Combining a polarizable force-field and a coarse-grained polarizable solvent model: Application to long dynamics simulations of bovine pancreatic trypsin inhibitor

Combining a Polarizable Force-Field and a Coarse-Grained Polarizable Solvent Model: Application to Long Dynamics Simulations of Bovine Pancreatic Trypsin Inhibitor

MICHEL MASELLA,[1] DANIEL BORGIS,[2] PHILIPPE CUNIASSE[1]
[1]Laboratoire de chimie du vivant, Service d'ingéniérie moléculaire des protéines, Institut de Biologie et de Technologies de Saclay, Commissariat à l'Energie Atomique, Centre de Saclay, 91191 Gif-sur-Yvette Cedex, France
[2]Laboratoire d'analyse et Modélisation pour la Biologie et l'Environnement - UMR 8587, Bâtiment Maupertuis, Université d'Evry-Val-d'Essonne, 91025 Evry Cedex, France

Received 30 May 2007; Revised 12 November 2007; Accepted 2 January 2008
DOI 10.1002/jcc.20932
Published online 19 March 2008 in Wiley InterScience (www.interscience.wiley.com).

Abstract: The dynamic coupling between a polarizable protein force field and a particle-based implicit solvent model is described. The polarizable force field, TCPEp, developed recently to simulate protein systems, is characterized by a reduced number of polarizable sites, with a substantial gain in efficiency for an equal chemical accuracy. The Polarizable Pseudo-Particle (PPP) solvent model represents the macroscopic solvent polarization by induced dipoles placed on mobile Lennard-Jones pseudo-particles. The solvent-induced dipoles are sensitive to the solute electric field, but not to each other, so that the computational cost of solvent–solvent interactions is basically negligible. The solute and solvent induced dipoles are determined self-consistently and the equations of motion are solved using an efficient iterative multiple time step procedure. The solvation cost with respect to vacuum simulations is shown to decrease with solute size: the estimated multiplicative factor is 2.5 for a protein containing about 1000 atoms, and as low as 1.15 for 8000 atoms. The model is tested for six 20 ns molecular dynamics trajectories of a traditional benchmark system: the hydrated Bovine Pancreatic Trypsin Inhibitor (BPTI). Even though the TCPEp parameters have not been refined to be used with the solvent PPP model, we observe a good conservation of the BPTI structure along the trajectories. Moreover, our approach is able to provide a description of the protein solvation thermodynamic at the same accuracy as the standard Poisson-Boltzman continuum methods. It provides in addition a good description of the microscopic structural aspects concerning the solute/solvent interaction.

© 2008 Wiley Periodicals, Inc. J Comput Chem 29: 1707–1724, 2008

Key words: polarizable force-field; coarse-grain approach; mesoscopic solvent; protein solvation; molecular dynamics

Introduction

In the field of biomolecular modeling, the development of so-called "second generation" force fields giving, in particular, a better description of the charge distribution and of the various charge flows occuring within a complex molecule, constitute a very active area of research. Various approaches to electronic charge redistribution and local electronic polarizability effects are followed by different groups: chemical potential equalization principle with variable atomic charges and dipoles,[1,2] drude oscillator models,[3,4] atomic induced dipole representation (cf. ref. 5 and the ff02 force-field cited in[6]), induced multipole expansions,[7] or frozen electronic density approaches.[8] In this context, two of us have proposed recently the TCPEp force field, in which induced dipoles are placed on a reduced number of atomic sites. This approach was shown to lead to substantial gain in efficiency at an equivalent level of chemical accuracy. It was shown also to provide an accurate description of proteins ionic chelation.[9–11]

In polarizable approaches, the solvent is treated usually at an equivalent all-atom level, in terms of a n-polarizable sites water model. Even more so than with the standard nonpolarizable force field, the integration of the water degrees of freedom constitutes an important limitation for long molecular dynamics simulations. For nonpolarizable force fields, several efficient

Correspondence to: M. Masella; e-mail: michel.masella@cea.fr

implicit solvent methods have been proposed to overcome that difficulty, among which may be quoted the Langevin dipole model,[12–15] the distance-dependent dielectric model,[16–18] and the Generalized-Born/Surface Area (GB-SA) method.[19–21] The latter method can be considered as efficient although approximate implementations of the widely used Poisson-Boltzmann equation.[22,23] Among all those methods, to our knowledge, only the Langevin Dipole approach, in its so called Protein Dipole/Langevin Dipole version,[13,15] is directly compatible with a polarizable force field. In a Poisson-Boltzmann formalism, the atomic electronic polarizability can be related to the local internal dielectric constant through the Clausius-Mossotti relation, but this relation does not seem convertible into a consistent analytical force-field expression that can be used for dynamics.

Recently, one of us and his collaborators have proposed an alternative strategy based on a semi-implicit, particle-based representation of the solvent.[24] Although remaining in the context of macroscopic density functional theory, the method does retain the notion of solvent particles; pseudo-particles are introduced that interact among themselves and with the solute atoms with a simple short-range potential. Using configuration space sampling techniques, such as molecular dynamics, it is possible to generate a series of different, sparse and disordered grids, whose nodes correspond to the particle centres, and which can be used to estimate any macroscopic free-energy functional (in particular, a polarization free-energy functional). The functional can then be minimized " on the fly" on those grids, and the resulting free-energy, averaged over different grids, provide the electrostatic component of the solvation free energy. These fictitious particles are a priori introduced to generate random grids, and to define the solute–solvent boundaries in a natural way. Their microscopic interaction properties can be thus defined arbitrarily. In the original paper of Haduong et al.,[24] the particles were treated as Lennard-Jones particles with two microscopic parameters (σ and ε) which can be related to two macroscopic properties of the solvent considered, for example compressibility and liquid–gas surface tension, in addition to the correct bulk dielectric constant, imposed by the electrostatic model. In the simplest version of the model, based on approximate local polarization free-energy functional, the pseudo-particles induced dipoles are sensitive to the solute electric field but not to each other, so that the solvent particles only see each other through a short-ranged Lennard-Jones potential.

The adequacy, accuracy, and numerical efficiency of this semi-implicit model, that has been called polarizable pseudo-particle (PPP) solvent model, has already been discussed and demonstrated in the case of proteins[25] and of nucleic acids.[26] Until now, only a classical pairwise atomistic force-field has been combined with this solvent model.[24–26] However, this class of force-fields, involving only fixed charges interacting via a classical Coulombic potential, are known to present strong deficiencies, especially in the presence of external charges, since the charge distribution inside a molecule depends on its environment. In this sense, polarizable force-fields represent to date a major improvement. In particular, they were shown to provide an accurate description of interactions with ionic species.[27]

In this article we present, for the first time to our knowledge, the dynamical coupling of a polarizable protein force field to an implicit solvent model. This concerns an updated version of the TCPEp force field[9] and the PPP solvent model described above. Compared to the original paper of Basdevant et al.,[25] the parameterization strategy has been modified: the short range potentials handling the solvent/solute interactions are now adjusted to reproduce the experimental hydration free energies of a learning set of organic molecules. Furthermore, the solute and solvent induced dipoles are treated consistently using a multiple time-step algorithm introduced recently.[28] The numerical efficiency of the multiple time-step algorithm with respect to a dynamical Car-Parrinello-like algorithm, and of the solvation model with respect to simulations *in vacuo*, will be both assessed. Finally, the model is tested by performing six 20 ns long MD simulations of a small protein, the Bovine Pancreatic Trypsin Inhibitor, which is in the literature a standard benchmark for assessing the stability of a force field and/or solvation model.[18,29,30] In the present report, we have mainly focused the trajectory analysis on evaluating the ability of our solvation model to reproduce realistic solvation free energies, as well as to provide reliable microscopic structural informations regarding the solvent/solute interactions. Regarding the protein, only basic structural properties concerning the protein backbone are discussed here.

Theoretical Methods

The Revised Polarizable TCPEp Model

We have recently developed a new many-body polarizable force-field for proteins, called TCPEp.[9,10] In our approach, in contrast to all other polarizable models, only heavy atoms are considered as polarizable centres, so that the number of induced dipoles to be determined self-consistently is drastically reduced. The deficiencies introduced by this approximation are corrected by damping the electric field at short distances and by introducing short-range many-body functions to model hydrogen bonds. Since the original article several changes have been introduced to improve TCPEp. First, the analytical form of the energy term modeling the interatomic repulsive interaction has been altered, as follows

$$U^{\text{rep}} = \sum_{ij,i\neq j}^{N} A_{ij}\exp(-B_{ij}r_{ij}) \qquad (1)$$

This does not lead to any major improvement of the model accuracy, but that new repulsive term agrees with a classical definition of repulsive interactions. Several improvements were also introduced in the TCPEp polarisation term. The polarization energy was originally defined as

$$U^{\text{pol}} = \sum_{i=1}^{N_{\text{pol}}} \frac{\mathbf{p}_i^2}{2\alpha_i} - \sum_{i=1}^{N_{\text{pol}}} \mathbf{p}_i \cdot \mathbf{E}_i^0 - \frac{1}{2}\sum_{ij,i\neq j}^{N_{\text{pol}}} \mathbf{p}_i T|\mathbf{r}_i - \mathbf{r}_j|\mathbf{p}_j \qquad (2)$$

Here, α_i, $T|\mathbf{r}_i - \mathbf{r}_j|$ and $\mathbf{E}_i^0$ correspond to the polarizability, the dipolar tensor and the electric field generated by the static charges and acting on the polarizable center i, respectively. The induced dipole moments are noted $\mathbf{p}_i$. Both the static electric field and the dipolar tensor are damped at short distances to prevent the so-called "polarization catastrophe." However, the damping function employed in the original version of TCPEp turned out to be inadequate for molecular dynamics simulations. This deficiency was overcome following the ideas of B.T. Thole,[5] by introducing a new damping term related to the following charge distribution

$$\rho(r) = \frac{3a}{4\pi} \times \exp(-ar^3) \tag{3}$$

In that case, the electric field and the dipole tensor have to be altered according to

$$\mathbf{E}_i^0 = \frac{1}{4\pi\varepsilon_0} \sum_{j=1, j\neq i}^{N} \lambda_3 q_j \frac{(\mathbf{r}_i - \mathbf{r}_j)}{|\mathbf{r}_i - \mathbf{r}_j|^3} \tag{4}$$

$$T|\mathbf{r}_i - \mathbf{r}_j| = 3\lambda_5 \frac{\mathbf{r}_i \cdot \mathbf{r}_j}{|\mathbf{r}_i - \mathbf{r}_j|^5} - \lambda_3 \frac{\tilde{\mathbf{1}}}{|\mathbf{r}_i - \mathbf{r}_j|^3} \tag{5}$$

here, $\tilde{\mathbf{1}}$ is the unit dyadic matrix, and λ_3 and λ_5 are defined as

$$\lambda_3 = 1 - \exp\left(-ar_{ij}^3\right) \tag{6}$$

$$\lambda_5 = 1 - \left(1 + ar_{ij}^3\right) \exp\left(-ar_{ij}^3\right) \tag{7}$$

The parameter a depends on the nature of the atom pair ij and was originally introduced to reproduce the experimental polarizabilities of a set of small molecules.[5] More recently, new strategies have been proposed in which interaction properties are also taken into account.[31]

In contrast to the Thole's model and to most polarizable force fields for proteins proposed so far, the interactions among polarizable centres in the so-called 1–2 and 1–3 positions are excluded in the TCPEp approach. Hence, only energetic considerations can be taken into account to derive the a parameters. a is taken equal to 0.3 for pairs of common atoms, and to 0.5 for pairs involving hydrogen. With such values, the incidence of the damping on the static electric field and on the dipole tensor is almost neglectible for interatomic distances corresponding to non-bonded atoms.

For atom pairs involving monoatomic ions, the parameter a was assigned to carefully reproduce both the interaction energies and geometries of small aggregates (in that case, a range from 0.1 to 0.5).

Lastly, both to be consistent with the mesoscopic solvent approach described below and to improve convergence, the induced dipole moments are now allowed to saturate in TCPEp, according to

$$\mathbf{p}_i = \frac{\mu^{\text{sat}}}{E_i} L\left(\frac{3\alpha_i E_i}{\mu^{\text{sat}}}\right) \mathbf{E}_i \tag{8}$$

$L(x) = \coth x - \frac{1}{x}$ denotes the Langevin function. At this stage of the TCPEp development, the dipole saturation value is fixed to a uniform and high value of 5 Debye, whatever the polarizable atomic center (this choice could be modified in further development).

Hence, after convergence, the polarization energy term considered now in TCPEp is defined as[24]

$$U^{\text{pol}} = -\frac{\mu^{\text{sat}}}{3\alpha_i} \sum_{i=1}^{N_{\text{pol}}} \ln\left(\frac{\sinh(3\alpha_i E_i/p_i)}{3\alpha_i E_i/p_i}\right) + \sum_{i=1}^{N_{\text{pol}}} \mathbf{p}_i T|\mathbf{r}_i - \mathbf{r}_j|\mathbf{p}_j \tag{9}$$

One of the major advantage of the TCPEp approach is its computational efficiency. As recently shown, it is well suited to be used in conjunction with a multiple time step procedure like r-RESPA.[28] In that new scheme, called r-RESPAp, the induced dipoles are computed using an iterative self-consistent field (SCF) method and the speed up factor provided by this integration scheme depends on the number of iterations, n_{SCF}, needed to achieve the SCF convergence. With a n_{SCF} value smaller than 8, MD simulations can be more than twice as fast as when performed with a Car-Parinello-like algorithm to handle dynamically the induced dipole moments.[28]

From our own computations, it appears that damping the electric field at short distances and allowing the dipoles to saturate permits to noticeably reduce the number of iterations necessary to achieve the SCF dipole convergence. Moreover, that procedure also improves the stability of long time molecular dynamics simulations substantially.

Furthermore, in TCPEp, only the atoms belonging to polar or charged groups generates the static electric field $\mathbf{E}_i^0$. This leads to particularly weak interactions among nonpolar groups such as alkyl chains or phenyl groups. To slightly reinforce those interactions, a dispersive terms has been introduced in the form

$$U^{\text{disp}} = -\sum_{i=1}^{N_{np}} \frac{D_{ij}}{r_{ij}^6} \tag{10}$$

The sum runs on nonpolar carbon atoms, and the U^{disp} parameters were defined to reproduce the interactions energies of small aggregates, such as the benzene dimer. All of the remaining TCPEp parameters have been newly refined, using the strategy proposed in ref. 9. Lastly, the intramolecular potential energy term, handling the stretching, bending and torsional energies, corresponds to that of CHARMM v. 22.[32]

In all the following, a polarizable solute is characterized by an atomic static charge set $\{q_i\}$ and by a set of N_{pol} polarizable centers, of polarizability $\{\alpha_i\}$, with induced dipoles labeled $\{\mathbf{p}_i\}$. The dielectric permittivity is set to 1 inside the solute boundaries, and the electric field generated outside is given by

$$\mathbf{E}_S(\mathbf{r}) = \sum_{i=1}^{N} q_i \lambda_3 \frac{\mathbf{r} - \mathbf{r}_i}{|\mathbf{r} - \mathbf{r}_i|^3} - \sum_{i=1}^{N_{\text{pol}}} \left(3\lambda_5 \frac{(\mathbf{p}_i \cdot \mathbf{r}_i)\mathbf{r}}{|\mathbf{r} - \mathbf{r}_i|^5} - \lambda_3 \frac{\mathbf{p}_i}{|\mathbf{r} - \mathbf{r}_i|^3}\right) \tag{11}$$

The PPP Solvent Model

Here, we briefly describe the rationale of the PPP solvent model introduced recently for the efficient simulation of proteins and nucleic acids.[24–26] In a continuum dielectric approach, the solute is immersed in a continuous solvent characterized by a dielectric permittivity $\varepsilon(\mathbf{r})$ (equal to ε_s outside the solute boundaries), by an electric susceptibility $\chi(\mathbf{r}) = (\varepsilon(\mathbf{r}) - 1)/4\pi$, and by a polarization density $\mathbf{P}(\mathbf{r})$. If the solute generates an electric field $\mathbf{E}_S(\mathbf{r})$ outside its boundaries, the solvation electrostatic problem can be treated in terms of the polarization free-energy functional[33]

$$F_P[\mathbf{P}] = \int \frac{\mathbf{P}(\mathbf{r})^2}{2\chi(\mathbf{r})} d\mathbf{r} - \int \mathbf{P}(\mathbf{r}) \cdot \mathbf{E}_s(\mathbf{r}) d\mathbf{r} - \frac{1}{2} \iint \mathbf{P}(\mathbf{r}) T(\mathbf{r} - \mathbf{r}') \mathbf{P}(\mathbf{r}') d\mathbf{r} d\mathbf{r}' \qquad (12)$$

T is here the dipolar tensor. As the long distance mutual induction of the solvent polarization densities are accounted for by the last term in Eq. (12), this functional is termed as "non-local." Minimizing F_P with respect to $\mathbf{P}(\mathbf{r})$ yields the thermodynamics equilibrium and the following fundamental equations

$$\mathbf{P}(\mathbf{r}) = \chi(\mathbf{r}) \mathbf{E}(\mathbf{r}) \qquad (13)$$

where $\mathbf{E}(\mathbf{r}) = \mathbf{E}_S(\mathbf{r}) + \mathbf{E}_P(\mathbf{r})$, the second term being defined as the polarisation electric field:

$$\mathbf{E}_P(\mathbf{r}) = \int T(|\mathbf{r} - \mathbf{r}'|) \mathbf{P}(\mathbf{r}') d\mathbf{r}' \qquad (14)$$

Those two equations can be shown to be equivalent to the well-known Poisson equation, and they have obviously to be solved self-consistently.

An approximate free-energy functional can be derived from F_P by considering that the solvent polarization density $\mathbf{P}(\mathbf{r})$ keeps a longitudinal character, i.e. may be written as the gradient of an arbitrary function $f(\mathbf{r})$. In that case, it can be shown that $\mathbf{E}_P(\mathbf{r}) = -4\pi \mathbf{P}(\mathbf{r})$, and this leads to the "local" functional[34]

$$F_P[\mathbf{P}] = \frac{1}{2} \int \frac{\varepsilon(\mathbf{r}) \mathbf{P}(\mathbf{r})^2}{\chi(\mathbf{r})} d\mathbf{r} - \int \mathbf{P}(\mathbf{r}) \cdot \mathbf{E}_S(\mathbf{r}) d\mathbf{r} \qquad (15)$$

The non-local dipole–dipole term in eq. (15) now enters in the definition of an effective local susceptibility. At thermodynamic equilibrium

$$\mathbf{P}(\mathbf{r}) = \frac{\chi(\mathbf{r})}{\varepsilon(\mathbf{r})} \mathbf{E}_S(\mathbf{r}) \qquad (16)$$

so that the solution is analytical, with no need of self-consistent iterations. It was proved that such a local approximation turns out to be very accurate for describing the solvation energy of complex molecules if one accept to simply renormalize the solute permanent charges, according to a rigourous variational principle. More pragmatically, the charge renormalization can be done as to reproduce exactly the electrostatic solvation free energy of charged and dipolar solutes.[35]

Recently, Haduong et al.[24–26] have proposed an original implementation of this local approximation in classical MD schemes. They introduced a set of N_s polarizable pseudo-particles, with an accessible volume per particle $\Delta\upsilon = \rho^{-1}$ and an intrinsic polarizability α_s. Relating the particle induced dipoles to the solvent polarization density according to $\mathbf{p}_i^s = \Delta\upsilon \mathbf{P}(\mathbf{r})$, a discrete version F_{dP} of the continuous functional F_P can be readily derived as follows

$$F_{dP} = \sum_{k=1}^{N_s} \frac{\mathbf{p}_k^{s\,2}}{2\alpha_s} - \sum_{k=1}^{N_s} \mathbf{p}_k^s \cdot \mathbf{E}_S'(\mathbf{r}_k) \qquad (17)$$

here, $\mathbf{E}_S'$ represents the renormalized solute electric field, and the effective polarizability α_s is defined by

$$\alpha_s = \frac{\varepsilon_s - 1}{4\pi\rho\varepsilon_s} \qquad (18)$$

At equilibrium, the induced dipoles located at the particle centres obey

$$\mathbf{p}_k^s = \alpha_s \mathbf{E}_S'(\mathbf{r}_k) \qquad (19)$$

and the resulting solvent polarization energy is equal to

$$F_{dP} = -\frac{1}{2} \sum_k \alpha_s \mathbf{E}_S'(\mathbf{r}_k)^2 \qquad (20)$$

The model can be further refined by allowing for a Langevin dipole saturation, which prevents the local polarization to go beyond microscopically realistic values. This is done by replacing the two previous relations by

$$\mathbf{p}_i^s = \mu_s \mathcal{L}\left(\frac{3\alpha_s E_{Sk}'}{\mu_s}\right) \frac{\mathbf{E}_S'(\mathbf{r}_k)}{E_{Sk}'} \qquad (21)$$

$$F_{dP} = -\frac{\mu_s^2}{3\alpha_s} \sum_{k=1}^{N_s} \ln\left[\frac{\sinh(3\alpha_s E_{Sk}'/\mu_s)}{3\alpha_s E_{Sk}'/\mu_s}\right] \qquad (22)$$

The discrete functional F_{dP} can be easily introduced in MD schemes by considering the hamiltonian H

$$H = \sum_{i=1}^{N} \left(\frac{1}{2} m_i \dot{\mathbf{r}}_i^2 + \Phi(\mathbf{r}_i)\right) + \sum_{k=1}^{N_s} \left(\frac{1}{2} m_s \dot{\mathbf{r}}_k^2 + \frac{1}{2} \sum_{l \neq k} \phi_s(r_{kl})\right) + \sum_{ik} \phi_a(r_{ik}) + F_{dP} \qquad (23)$$

The first two terms of H correspond to the Hamiltonian handling the interactions among the solute atoms alone and among the solvent pseudo-particles alone, respectively, whereas the last two terms correspond to the interactions between the solute atoms and the solvent pseudo-particles. In practice, ϕ_a is introduced to define a solute volume unreachable to the pseudo-particles. The associated surface could be used to compute the hydrophobic part of the solvation free-energy not accounted for in F_{dP} (this is not done at this stage;

see below). Both the ϕ_a and ϕ_s are taken as Lennard-Jones-like potentials, for computational efficiency reasons. Lastly, the pseudo-particles dipoles are computed according to the Born-Oppenheimer approximation: they relax instantaneously to their equilibrium value as the particles evolve. They are thus taken at "orientational" thermodynamics equilibrium while they explore their translational phase space.[24]

In the local approximation, there is no need to compute the induced-dipole/induced-dipole interactions for the solvent and thus the solvent pseudo-particles see each other as mere Lennard-Jones particles. The computational time needed to compute their interactions is small compared to explicit solvent approaches. As compared to implicit solvation models, a particle-based description presents also several advantages, as it permits to investigate the dynamics of a solute in solution in a consistent way, with well-established all-particle algorithms. Numerical efficiency will be discussed in more details further on.

The solvation free-energies can be estimated by averaging the value of F_{dP} along the trajectories, which can be interpretated as averaging the results obtained on a set of sparse and disordered grid. Obviously, the system under consideration has to be fully relaxed before collecting the data for statistical averages. For nonpolarizable systems, the solvation free-energy are estimated directly by averaging F_{dP} along a trajectory. In the case of polarizable systems, the average F_{dP} value has to be corrected by $\Delta G_{v\rightarrow s}$ to account for the incidence of the solvent on the solute induced dipole moments. This is achieved by performing the difference between the polarization energy U^{pol} of the solute in solution and *in vacuo*.

Furthermore, the model was parametrized in ref. 25 for a nonpolarizable solute, and it was shown there that a local dipole renormalization factor of 1.185 (close to the theoretical value of $\sqrt{3/2} = 1.225$ for sharp dielectric boundaries) was appropriate. This parametrization has to be examined again in the case of a polarizable solute. This will be done in Solvent Parameterization of TCPEp.

Efficiently Mixing the Polarizable Solute and the Coarse-Grained Polarizable Solvent

Principles

The details of the potential ϕ_s and ϕ_a entering in the definition of the Hamiltonian H [eq. (23)] will be discussed in the next section. We begin by describing how the polarizable TCPEp force field can be efficiently mixed with the PPP solvent model. It should be noted first that, in the local version described above, the solvent induced dipoles ignore each other and are determined solely by the solute permanent charges and induced dipoles through eqs. (11) and (21). The variables to be optimized are thus the solute dipoles only. In a practical implementation, however, it is easier and more efficient to conserve the coupled optimization equations for the $\mathbf{p}_i$ and the $\mathbf{p}_k^s$ [eqs. (8) and (21)]. With respect to the regular self-consistent procedure for polarizable atoms, one has simply to set the dipolar tensor between solvent particles to zero, and renormalize the solute permanent charge when computing solute-solvent interactions. The solvent induced dipoles appear thus explicitly as "slave" variables. In our implementation, there is no cut-off for solute-solute electrostatic interactions, and we have introduced a smooth cut-off between 11.5 and 12 A for solute-solvent interactions.

The Propagator: The r-RESPAp Scheme

To date, the most efficient procedure to handle the induced dipoles in MD simulations is a Car-Parinello like procedure (CP), where the dipoles are considered as adiabatic dynamic variables. Recently, we have introduced a multiple time steps algorithm for polarizable force-fields based on induced dipoles, derived from the r-RESPA procedure.[36] r-RESPA is based on the splitting the forces $\mathbf{F}_i$ acting on each particle i in a set of components, which vary in time more and more slowly

$$\mathbf{F}_i = \mathbf{F}_{fv,i} + \mathbf{F}_{sr,i} + \mathbf{F}_{l,i} \tag{24}$$

In the case of classical pairwise force-fields, the component $\mathbf{F}_{fv,i}$ corresponds to the fast varying terms (i.e. the forces occurring among covalently bonded atoms, the so-called stretching and bending terms), $\mathbf{F}_{sr,i}$ to the short range forces occurring among nonbonded atoms (i.e. the torsional, the repulsive and the short range electrostatic forces), and $\mathbf{F}_{l,i}$ to the long range electrostatic forces. The computationally most expensive forces correspond to the latter ones. In the r-RESPA scheme, the most expensive forces $\mathbf{F}_{l,i}$ are thus computed less often than the others, which speeds up the computations by an important factor compared to updating all forces at each time step.

In classical force-fields, the electrostatic interactions are handled using a classical Coulombic pairwise term. Because of its pairwise nature, the Coulombic term can be easily split into long and short range components: the short one corresponds to atoms that are closer than a reference distance R_{ref}, and the opposite for the long component. Because of the nonpairwise character of the polarization phenomena, the latter splitting cannot be applied to a polarizable force-field based on induced dipole moments (such as TCPEp). Recently, we have proposed a generalization of the r-RESPA scheme for that case by defining a set of "short-range" dipole moments $\{\mu_i^{sr}\}$, solution of

$$\mu_i^{sr} = \alpha_i \left(\mathbf{E}_i^{sr} + \sum_{i \neq j, r_{ij} < R_{\mathrm{ref}}} T_{ij} \mu_j^{sr} \right) \tag{25}$$

The last two terms correspond here to the electric fields generated by the static charges and by the dipole moments located both at a shorter distance than R_{ref} from the i^{th} polarizable center. This dipole set is associated to a short range polarization energy based on eq. (9), from which the short range component $\mathbf{F}_{sr,i}^{\mathrm{pol}}$ of the polarization forces can be computed. The full polarization forces $\mathbf{F}_i^{\mathrm{pol}}$ are computed from the "exact" dipole set $\{\mu_i\}$, which is the solution of the above equation without any distance limit. The long range polarization forces can be then readily computed according to

$$\mathbf{F}_{l,i}^{\mathrm{pol}} = \mathbf{F}_i^{\mathrm{pol}} - \mathbf{F}_{sr,i}^{\mathrm{pol}} \tag{26}$$

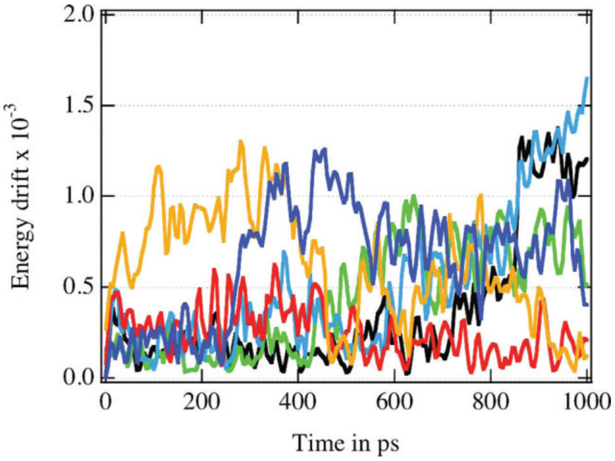

Figure 1. Time evolution of the dimensionless index $\Delta\bar{E}$ (the results corresponding to the trajectories labeled **T1** to **T6** are shown respectively in black, red, orange, green, blue and dark blue).

As both the short range and the "exact" dipole sets are computed using a SCF procedure, the efficiency of the r-RESPAp scheme depends obviously on the number of iterations SCF needed to achieve the convergence of the "exact" dipoles (and thus on the convergence criterion chosen). However, it depends also on the ratio $\beta = N_{\text{pol}}/N$, with N_{pol} the number of polarizable centers of the molecular system. For n_{SCF} values ranging from 6 to 10 and a β value smaller than an half, speed up factors greater than two can be obtained with respect to the CP scheme. For a typical bio-organic molecule, about half of the atoms are considered as polarizable in the TCPEp approach, which suggests that combining TCPEp with the r-RESPAp scheme can provide an efficient method to perform MD simulations of such systems. This is proved below.

Numerical Stability and SCF Convergence Criterion

The numerical stability of an integration scheme used to solve the newtonian equations of motion is commonly assessed by monitoring the conservation of the system total energy. To this end, the dimensionless index

$$\Delta\bar{E} = \frac{1}{N_{\text{iter}}}\sum_{i=1}^{N_{\text{iter}}}\left|\frac{E^{\text{tot}} - E^{\text{tot}}(0)}{E^{\text{tot}}(0)}\right| \qquad (27)$$

can be used.[37] In particular, it is assumed that a value of $\Delta\bar{E}$ smaller than 3.10^{-3} gives an acceptable numerical accuracy for, typically, a MD run of 100 ps.

In the case of methods based on iteratively solving the induced dipoles, most of the available studies[28,38] concluded that a tight SCF criterion is needed to maintain a good conservation of the system total energy along a trajectory. In our previous article presenting the r-RESPAp procedure,[28] a very good energy conservation was obtained by considering a mean convergence criterion per dipole of 10^{-6} Debye, with the iterations continuing until the largest difference between two successive iterations for a single dipole was smaller than five times that value - such a criterion is noted $(5, 10^{-6})$ Debye.

To reduce the number of iterations needed for SCF convergence in our r-RESPAp scheme, the initial guesses for the dipole moments at each time step are provided using efficient predictors for the "short range" dipoles. In the present study, we also used a more efficient SCF algorithm for the "exact" dipoles, where the "short range" and the "exact" ones are alternatively considered during the iterative procedure (the details of this algorithm will be presented elsewhere). This leads to increase the number of evaluations of the short range dipoles, but the number of evaluations of the "exact" dipoles is usually halved. Lastly, to maintain a good efficiency of our MTS scheme, we considered here a slightly less stringent criterion of $(10, 10^{-6})$ Debye, which is usually reached after 8 iterations during our MD runs. Such a criterion permits one to still keep a good conservation of the total energy, as shown by computing the dimensionless index $\Delta\bar{E}$ on a 1 ns time window along the six trajectories of solvated BPTI that will be discussed below. Those runs were generated using the r-RESPAp scheme (the short range dipoles are computed every 1 fs and the exact ones every 5 fs). The energy fluctuations $\delta E(t) = |(E^{\text{tot}} - E^{\text{tot}}(0))/E^{\text{tot}}(0)|$ are plotted vs time in Figure 1. The total energy of the six trajectories is well preserved on the nanosecond time scale: the computed $\Delta\bar{E}$ values are all within 0.2 and 0.7×10^{-3} after 1 ns. However, along the full 20 ns trajectories, the total energy appears to be slowly drifting downwards at a rate smaller than 0.005 kcal mol^{1} ps^{-1}, which represents a total energy drift of 100 kcal mol^{-1} after 20 ns (value to be compared to the total potential energy along the trajectories: about $-10, 100$ kcal mol^{-1}).

Algorithm Efficiency

In Figure 2, the relative CPU times needed to perform short MD simulations of solvated proteins using our mesoscopic solvent model and three integration schemes (the r-RESPAp, the CP and the full iterative scheme, based on a Verlet algorithm with the "exact"

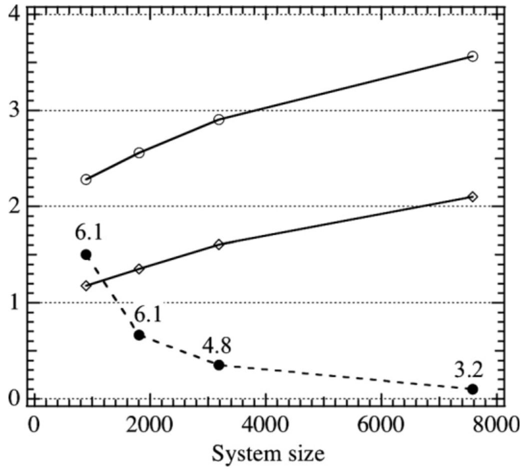

Figure 2. CPU ratio between the Car-Parinello like procedure and the r-RESPAp scheme (empty lozanges) and between the standard Verlet algorithm and the r-RESPAp scheme (empty circles). Full circles: computational cost R_s of our solvation approach vs the protein size (the ratio between the number of solvent pseudo-particles and the number of the protein atoms is also shown).

dipoles computed at each time step) are plotted versus the atomic size of the proteins (namely, the bovine trypsin inhibitor BPTI, the phospholypase A2 from Naja-Naja sagittifera, the streptomyces griseus trypsin and a human phosphatase, their respective PDB code names are 1BPI, 1MF4, 1SGT and 1EW2). These proteins possess 58, 125, 223, and 475 residues, respectively. During those short runs, the number of iterations needed to achieve the dipole SCF convergence was fixed to 12 ("short-range" dipoles) and 8 ("exact" dipoles) when using r-RESPAp, and to 12 with the full iterative scheme. Theses values correspond to those needed to reach a SCF criterion of $(10, 10^{-6})$ Debye for solvated BPTI.

Even if r-RESPAp is the most efficient integration scheme among those tested, it only permits speed up factors of about 20% compared to CP in the case of the small protein BPTI. However, its efficiency is improved as the system size increases: in the case of the phosphatase alcaline 1EW2, a speed-up factor of about 2.1 is observed compared to CP. Those factors are smaller than those previously reported when applying the r-RESPAp and the CP schemes to a water aggregate containing 1000 molecules (they were shown to range from 66% to 90 %).[28] The reduced efficiency of r-RESPAp observed in the present study results from the fact that: (1) the ratio between the number of polarizable centers and the total number of atoms and solvent pseudo-particles is greater than 0.5; and (2), the interactions among the solvent pseudo-particles and the protein atoms are cut off for distances greater than 12 Å.

In Figure 2, the computational overhead R_s introduced by our coarse-grained solvent approach is also plotted as function of the size of the solvated protein. It is defined as

$$R_s = \frac{(t_{\text{solv}} - t_{\text{vacuo}})}{t_{\text{vacuo}}} \tag{28}$$

where, t_{vacuo} and t_{solv} are the cpu times needed to perform a short MD run of a given protein, using the r-RESPAp scheme, *in vacuo* and in aqueous phase, respectively. Regarding the solvated case, the protein is complemented with solvent particles in a cubic box defined such that the distance between the box edges and any protein atoms is at least of 12 Å (this is refered as the size condition in the following). The ratio between the number of solvent particles needed to meet that condition (with a solvent density of 1.01 g cm^{-3}) and the number of protein atoms are reported for reference in Figure 2. The solvation computational cost decreases as the size of the protein increases: compared to vacuum computations, a simulation in our water model is 2.5 times more expensive for BPTI (892 atoms), whereas this factor drops to 1.15 for 1EW2 (7572 atoms). This is simply explained by the fact that the number of solvent particles needed to meet the size condition defined above is proportionally smaller as the protein size increases: for instance, the ratio between the number of particles and the protein size is twice smaller for 1EW2 than for BPTI. The observed speed-up factors are quite competitive with respect to those quoted for fast implicit solvent methods such as the distance-dependent dielectric model, or the Generalized Born method.

To conclude, the coupling of the PPP solvent model, the TCPEp force-field and the r-RESPAp procedure leads to a particularly efficient method to investigate the properties of a polarizable macromolecule in solution. For instance, the MD runs discussed below concerning solvated BPTI (892 atoms and 5441 pseudo-particles) were performed at a rate of 600 ps per day on a single processor Xeon-based workstation.

Solvent Parameterization for TCPEp

The parametrization of the PPP solvent was introduced in refs. 24–26 for a nonpolarizable solute, and it has to be adjusted for the TCPEp force-field in order to accomodate for the solute polarizability. This concerns the value of the local charge renormalization factor λ_q implied by the local approximation described above, and also the solute–solvent potential ϕ_a.

To test the new parametrization, we have carried out a series of MD simulations for various ions and organic molecules dissolved in the PPP solvent. All those preliminary MD runs were performed in the NVT ensemble at 300 K using cubic periodic conditions. The temperature was monitored using a second order GGMT scheme,[39] with a coupling constant of 100 fs. The solutes were embedded at the center of a solvent cubic box, whose dimensions were $24.76 \times 24.76 \times 24.76$ Å^3. The solvent was set at a density of 1.01 g cm^{-3}, and all solvent particles located at less than 1.8 and 2.8 Å respectively from any solute hydrogen and heavy atom were discarded at the beginning of the run. The interactions between the solute and its periodic images were ignored, and the solute/solvent electrostatic interactions were smoothly cut between 11.5 and 12 A. The equations of motion were solved using the generalized multiple time steps r-RESPAp procedure[28] (the short-range electrostatic and polarization forces corresponding to interactions among atoms located at less than 6 Å from each other). The covalent bonds involving hydrogens were restraint during the simulations using the SHAKE algorithm.[40] The data collection along the trajectories starts after an equilibration period of 60 ps.

Local Dipole Renormalization Factor

For a nonpolarizable, polar spherical solute, the local approximation can be shown to underestimate the electrostatic solvation energy by a factor 2/3. This can be corrected by renormalizing the solute dipole by a factor $\lambda_q = \sqrt{3/2} \simeq 1.225$. In their original paper, Haduong et al.[24] thus proposed to correct this systematic error for complex solutes by scaling the electrostatic charges of polar groups by the same factor λ_q. The numerous tests they performed led to the conclusion that a slightly smaller λ_q factor of 1.185 (accounting for the dielectric boundary smoothness) provides more accurate results, in terms of both structural stability and free energy correlations with continuum solvent approaches. It is shown in Appendix A that, for a polar, polarizable solvated hard sphere, a similar λ_q factor can be defined. It depends on the solute polarizability and ranges from 1.34 to 1.44 for typical polarizability values. An approach similar to that of Haduong et al. was thus considered to overcome the deficiencies of the local model for polar groups: their charges were scaled by a factor λ_q, which has to depend, *a priori*, on the nature of the groups. For simplicity we opted for a unique average value of 1.36. Note that this effective charge set is only taken into account when computing the solute/solvent interactions, and not for the interactions among the solute itself.

The Short Range Potentials

As in Refs. 24–26, the ϕ_s potential handling the interactions among the solvent pseudo-particles is taken as a standard Lennard-Jones potential, with the following parameters: $\varepsilon = 0.766$ kcal mol^{-1} and $r^* = 2.88$ Å. The mass and the dipole saturation of the solvent particles are set to 18 g mol^{-1} and to 1.5 Debye. The solvent density ρ_s is taken equal to 1.01 g cm^{-3} (the incidence of ρ_s on the model will be briefly discussed further on). With those parameters, the self-diffusion coefficient for the solvent particles is computed to be 1.810^{-4} cm^2 s^{-1} at 300 K, which is about an order of magnitude greater than that of the pure water at 300 K (2.410^{-5} cm^2 s^{-1}). It is important to emphasize that, with respect to an explicit solvent description, the solvent dynamics is greatly accelerated and thus allows for a much more efficient exploration of the solute phase space.

In the original article of Haduong et al.,[24] ϕ_a corresponds to a classical Lennard-Jones potential. However, it appeared that such a potential cannot define hard enough boundaries to prevent polarization catastrophes between the solute and the solvent, especially for charged solutes. The incidence of the repulsive wall hardness of ϕ_a on the prediction quality of the local model has been investigated by testing three ϕ_a^n potentials derived from the Lennard-Jones one according to

$$\phi_a^n(r_{ij}) = \left(\frac{r_{ij}^*}{r_{ij}}\right)^{6n} - n\left(\frac{r_{ij}^*}{r_{ij}}\right)^6 \tag{29}$$

The classic Lennard-Jones potential corresponds to $n = 2$, and the minimum of ϕ_{ij}^n is located at r_{ij}^*, with $\phi_a^n(r_{ij}^*) = -\varepsilon_{ij}/(n-1)$. The potentials tested correspond to $n = 2, 3$, and 4 respectively. For a complex solute such as the bovine pancreatic inhibitor BPTI, the most promising results in terms of structural stability were obtained with ϕ_a^4. This is merely due to the much stronger repulsive wall of this potential, which prevents a sizeable part of the solvent particles to diffuse inside the solute and to disrupt critical interactions among the solute sub-elements (such as the protein intramolecular hydrogen bond network).

To understand the improvement introduced by ϕ_a^4 compared to a standard Lennard-Jones term, we consider the theoretical results for a monoatomic dication, with a diameter r^* of 3.1 Å. The dication solvation was investigated using the local model and the three ϕ_a^n potentials. MD simulations were performed according to the protocol described at the beginning of this section. The solvent structure is described in terms of radial distribution functions $g^n(r)$, whose profiles, together with the corresponding ϕ_a^n, are shown in Figure 3. Regardless of n, the position of the first peak maximum appears located at distances shorter than r^* by 0.2 to 0.3 Å. However, as n increases, the distribution is more and more sharpened and its maximum is pushed at greater distances.

Hence, the use of ϕ_a^4 allows to define a stronger boundary between the solute atoms and the solvent particles, preventing a large incursion of solvent particles into the structure of a complex solute. ϕ_a^n potentials corresponding to n values greater than four have been also tested with, however, no significant improvement regarding the structural stability of our benchmark system (i.e. the BPTI protein).

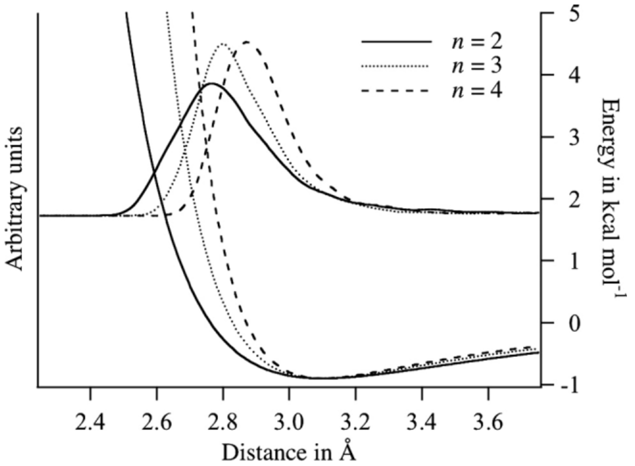

Figure 3. Profile of the potential ϕ_a^n and of their corresponding radial distribution functions $g^n(\Omega)$ (see text for details). Plain line: $n = 2$; dotted line: $n = 3$; dashed line: $n = 4$.

Parameterization Strategy

In continuum solvent approaches, the solvation free energy is usually expressed as a sum of separate electrostatic, ΔG_{elec}, and non-polar, ΔG_{np}, solvation contributions. The non-polar contribution can also be partitioned into the reversible work ΔG_{sr} associated to the short-range repulsive and dispersive (van-der-Waals) solute/solvent interactions, and into the cavitation work ΔG_{cav}, corresponding to the creation of the excluded volume for the solvent around the solute. However, in popular PB/SA continuum approaches,[41] only one single term is considered to handle the two non-polar contributions. All the parameters corresponding to the electrostatic and non-polar contributions are calibrated to reproduce the experimental solvation free energies of a set, M, of small organic molecules (such as methanol or benzene). For those molecules, regardless of the solvation approach considered, the non-polar contribution ranges from 1 to 2.5 kcal mol^{-1} (see for instance[41–43]).

The spirit of our coarse-grained solvent model is the ability to provide electrostatic solvation free energy estimates “on-the-fly.” We extend here this idea to the nonpolar contribution and ask that the sum of the solute/solvent electrostatic and van-der-Waals energies, averaged “on-the-fly” over short time windows, reproduce the experimental solvation free-energies of a learning set of molecular solutes. Cavity contributions could be included too through a surface term, but they are discarded at this stage since they fit less naturally into the spirit of the model. The ϕ_a^4 potentials are thus constructed to fullfill

$$\Delta G_{\text{solv}}^{\text{exp}} = \Delta G_{\text{elec}} + \Delta G_{\text{sr}} = \langle F_{dp}\rangle + \langle \Phi_a\rangle \tag{30}$$

for the solute learning set M. Here Φ_a stands for the sum of the individual $\phi_a^4(r_{ik})$, see Eq. 23. In practice, the average were computed

Table 1. Atomic Radii in Å.

Atom	Chemical groups	Radius r_i^*
C	Guanidium ion	1.5
C	Alcohols	1.6
C	Other types	1.7
O	Aldehyde, ketone	1.7
O	Carboxylic acid	1.6
O	Water, alcohols	1.4
N	Amide	1.4
N	Imidazole	1.8
N	Ammonium ion	1.8
H	Alcohol OH, amide NH	0.4
H	Alkyl, aromatic	0.6
H	Ammonium ion	0.8
Li	Monoatomic cation	1.2
Na	Monoatomic cation	1.8
Mg	Monoatomic dication	1.2
Ca	Monoatomic dication	1.8

The parameters r_{ij}^* of the ϕ_a^4 potential are defined as : $r_{ij}^* = (r_i^* + r_j^*)$. The radius of the solvent particles is of 1.4 Å.

during the last 50 ps of 110 ps long MD simulations, performed according to the above protocol. In the case of a charged solute (of charge Q), we ask the sum $\Delta G_{elec} + \Delta G_{\text{sr}}$ to reproduce the experimental $\Delta G_{\text{solv}}^{\text{exp}}$ value, to which the following contribution is substracted

$$\Delta G_{\text{elec}}^c = -\left(1 - \frac{1}{\varepsilon}\right)\frac{Q^2}{2R_{\text{cut}}} \tag{31}$$

Here, R_{cut} is the cutoff distance imposed to the solute/solvent interaction (see Principles), and the latter contribution corresponds to the long-range electrostatic solute/solvent interaction not accounted for because of that cutoff.

The model parameters were also selected in order to keep the r_{ij}^* solute/solvent distances (see Table 1) close to the van der Waals radii proposed by earlier studies.[41,44,45] For neutral molecules, the electric field damping parameters a introduced in the PPP Solvent Model were set to 0.3 and 0.5 for solute/solvent atom pairs involving a hydrogen and a heavy atom, respectively. For a charged entity, that parameter, as well as those of ϕ_a^4, were assigned to reproduce its hydration free energy. In the particular case of mono-atomic ions, that strategy permitted to define a set of distances r_{ij}^*, that are compatible with the known ion-water distance in the first hydration shell (for instance, for monoatomic ions Li^+, Na^+ and K^+, the r_{ij}^* values are 2.1, 2.6 and 3.1 Å respectively).

Once the parameters assigned, the model was tested by applying it to a larger test set M' of small molecules (the set M and M' are defined in Appendix B). For the neutral molecules of M', the computed ΔG_{solv} energies and ΔG_{elec} contributions are both plotted in Figure 4 versus their experimental values, $\Delta G_{\text{solv}}^{\text{exp}}$, as well as the theoretical electrostatic contributions computed from a PB/SA continuum approach (using the DELPHI program[46] with the PARSE parameter set[41]). The same plots were drawn in the case of charged molecules. Lastly, some results regarding the ΔG_{elec} and ΔG_{sr} contributions, and the solvated dipole moments for a selected set of neutral molecules are listed in Table 2.

For neutral molecules, a good correlation between the predicted ΔG_{solv} and the experimental $\Delta G_{\text{solv}}^{\text{exp}}$ is observed for the full set M': the mean and maximum absolute difference are 0.3 and 1.1 kcal mol^{-1}, respectively. Compared to the continuum approach, our local model systematically underestimates the electrostatic ΔG_{elec} contributions by about 20%. Nevertheless, a very good linear correlation is observed. This systematic underestimation is related in part to the atomic electrostatic charge set considered in the TCPEp force-field: in contrast to PARSE, they are located on all atomic centers (including hydrogens covalently bonded to carbons), and they are usually smaller. For instance, for alcohol groups, the oxygen and hydrogen charges are -0.50 and $+0.30e$ for TCPEp and of -0.49 and $+0.49e$ for PARSE, respectively.

ΔG_{sr} corresponds to a stabilizing contribution weaker than ΔG_{elec}. Its mean and maximal values are of -1.3 and -2.5 kcal mol^{-1} for the molecules of the set M', and it plays its strongest role in the case of alcohols (Table 2).

Regarding the dipole moment values, a strong increment is observed between gas phase and solution, ranging from 19 to 44%, except for phenol, for which the increment reaches a value of 66 %. The predicted dipole moment in solution for water is of 2.30 Debye. Incidentally, this value is quite similar to the effective value appearing in explicit models such as SPC or TIP3P. It appears somewhat smaller, however, than the value determined in explicit polarizable water simulations: 2.5 to 2.9 Debye.[27,47] Compared to the continuum approach of Sharp et al.[48] which does account for solute polarizability, the dipole increments are of the same order of magnitude: for instance, in the case of methanol and phenol, that approach predicts an increment of about 0.61 and of 0.85 Debye, whereas our particle-based model predicts an increment of 0.43 and 1.09 Debye.

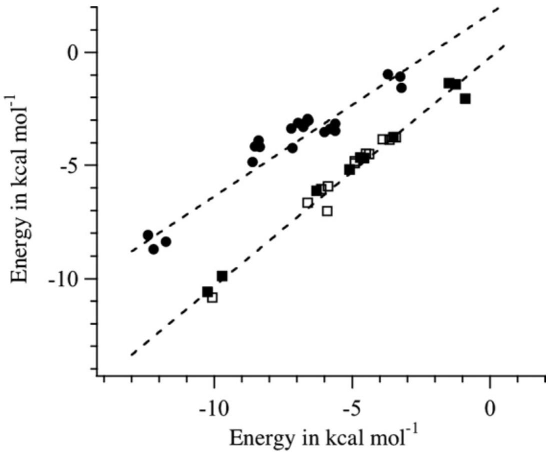

Figure 4. Comparison between the solvatation free energies predicted by our polarizable mesoscopic approach and experiment (squares), and comparison between our approach and a Poisson-Boltzmann one concerning the solvation electrostatic free energy component ΔG_{elec} (full circles). Full squares: results concerning the learning set M of the small molecules used during the parameter definition; empty squares: results concerning the remaining molecules of the set M'.

Table 2. $\bar{\mu}_{\text{vac}}$, $\bar{\mu}_{\text{solv}}$.

Solute	$\bar{\mu}_{\text{vac}}$	$\bar{\mu}_{\text{solv}}$	$\Delta\mu$	ΔG_{elec}	ΔG_{sr}	ΔG_{solv}	$\Delta G_{\text{solv}}^{\text{exp}}$	Error
Water	1.86	2.27	0.41 (22)	−5.35	−0.78 (13)	−6.13	−6.30	0.17
Methanol	1.75	2.18	0.43 (25)	−3.36	−1.82 (35)	−5.18	−5.10	0.08
Phenol	1.64	2.73	1.09 (66)	−4.85	−1.81 (27)	−6.66	−6.62	0.04
Acetone	2.95	3.54	0.59 (20)	−3.52	−0.33 (7)	−3.85	−3.90	0.05
Methylamine	1.46	2.07	0.61 (42)	−3.30	−1.38 (29)	−4.68	−4.57	0.11
Formamide	3.70	4.85	1.15 (31)	−8.37	−1.52 (15)	−9.89	−9.72	0.17
Dimethylamide	3.86	4.73	0.87 (23)	−8.71	−2.13 (20)	−10.84	−10.08	0.72
Imidazole	3.23	4.64	1.41 (44)	−8.08	−2.51 (24)	−10.59	−10.25	0.34
Pyridine	2.24	2.91	0.67 (30)	−4.23	−0.43 (10)	−4.66	−4.70	0.04
Propane	0.00	0.15	0.15 (na)	−0.09	−0.32	−0.41	1.96	na
2-Methylbutane	0.00	0.16	0.16 (na)	−0.12	−0.52	−0.64	2.38	na
Benzene	0.00	0.34	0.34 (na)	−1.57	−0.47	−2.04	−0.90	na

Dipole moment value *in vacuo* and in solution; $\Delta\mu$: dipole moment difference between gas phase and solution (in parentheses, the dipole moment increment as compared to the *vacuum* value, in percent); all latter values are given in Debye; ΔG_{elec} and ΔG_{sr}: model contributions to the solvation free energies; ΔG_{solv} and $\Delta G_{\text{solv}}^{\text{exp}}$: model and experimental free solvation energies, all values in kcal mol^{-1}.
Error: difference between the computed and experimental free solvation energies, in kcal mol^{-1}.

The strongest disagreement between our model and experiment is found for hydrophobic molecules, such as propane, 2-methylbutane and benzene. In the case of the first latter two, the predicted ΔG_{solv} is slightly negative (about −0.5 kcal mol^{-1}), while the experimental value is positive by about 2.0 kcal mol^{-1}. For benzene, both predicted and experimental values are negative, but our predicted value is stronger by 1 kcal mol^{-1} compared to experiment. That deficiency results from the lack of an explicit cavitation term in our solvation free-energy estimator. However, as it will be shown in Section IV, this does not seem to introduce an important drawback when applying the model to the study of complex molecules, such as proteins.

Regarding the charged molecules of the set M', a very good correlation is observed between our predicted ΔG_{solv} and experiment: the slope and the correlation coefficient of the linear relationship are both equal to 0.99. For charged entities, as the ΔG_{sr} values are particularly weak (ranging from −0.9 to −3.4 kcal mol^{-1}), ΔG_{solv} reduces mostly to the ΔG_{elec} contribution. This explains the good correlation between our predicted ΔG_{elec} and those computed with a Poisson-Boltzmann method using the PARSE parameter set: the slope and correlation factor are equal to 0.96 and 0.99, respectively. It has to be noted here that the ΔG_{solv} values computed from our approach have to be corrected for charged entities by adding the long-range contribution $\Delta G_{\text{elec}}^{c}$ [see eq. (31)]. In the particular case of the solvated ammonium ion, we performed eight simulations with a cutoff distance R_{cut} for the solute/solvent interaction ranging from 12 to 26 Å (the dimensions of the solvent box considered for that study were of 50 × 50 × 50 Å^3, the equilibration period was 100 ps and the total simulation length was 500 ps). In Figure 5, both the uncorrected ΔG_{solv} and $\Delta G_{\text{solv}} + \Delta G_{\text{elec}}^{c}$ values are plotted vs. R_{cut}. As observed, the uncorrected ΔG_{solv} slowly converges toward the experimental value (for $R_{\text{cut}} = 26$ Å, the computed ΔG_{solv} only represents about 93% of the experimental one). However, regardless of R_{cut}, the computed $\Delta G_{\text{solv}} + \Delta G_{\text{elec}}^{c}$ value matches the experimental one within less than 0.2 kcal mol^{-1}. This demonstrates that our parameterization strategy does not depend on R_{cut}. Lastly, note that the computed ΔG_{solv} values of all the neutral entities of the set M' have already converged for $R_{\text{cut}} = 12$ Å.

Finally, since the shape of a complex solute can evolve during the dynamics, and solvent particles can somehow penetrate into the solute excluded volume, there is some uncertainty concerning the choice of the solvent density and its influence on the results. This factor was investigated by performing MD simulations of two solvated molecules, a neutral one (formamide) and a charged one (deprotonated propanoic acid). The MD protocol corresponds to that described earlier, except for a longer simulation time of 2 ns. For the two solutes, five trajectories were generated with different solvent densities ρ_s equal to 0.90, 0.95, 1.01, 1.05, and 1.10 g cm^{-3}, respectively. The solvent self-diffusion coefficient is found to decrease from 2.4 to 1.610^{-4} cm^2 s^{-1} as ρ_s increases. Nevertheless, the solvent clearly remains in the liquid state in all

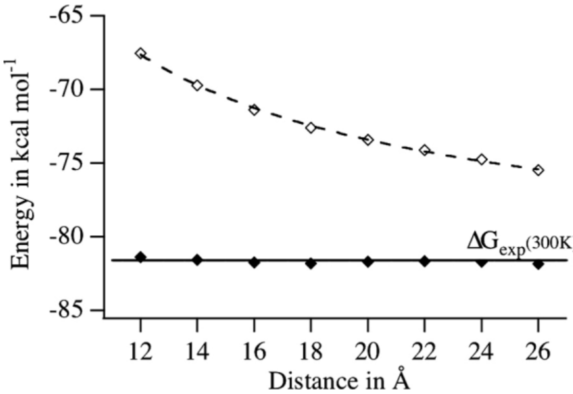

Figure 5. Theoretical solvation free energies ΔG_{solv} of NH_4^+ computed using the mesoscopic solvent approach for solute/solvent interaction cutoff distance R_{cut} ranging from 12 to 26 Å. Empty lozanges: uncorrected ΔG_{solv} values; full lozanges: ΔG_{solv} values corrected by adding the corresponding $\Delta G_{\text{solv}}^{c}$ contribution [see eq. 31]. Plain line : the experimental solvation free energy of NH_4^+ at 300 K (i.e. −81.5 kcal mol^{-1}).

cases. Analysis of the trajectories shows that a density fluctuation of 10^{-2} g cm^{-3} leads to an uncertainty in the predicted ΔG_{elec} ranging from 0.4% (propanoic acid) to 1.0 % (formamide). Thus, although using pseudo-particles with the size and the density of liquid water is a sensible choice to provide consistent thermodynamic properties (e.g. compressibility), the electrostatic energy itself appears to be quite insensitive to the chosen density.

Energy Surface of the Alanine Dipeptide

The parameters of the TCPEp force-field and of the mesoscopic model were assigned by considering only small molecules or molecular aggregates. However, the analytical form and the parameters of the intramolecular interaction terms considered in TCPEp (i.e. the stretching, bending and torsional ones) correspond to those of the CHARMM force-field v. 22. That suggests a priori that our approach can be used to simulate solvated proteins. However, the treatment of the conformational energy differences corresponding to the protein backbone is a key point in simulating proteins using empirical force-fields, as shown by the continuing efforts devoted to improve the treatment of the protein backbone dihedral terms ϕ/ψ (cf ref. 49 and the references cited therein).

In Figure 6, the energy surfaces of the alanine dipeptide in the (ϕ, ψ) space calculated using our approach *in vacuo* and in solution are shown. *In vacuo*, the surface was generated by optimizing the dipeptide structure with the (ϕ, ψ) dihedral angles constrained. In solution, the surface was generated by performing 36 molecular simulations of the unconstrained alanine dipeptide for a total duration of 108 ns (the dipeptide starting structures correspond to structures with the (ϕ, ψ) angles regularely spaced from -180 to $180°$). For each trajectory, the (ϕ, ψ) probability distributions were computed, and the surface was generated from them, based on a Boltzmann distribution.

Both *in vacuo* and in solution, the positions and energies of the surface minima predicted by our approach are in good agreement with the reported results concerning the CHARMM v. 22 force-field.[18,49] For instance, in solution, the surface generated from the (ϕ, ψ) probability distributions is mainly populated in the $-180° \leq \phi \leq 0°$ region. The main difference between our approach and CHARMM results in solution concerns the position of one of the minima located in the $0° \geq \phi \geq 180°$ region (and usually labelled α_L): we predict it to be located at about (70°, 130°), whereas it is commonly predicted to lie at about (60°, 40°).[18]

Hence, the above results exhibit that our approach for modeling the atomic nonbonded interactions and the solvent can be used in conjunction with the CHARMM intramolecular energy terms (and parameters) to simulate solvated proteins or peptides.

Applications to the Study of BPTI in Aqueous Phase

The structure of the Bovine Trypsin Inhibitor (BPTI, code name 1BPI in the Protein Data Bank) was resolved using X-ray cristallography[50] and NMR spectroscopy[51] methods. This protein possesses 58 residues and it contains two strands of antiparallel β-sheet and two short segments of α-helix, connected by two long loops. The BPTI structure is stabilized by three disulfide bridges linking the cysteines 5 and 55, 14 and 38, and 30 and 51, and by a small core of hydrophobic residues. Its total charge is $+6e$ for neutral pH values.

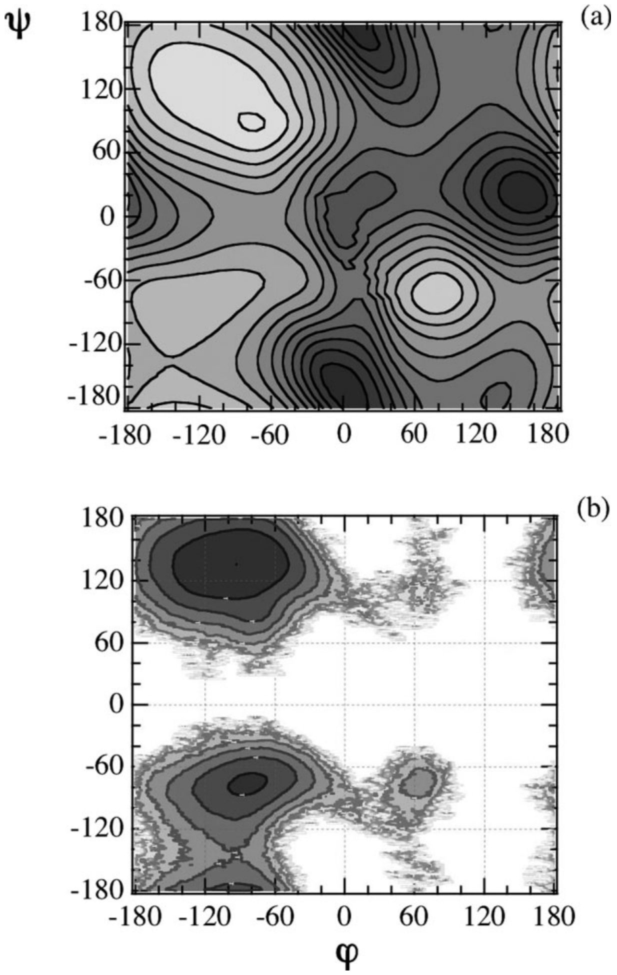

Figure 6. (ϕ, ψ) contour plots of the energy surface of the alanine dipeptide calculated with the TCPEp force-field in vacuo (a) and solvated using the solvent polarizable mesoscopic approach (b). The countours are in increments of 1 kcal mol^{-1}.

Because of its size and as NMR experiments showed a remarkable stability of its structure in solution,[51] the accuracy of MD simulation methods devoted to proteins is commonly assessed by monitoring the time evolution of the BPTI structure. Among the explicit all atoms approaches tested in such a way, we may quote the work of Kim et al.,[29] who have tested a set of polarizable force-fields.

The ability of our approach to describe proteins was investigated accordingly, by performing long dynamics simulations (running up to 20 ns) of BPTI in water. Moreover, as the mesoscopic approach keeps the notion of particules, we have also investigated the local properties of the solvent in the vicinity of the solute. The goal is here to evaluate if our approach can provide useful microscopic informations concerning the solute/solvent interactions, which are unreachable with continuum based solvent models.

As MD simulations are highly sensitive to the initial conditions (see for instance Caves et al.[52]), we have performed several runs of the solvated BPTI, each of them corresponding to different initial conditions. An improvement of the conformational space sampling is expected using such a methodology, and this should lead to more reliable indications on the accuracy of a numerical method.

Simulation and Trajectory Analysis Details

Simulation Details

Two series of six molecular dynamics simulations of solvated BPTI were performed (the trajectories are labelled **T1** to **T6**). In the first series, the protein was considered as embedded alone in the solvent, whereas in the second series, it was considered together with six explicit Cl^- counter ions, introduced to neutralize its charge. In the latter case, the six counter-ions were added successively in order to optimize the electrostatic interaction energy between the newly added ion and the protein (considered together with the previously added ions).

The simulation protocol is close to that described at the beginning of Solvent-Parametrization of TCPEp. The solute is now burried into a cubic box, of dimension 55.7 × 55.7 × 55.7 Å ^{3}. The equations of motion were solved here also using the multiple time steps r-RESPAp procedure, the short-range electrostatic and polarization forces corresponding to interactions among atoms located at less than 8 Å. The time steps were 0.25 fs (updates of the intramolecular stretching and bending forces), 1 fs (updates of the torsional, the hydrogen bonding, the repulsive, and the short range electrostatic and polarization forces), and 5 fs (updates of the long range electrostatic and polarization forces). The induced dipoles were solved using a SCF scheme, until a convergence criterion of $(10, 10^{-6})$ was reached (cf. Theoretical Methods).

The first step of each simulation corresponds to a solvent relaxation phase, during which the protein backbone C_α's are restrained to their crystallographic positions. The simulations of each series differ in the duration of this phase (from 60 to 200 ps), and in the nature of the constraints imposed to the C_α's. The first type of constraints is defined as

$$U^{c,\text{pos}} = \sum_{i=1}^{N_c} k_{\text{pos}} \left(\mathbf{r}_i^c - \mathbf{r}_{i,\text{crist}}^c\right)^2 \tag{32}$$

here, $\mathbf{r}_i^c$ and $\mathbf{r}_{i,\text{crist}}^c$ correspond to the position vector of the i^{th} C_α at time t in the simulation and in the crystallographic structure, respectively. The origin of these vectors is set to the center of mass of the restrained C_α's. The second type of constraints allows more flexibility. It is defined as

$$U^{c,\text{dir}} = \sum_{i=1}^{N_c} k_{\text{dir}} (1 - \cos\psi)^2 \tag{33}$$

where, ψ corresponds to the angle between the vectors $\mathbf{r}_i^c$ and $\mathbf{r}_{i,\text{crist}}^c$ defined above. For both kind of constraints, the force constant k was 100 kcal mol^{-1}. Once the relaxation phase achieved, the RMSDs concerning the C_α's were about 0.1 with $U^{c,\text{pos}}$ and about 0.6 with $U^{c,\text{dir}}$. The constraints were then removed and the system evolved freely, up to 6 ns for the first series of simulations, and up to 20 ns for the second series.

Structural Analysis Details

The time evolution of the protein structure along the trajectories was monitored via the root mean square deviation (RMSD) of the protein backbone heavy atoms (i.e. the N, O, and C_α atoms) computed with respect to the crystallographic data. The root mean square fluctuations (RMSF) of the protein C_α's were computed with respect to the average protein structure obtained along the trajectories. Regarding the RMSD, the C_α's corresponding to the terminal residues 1 to 3 and 56 to 58 were omitted, as their motion was shown to be of large magnitude. We have also considered the local index, RMSD$_i$, which corresponds to the mean deviation of the position of the i^{th} C_α atom with respect to its position in the crystallographic structure.

We have also estimated the relative orientation of the two helices and of the β-sheet by computing the angle between the axes corresponding to the greatest moment of inertia of the latter elements (calculated by considering only their C_α atoms). In the following, the helices corresponding to the residues 3 to 8 and 46 to 56 are respectively labelled $h1$ and $h2$.

Two kind of quantities were considered to investigate the local structure of the solvent at the protein surface: the local density ρ_i^{local} of the solvent close to the i^{th} C_α, and the solvent radial distribution function (RDF), $g_i(r)$, with respect to an atom i. The local densities were computed according to Brooks et al[53]

$$\rho_i^{\text{local}} = \frac{\overline{N}_s}{V_R} \tag{34}$$

here, $\overline{N}_s$ is the mean number of pseudo-particles within a radius R of a given C_α. The radius R is set to a value of 6 Å and the ρ_i^{local} values are normalized with respect to the density of the pure solvent (3.310^{-2} particles per Å ^{3}).

Results and Discussion

Regarding the first series of simulations (corresponding to the charged BPTI system), the protein evolves quickly toward structures whose RMSDs are included between 2.3 and 3.2 Å (with an average value of 2.6 Å), and these structures remain then stable for 5 ns. The most stable structure in terms of energy (by at least 60 kcal mol^{-1}) corresponds to a RMSD value of 2.6 Å, and it is also characterized by a strong alteration of the helix $h2$ (which is observed only in another trajectory). However, the structure of the helices and the β-sheet is usually well preserved along the six trajectories. Concerning the orientation of the two helices and of the β-sheet and as compared to the crystal structure, the angle between the helix $h1$ and the β-sheet is well preserved along all the trajectories (the difference is at most 7° on average), whereas a stronger difference is observed concerning the angle between this helix and the second one (about 20°) and between the β-sheet and the second helix (about 26°).

Regarding the neutral system (second series of simulations), the time evolutions of the protein RMSD are given in Figure7. As shown on that figure, the structure of BPTI is stabilized after 5 ns for four trajectories (i.e. **T1**, **T4**, **T5** and **T6**) and then it remains stable with average RMSD values included within 1.8 and 2.2 Å. Concerning the trajectory **T3**, the structure of BPTI fluctuates more noticeably than along the five other trajectories. However, along the full **T3**

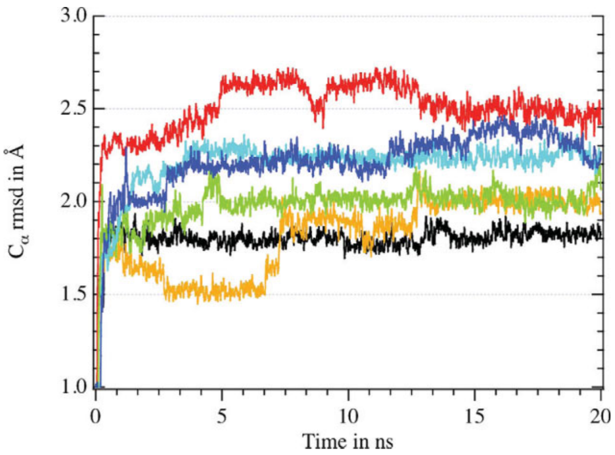

Figure 7. Time evolutions of the C_α RMSD for the six trajectories corresponding to the neutral BPTI system (same notations as for Fig. 1).

trajectory, the cristallographic BPTI structure is well preserved: the RMSD value evolves from 1.5 to 2.0 Å. Lastly, for trajectory **T2**, as the contrains imposed to the BPTIstructure during the relaxation phase are removed, the structure of the β-sheet is disrupted, which explains the large RMSD value observed along this trajectory, about 2.5 Å after 20 ns. The most stable structure in terms of energy is observed along the trajectory **T1**, which corresponds to the smallest RMSD value (about 1.8 Å). Note that the potential energy along this trajectory is fully converged after 2 ns, and the structure of BPTI remains then stable for 18 ns. The mean potential energy $\bar{U}_{\text{pot}}$ along the trajectories **T5** and **T6** lie 4 and 12 kcal mol^{-1} higher in energy than for the trajectory **T1**, whereas for the three remaining trajectories, their energies $\bar{U}_{\text{pot}}$ are less stable by 50 to 220 kcal mol^{-1}. Lastly, concerning the relative orientations of the two helices and of the β-sheet and as compared to the crystal structure, the angles between those three elements are particularely well preserved along the trajectory **T1** (the difference is at most of 6° for the angle between $h1$ and $h2$, and about 1° for the remaining two angles). Note that, for the five other trajectories, these angles are smaller on average than for the charged system: the angles between $h1$ and $h2$ and between $h1$ and the β-sheet are about 16 and 6° respectively; however, the angle between $h1$ and the β-sheet is close to that observed in the charged system, about 8°.

The energetic results concerning the neutral system are discussed in details below (see also the results summarized in Table 3). Concerning the charged system and as compared to the neutral one, we note that the intramolecular energy of the protein $\bar{U}_{\text{BPTI}}$ and the ΔG_{elec} contributions are respectively smaller and greater by about 100 kcal mol^{-1}, difference which matches the estimate off the electrostatic solvation free energy not accounted for due to the solute/solvent cutoff distance (cf. Incidence of the Cutting of the Solute/Solvent Interaction). We note also that the intramolecular energy $\bar{U}_{\text{solv}}$ is $-8150\pm$ 20 kcal mol^{-1} whaveter the system and the trajectory, and that accounting for the counter-ions stabilized the neutral system by about 1000 kcal mol^{-1} compared to the charged one. Hence, considered together, the latter results suggest that the stronger deviation of the protein BPTI structure compared to the crystal one in the case of the charged system can be due to the solute/solvent interaction cutoff, and that accounting for the counter-ions allows to balance out this effect.

Most of the NMR studies devoted to BPTI exhibited a remarkable stability of its structure in aqueous solution, structure which is close to the crystal one (see for instance[51]). From our above results, accounting for the counter-ions appears to be a key point to perform reliable simulations of this protein with our polarizable approach. As the crystallographic structure of BPTI is better preserved along the trajectories of the second series of simulations, we will discuss in the following only the results concerning the neutral solvated system. As all trajectories are stable over the last 5 ns of simulations, the averages discussed below were computed over that period. The trajectories were sampled each 5 ps, which leads to a statistical ensemble of 1000 points.

Backbone Local Structural Analysis

The average values computed along the most stable trajectories **T1**, **T5**, **T6** of the RMSD$_i$, the RSMF, the ρ_i^{local} and the mean dipole moment value of the solvent particles located at less than 6 Å from any C_α are shown in Figure 9. The RMSD's corresponding

Table 3. Statistical Averages.

Simulation	Therm.	$\overline{RMSD}$	$\overline{U}_{\text{pot}}$	$\overline{U}_{\text{BPTI}}$	$\overline{U}_{\text{solv}}$	ΔG_{sr}	ΔG_{elec}	$\Delta G_{\text{elec}}^{\text{APBS}}$
T1	pos,60	1.82 (1.36)	−10247.6	945.0	−8163.6	−63.0	−1865.6	−1811.6
T2	dir,60	2.49 (2.62)	−10145.7	633.4	−8148.8	−56.4	−1485.3	−1544.8
T3	pos,100	2.00 (1.89)	−10195.6	875.4	−8167.6	−59.1	−1784.7	−1813.9
T4	dir,100	2.01 (1.91)	−10025.7	573.2	−8153.2	−60.1	−1290.6	−1536.2
T5	pos,200	2.22 (1.93)	−10235.4	874.0	−8168.4	−60.8	−1800.0	−1842.1
T6	dir,200	2.30 (2.16)	−10243.2	894.9	−8164.7	−54.4	−1793.3	−1773.0

Therm., thermalization conditions (nature of the constraints and total length in ps of that step). $\overline{RMSD}$, mean RMSD values concerning the backbone C_α (in parentheses, the $\overline{RMSDe}$ value, see text). $\overline{U}_{\text{pot}}$, total potential energy; $\overline{U}_{BPTI}$, intramolecular potential energy of BPTI; $\overline{U}_{\text{solv}}$, total potential energy of the solvent; ΔG_{elec} and ΔG_{sr}, electrostatic and short-range solvation free energies. $\Delta G_{\text{elec}}^{\text{APBS}}$, electrostatic solvation free energies computed from the APBS software. All energies in kcal mol^{-1}.

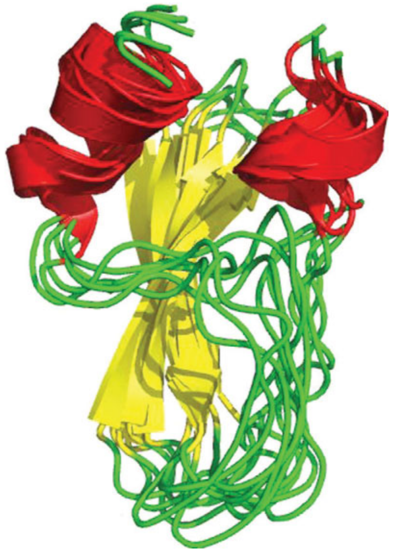

Figure 8. Superposition of the average structures of BPTI corresponding to the trajectories **T1**, **T3**, **T4**, **T5**, and **T6**. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

to the three main elements of BPTI considered together (i.e. the two α -helices and the β-sheet) and computed with respect to the crystal structure are summarized in Table 3 for each trajectory (they are labelled RMSDe below).

As already mentioned, the global RMSD values obtained at the end of the 20 ns trajectories are located between 1.8 to 2.6 Å. The strongest RMSD is observed for **T2**, and it corresponds to a local destructuration of the protein β-sheet, which explains the strong RMSDe value observed for this trajectory: about 2.6 Å. Concerning the five remaining trajectories, their RMSD's are included within 1.8 and 2.2 Å, whereas their RMSDe values are located between 1.4 and 1.9 Å, except for **T6**, 2.2 Å. This suggests that the structure of the two loops of BPTI are much more altered during the simulations than the other elements of BPTI. This is confirmed by the superposition of average structures of BPTI corresponding to the trajectories **T1**, **T3**, **T4**, **T5** and **T6** shown in Figure 8 and by the RMSD$_i$ values shown in Figure 9 (whose mean values are smaller by about 0.2 Å than the global RMSD's).

Concerning the RMSF fluctuations, they range from 0.25 to 0.95 Å, with a mean value of 0.44 Å. These values are in good agreement with the fluctuations computed by Li et al.[18] from the experimental crystallographic B-factors: the mean difference between both series of values is 0.15 Å. Even if the B-factors are affected by additional effects besides the internal atomic fluctuations, the individual motion of the backbone C$_\alpha$'s predicted by our simulations displays an overall good agreement with experiment.

Solvent Structure

The solvent structure around the protein BPTI is here discussed in terms of local RDF, $g_i(r)$. In particular, we have investigated the local solvent structure around fives types of atom: the carboxylic oxygens, the nitrogens of the lysine and of the arginine side chains, the counter-ions Cl$^-$, and the backbone carbonyl oxygens belonging to hydrophilic residues (which are defined as corresponding to normalized ρ_i^{local} value greater than 0.5).

The five types of RDF computed along the six trajectories are plotted in Figure 10. Regardless of their type and of the trajectory, all RDF's present a sharp first peak at distances between 2.7 and 3.2 Å, and a second less pronouced peak located at about 5.5 Å. In Table 4, the properties of the RDF first peak concerning four types of atoms are compared to those computed by Kim et al.[29] from explicit water simulations of BPTI with an all-atom polarizable approach. In that table, these properties (position and coordination number CN) are also compared to the available experimental data reported by Ohtaki and Radnai[54] concerning respectively the NH$_4^+$, CH$_3$COO$^-$ and Cl$^-$ ions solvated in water.

Regarding the simulations results, there is an overall good agreement between the polarizable all-atom approach and our,

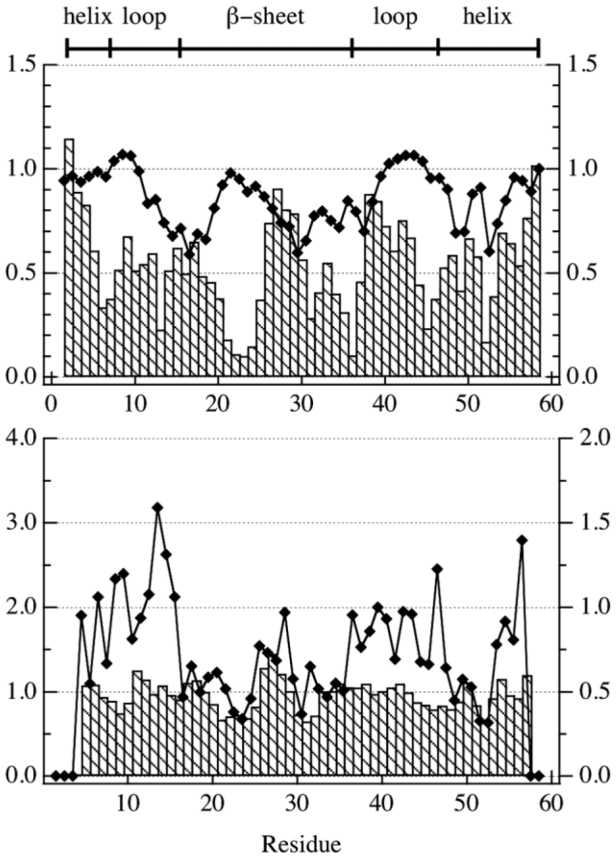

Figure 9. (a) Local solvent density at the vicinity of the C$_\alpha$'s (dashed histogram, the values are normalized with respect to the density of the pure solvent, left axis) and mean dipole moment in Debye of the solvent particles located within 6 Å of the C$_\alpha$'s (black lozanges and right axis). (b) Local RMSD$_i$ (dashed histogram and left axis) and RMSF$_i$ (full lozanges and right axis), both values in Å.

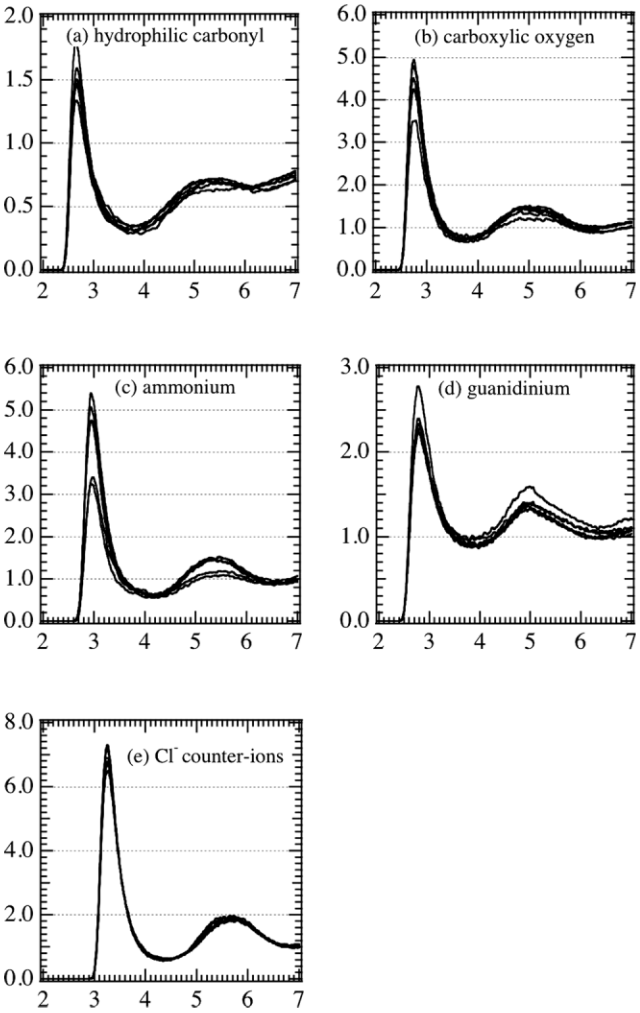

Figure 10. Local radial distribution functions of the solvent with respect to the carbonyl oxygens of hydrophilic residues (a); to the carboxylic oxygens (b); to the nitrogens of the side chain of the lysine (c) and arginine (d) residues; and to the counter-ions Cl^- (e). For all plots, the bottom axe is expressed in Å.

both in terms of position and height, except for the RDF's corresponding to the nitrogens: the first peaks are located at about 2.5 Å with a height of about 6 for the polarizable all-atom model, whereas our approach predicts 2.9 Å and a height included between 1.3 and 5.4. Compared to experiment, our results are in better agreement than those of Kim et al. Note that the latter authors obtained results close to experiment when they computed the nitrogen RDF's along a trajectory generated with a classical pairwise force-field.

Regarding the CN's, our approach predict values corresponding to the upper limit of the experimental domain values, except for the counter-ions Cl^-, for which a strong discrepancy is observed: our approach predicts a CN of about 15, which is more than twice greater than the experimental estimate (about 8).

The Counter-Ion Cloud

Whatever the trajectory, three protein regions are clearly highlighted. If we consider the 6 × 6 counter ions together, most of them (about 20) are statiscally highly localized into a large region extending from the arginine 39 to the arginine 46 of the protein second loop; height of them are observed into a smaller volume located at short distance from the Lys26 (involved in the 25-28 β turn); and four into a small volume located within the residue 19 of the protein β-sheet. Lastly, only for three trajectories, a single counter ion is observed outside of these regions.

The dynamics of ions in solution, especially around a charged solute, is known to be in the multi-nanosecond range (cf. for instance[55,56]). Hence, identifying the solute regions with a high affinity for the counter ions using MD simulations at the nanosecond scale is questionable. However, as our six trajectories sample different conformational regions and as they correspond to a cumulated simulation time of 120 ns, the three regions discussed above can be considered as physically meaningful.

Energetic Analysis

The main average energy values are summarized in Table 3. They correspond to the total potential energy, $\bar{U}_{\rm pot}$, to the intramolecular energy, $\bar{U}_{\rm BPTI}$ of the protein, to the intermolecular solvent energy, $\bar{U}_{\rm solv}$, and to the protein hydration electrostatic and nonpolar free energies, $\Delta G_{\rm elec}$ and $\Delta G_{\rm sr}$. As mentioned at the beginning of that section, the statistical ensemble considered here is made of 1000 points. The root mean square fluctuation, σ_E, corresponding to the total potential energy computed on that ensemble within each trajectory is about 35 kcal mol^{-1}, and the σ_E corresponding to the other energy terms are all smaller. The statistical uncertainty affecting our energy averages is thus of about 1 kcal mol^{-1}, which is neglegible for our purpose.

The short-range free energy $\Delta G_{\rm sr}$ values are very close for all trajectories: they range from −54 to −63 kcal mol^{-1}. In contrast, a much more pronounced difference is observed for the $\Delta G_{\rm elec}$ values:

Table 4. Position $r^*_{\rm MO}$ and height of the first peak corresponding to the solvent RDF's with respect to the carbonyl oxygens of the hydrophilic region (Carbonyl), to the carboxylic oxygens (Carboxylic), to the nitrogens of the side chains of positively charged residues (Lysine and Arginine), and to the counterions Cl^-.

	Explicit water		Pseudo-solvent		Experiment	
	$r^*_{\rm MO}$	Height	$r^*_{\rm MO}$	Height (CN)	$r^*_{\rm MO}$	CN
Carbonyl	2.9	2.4–2.6	2.7	1.3–1.8		
Carboxylic	2.6	4.0–4.4	2.8	3.4–4.8 (6–9)	3.0	4–6
Lysine	2.5	5.0–6.6	2.9	3.1–5.4 (8–11)	2.9	4–8
Arginine	2.5	5.0–6.6	2.8	2.3–2.8 (7–8)		
Cl^-	–	–	3.2	6.4–7.2 (∼ 15)	3.2	∼ 6

CN corresponds to the coordination number. The results corresponding to explicit water simulations are taken from,[29] and the experimental results from.[54]

they range from -1291 to -1866 kcal mol^{-1}. The strongest ΔG_{elec} correspond to the trajectory with the smallest RMSD (namely, **T1**, cf. Table 3), and this value is at least 65 kcal mol^{-1} more stable than those corresponding to the five other trajectories. Concerning the intramolecular energies, $\bar{U}_{\text{BPTI}}$, they range from 573 to 945 kcal mol^{-1}, and a linear relation exists between them and the ΔG_{elec} contributions (the correlation factor is 0.98), exhibiting that a subtle balance exists between intra and intermolecular interactions along the trajectories. The magnitude of the $\bar{U}_{\text{BPTI}}$ energies are related to the number of stable intramolecular salt-bridge observed whithin BPTI along the trajectories: the weakest $\bar{U}_{\text{BPTI}}$ energies are related to structures presenting from two to three salt-bridges (ie. **T2** and **T4**), whereas no salt-bridge is observed within the BPTI structure along the remaining trajectories.

The smallest RMSD value (1.82 Å) is observed along the most stable trajectory in terms of energy (ie. **T1**). However, we have to keep in mind that the different energy terms considered in our approach are not of comparable kind. Most of them, in particular those handling the interactions among explicit atoms (making up the solute and the counter-ions), correspond to microscopic energies, whereas those handling the interactions among the explicit atoms and the solvent pseudo-particles correspond to macroscopic free energies. Moreover, the counter-ions are accounted for explicitly and, as already noted, their dynamics are in the multi nanosecond range. Hence, the counter ion cloud can not be still fully relaxed, even after 20 ns of simulation. Lastly, note also that we used a parameter set for the energy terms handling the classical intramolecular stretching-bending-torsional interactions, which has not been refined for our polarizable intermolecular force-field, and this may have an important incidence on the evolution of the protein structure along the trajectories, as shown for instance in ref. 18. However, at this stage of development of our approach, all the above results are particularely encouraging.

Now, the key-point is to evaluate if our polarizable approach is able to provide a thermodynamic description of BPTI in solution equivalent to that provided by a continuum approach. To this end, we have considered the ΔG_{elec} contributions corresponding to a set of equally spaced BPTI structures extracted from the six trajectories and compared them to those computed by the resolution of the Poisson-Boltzman equation using the APBS[57] software. The cubic grid for the finite differences algorithm of APBS was set so that the grid size was 0.25 Å, which represents about 200 nodes in each direction. The comparison was made on a set of 200 structures (extracted from the last nanosecond of each trajectory). Both series of values are compared in Table 3. As observed, there is a good agreement between both: except for the trajectory **T4**, the difference between both series of value is of 41 kcal mol^{-1}, which represents about 2.3 % of the ΔG_{elec} contributions. Note that both methods predict two groups of value: for the trajectories **T1**, **T3**, **T5**, and **T6**, their ΔG_{elec} contributions are about -1800 kcal mol^{-1}, whereas for the trajectories **T5** and **T6**, they are greater at least by 300 kcal mol^{-1}. This is related with the number of intramolecular salt-bridge observed within the protein BPTI along the trajectories: none for the first group, and from 2 to 3 for the second. To conclude, the above results exhibits that the thermodynamic properties of the BPTI solvation predicted by our polarizable mesoscopic approach are equivalent, on average, to those predicted by a continuum solvent model.

Incidence of the Cutting Off the Solute/Solvent Interactions

To evaluate the incidence of cutting off the solute/solvent interactions in the case of a complex solute such as the protein BPTI, we have performed a new series of computations on the 6 × 200 structures considered above. Each of these structures (protein + counter − ions + solvent particules) are embedded in a greater solvent box, whose dimensions are $92.85 \times 92.85 \times 92.85$ Å ^{3}. Hence, about 20 000 solvent particules were added to the previous structures, which are located on the nodes of a regular cubic grid. The ΔG_{elec} contributions were then computed from single energy point computations and for cutoff distances R_{cut} ranging from 12 to 60 Å and regularely spaced.

On Figure 11, the ΔG_{elec}'s are plotted vs. R_{cut}. As observed and whatever the trajectrory, these contributions increase between 12 and 24 Å, and then decrease until to reach values for $R_{\text{cut}} = 60$ Å close to those corresponding to $R_{\text{cut}} = 12$ Å. A remarkable result is that the difference in the ΔG_{elec} contributions among the trajectories is almost constant whatever the cutoff distance.

The trend of the ΔG_{elec} for long cutoff distances was expected because of the strong charge of the protein BPTI ($+6e$). Note that, according to eq. (31), if we consider the structure of the (protein+counter-ions+solvent particules) systems for a cutoff distance of 60 Å as being almost spherical, the ΔG_{elec} contributions have still to be corrected by -99.6 kcal mol^{-1} to account for the cutoff. Lastly, the trend observed for distances included within

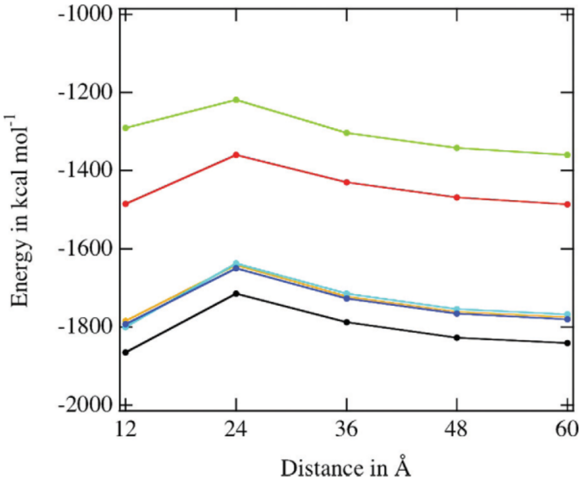

Figure 11. ΔG_{elec} contributions with respect to the solute/solvent cutoff distances, in the case of the protein BPTI, whose electrostatic charge is neutralized by six counter ions. The contributions are computed by averaging the results corresponding to the last nanosecond of the six trajectoires. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

12 and 24 Å has to result from the charge dispersion at the protein surface.

Hence, the latter results suggest that the incidence of cutting off the solute/solvent interactions has to be weak in the case of the protein BPTI (considered together with its counter ions). However, other studies are needed to check if the present observation holds for other charged proteins.

Conclusion

The results presented here show that combining a solvent mesoscopic approach with a polarizable force-field represents an efficient way to perform reliable simulations of proteins in solution. In particular, it provides an accurate description of the thermodynamic as well as of numerous local structural aspects of the protein solvation, which are unreachable using continuum solvation models. Furthermore, the computational cost of our particle-based solvation model is weak for a small protein like BPTI (the multiplicative factor is about 2.5 with respect to vacuum simulations), and it becomes essentially negligible for larger molecular edifices.

Regarding the solvent model, compared to the original articles of Haduong and co-workers,[24–26] we have here tentatively adressed some questions, concerning in particular the possible incidence of the analytical form of the energy term handling the local interactions between the solvent and the solute, as well as the influence of small fluctutations of the solvent density on the model accuracy. We have also tentatively adressed the possible drawbacks introduced by cutting off the long range solute/solvent interactions. However, this point will have to be more thoroughly discussed, by considering for instance the results corresponding to different charged proteins. Moreover, because of the pivotal role played by the counter-ions during the simulations, the extension of the mesoscopic approach to include ions will represent a noticeable improvement of our approach (in particular, that will permit to accelerate the relaxation of the counter ion cloud around the solute). Even if the ability of our methods in carefully describing the properties of solvated proteins will needs further investigations, the present results are particularly encouraging, and it appears that most of the parameters handling the solute/solute and solvent/solute interactions are already of a good quality.

Finally, concerning the thermodynamic and structural aspects of protein solvation, reliable results can also be obtained by combining a classical pairwise force-feld with a solvent mesoscopic approach.[25,26] As polarizable force-fields are much more computational demanding compared to the paiwise ones, one may thus wonder why to use polarizable force-fields to model proteins. This obviously depends on the protein property under investigation. In particular, the ability of a protein in chelating a metal ion cannot be accurately investigated using a pairwise approach, as the electric fields generated by metal ions are at the origin of strong polarization effects. Such studies are of importance, as it is inferred that about one third of the proteins needs an ion as a cofactor to their biological activity.[58] Hence, the approach reported here, combining a mesoscopic solvent model and a polarizable force-field, represents an interesting way to investigate the chelation properties of proteins. At the present time, we are using it to identify potential chelation sites within or at the surface of a protein.

Appendix A: Charge Renormalization Factor for a Polarizable Solute

The local functional F_{dP} of eq. 17 can be shown to be exact for spherical symmetries. However, it presents already a deficiency in the case of a point dipole **p** embbeded in a spherical cavity of radius R_c, as it leads to a solvation energy $\Delta G^{\mathrm{local}}_{\mathrm{elec}}(\mathrm{dipole})$

$$\Delta G^{\mathrm{local}}_{\mathrm{elec}}(\mathrm{dipole}) = \frac{1}{3}\left(\frac{\varepsilon_P - 1}{\varepsilon_P}\right)\frac{\mathbf{p}^2}{R_c^3} \tag{35}$$

which is underestimated by a factor 2/3 as compared to the classical Kirkwood formula for ε_P value $\gg 1$

$$\Delta G^{\mathrm{Kirkwood}}_{\mathrm{elec}}(\mathrm{dipole}) = \frac{1}{2}\left(\frac{\varepsilon_P - 1}{\varepsilon_P + 1/2}\right)\frac{\mathbf{p}^2}{R_c^3} \tag{36}$$

It has also to be noticed that the reaction field **R** generated by the solvent and acting on a point dipole is also underestimated by the local model

$$\mathbf{R}^{\mathrm{local}} = g_{\mathrm{local}}\mathbf{p} = \frac{2}{3}\left(\frac{\varepsilon_P - 1}{\varepsilon_P}\right)\frac{\mathbf{p}}{R_c^3} \tag{37}$$

$$\mathbf{R}^{\mathrm{Kirkwood}} = g_{\mathrm{Kirkwood}}\mathbf{p} = \left(\frac{\varepsilon_P - 1}{\varepsilon_P + 1/2}\right)\frac{\mathbf{p}}{R_c^3} \tag{38}$$

Hence, for a polarizable point dipole, the above deficiency concerning $\Delta G^{\mathrm{local}}_{\mathrm{elec}}$ is more accented, as its total dipole is related to the solvent reaction field **R** according to

$$\mathbf{p}^{\mathrm{tot}}_{\mathrm{model}} = \mathbf{p} + \alpha\mathbf{R}^{\mathrm{model}} = \frac{1}{1 - \alpha g_{\mathrm{model}}}\mathbf{p} \tag{39}$$

where α is the dipole polarisability. Hence, as compared to the macroscopic laws of electrostatic, the local model underestimates the free-energy ΔG_{el} of a polarizable point dipole immersed in water by a factor f_{local}

$$f_{\mathrm{local}} = \frac{2}{3}\left(\frac{1 - \alpha/R_c^3}{1 - 2\alpha/3R_c^3}\right)^2 \tag{40}$$

To give an approximate calculation of f_{local}, the Lorentz equation can be considered[59]

$$\frac{\alpha}{R_c^3} = \frac{n_D^2 - 1}{n_D^2 + 2} \tag{41}$$

n_D is the refractive index of the solute. The usual n_D values (corresponding to the sodium D line) range from 1.3 to 1.6, that leads to f_{local} values ranging from 0.48 to 0.56.

Appendix B: Learning and Test Sets of Molecules

Set M of molecules: water, methanol, formaldehyde, benzene, methylamine, formamide, pyridine, methylindole, methylthiol, dimethyl etherthiol, protonated and unprotonated acetic acid, ammonium ion, imidazolium ion, N-p-guanidinium ion, Li^+, Na^+, K^+, Cl^-.

Set M' of molecules, which includes the set M: ethanol, 1-methyl-butanol, isopropanol, phenol, *p*-cresol, 2-methyl-phenol, dimethylamine, trimethylamine, methyl-imidazole, propionaldehyde, acetone, 2-butanone, 3-pentanone, *N*-methyl-acetamide, protonated and unprotonated propionic acid, protonated butyric acid.

References

1. Stern, H. A.; Kaminski, G.; Banks, J. L.; Zhou, R.; Berne, B.; Friesner, R. A. J Phys Chem B 1999, 103, 4730.
2. Chelli, R.; Proccacci, P. J Chem Phys 2002, 117, 9175.
3. Anisimov, V. M.; Lamoureux, G.; Vorobyov, I. V.; Huang, N.; Roux, B.; MacKerell, A. D. J Comp Theor Comp 2005, 1, 153.
4. Vorobyov, I. V.; Anisimov, V. M.; MacKerell, A. D. J Phys Chem B 2005, 109, 18988.
5. Thole, B. Chem Phys 1981, 59, 341.
6. Vorodymyr, B.; Baucom, J.; Darden, T.; Sagui, C. J Phys Chem B 2006, 110, 11571.
7. Gresh, N.; Claverie, P.; Pullman, A. Theor Chim Acta 1984, 66, 1.
8. Piquemal, J.-P.; Cisneros, G.; Reinhardt, P.; Gresh, N.; Darden, T. J Chem Phys 2006, 124, 104101.
9. Masella, M.; Cuniasse, P. J Chem Phys 2003, 119, 1866.
10. Cuniasse, P.; Masella, M. J Chem Phys 2003, 119, 1874.
11. Llinas, P.; Masella, M.; Stigbrand, T.; Menez, A.; Stura, E. A.; Du, M.-H. L. Protein Sci 2006, 15, 1691.
12. Warshel, A.; Levitt, M. J Mol Biol 1976, 103, 227.
13. Warshel, A.; Russell, S. T. Quat Rev Biophys 1984, 17, 283.
14. Florian, J.; Warshel, A. J Phys Chem B 1997, 101, 5583.
15. Muegge, I.; Schweins, T.; Warshel, A. Proteins 1998, 30, 407.
16. Hassan, H.; Guarnieri, F.; Mehler, E. J Phys Chem B 2000, 104, 6478.
17. Hassan, S. A.; Mehler, E. L. Proteins 2002, 47, 45.
18. X. Li, S. H.; Mehler, E. Proteins 2005, 60, 464.
19. Bashford, D.; Case, D. Annu Rev Phys Chem 2000, 51, 129.
20. Kollman, P.; Massova, I.; Reynes, C.; Kuhn, B.; Huo, S.; Chong, L.; Lee, M.; Lee, T.; Duan, Y.; Wang, W.; Domini, O.; Cieplak, P.; Srinivasan, J.; Case, D.; Cheatham, T. E., III. Acc Chem Res 2000, 33, 889.
21. Gohlke, H.; Kiel, C.; Case, D. J Mol Biol 2003, 330, 891.
22. Gilson, M.; M. Davis, B. L.; McCammon, A. J Phys Chem B 1993, 97, 3591.
23. Honig, B.; Nichols, A. Science 1995, 268, 1144.
24. Haduong, T.; Phan, S.; Marchi, M.; Borgis, D. J Chem Phys 2002, 117, 541.
25. Basdevant, N.; Haduong, T.; Borgis, D. J Comput Chem 2004, 25, 1015.
26. Basdevant, N.; Haduong, T.; Borgis, D. J Chem Theory Comput 2006, 2, 1646.
27. Dang, L. X.; J. E. Rice, J. C.; Kollman, P. J Am Chem Soc 1991, 113, 2481.
28. Masella, M. Mol Phys 2006, 104, 415.
29. B. Kim, T. Y.; Harder, E.; Friesner, R.; Berne, R. J Phys Chem B 2005, 109, 16259.
30. Harder, E.; Kimm, B.; Friesner, R. A.; Berne, B. J. J Comp Theor Comp 2005, 1, 169.
31. Ren, P.; Ponder, J. J Phys Chem B 2003, 107, 5933.
32. MacKerell, A. D., Jr.; D. B.; Bellott, M.; Dunbrack, R. L.; Jr., J. D. E.; Field, M. J.; Fischer, S.; Gao, J.; Guo, H.; Ha, S.; Joseph-McCarthy, D.; Kuchnir, L.; Kuczera, K.; Lau, F. T. K.; Mattos, C.; Michnick, S.; Ngo, T.; Nguyen, D. T.; Prodhom, B.; Reiher, W. E.; III; Roux, B.; Schlenkrich, M.; Smith, J. C.; Stote, R.; Straub, J.; Watanabe, M.; Wiorkiewicz-Kuczera, J.; Yin, D.; and Karplus, M. J Phys Chem B 1998, 102, 3586.
33. Marcus, R. J Chem Phys 1956, 24, 979.
34. Calef, D.; Wolines, P. J Phys Chem 1983, 87, 3387.
35. Borgis, D.; Levy, N.; Marchi, M. J Chem Phys 119, 3516.
36. Tuckerman, M.; Berne, B.; Martyna, G. J Chem Phys 1992, 94, 6811.
37. Watanabe, M.; Karplus, M. J Chem Phys 1993, 99, 8063.
38. Toukmaji, A.; Sagui, C.; Board, J.; Darden, T. J Chem Phys 2000, 113, 10913.
39. Liu, Y.; Tuckerman, M. J Chem Phys 2000, 112, 1685.
40. Rychaert, J.; Cicotti, G.; Berendsen, H. J Comput Chem 1977, 23, 327.
41. Sitkof, D.; Sharp, A.; Honig, B. J Phys Chem 1994, 98, 1978.
42. Chambers, C.; Hawkins, G.; Cramer, C.; Truhlar, D. J Phys Chem 1996, 100, 16385.
43. Goncalves, P.; Stassen, H. Pure Appl Chem 2004, 76, 231.
44. Chothia, C. J Mol Biol 1976, 105, 1.
45. Ramachandran, G.; Ramakrishnan, C.; Sasisekharan, V. J Mol Biol 1963, 7, 95.
46. Nicholls, A.; Honig, B. J Comp Chem 1991, 12, 435.
47. Sprik, M.; Klein, M. J Chem Phys 1988, 89, 7556.
48. Sharp, K.; Arald, J.-C.; Honig, B. J Phys Chem 1992, 96, 3822.
49. MacKerell, A. D., Jr; Feig, M.; Brooks, C. L., III. J Am Chem Soc 2003, 126, 68.
50. Marquart, M.; Deisenhofer, J.; Bode, W.; Huber, R. Acta Crys Sect B 1983, 39, 480.
51. Berndt, K.; Guntert, P.; Orbons, L.; Wuthrich, K. J Mol Biol 1992, 227, 757.
52. Caves, L.; Evanseck, J.; Karplus, M. Proteins 1998, 7, 649.
53. Brooks, C.; Karplus, M. J Mol Biol 1989, 208, 159.
54. Ohtaki, H.; Radnai, T. Chem Rev 1993, 93, 1157.
55. Auffinger, P.; Westhof, E. J Mol Biol 2000, 300, 1113.
56. McConnell, K.; Beveridge, D. J Mol Biol 2000, 304, 803.
57. Baker, N.; Sept, D.; Joseph, S.; Holst, M.; MacCammon, J Proc Natl Acad Sci USA 2001, 98, 10037.
58. Holm, R.; Kennepohl, P.; Solomon, E. Chem Rev 1996, 96, 2239.
59. Böttcher, C. Theory of Electric Polarization; Elsevier: Amsterdam, 1973.