution $<2$ Å) in PDB release 56 [45] and are shown in Table 1. Hydrogen atoms were ignored in all cases.

Side-chain analogue free energies of solvation for all residues except glycine and proline were taken from [44] and corrected for the entropy of mixing according to Sharp *et al.* [13]. Using (6), these can be expressed in atomic solvation parameters. After calculation of the actual and maximum occupancies the only unknown quantities are the atomic solvation parameters. Wesson and Eisenberg [14] fitted atomic solvation parameters for five atom types, C, neutral N or O, $N^+$, $O^-$ and S (see Table 2). Their approach is based on accessible surface areas, so the numbers cannot be compared directly to ours. Several sets of parameters were investigated here in order to strike the best balance between a low number of parameters and a good correspondence between observed and calculated free energies of solvation. Parameters were determined through a standard linear least-squares procedure with 18 equations (all available side-chain analogues) with the assumption that errors in the occupancies are negligible and errors in the corrected free energies of solvation are of uniform magnitude. Table 2 shows the resulting values for two parameter sets. Our 5-parameter set contains C, N/O, $N^+$, $O^-$, S, while the 6-parameter set in addition distinguishes between aliphatic and aromatic carbon atoms. Following Wesson and Eisenberg [14], in case of potentially delocalized charges (Glu, Asp, Arg and charged His) the charge was assigned to the most exposed atom, i.e., the atom with the largest difference between average and maximum occupancy. Under the experimental conditions [44] His has a nominal charge of about +0.1 and its ND1 atom was taken to be a linear combination of 10% $N^+$ and 90% neutral N. In case of the 6-parameter fitting, all carbon atoms in the rings of Phe, Tyr and Trp (CG and beyond) were considered aromatic; CG, CD2 and CE1 of His were arbitrarily regarded "half-aromatic" and considered to be an equimolar mixture of aliphatic and aromatic carbon. No attempt was made to optimize this ratio.

The atomic solvation free energy contributions for the 18 side-chain analogues that are calculated with the six-parameter set are shown in Table 1. Table 3 shows the corrected experimental free energies of solvation of the side-chain analogues and

**Table 2** Atomic solvation parameters for several parameter sets. Atomic solvation parameters ("SolPar") according to the Wesson-Eisenberg accessible surface area approach ("Wesson") [14], and to our 5- and 6-parameter occupancy approach. The Wesson parameters are in cal mol$^{-1}$ Å$^{-2}$, ours in cal mol$^{-1}$ Å$^{-3}$. "C ar" stands for aromatic carbon.

| | *Wesson* | *5 parameters* | *6 parameters* |
|---|---|---|---|
| C | 12 | 1.7 | 4.0 |
| C ar | | | 0.6 |
| N/O | −116 | −15.5 | −17.4 |
| $N^+$ | −186 | −31.7 | −39.2 |
| $O^-$ | −175 | −17.3 | −20.5 |
| S | −18 | −1.9 | −6.4 |

**Table 3** Experimental and fitted solvation free energies of side-chain analogues. All energies are in kcal/mol. The first column ("Sharp Absolute") contains the experimental free energies of solvation [44], corrected for the entropy of mixing [13]. The other columns contain fitted energies ("Absolute") and the deviations from the experimental values ("Difference") for three sets of parameters: the Wesson and Eisenberg set ("Wesson") [14], and our 5- and 6-parameter sets.

| | *Sharp* | *Wesson* | | *5 parameters* | | *6 parameters* | |
|---|---|---|---|---|---|---|---|
| | *Absolute* | *Absolute* | *Difference* | *Absolute* | *Difference* | *Absolute* | *Difference* |
| Ala | 2.63 | 1.64 | −0.99 | 0.58 | −2.05 | 1.37 | **−1.26** |
| Arg | −17.46 | −16.55 | 0.91 | −16.88 | 0.58 | −17.46 | 0.00 |
| Asn | −8.31 | −9.96 | −1.65 | −8.98 | −0.67 | −8.73 | −0.42 |
| Asp | −9.64 | −9.61 | 0.03 | −9.55 | 0.09 | −9.74 | −0.10 |
| Cys | 0.01 | −0.33 | −0.34 | −0.07 | −0.08 | −0.79 | **−0.80** |
| Gln | −7.35 | −9.00 | −1.65 | −7.88 | −0.53 | −6.90 | 0.45 |
| Glu | −8.36 | −8.47 | −0.11 | −8.84 | −0.48 | −8.26 | 0.10 |
| His | −8.25 | −3.96 | 4.29 | −8.21 | 0.04 | −8.31 | −0.06 |
| Ile | 4.89 | 2.71 | −2.18 | 2.11 | −2.78 | 4.95 | 0.06 |
| Leu | 5.20 | 2.65 | −2.55 | 2.14 | −3.06 | 5.00 | −0.20 |
| Lys | −6.84 | −8.94 | −2.10 | −7.43 | −0.59 | −6.84 | 0.00 |
| Met | 0.93 | 1.57 | 0.64 | 1.01 | 0.08 | 1.78 | **0.85** |
| Phe | 2.18 | 3.12 | 0.94 | 3.55 | 1.37 | 2.24 | 0.06 |
| Ser | −4.31 | −3.73 | 0.58 | −5.32 | −1.01 | −4.58 | −0.27 |
| Thr | −3.54 | −2.96 | 0.58 | −4.06 | −0.52 | −3.19 | 0.35 |
| Trp | −2.40 | 0.33 | 2.73 | −0.09 | 2.31 | −2.61 | −0.21 |
| Tyr | −3.17 | −1.61 | 1.56 | −1.58 | 1.59 | −2.93 | 0.24 |
| Val | 4.07 | 2.35 | −1.72 | 1.63 | −2.44 | 3.82 | −0.25 |
| RMS deviation | | | 1.77 | | 1.49 | | 0.46 |
| Maximum deviation | | | 4.29 | | 3.06 | | 1.26 |

values that were calculated on the basis of the three different sets of fitted atomic solvation parameters. In terms of RMS and maximum deviations our five-parameter set is somewhat more successful in reproducing the observed solvation free energies than Wesson and Eisenberg's. When carefully looking at the 5-parameter based results in Table 3 it becomes evident that the solvation free energies of the hydrophobic residues Ala, Ile, Leu and Val are systematically too small, while the values for the aromatic residues Phe, Tyr and Trp are systematically too large. Discrimination between aliphatic and aromatic carbon seems, therefore, logical. In doing so the correspondence between observed and calculated values improves dramatically. In case of five parameters the system of equations is overdetermined by a factor of 3.6; in case of six parameters this factor drops to 3.0. This is quite acceptable since the RMS deviation decreases from 1.49 to 0.46 kcal/mol. With the 6-parameter set the only considerable deviations occur for Ala, Cys and Met (boldface in Table 3). The latter two are the only sulphur-containing compounds and their deviations from the observed values are of equal magnitude, but opposite sign. This indicates that the description of one sulphur-type is not adequate and/or that data on more sulphur-containing compounds are required to reliably determine the atomic solvation parameter of sulphur. We also tested a 4-parameter set with one value for $N^+/O^-$ (which led to a marked RMS increase) and a 6-parameter set with separate values for neutral N and O (which did not lead to a substantial decrease in RMS). In all subsequent work only the 6-parameter set (with different parameters for the two types of carbon atoms) was used to describe protein-solvent interactions.

## 4 TESTING THE EFFECTIVE SOLVATION TERM: BPTI

### 4.1 *Methods*

The choice of tests to assess the performance of the solvation term depends on the properties that it should reproduce. As stated above, our main aim is to reproduce structural features. Secondly (and related to our main aim) the method should allow realistic exploration of conformational space especially for surface residues. Therefore, we carry out SD simulations with the solvation term and compare the results with crystal structure data and with full solution MD simulations. We focus on crystal rather than NMR data because we want to compare our results with one unique structure. Since our fitting procedure did not include data on backbone atomic solvation free energies, we also test the performance of the solvation term with solvation parameters set to zero for all backbone atoms. We also carry out control simulations *in vacuo* without any solvation term. BPTI was chosen as test compound for three reasons: a reliable crystal structure is available, it is a small protein, and it has rigid as well as very flexible regions (specifically, residues 1–15). The reason the protein must be small is that MD simulations with explicit solvent are carried out for comparison. The advantage of having flexible and rigid regions is that we can test both whether rigid regions preserve their rigidity and whether flexible regions are flexible in our simulations. This provides a means of assessing whether on the average the effective solvation term underestimates, overestimates or reproduces the effect of solvent. Also, if the framework of the protein is basically rigid then the crystal and solution structures for the greater part will be very similar. The structures produced by the SD simulations with solvation term do not have to be nearly identical to the crystal structure, but they must be realistic. We routinely use the following methods to assess the plausibility of (modelled) proteins: (1) Polar-Diagnostics 88 [46, 47], which calculates the polar fractions of exposed and buried surfaces and compares them with the distributions of these quantities for proteins of known structure; and (2) analysis and comparison of secondary structure elements, using the DSSP program [48]. We also compare flexibilities, hydrogen bonds, accessibilities, calculated (free) energies and RMS deviations between crystal structures and models.

**Table 4** Simulation protocol of the MD and SD simulations and EM calculations.

| | |
|---|---|
| *Initialization* | |
| • Coordinates from 5PTI PDB file. Protein and all 63 biowaters retained | |
| • Energy Minimization of protein and biowater with backbone and water oxygen restraints | |
| • Creation of truncated octahedral box with protein, biowater and 2301 solvent molecules | |
| • Replacement of 6 water molecules at positions of highest electrostatic potential by 6 Cl$^-$ ions<br>Box contains 7648 atoms: 568 protein (58 residues) + 7074 water (2358 molecules) + 6 ions | |
| • Restrained Energy Minimization of the complete system with periodic boundary conditions. | |
| *Settings of the MD protocol* | |
| General | |
| • Box | periodic truncated octahedron |
| • All bond distances | conserved by the SHAKE procedure |
| • SHAKE relative geometrical tolerance | 0.001 |
| • Isothermal simulation | separate scaling for solute and solvent |
| • Isobaric simulation | constant isotropic pressure: 1 atm |
| • Isothermal compressibility | $4.6 \times 10^{-10}$ m$^2$ N$^{-1}$ |
| • Updating non-bonded pair list | every 10 steps |
| First equilibration | |
| • Time span | 26 ps |
| • Temperature | increasing from 0 to 274 K |
| • Temperature coupling time constant | increasing from 0.001 to 0.1 ps |
| • Pressure coupling time constant | increasing from 0.2 to 0.5 ps |
| • Time step | increasing from 0.4 to 1.6 fs |
| • Non-bonded cut-off radius | increasing from 9 to 11 Å |
| Second equilibration and data collection | |
| • Time span (second equilibration) | 20 ps |
| • Time span (data collection) | 200 ps |
| • Temperature | 274 K |
| • Temperature coupling time constant | 0.1 ps |
| • Pressure coupling time constant | 0.5 ps |
| • Time step | 2 fs |
| • Non-bonded cut-off radius | 12 Å |
| • Data collection rate (coordinates) | every 0.1 ps; 2000 frames in total |
| *Settings of the SD protocol* | |
| • Where applicable the same as those of the MD data collection protocol | |
| • Protein coordinates and velocities taken from solution MD after the second equilibration | |
| • Friction constant | 91 ps$^{-1}$ |
| • Time step | 1–2 fs |
| • Time span (constant temperature SD) | 200 ps |
| • Temperature (constant temperature SD) | 274 K |
| • Time span (annealing SD) | 25 ps |
| • Temperature (annealing SD) | decreasing from 274 to 4 K |
| *Settings of the EM protocol* | |
| • Coordinates taken from SD after annealing | |
| • Updating non-bonded pair list | every 10 steps |
| • Non-bonded cut-off radius | 12 Å |
| • All bond distances | harmonic potential; SHAKE not used |
| • Span (steepest descent) | 200 steps |
| • Span (conjugate gradient) | until no significant energy change occured |

## 4.2 *Simulations*

All simulations used for the comparisons were carried out using the GROMOS program suite and with the standard GROMOS force field for proteins [42]. In the simulations with explicit solvent the charged force field version was employed (with net charges on formally charged residues). In all *in vacuo* simulations and in the simulations with the effective solvation term the neutral GROMOS force field was used (with reduced charges on formally charged residues to ensure electroneutrality of all individual residues). The neutral set is generally used in cases where there is no explicit water to shield charges from each other. For water the SPC/E potential function was used [49]. Non-bonded parameters for atom pair $(i,j)$ were taken as the geometrical mean of the parameters for $(i,i)$ and $(j,j)$. A constant relative dielectric of 1.0 was employed throughout. The explicit solvent simulations were carried out on the Fujitsu VP2600 super computer at PERI, Osaka, while all other simulations were run on the Vax cluster at EMBL, Heidelberg. The complete simulation protocols are detailed in Table 4. Initial coordinates and velocities for the vacuum SD, the full effective solvation SD and the side-chains-only solvation SD were taken from the explicit water simulations after the two equilibration periods (46 ps in total). Only water and chloride ions were removed; no other adjustments were made. Data collection started immediately afterwards and spans in all four cases exactly the same period in time. The SHAKE procedure was occasionally not able to satisfy all distance constraints simultaneously (both with and without solvation term). In these cases the time step was temporarily reduced from 2 to 1 fs.

Data obtained or derived from the simulations are presented in Tables 5 through 8. The four different simulation types detailed in these tables are "Water" (explicit water), "Vacuum", "SC solvation" (solvation term for side-chain atoms only) and "Full solvation" (solvation term for all atoms). "Average" stands for the average structure of an SD or MD simulation, and "final" for the structure at the end of an SD or MD simulation or EM calculation. The energetic data derived from the simulations are detailed in Table 5. Irrespective of the solvent description that was actually used (explicit water, none, side-chain solvation, full solvation), the total energy in the table comprises the full solvation contribution calculated for the final structure. Table 6 gives the position fluctuations over the SD and MD simulations and the RMS differences between simulation structures and the 5PTI crystal structure that was used as a starting point for all simulations. It also lists the number of hydrogen bonds for the final SD, MD and EM structures. Relative accessibilities and polar fractions are provided in Table 7 and the secondary structure assignments in Table 8. The latter three tables show data on more than one BPTI crystal structure in order to give an impression of the distribution of relevant properties for these structures. It should be noted that 7PTI is a C30A/C51A mutant (and therefore lacks the Cys30–Cys51 disulfide bond), 8PTI has glycine instead of Tyr35, and 9PTI has Met52 oxidized. C$\alpha$ traces of the 5PTI structure and the final "Vacuum", "SC solvation" and "Full solvation" structures after Energy Minimization are shown in Figure 1.

## 4.3 *Analysis*

Analysis of the following properties provides a very powerful means of assessing the quality of model (or other) protein structures: (1) accessibility, (2) polar surface

**Table 5** Energies derived from the MD and SD simulations and EM calculations. All energies are in kcal/mol. The bonded energy comprises bond, angle, and proper and improper torsion terms. LJ stands for Lennard-Jones energy, and bb, sc, bb+sc stand for backbone, side-chain and full (backbone plus side-chain) solvation energies, respectively. For comparison, the energies of 5PTI before and after energy minimization with the full solvation term are also given.

| | *Total* | *Bonded total* | *Non-bonded* | | *Solvation* | | |
|---|---|---|---|---|---|---|---|
| | | | *Coulomb* | *LJ* | *bb* | *sc* | *bb + sc* |
| 5PTI crystal | −575 | +500 | −514 | −287 | −148 | −126 | −274 |
| 5PTI EM (Full solvation) | −1172 | +156 | −610 | −442 | −146 | −130 | −276 |
| Water (MD final) | −715 | +380 | −470 | −349 | −165 | −111 | −276 |
| Vacuum (SD final) | −860 | +390 | −618 | −429 | −121 | −82 | −203 |
| SC solvation (SD final) | −865 | +414 | −614 | −401 | −125 | −139 | −264 |
| Full solvation (SD final) | −883 | +371 | −584 | −376 | −160 | −134 | −294 |
| Vacuum (EM final) | −1186 | +188 | −680 | −515 | −105 | −74 | −179 |
| SC solvation (EM final) | −1215 | +180 | −671 | −478 | −112 | −134 | −246 |
| Full solvation (EM final) | −1222 | +162 | −653 | −461 | −143 | −127 | −270 |

**Table 6** Deviations from 5PTI positions, fluctuations and hydrogen bonds. In the first two columns RMS C$\alpha$ deviations of the models from the 5PTI crystal structure are given for all residues and for the residues 16–58, respectively. The next two columns show atomic position fluctuations during the SD and MD simulations for C$\alpha$ and all atoms, respectively. The final two columns list the total number of hydrogen bonds ("all") and the number of hydrogen bonds between backbone atoms ("bb"). Hydrogen bond criteria are: donor ··· acceptor distance <3.5 Å, hydrogen ··· acceptor distance <2.5 Å and donor–hydrogen–acceptor angle between 120 and 180°.

| | *5PTI-RMS ($C_\alpha$)* | | *Fluctuations* | | *Hydrogen bonds* | |
|---|---|---|---|---|---|---|
| | *All residues* | *16–58* | *$C_\alpha$* | *All atoms* | *All* | *bb* |
| *BPTI crystal structures* | | | | | | |
| 4PTI crystal | 1.24 | 1.40 | | | 34 | 23 |
| 5PTI crystal | | | | | 34 | 22 |
| 7PTI crystal | 0.20 | 0.21 | | | 34 | 22 |
| 8PTI crystal | 1.82 | 1.97 | | | 41 | 24 |
| 9PTI crystal | 0.09 | 0.09 | | | 34 | 22 |
| *Simulations* | | | | | | |
| Water (average) | 1.91 | 1.94 | 0.76 | 1.14 | | |
| Water (MD final) | 2.13 | 2.09 | | | 27 | 14 |
| Vacuum (average) | 2.47 | 2.36 | 0.92 | 1.23 | | |
| Vacuum (SD final) | 2.82 | 2.82 | | | 41 | 20 |
| Vacuum (EM final) | 2.94 | 2.90 | | | 57 | 27 |
| SC solvation (average) | 2.06 | 1.83 | 0.99 | 1.52 | | |
| SC solvation (SD final) | 2.38 | 2.07 | | | 32 | 17 |
| SC solvation (EM final) | 2.40 | 2.07 | | | 48 | 26 |
| Full solvation (average) | 2.45 | 1.66 | 1.28 | 1.86 | | |
| Full solvation (SD final) | 2.71 | 1.49 | | | 29 | 18 |
| Full solvation (EM final) | 2.81 | 1.48 | | | 39 | 22 |

fractions, (3) secondary structure, (4) hydrogen bonds, (5) RMS deviations from the crystal structure, and (6) energy. Here the analysis is carried out to compare structures resulting from explicit water, vacuum and solvation term simulations in order to judge the performance of SD simulations with the effective solvation term.

(1) The relative total accessibility is defined as the surface area of a protein that is exposed divided by the maximum exposed surface area. The accessibilities in Table 7 clearly show that after energy minimization the "Vacuum" simulations give values that are lower than normally observed for proteins. The values derived from the "SC solvation" and "Full solvation" simulations appear to be approximately correct. Especially the value of the energy minimized "Full solvation" structure (0.418) is very close to the accessibility of the 5PTI crystal structure (0.423). The average and final accessibility values for the MD simulations with explicit water are somewhat higher than the values for the "Full solvation" SD simulations.

**Table 7** Relative accessibilities and exposed and buried polar fractions. Internal and external polar stand for the polar fractions of buried and exposed atomic surfaces. SC internal and SC external stand for the equivalent quantities of side-chain atoms only. Polar fractions are defined as the product of the absolute values of the partial atomic charges and the corresponding (exposed or buried) atomic surface areas divided by the total (exposed or buried) atomic surface area [47]. The relative accessibility is dimensionless; the four types of polar fractions are in $|e|$ (absolute electron charges). Reference values are based on values for all 150 known protein structures at the time with the exclusion of atypical structures [47]. Unusual high and low values are in boldface and marked + and −, respectively.

| | *Relative accessibility* | *Internal polar* | *External polar* | *SC internal polar* | *SC external polar* |
|---|---|---|---|---|---|
| *Normal range* | | | | | |
| average in 150 protein | 0.400 | 0.172 | 0.172 | 0.098 | 0.143 |
| standard deviation | 0.055 | 0.010 | 0.015 | 0.011 | 0.019 |
| minimum in 150 proteins | 0.320 | 0.158 | 0.147 | 0.074 | 0.115 |
| maximum in 150 proteins | 0.520 | 0.190 | 0.204 | 0.118 | 0.183 |
| *BPTI crystal structures* | | | | | |
| 4PTI crystal | 0.429 | 0.166 | 0.175 | 0.091 | 0.140 |
| 5PTI crystal | 0.423 | 0.167 | 0.173 | 0.097 | 0.134 |
| 7PTI crystal | 0.426 | 0.172 | 0.170 | 0.101 | 0.132 |
| 8PTI crystal | 0.400 | 0.172 | 0.170 | 0.098 | 0.135 |
| 9PTI crystal | 0.419 | 0.170 | 0.169 | 0.099 | 0.132 |
| *Simulations* | | | | | |
| Water (average) | 0.442 | **0.195+** | 0.151 | **0.122+** | 0.115 |
| Water (MD final) | 0.482 | **0.192+** | **0.144−** | **0.121+** | **0.107−** |
| Vacuum (average) | 0.328 | **0.196+** | **0.140−** | **0.131+** | **0.101−** |
| Vacuum (SD final) | 0.345 | 0.185 | **0.144−** | **0.122+** | **0.101−** |
| Vacuum (EM final) | **0.319−** | 0.189 | **0.134−** | **0.125+** | **0.095−** |
| SC solvation (average) | 0.380 | 0.180 | 0.168 | 0.102 | 0.139 |
| SC solvation (SD final) | 0.416 | 0.171 | 0.168 | 0.089 | 0.142 |
| SC solvation (EM final) | 0.386 | 0.174 | 0.164 | 0.095 | 0.137 |
| Full solvation average) | 0.405 | 0.181 | 0.168 | 0.108 | 0.132 |
| Full solvation (SD final) | 0.462 | 0.171 | 0.168 | 0.096 | 0.133 |
| Full solvation (EM final) | 0.418 | 0.169 | 0.171 | 0.098 | 0.135 |

(2) Baumann *et al.* [47] observed that buried and exposed polar fractions of the side-chains (Table 7, last two columns), which are independent quantities, provide a sensitive test for "normality" of protein structures. The buried polar fractions are systematically too high for the "Vacuum" structures and the exposed polar fractions too low. Together with the low accessibility (criterion 1) this indicates that vacuum simulations cause polar side-chains to fold back onto the protein. In the "Water" case much of the protein's polar surface is buried, but the total accessibility (criterion 1) is relatively high. This suggests that the protein as a whole has opened up and that consequently more apolar atomic surface gets exposed. Both the "SC solvation" and "Full solvation" cases have normal buried and exposed polar fractions. There is almost perfect agreement in polar fraction values between the 5PTI crystal structure and the "Full solvation" final EM structure.

(3) In a stable solution structure the secondary structure [48] as observed in the crystal structure should be preserved as much as possible and not very much additional secondary structure should form. The secondary structure assignment in Table 8 shows that the "SC solvation" and "Full solvation" simulations preserve the major secondary structure elements (residues 18–35 and 48–55). In contrast, the $\beta$-strand-turn-$\beta$-strand element (residues 18–35) progressively dissolves during the "Vacuum" and "Water" SD/MD simulations. In Figure 1 the disruption of the two hydrogen bonded strands in the "Vacuum" case (yellow structure) is visible. In case of "SC solvation" the $\beta$-strand 18–24 gets extended at the N-terminus until residue 15. In the crystal structures and in the structures resulting from all simulations there is a small segment of $3_{10}$-helix or hydrogen-bonded turn near Cys5 (which bridges to Cys55 and connects beginning and end of the protein chain). Apart from the stretch 3–6 the residue range 1–17 generally does not contain much secondary structure in any of the simulations. The flexibility implied by the absence of secondary structure in the simulations is consistent with NMR data on mutants of BPTI [50] and with the fact that residues 14–17 are involved in inhibition.

(4) The number of backbone-backbone hydrogen bonds and the number of hydrogen bonds involving side-chain atoms should be approximately the same in simulations and crystal structures. The analysis of hydrogen bonds (Table 6) supports the picture that emerges from the previous paragraph. The "Vacuum" and "SC solvation" structures have the highest number of backbone-backbone hydrogen bonds, which is a logical consequence of the fact that in both cases the solvation parameters for the backbone are zero. The "Vacuum" structure also has the highest total number of hydrogen bonds by far, again a clear indication that the side-chains have folded back onto the protein. In the "SC solvation" case polar side-chains are drawn into solvent and cannot compete with backbone atoms to form hydrogen bonds, which may explain the observed formation of more secondary structure at residues 15–17 (criterion 3). The "Water" final MD structure again is unusual in that it has a low number of backbone-backbone hydrogen bonds, but the number of hydrogen bonds involving side-chain atoms (13) is comparable with the "SC solvation" and "Full solvation" cases (15 and 11, respectively, after SD). The number of backbone-backbone hydrogen bonds in the "Full solvation" EM structure is in perfect agreement with the crystal structures, although the total number of hydrogen bonds is somewhat high.

| | |
|---|---|
| | 1 2 3 4 5 |
| Sequence number | 1234567890123456789012345678901234567890123456789012345678 |
| Sequence | RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA |
| *BPTI crystal structures* | |
| 4PTI crystal | --GGGGS-----S----EEEEEEETTTTEEEEEEE-SSS--SS--SSHHHHHHHHS-- |
| 5PTI, 7PTI and 9PTI crystals | --GGGGS-----S----EEEEEEETTTTEEEEEEE-SSS--SS--SSHHHHHHHH--- |
| 8PTI crystal | --GGGGS-----S--S--EEEEEETTTTEEEEEES-TT--SSS--SSHHHHHHHHTT- |
| *Simulations* | |
| Water (average) | --STT-------S----EEEEE--STT-S-EEEEE-SS---SS--SSHHHHHHHH--- |
| Water (MD final) | --S---------S---SSEEEE--STT-S-EEEE--SS---TTS-SSHHHHHHHT--- |
| Vacuum (average) | --GGG-------SS--SSEEEEE-GGGSSEEEEE--SS---TT-EESHHHHHHHTT-- |
| Vacuum (SD final) | --TT--------SS--S-EEEE--GGGSS---EE--SS---TTEEESHHHHHHHTT-- |
| Vacuum (EM final) | --GGG-S-----TT--SSEEEE--GGGSS-EEEE--SS---TT-EEHHHHHHHHTT-- |
| SC solvation (average) | --GGG--TT--TT-EEEEEEEEE-GGGSSEEEEEE-SS--SS----SHHHHHHH---- |
| SC solvation (SD final) | --GGG--TT--GGGEEEEEEEEEETTTEEEEEEEE-SS--SS----SHHHHHHH---- |
| SC solvation (EM final) | --GGG--TT--SSSEEEEEEEEEETTTEEEEEEEE-SS--SS----SHHHHHTT---- |
| Full solvation (average) | --GGG------SSS---EEEEEEEGGGTEEEEEEE-SS---TT--SSHHHHHHHH--- |
| Full solvation (SD final) | --TTT-----SSSS--SEEEEEEETTTTEEEEEEE-SSS--S----SHHHHHHHH--- |
| Full solvation (EM final) | --TTT-----SSSS--SEEEEEEETTTTEEEEEEE-SSS--TT--SHHHHHHHHH--- |

**Table 8** Secondary structure of the various BPTI crystal and model structures. The assignment is according to the DSSP program [48], based on a generous hydrogen bond energy criterion: $E_{HBond} < -0.5$ kcal/mole. Codes are: H-helix, G-$3_{10}$-helix, E-extended ($\beta$-strand), T-hyrdogen-bonded turn, S-bend. The residues that are part of the rigid secondary structure elements in the crystal structure (segments 18–35 and 48–55) are underlined.

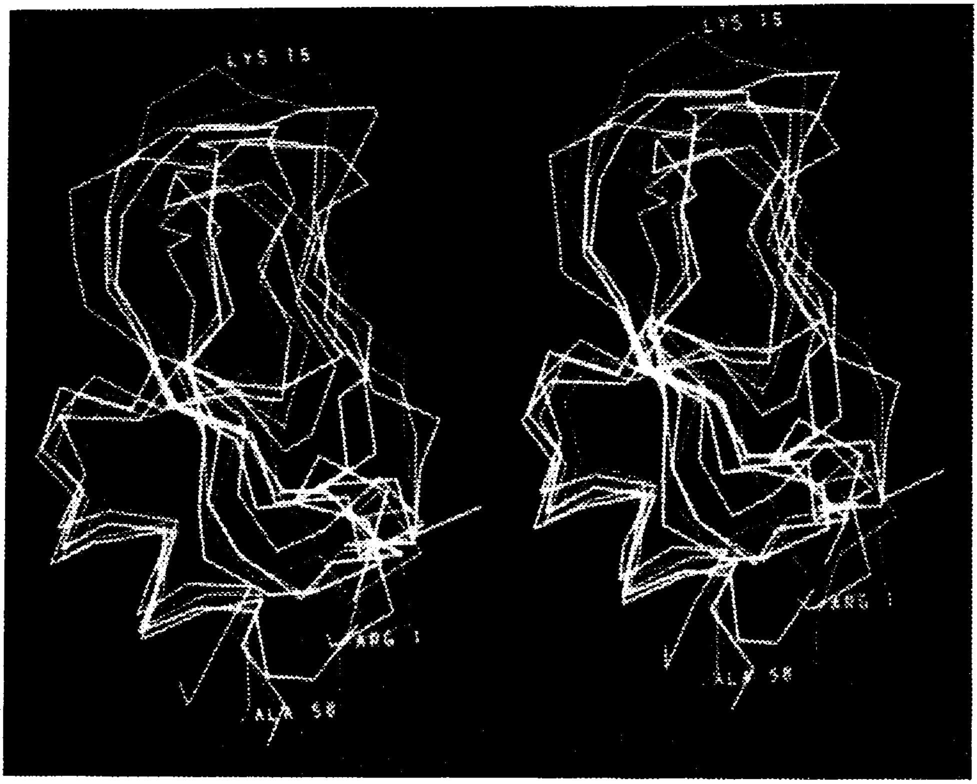

**Figure 1** Superimposed $C\alpha$ traces of the 5PTI crystal and the model structures after MD/SD simulations. The 5PTI crystal structure is purple, the "Water" model structure red, the "Vacuum" one yellow, the "SC solvation" one green and the "Full solvation" one blue. This picture was prepared with the WHAT IF modelling package [54]. (See colour plates)

(5) The position fluctuations ($C_\alpha$ and all atoms separately) observed in the MD and SD simulations are detailed in Table 6. The fluctuations increase in the order "Vacuum" → "SC solvation" → "Full solvation," which is due to the fact that drawing atoms outward into solvent leads to a higher average mobility. In the "Water" case the fluctuations are smallest because protein motion requires considerable rearrangement of water molecules leading to a relatively low mobility. Table 6 also shows the $C\alpha$ RMS position deviations of the model structures and crystal structures with respect to the 5PTI crystal structure. Small RMS deviations between model and crystal structures in regions that are known to be rigid indicate plausible model structures. In general the model structures show large deviations from the 5PTI structure (in the order of 2–3 Å), but large differences (up to 1.8 Å) occur between the various BPTI (mutant) crystal structures as well. However, there is a striking correspondence between the structures resulting from both solvation simulations and 5PTI for the rigid part of the molecule (residues 16–58), especially when considering that considerable positional fluctuations occur during the simulations.

Figure 1 shows $C_\alpha$ traces of the final models after MD or SD superimposed on the 5PTI structure. It is obvious that, apart from the "Water" structure (red) there are considerable deviations between the model structures and the crystal structure (purple) over the first 17 residues. It is unclear whether the similarity between the "Water" and crystal structures in this region is due to the low intrinsic mobility of BPTI in explicit water or to the inherently more realistic explicit water description. It is also unclear whether the large movement of Gly12 in the "Full solvation" structure (blue) is physically realistic or caused by overestimating the solvation free energy of glycines. Apart from this first segment of 17 residues, however, the "Full solvation" and crystal structures are very similar, even the C-termini (with two glycines and an alanine) match well. Considerable structural differences between the final SD "SC solvation" (green) and "Full solvation" (blue) structures can be observed in Figure 1. This indicates that even in folded proteins the (structural) effect of backbone solvation should not be underestimated.

(6) A careful analysis of various contributions to the total energy of model structures is generally very important in distinguishing "good" and "bad" structures. However, here our intention is to test the occupancy-based effective solvation term and for that purpose the force field-independent analysis methods (see above) are more suitable. The energies of the final MD, SD and EM structures are detailed in Table 5. In all cases they have been calculated using the uncharged GROMOS parameters and the full solvation term. In the "Water" case this can have consequences since the "Water" simulations were carried out using the charged GROMOS parameters. Indeed, in the "Water" case the (intra-molecular) Coulomb energy is much lower than in the other cases. The solvation energy of the final "Water" MD structure is comparable to the "SC solvation" and "Full solvation" energies, although in the "Water" case the side-chain contribution is somewhat small. The "Vacuum" and "SC solvation" simulations were not carried out with the full solvation term, and consequently the resulting structures have relatively high solvation free energies. The solvation energies of the final "Full solvation" EM structure agree well with those of the 5PTI crystal structure. For comparison, an energy minimization of 5PTI with full solvation was carried out as well. Although the total energy decreases by 600 kcal/mol, the solvation energy stays roughly the same. An energy analysis is generally quite useful, but in this case the other analysis tools provide a better measure of performance of the solvation term.

## 5 DISCUSSION AND CONCLUSIONS

### 5.1 *Limitations*

The effective solvation term proposed here and its implementation have several limitations.

- A small set of atomic solvation parameters is used. We only distinguish six different atom types (aliphatic C, aromatic C, uncharged N/O, S, $N^+$ and $O^-$) and consequently have only six different solvation parameters. Especially in the case of alanine (and cysteine and methionine to a lesser degree) this leads to discrepancies with experimental values (Table 3).

- Only side-chain analogues are used to determine solvation parameters. It is not *a priori* obvious that parameters derived on the basis of side-chain data are directly transferable to backbone atoms. The derived solvation free energies, especially of small residues such as Gly and Ala, may therefore differ from the physically correct values.
- There are experimental errors in the data used to fit the atomic solvation parameters: Wolfenden's free energies of solvation [44], Sharp's correction for the entropy of mixing [13], the fragmental volumes [41] and the extrapolated maximum occupancies.
- Contrary to our linear description, in reality a minimum empty volume is required for a water molecule to engage in direct interaction with the solute. This means that holes that are too small to contain a single water molecule in our description still contribute to the total solvation free energy.
- Experimental free energies comprise side-chain entropic effects (rotational freedom) and bulk entropic effects. In our force field the non-solvation part already explicitly describes rotation of side-chains and the side-chain entropic contributions to the experimental solvation free energies may cause an exaggerated preference for high solvent exposure.
- Subtle interactions such as water-mediated hydrogen bonds between two side-chains cannot be described by the effective solvation term. This situation arises when one water molecule is positioned such that it can form very favourable hydrogen bonds with two nearby polar or charged side-chain atoms, effectively fixing the relative position of these side-chain atoms. Our solvation description would tend to push these polar/charged atoms away from each other instead of fixing them in one position.
- It has been observed that the total solvation free energy of two polar or charged groups that are close together in space is less negative than the sum of the individual solvation free energies [e.g. 51, 52]. This relative increase in free energy of solvation for two nearby charged/polar side-chains is not reproduced by our solvation model at its current level of sophistication.

Improvements can be made that deal with part of these limitations. Including more experimental solvation data will lead to more reliable parameters. Using more parameters should lead to a better force field. Effective volumes that take into account specific geometrical features of proteins [53] may prove to be better than using fragmental volumes.

However, these limitations pose no serious problems as the method does not pretend to be more than a first-order approximation of part of physical reality. Whether the approximation is valid is primarily judged on the basis of its performance in simulations. In the context of our research this means that results of protein simulations should be compared with experimental data, mainly focusing on structural properties, and with the results of other simulation protocols.

### 5.2 *Performance*

When assessing the performance of the solvation term the crucial questions are: (1) how well does it reproduce relevant structural features, (2) to what extent can we expect these features to be reproduced, and (3) how important are the observed differences? Another important issue is the efficiency of effective solvation simulations in terms of (4) sampling conformational space and (5) CPU usage.

(1) How well does SD with effective solvation reproduce structural features observed in the crystal structure of 5PTI, and how normal are the model structures compared to a representative set of protein structures?

Despite large position fluctuations during the simulations, the solvation simulations produced models that are very much like the 5PTI crystal structure (in terms of $C_\alpha$ positions) in the rigid regions of the protein (segments 18–35 and 48–55). The secondary structure of these models is in good agreement with the crystal structure as well. The values for the buried and exposed polar and total surfaces of these models are normal and very close to those derived from the crystal structure.

The structures stemming from the explicit water MD simulations have the lowest overall deviations from the 5PTI crystal structure, but these simulations also have the lowest position fluctuations. The explicit water simulations caused loss of secondary structure in a rigid region. The buried polar surface fractions are higher than normal and the total accessibility is also rather high, suggesting that during these simulations hydrophobic surface tends to get exposed more.

The vacuum SD simulations failed entirely in producing plausible structures. In the course of these simulations part of the secondary structure dissolved and large deviations from the 5PTI structure were observed over the entire residue range despite a low intrinsic mobility. Far too many hydrogen bonds involving side-chains were observed and the relative total accessibility and the buried and exposed polar fractions fall far outside the "normal" ranges.

The above observations are not all independent, but they show clearly that vacuum simulations lead to very unrealistic structures, while simulations with an effective solvation term do very well with respect to all our criteria.

(2) To what extent can we expect our simulations to reproduce structural features?

One cannot expect vacuum, solution and effective-solvation simulations to reproduce crystal properties in great detail. Without the presence of crystal forces differences will arise. However, folded proteins in solution and in the crystal contain large numbers of stable secondary structure elements and any realistic simulation should produce structures that preserve secondary structure. On the other hand, production of much more secondary structure than observed in the crystal structure may also hint at defects in the simulation. In the case of BPTI the secondary structure of the residue ranges 18–35 and 48–55 should be preserved. In this respect the effective solvation simulations perform better than the vacuum simulations. We do not have a good explanation for the fact that the explicit water simulations produce unrealistic structures in that region, while its overall conservation of structure is quite adequate. It may be that the simulations were too short and that extending them will lead to more plausible structures (see point 4 below), or that the uncharged GROMOS force field is better in reproducing crystal properties than the charged force field, or that the combination of the SPC/E water and charged GROMOS protein force fields does not describe a protein in water well. A rigorous investigation is required, but this goes beyond the scope of this paper.

(3) How important are the observed differences between the model and crystal structures?

The observed differences between the models and crystal structures in the N-terminal region (residues 1–17) are quite large. Figure 1 shows that none of the final MD/SD structures has a conformation that is the same as any of the others in that region. After superposition of the 16–58 regions on each other the Cα RMS deviations in the 1–15 region vary from 2.3 (between 5PTI and "Water") to 6.1 Å (between "Vacuum" and "Full solvation"). How realistic and how important the differences are, is not completely clear. However, the N-terminal section is involved in inhibition, has no fixed secondary structure in the crystal, and, based on NMR data on BPTI mutants, has been shown to be very flexible [50]. Therefore, the structures resulting from our simulations may well represent low-energy conformers that are present in solution.

(4) How well do the effective solvation simulations sample conformational space?

Having a procedure that samples conformational space adequately is important, since it provides a fast means of generating relevant structures. This sampling is done best during "Full solvation" simulations, as judged from the position fluctuations of all atoms and Cα atoms only in Table 6. So, in addition to providing an adequate description of a protein in solution, the effective solvation term also enhances sampling. The explicit water simulations are least efficient in scanning conformational space due to the damping effect that the water molecules have on protein motion. Although long solution simulations may provide a more realistic description of physical reality than vacuum or solvation simulations, the "effective time step" for explicit water simulations is smaller than for vacuum and effective solvation simulations due to slow water rearrangement. The explicit water simulations effectively span a smaller time range than the other simulation and the results are not directly comparable. It may be that longer explicit water simulations will result in different, more plausible structures.

(5) How well do effective solvation simulations perform in terms of CPU time?

In terms of CPU requirements simulations with the effective solvation term have a clear advantage. 200 ps of the full water MD simulation took 95 hours CPU on the Fujitsu VP2600 and would have taken 7.8 months (5,700 hours) on a Vax 8810. On a Vax 8810 200 ps of SD (with the mixture of 1 and 2 fs step sizes actually used) would have taken 190 and 290 hours CPU for the vacuum and solvation simulations, respectively. The SD simulations with solvation are 50% slower than the corresponding vacuum simulations. The explicit water MD simulations take 30 times more time than the vacuum SD simulations.

### 5.3 *Future Extensions*

Our method can be easily extended to capture not only the hydrophobic effect, but also the electrostatic shielding effect of sovlent. At present we use the uncharged GROMOS protein force field in order to have at least some degree of shielding. The often used distance-dependent dielectric does not constitute a very good solution to the shielding problem since interactions through the protein are reduced just as much as interactions through empty space (i.e., solvent). An occupancy-dependent relative dielectric would describe shielding better and partly solve the problem. The GROMOS SD simulation program now calculates the number of atoms in a shell around any given atom and used that number as proportionality constant in the friction term for that atom. This description could easily be refined by making the friction coefficient occupancy-dependent. However, these two improvements may lead either to a force field that is no longer conservative or to a considerable increase in CPU usage due to the evaluation of more complex expressions, but they are definitely worth testing.

### 5.4 *Conclusions*

Our aim was to develop a tool to describe solvation effects in simulations in a very simple and inexpensive fashion which is accurate enough for most purposes. Despite the simplicity of the effective solvation term (which essentially is a sum of two-body terms) and despite inherent limitations, remarkably plausible structures are obtained. In fact, simulations with the solvation term in some respects seem to perform better than simulations with explicit water, although the evidence is based on one test case only. Comparable vacuum simulations do not produce realistic structures at all. Since CPU requirements for effective solvation simulations are only 50% higher than for vacuum simulations (and a factor of 20 lower than for explicit water simulations) we propose use of out method in all cases where presently vacuum simulations are employed.

## *Acknowledgements*

We thank our colleagues Wilfred van Gunsteren, Herman Berendsen and Gerrit Vriend for placing their programs at our disposal. We also thank S. Kawakita for valuable assistance in using the Fujitsu VP2600 and Jolanta Stouten for critically reading the manuscript. CS and PS thank the EC BRIDGE program for financial support. PS gratefully acknowledges a short-term fellowship to work at the Protein Engineering Research Institute (where part of this work was carried out), awarded by the Human Frontier Science Program Organization.

## *References*

[1] J.N. Israelachvili, *Intermolecular and surface forces*, Academic Press, London (1985).

[2] J.N. Israelachvili and R. Pashley, "The hydrophobic interaction is long range, decaying exponentially with distance", *Nature*, **300**, 341-342 (1982).

[3] P.L. Privalov and S.J. Gill, "Stability of protein structure and hydrophobic interaction", *Adv. Protein Chem.*, **39**, 191-234 (1988).

[4] W. Kauzmann, "Some factors in the interpretation of protein denaturation", *Adv. Protein Chem.*, **14**, 1-63 (1959).

[5] M.F. Perutz, "Electrostatic effects in proteins", *Science*, **201**, 1187-1191 (1978).

[6] D. Stigter, D.O.V. Alonso and K.A. Dill, "Protein stability: electrostatics and compact denaturated states", *Proc. Natl. Acad. Sci.*, **88**, 4176-4180 (1991).

[7] J.A. Rupley, E. Gratton and G. Careri, "Water and globular proteins", *Trends in Biochemical Sciences*, **8**, 18-22 (1983).

[8] E.N. Baker and R.E. Hubbard, "Hydrogen bonding in globular proteins", *Prog. Biophys. Molec. Biol.*, **44**, 97-179 (1984).

[9] C. Chothia, "Hydrophobic bonding and accessible surface area in proteins", *Nature*, **248**, 338-339 (1974).

[10] D. Eisenberg and A.D. McLachlan, "Solvation energy in protein folding and binding", *Nature*, **319**, 199-203 (1986).

[11] T. Ooi, M. Oobatake, G. Nemethy, H.A. Scheraga, "Accessible surface area as a measure of the thermodynamic parameters of hydration of peptides", *Proc. Natl. Acad. Sci. USA*, **84**, 3086-3090 (1987).

[12] T. Ooi and M. Oobatake, "Prediction of the thermodynamics of protein unfolding: the helix-coil transition of poly(L-alanine)", *Proc. Natl. Acad. Sci.*, **88**, 2859-2863 (1991).

[13] K.A. Sharp, A. Nicholls, R. Friedman and B. Honig, "Extracting hydrophobic free energies from experimental data: relationship to protein folding and theoretical models", *Biochemistry*, **30**, 9686-9697 (1991).

[14] L. Wesson and D. Eisenberg, "Atomic solvation parameters applied to molecular dynamics of proteins in solution", *Protein Science*, **1**, 227-235 (1992).

[15] Y.K. Kang, K.D. Gibson, G. Némethy and H.A. Scheraga, "Free energies of hydration of solute molecules. 4. Revised Treatment of the hydration shell model", *J. Phys. Chem.*, **92**, 4739-4742 (1988).

[16] F. Colonna-Cesari and C. Sander, "Excluded volume approximation to protein-solvent interaction. The solvent contact model", *Biophys. J.*, **57**, 1103-1107 (1990).

[17] T. Ooi and M. Oobatake, "Effects of hydrated water on protein folding", *J. Biochem. (Tokyo)*, **103**, 114-120 (1988).

[18] M.K. Gilson, K.A. Sharp and B.H. Honig, "Calculating the electrostatic potential of molecules in solution: method and error assessment", *J. Comput. Chem.*, **9**, 327-335 (1988).

[19] J. Warwicker and H.C. Watson, "Calculation of the electric potential in the active site cleft due to α-helix dipoles", *J. Mol. Biol.*, **157**, 671-679 (1982).

[20] M.K. Gilson and B. Honig, "Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis", *Proteins*, **4**, 7-18 (1988).

[21] H. Nakamura and S. Nishida, "Numerical calculations of electrostatic potentials of protein-solvent systems by the self-consistent boundary method", *J. Phys. Soc. Jap.*, **56**, 1609-1622 (1987).

[22] A. Jean-Charles, A. Nicholls, K. Sharp, B. Honig, A. Tempczyk, T.F. Hendrikson and W.C. Still, "Electrostatic contributions to solvation energies: comparison of free energy perturbation and continuum calculations", *J. Am. Chem. Soc.*, **113**, 1454-1455 (1991).

[23] D. Bashford and M. Karplus, "pKa's of ionizable groups in proteins: atomic detail from a continuum electrostatic model", *Biochemistry*, **29**, 10219-10225 (1990).

[24] M. Schäfer and C. Frömmel, "A precise analytical method for calculating the electrostatic energy of mactomolecules in aqueous solution", *J. Mol. Biol.*, **216**, 1045-1066 (1990).

[25] T. Takahashi, H. Nakamura and A. Wada, "Electrostatic forces in two lysozymes: calculations and measurements of histidine pKa values", *Biopolymers*, **32**, 897-909 (1992).

[26] W.F. van Gunsteren and A.E. Mark, "On the interpretation of biochemical data by molecular dynamics computer simulation", *Eur. J. Biochem.*, **204**, 947-961 (1992).

[27] J.A. McCammon, B.R. Gelin and M. Karplus, "Dynamics of folded proteins", *Nature*, **267**, 585-590 (1977).

[28] M. Levitt and R. Sharon, "Accurate simulation of protein dynamics in solution", *Proc. Natl. Acad. Sci.*, **85**, 7557-7561 (1988).

[29] V. Daggett and M. Levitt, "Molecular dynamics simulations of helix denaturation", *J. Mol. Biol.*, **223**, 1121-1138 (1992).

[30] J. Anderson, J.J. Ullo and S. Yip, "Molecular dynamics simulation of dielectric properties of water", *J. Chem. Phys.*, **87**, 1726-1732 (1987).

[31] P.A. Kollman and K.A. Dill, "Decisions in force field development: an alternative to those described by Roterman *et al.* (J. Biomol. Struct. & Dyn. 7, 421 (1989))", *J. Biomol. Struct. & Dyn.*, **8**, 1103-1107 (1991).

[32] J.A.C. Rullmann, M.N. Bellido and P.Th. van Duijnen, "The active site of papain, all-atom study of interactions with protein matrix and solvent", *J. Mol. Biol.*, **206**, 101-118 (1989).

[33] G. King and A. Warshel, "A surface contrained all-atom solvent model for effective simulations of polar solutions", *J. Chem. Phys.*, **91**, 3647-3661 (1989).

[34] H. Nakamura, "Theoretical studies of electrostatic aspects of proteins". In *Recent Advances in*

*Biochemistry* (eds. S.M. Byun, S.Y. Lee and C.H. Yang), The Biochemical Society of the Republic of Korea, Seoul, pp. 29–42 (1991).

[35] W.C. Still, A. Tempczyk, R.C. Hawley and T.F. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics", *J. Am. Chem. Soc.*, **112**, 6127–6129 (1990).

[36] K. Sharp, "Incorporating solvent and ion screening into molecular dynamics using the finite-difference Poisson-Boltzmann method", *J. Comput. Chem.*, **12**, 454–468 (1991).

[37] C. Niedermeier and K. Schulten, "Molecular dynamics simulations in heterogeneous dielectrica and Debye-Hückel media - application to the protein bovine pancreatic trypsin inhibitor", *Molec. Simulation*, **8**, 361–387 (1992).

[38] C.A. Schiffer, J.W. Caldwell, R.M. Stroud and P.A. Kollman, "Inclusion of solvation free energy with molecular mechanics energy: alanyl dipeptide as a test case", *Protein Science*, **1**, 396–400 (1992).

[39] W.F. van Gunsteren and H.J.C. Berendsen, "A leap-frog algorithm for stochastic dynamics", *Molec. Simulation*, **1**, 173–185 (1988).

[40] S. Yun-yu, W. Lu and W.F. van Gunsteren, "On the approximation of solvent effects on the conformation and dynamics of cyclosporin A by stochastic dynamics simulation techniques", *Molec. Simulation*, **1**, 369–388 (1988).

[41] I. Motoc and G.R. Marshall, "van der Waals volume fragmental constants", *Chem. Phys. Lett.*, **116**, 415–419 (1985).

[42] W.F. van Gunsteren and H.J.C. Berendsen, GROMOS: Groningen molecular simulation computer program package, University of Groningen, The Netherlands (1987).

[43] K. Morikami, T. Nakai, A. Kidera, M. Saito and H. Nakamura, "PRESTO (PRotein Engineering SimulaTOr): a vectorized molecular mechanics program for biopolymers", *Computers Chem.*, **16**, 243–248 (1992).

[44] R. Wolfenden, L. Andersson, P.M. Cullis and C.C.B. Southgate, "Affinities of amino acid side chains for solvent water", *Biochemistry*, **20**, 849–855 (1981).

[45] F.C. Bernstein, T.F. Koetzle, G.J.B. Williams, E.F. Meyer Jr., M.D. Brice, J.R. Rodgers, O. Kennard, T. Shimanouchi and M. Tasumi, "The protein data bank: a computer-based archival file for macromolecular structures", *J. Mol. Biol.*, **112**, 535–542 (1977).

[46] C. Frömmel, "The apolar surface area of amino acids and its empirical correlation with hydrophobic free energy", *J. Theor. Biol.*, **111**, 247–260 (1984).

[47] G. Baumann, C. Frömmel and C. Sander, "Polarity as a criterion in protein design", *Prot. Eng.*, **2**, 329–334 (1989).

[48] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features", *Biopolymers*, **22**, 2577–2637 (1983).

[49] H.J.C. Berendsen, J.R. Grigera and T.P. Straatsma, "The missing term in effective pair potentials", *J. Phys. Chem.*, **91**, 6269–6271 (1987).

[50] J. Kemmink and T. Creighton, private communication.

[51] S. Cabani and P. Gianni, "Thermodynamic functions of hydration of saturated uncharged organic compounds. Free energies, enthalpies and entropies at 25°C", *J. Chem. Soc. Faraday Trans. I*, **75**, 1184–1195 (1979).

[52] B.H. Honig and W.L. Hubbell, "Stability of salt bridges in membrane proteins", *Proc. Natl. Acad. Sci.*, **81**, 5412–5416 (1984).

[53] I. Grötzinger and C. Frömmel, to be published.

[54] G. Vriend, "WHAT IF: a molecular modeling and drug design program", *J. Mol. Graph.*, **8**, 52–55 (1990).