Volume 1 | Number 1 | Jan 2013 | Pages 1–100
PCCP
Physical Chemistry Chemical Physics
www.rsc.org/pccp
ROYAL SOCIETY OF CHEMISTRY

Journal Name

ARTICLE TYPE

Cite this: DOI: 10.1039/xxxxxxxxxx

Received Date
Accepted Date

DOI: 10.1039/xxxxxxxxxx

www.rsc.org/journalname

In silico characterization of protein partial molecular volumes and hydration shells

Sara Del Galdo,[a] Paolo Marracino,[b] Marco D'Abramo,[c] and Andrea Amadei*[a]

In this paper we present a computational approach, based on NVT molecular dynamics trajectories, that allows the direct evaluation of the protein partial molecular volume. The results obtained for five different globular proteins demonstrate the accuracy of this computational procedure in reproducing protein partial molecular volumes, providing a quantitative characterization of the hydration shell in terms of the protein excluded volume, hydration shell ellipsoidal volume and related solvent density. Remarkably, our data indicate for the hydration shell a $\approx$ 10% solvent density increase with respect to the liquid water bulk density, in excellent agreement with the available experimental data.

1 Introduction

It is generally accepted that the hydration shell of a protein in aqueous solution is of higher density than bulk water, but how much higher is it? And how far from the protein surface this density increment is extended? The importance of answering to these questions is clearly related to the fundamental role of the hydration shell in determining the protein thermodynamic properties including, possibly, the folding-unfolding equilibrium.

By means of x-ray and neutron scattering studies, Svergun *et al*[1] showed the existence of protein first hydration layers with an average density $\approx$10% higher than that of the bulk water, in line with molecular dynamics (MD) simulation results[2]. However, which is the density and the extension of the complete hydration shell (i.e. the collection of all the water molecules surrounding the protein characterized by a significantly different density from the bulk water one) is still unclear.

In recent years much effort was spent in understanding how the presence of the protein affects the water dynamics close to the protein surface, essentially by means of MD simulations[3–5] and terahertz spectroscopy data[6–9]. From these studies it was possible to estimate how far the presence of the protein influences the water dynamics, thus obtaining an evaluation of the hydration shell size, as provided by the solvent dynamical behaviour. However the results obtained, although basically consistent, do not provide yet a univocally determined answer to the problem leading to different inferred shell sizes. Moreover, the characterization of the hydration shell size might be dependent on the choice of the physical observables under investigation, and hence the hydration shell thickness from the protein surface provided by water dynamics analysis could be different from that obtained according to the solvent density increase.

The main aim of this study is to characterize the hydration shell of five (native state) globular proteins in aqueous solution (providing a good coverage of fold space) in terms of solvent density and, once verified in each simulation box the inclusion of the whole hydration shell and a significant number of bulk solvent molecules, the evaluation of the proteins partial molecular volumes. Typically, computational studies of protein hydration layers are focused on the protein-water interface region in order to obtain information on the arrangement of the water molecules close to the protein surface[2,10,11], thus requiring to accurately model the details of protein shape and surface. In this work, where we consider the whole hydration shell, we focus on the hydration shell-bulk water interface and hence we have no need to model with accuracy the protein surface. It is well-known that a simple ellipsoid-based model is useful for providing a reasonable idea of the molecular shape and corresponding hydrody-

[a] *Department of Chemical Science and Technology, University of Roma Tor Vergata, via della Ricerca Scientifica 00133 Roma, Italy; E-mail: andrea.amadei@uniroma2.it*
[b] *Department of Information Engineering, Electronics and Telecommunications, University of Roma Sapienza, via Eudossiana 18 00184 Roma, Italy*
[c] *Department of Chemistry, University of Roma Sapienza, piazzale Aldo Moro 5 00185 Roma, Italy*

namic properties[12–14]. We then made use of this simple model to conceive the protein hydration shell as an ellipsoid, thus loosing the protein surface details, irrelevant in our investigation, but gaining a very easy-to-employ model to properly reconstruct the protein hydration shell shape and volume. It is worth to note that within our knowledge, only a few theoretical-computational studies on protein partial molecular volumes based on atomistic simulations have been attempted[15–19]. Such computational works were based on isobaric-isothermal simulations (i.e. NpT ensemble) thus making hopeless[15] any reliable direct measure of the protein partial molecular volume. This is due to the relevant volume fluctuations of the MD simulation box providing a large noise compared to the difference between the mean volume of the protein-solvent and the pure solvent systems to be estimated. Therefore in these theoretical-computational studies the use of indirect measurements based on theoretical models involving approximations and extrapolations (typically the Kirkwood-Buff theory[20] coupled with the 3D-RISM method[21]) is mandatory. In this paper we present an efficient and accurate computational procedure, based on isochoric-isothermal MD simulations (i.e. NVT ensemble), allowing a reliable direct evaluation of the protein partial molecular volume from the protein-solvent and pure solvent MD simulations and providing an explicit and quantitative description of the protein hydration shell and its relation to the protein partial molecular volume.

2 Methods

2.1 MD simulations

The MD simulations were performed using the Gromacs software package[22]. We utilized the amber99sb force field[23] for the simulation of red blood cell bovine Ubiquitin (PDB code 3M3J), hen egg white Lysozime (1LZT), bovine pancreatic Ribonuclease A (7RSA) and *Bacillus Amyloliquefaciens* Barnase (2KF3) and the gromos96 force field[24] for sperm whale Myoglobin (5MBN). The SPC model[25] was used in the simulations to mimic the water. Four, nine and two chloride ions were included in the simulation boxes in order to neutralize the charge of Ribonuclease, Lysozyme and Barnase, respectively. We used for all the simulations cubic boxes of different size. All the systems were simulated with periodic boundary conditions in the isothermal-isochoric ensemble (NVT), using an integration step of 2 fs and keeping the temperature constant (300 K) by the isokinetic temperature coupling[26], ensuring proper equilibrium distributions in configurational space. All bonds were constrained using the LINCS algorithm[27] and for short range interactions a cut-off radius of 1.1 nm was employed. Note that the size of the cut-off radius must be considered as a parameter of the atomistic force field and hence its value affects the system observables, in particular for the pressure evaluation. The particle mesh Ewald method[28] was used to compute long range interactions with grid search and cut-off radii of 1.1 nm. We performed an initial 50 ns NVT simulation of 1219 SPC molecules at 300 K, with a density corresponding to the experimental liquid water density at the same temperature ($\approx$ 33.3 molecules/nm^3), that we used as reference SPC simulation (we used 33.321 molecules/nm^3 providing, within our simulation conditions, a pressure of about 560 bar). We then calibrated the density of the boxes containing the SPC-protein solutions in order to obtain within the NVT MD simulations a pressure identical, within the noise, to the one provided by the MD simulation of the reference SPC box. In this way the SPC-protein systems could be considered as obtained by inserting, isobarically, the protein molecule into a SPC box at the same temperature and pressure of the reference SPC box, thus mimicking the experimental conditions of solvating protein molecules into liquid water (in the following subsection we discuss the accuracy and limitations of such an approximation). Note that for all the proteins the simulation box was large enough to ensure at least 1.2-1.3 nm distance between the protein surface and the box faces. After pressure calibration for each SPC-protein solution we performed one productive MD simulation lasting 20-40 ns. In addition, we performed four 10 ns long NVT simulations of SPC molecules at 300 K with different liquid state densities (i.e. 33.09, 33.22, 33.39 and 35.10 molecules/nm^3). Such pure SPC simulations, including the reference one, provided a basic linear dependence of the pressure versus the density which we used to estimate the pressure noise ($\approx$ 10 bar) in the simulations. By propagating this error we obtained the density standard error of the pure SPC box at the same temperature and pressure of the SPC-protein simulation and containing the same number of SPC molecules (i.e. the box ideally used to insert isobarically the protein). Such a standard error (0.014 molecules/nm^3) was then used as the standard error of the bulk SPC water and to estimate the standard error of the protein partial molecular volumes. Note that in all the SPC-protein MD simulations performed we did not observe, as expected, any unfolding process and hence our simulation results are fully consistent with characterizing the proteins native state behaviour.

2.2 Nanoscopic size effects on NVT versus NpT equilibrium distributions

In an isothermal-isochoric ensemble, the free energy function describing the thermodynamics of the system is defined by the numbers of particles $\mathbf{N}$, the temperature of the system T and the volume V of the system, i.e. the Helmholtz free energy $A = A(\mathbf{N}, V, T)$. Considering the i-th state defined by the i-th value/interval of a generic observable of the solute-solvent molecules in a NVT system with a volume V_0 and equilibrium pressure p_0, the corresponding Helmholtz free energy can be written as $A_i = A_i(\mathbf{N}, V_0, T)$. Removing the volume constraint, i.e. making the system in its i-th state free to expand (or compress) to reach the equilibrium pressure p_0, the corresponding Gibbs free energy can be expressed by

$$G_i(\mathbf{N},p_0,T)=A_i(\mathbf{N},V_0,T)-\int_{V_0}^{V_i} p_i(V')\,dV'+p_0V_i \tag{1}$$

where $p_i(V')$ is the pressure of the i-th state and V_i is the corresponding equilibrium volume, i.e. the volume such that $p_i(V_i)=p_0$.

For macroscopic systems, regardless of their chemical composition, we always have $V_i=V_0 \pm dV$ (i.e. $V_0=V_i$) thus leading to neglect the integral in eq. 1 and hence $G_i=A_i+p_0V_0$ providing $G_i-G_{ref}=\Delta G_i=\Delta A_i=A_i-A_{ref}$ with G_{ref} and A_{ref} the free energies of a reference state. Therefore, for macroscopic systems the equilibrium properties in the NVT or NpT ensemble are indistinguishable as it follows from the definition of the NVT and NpT equilibrium distributions for any observable χ

$$\rho_{NVT}(\chi_i) \quad = \quad \frac{e^{-\beta\,\Delta A_i}}{\sum_j e^{-\beta\,\Delta A_j}} \tag{2}$$

$$\rho_{NpT}(\chi_i) \quad = \quad \frac{e^{-\beta\,\Delta G_i}}{\sum_j e^{-\beta\,\Delta G_j}} \tag{3}$$

Conversely, when the system has nanoscopic size and therefore it cannot be considered at full thermodynamic convergence (i.e. the number of particles cannot be considered virtually infinite), a finite variation between V_i and V_0 must be taken into account and it follows that at least a first order correction term should be included in eq. 1. However, if the system is large enough to ensure that only small variations from the equilibrium values are to be considered, a linear relationship between pressure and volume can be assumed leading to

$$G_i(\mathbf{N},p_0,T)\cong A_i(\mathbf{N},V_0,T)-\frac{p_i(V_0)+p_0}{2}\,(V_i-V_0)+p_0V_i \tag{4}$$

By choosing as reference state the one corresponding to the mean observable value and hence reasonably assuming that for such a state $p_{ref}(V_0)\cong p_0$ and $V_{ref}\cong V_0$, the Gibbs free energy change with respect to the reference state is readily provided by

$$\begin{aligned}\Delta G_i \quad &\cong \quad \Delta A_i-\frac{p_i(V_0)+p_0}{2}\,(V_i-V_0)+p_0\,(V_i-V_0)\\ &\cong \quad \Delta A_i-\frac{\Delta V_i\Delta p_i}{2}\end{aligned} \tag{5}$$

with $\Delta V_i=V_i-V_0$ and $\Delta p_i=p_i(V_0)-p_0$. From equation 5 it is evident that ΔG_i and ΔA_i differ only for the term $(\Delta V_i\Delta p_i)/2$ given by the product of two first order corrections (ΔV_i and Δp_i) and hence corresponding to a second order correction term. Therefore, considering that for reasonably large simulation boxes only the first order correction should be significant we readily obtain (neglecting the second order correction) $\Delta G_i\cong\Delta A_i$, thus ensuring the equivalence between the equilibrium distributions on the NVT and NpT ensembles. Such considerations lead to the conclusion that the use of a NVT simulation box including the protein hydration shell and a significant amount of bulk solvent molecules, once it has calibrated to provide the same pressure of the SPC reference box, should properly reproduce the equilibrium behaviour of the same SPC-protein system in the NpT ensemble (i.e. the system obtained inserting the protein isobarically into a large solvent box at the same temperature and pressure of the SPC reference box).

2.3 Protein volume and ellipsoidal layers

The protein excluded volume, i.e. the volume enclosed by the solvent-accessible surface was obtained by Gromacs using a probe radius of 0.14 nm, according to the method reported in Eisenhaber *et al*[29]. By calculating the protein excluded volume at each MD time frame of the simulations we obtained the protein mean volume and the corresponding thermal distribution. In order to characterize the solvent density around the protein we used the approximation of treating the protein molecule as an ellipsoid defined, at each MD time frame, by the eigenvectors and eigenvalues of the 3x3 geometrical covariance matrix of the x,y,z atomic coordinates as described in recent papers[30,31]. In fact, the instantaneous protein ellipsoid axes are defined by the three eigenvectors of the covariance matrix with the corresponding lengths provided by the eigenvalues (considering a Gaussian atomic positional distribution along each eigenvector, we used as semi-axis $a_i=2\sqrt{\lambda_i}$ with i=1,2,3 and λ_i the eigenvalue of the i-th eigenvector). We then considered a set of ellipsoidal layers around the protein defined by the consecutive ellipsoids with semi-axes $a_i^{(n)}=a_i+n\delta$ with fixed increment δ=0.03 nm. By calculating at each MD frame the instantaneous SPC density within each layer (disregarding the possible presence of protein atoms and/or counterions) and averaging over the MD trajectory we obtained the solvent density profile around the protein, within layers of increasing distance from the protein ellipsoid surface (i.e. the layer solvent density profile). Note that such a solvent density profile does not account for the effect of non solvent atoms excluded volume, thus providing always low density values within layers including a significant number of protein atoms (typically, single protein atoms can be still present in layers at 0.7-0.8 nm from the protein ellipsoid surface while for layers beyond 1 nm virtually no protein atoms are detected, data not shown). Therefore, the differences among the protein SPC density profiles within the first hydration layers (up to $\approx$ 0.3 nm) largely reflect the differential protein atoms spatial arrangement and compactness.

3 Results and discussion

3.1 Protein hydration shell

In figure 1 we show the layer solvent density profile for the different simulated proteins, as a function of the distance from the

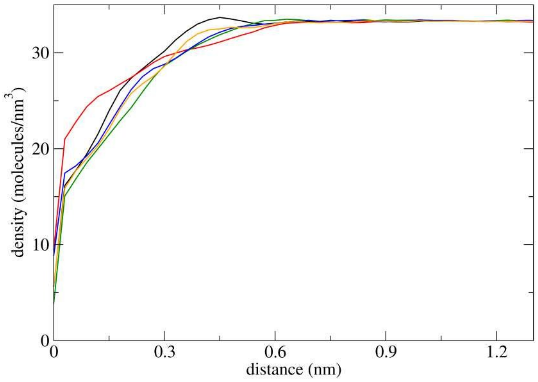

Fig. 1 Layer density profile of the solvation SPC molecules of Ubiquitin (black line), Ribonuclease A (red line), Myoglobin (green line), Barnase (blue line) and Lysozyme (yellow line) as a function of the distance from the protein ellipsoid surface.

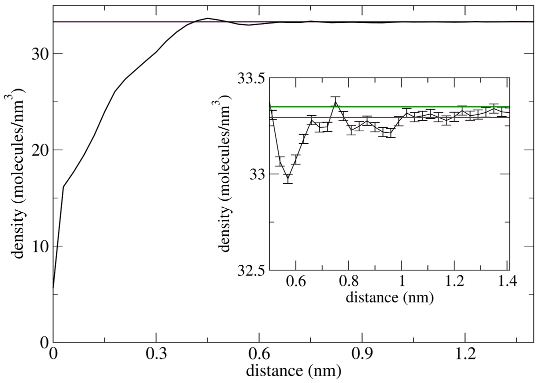

Fig. 2 Layer density profile of the solvation SPC molecules of Ubiquitin. The inset shows the convergence of the density profile, with its standard error, within the bulk density noise interval 33.293–33.349 molecules/nm^3 corresponding to the 95% confidence.

protein ellipsoid surface as obtained *via* the procedure reported in the methods section. For each of the protein under investigation it is evident that the layer solvent density tends to the bulk plateau value of $\approx$ 33.3 molecules/nm^3 (i.e. the experimental density of the liquid water at room temperature and $\approx$ 1 bar of pressure corresponding to our reference pure SPC density, see the methods section). Each layer density profile approaches the plateau at about 1 nm from the protein ellipsoid surface, thus indicating that beyond such a distance the SPC molecules of the protein-solvent system behave equivalently to the reference pure SPC ones (note that beyond 1 nm from the protein ellipsoid surface virtually no protein atoms are present). In figure 2 we report as an example the layer solvent density profile of Ubiquitin including, in the inset, its error bars (a single standard error) and the 95% confidence interval for the bulk density (i.e. 33.293–33.349 molecules/nm^3). As it is clear from the figure, the solvent density of the ellipsoidal layers placed beyond 1 nm or more from the protein ellipsoid surface is, within the noise, indistinguishable from the reference bulk density.

It is worth to note that the convergence of the layers densities to the reference SPC box density, demonstrates that the SPC-protein simulation boxes used fit into the conditions discussed in the methods section to reasonably ensure the equivalence between NVT and NpT equilibrium distributions/properties.

By using the hydration shell volume V_{shell} corresponding to the volume of the ellipsoid defined by the boundary layer of the hydration shell (as obtained by adding 1 nm to each semi-axis of the protein ellipsoid), the number of SPC molecules within such an ellipsoid n_{shell} and the protein excluded volume V_{ex} we can readily obtain the mean solvent density within the accessible volume of the protein hydration shell, related to x-ray or neutron scattering experiments, *via*

$$\left\langle \frac{n_{shell}}{V_{shell} - V_{ex}} \right\rangle \cong \frac{\langle n_{shell} \rangle}{\langle V_{shell} \rangle - \langle V_{ex} \rangle} = \rho_{shell} \tag{6}$$

and compare it with the solvent bulk density ρ_{bulk}. As it is shown in table 1 the simulations of all the proteins we considered provided, by means of eq. 6, a hydration shell mean density of about 10 % higher then the bulk one, in excellent agreement with the few experimental estimates of such a property[1,32].

It is worth to remark that in order to characterize the protein hydration shell no need of a detailed description of the protein surface is required and thus the ellipsoidal approximation of the protein geometrical shape is sufficiently accurate for our purposes. Moreover, the use of the simple ellipsoid model to estimate the protein excluded volume provides a rather reasonable reproduction of the protein mean excluded volumes (as obtained by the MD simulations), thus indicating the good quality of this approximation (see table 2). The excluded volume, as provided by the protein ellipsoidal model, is obtained by means of adding to each protein ellipsoid semi-axis the solvent radius plus the effective mean thickness due to the excluded volumes of the protein atoms over the ellipsoid surface (we used 0.14 nm as solvent radius and we estimated the ellipsoid surface mean thickness *via* a numerical fit to protein partial molecular volumes, *vide infra*, providing an effective value of 0.06 nm).

Table 1 Mean solvent density within the accessible volume of the hydration shell and relative density increment with respect to the bulk density. The standard error for the relative density increment is about 0.5%.

	Ubiquitin	Myoglobin	Ribonuclase A	Lysozyme	Barnase
Shell density (molecules/nm^3)	36.355	36.848	36.711	36.757	36.733
Relative density increment (%)	9.1	10.6	10.2	10.3	10.2

Table 2 Comparison of the mean protein excluded volume $\langle V_{ex} \rangle$ with its estimate as provided by the protein ellipsoid model $\langle V_{ex}^{ell} \rangle$ and their relative deviations. All the volume values are reported in nm^3/molecule.

	$\langle V_{ex} \rangle$	$\langle V_{ex}^{ell} \rangle$	
Ubiquitin	13.99	14.74	5.36%
Lysozyme	22.15	22.66	2.25%
Ribonuclease A	21.48	24.66	14.80%
Barnase	19.37	21.19	9.39%
Myoglobin	27.04	26.68	1.33%

3.2 Volume fluctuations

From our productive MD simulations we were able to reconstruct the probability distributions of the protein excluded volume, the protein ellipsoid volume and the hydration shell volume.

In table 3 we report the mean values (V), the standard deviations (σ) and the relative standard deviations ($\sigma_{rel} = \sigma/V$) of the protein excluded volume (V_{ex}), the protein ellipsoid volume (V_{ell}) and the hydration shell volume (V_{shell}). Note that V_{ell} is the volume of the protein ellipsoid as defined by the three semi-axes (not to be confused with the ellipsoid based excluded volume estimate reported in table 2) and V_{shell} is the volume of the ellipsoid obtained by adding 1 nm to each semi axis of the protein ellipsoid. The small values of both the standard and relative standard deviations of the protein ellipsoid volume indicate that for all the proteins investigated the protein shape does not fluctuate much during the simulation, i.e. the ellipsoid best describing the protein can be considered as almost a rigid body (the corresponding three semi-axes have negligible relative fluctuations, data not shown). It reasonably follows that the larger excluded volume fluctuations must be essentially due to cavities and grooves becoming alternatively much or less accessible to the SPC probe. However, such cavities fluctuations have no significant correlation with protein inward/outward SPC fluxes (data not shown), indicating that the cavities and grooves involved must be essentially hydrophobic, thus avoiding any relevant increase of the number of SPC molecules inside the protein even when a larger accessible volume is present. From table 3 it is also clear that for all the proteins under investigation the hydration shell relative volume fluctuations (i.e. the relative standard deviations) are rather small, being always $\leq 1.4\%$. Therefore, it follows that the hydration shell can be reasonably conceived as an ellipsoid with virtually fixed shape, volume and solvent density translating and rotating according to the protein roto-translational motion, see figure 3 (note that the water molecules residence mean time within the hydration shell is much shorter then protein roto-traslational mean time). In figure 4 we report, as an example, the distributions of the relative protein excluded volume, the relative protein ellipsoid volume and the relative hydration shell volume shifts with respect to the mean values for the SPC-Lysozyme system.

Interestingly, from the equilibrium probability distribution of the protein excluded volume it is possible to calculate the free energy variation due to the change of the protein excluded volume, corresponding to the protein chemical potential change. Considering the Gaussian shape of the excluded volume distributions, we can express such a chemical potential change *via* $\Delta\mu(V_{ex}) \approx kT(V_{ex} - \langle V_{ex} \rangle)^2/(2\sigma^2_{V_{ex}})$ where the angle brackets indicate averaging over the equilibrium ensemble. In table 4 we report the chemical potential increase to reach the excluded volume of the protein crystal structure ($V_{ex,crystal}$) from the simulation mean excluded volume, within the solution equilibrium ensemble, and in table 5 we compare the crystal structure protein ellipsoid volume with the corresponding mean ellipsoid volume as provided by the MD simulations. From these tables it is evident that although for all the proteins investigated the overall ellipsoidal shape and size in the crystal structure are very close to the simulation equilibrium values (relative deviations about 5-6%), for the largest proteins the crystal structure excluded volumes are rather different from the corresponding simulation mean values (see table 2) being hence inaccessible by the thermal fluctuations of the solution equilibrium ensemble. Such data indicate that hydration, although not relevantly altering the protein shape and size, induces a significant reduction of the water accessible cavities probably due to hydrophobic compacting. Such results, illustrating the possible effects of the crystallization interactions, suggest that the crystal structure should be used with care when evaluating typical observables of solvated proteins.

3.3 Protein partial molecular volume

On the basis of the results reported in the previous subsections, we can safely consider that our productive SPC-water simulations provide a reliable model of the solvated protein in typical experimental conditions, thus allowing their use to evaluate subtle ther-

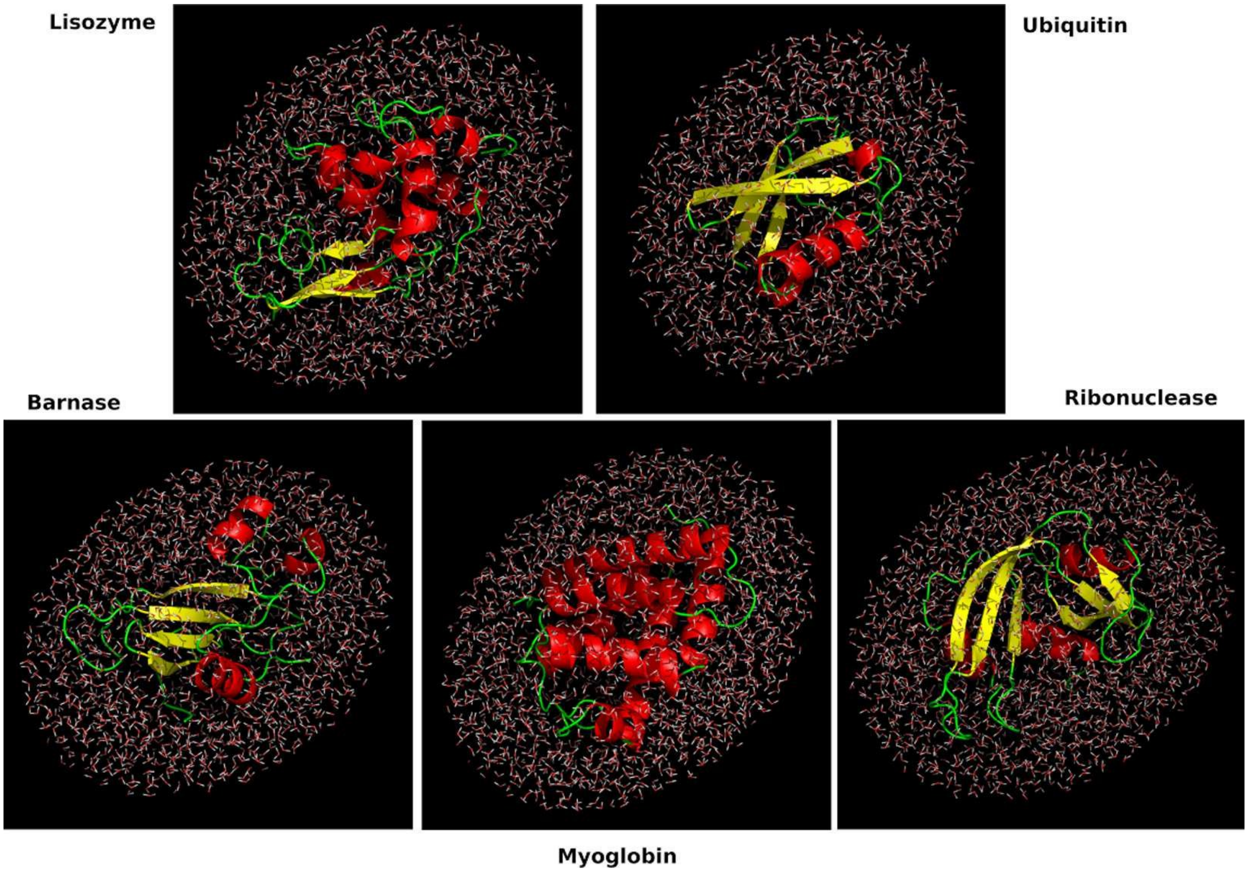

Fig. 3 Representative hydration shells for the five proteins as obtained by using the ellipsoidal model.

Table 3 Mean values V (nm^3/molecule), standard deviations σ (nm^3/molecule) and relative standard deviations σ_{rel} of the protein excluded volume V_{ex}, protein ellipsoid volume V_{ell} and hydration shell volume V_{shell}.

		Ubiquitin	Myoglobin	Ribonuclease A	Lysozyme	Barnase
V_{ex}	V	13.99	27.04	21.48	22.15	19.37
	σ	0.58	0.98	0.72	0.81	0.62
	σ_{rel}	0.041	0.036	0.035	0.037	0.032
V_{ell}	V	9.58	18.73	17.22	15.64	14.26
	σ	0.19	0.21	0.39	0.17	0.20
	σ_{rel}	0.020	0.011	0.023	0.011	0.014
V_{shell}	V	58.88	79.60	74.95	70.72	66.78
	σ	0.61	056	1.06	0.50	0.55
	σ_{rel}	0.011	0.0070	0.014	0.0071	0.0082

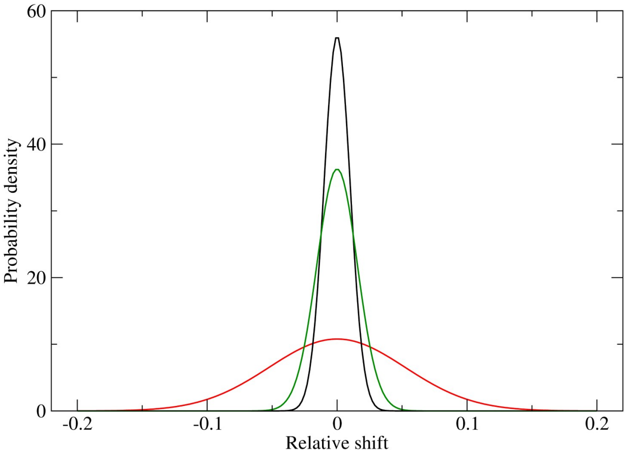

Fig. 4 SPC-Lysozyme distributions of the relative shift from the mean value for the protein excluded volume (red line), the protein ellipsoid volume (green line) and the hydration shell volume (black line).

Table 4 Chemical potential change for the $\langle V_{ex}\rangle \rightarrow V_{ex,crystal}$ transition in the solution equilibrium ensemble with $V_{ex,crystal}$ the excluded volume of the crystal structure.

	$V_{ex,crystal}$ (nm^3/molecule)	$\Delta\mu$ (kJ/mol)
Ubiquitin	13.27	1.91
Lysozyme	12.98	153.42
Ribonuclease A	19.24	11.97
Barnase	18.62	1.49
Myoglobin	21.38	42.42

Table 5 Comparison between the crystal structure protein ellipsoid volume $V_{ell,crystal}$ with the corresponding mean protein ellipsoid volume as provided by the MD simulations $\langle V_{ell}\rangle$.

	$V_{ell,crystal}$ (nm^3/molecule)	$\langle V_{ell}\rangle$ (nm^3/molecule)
Ubiquitin	9.97	9.58
Lysozyme	14.78	15.64
Ribonuclease A	16.67	17.22
Barnase	13.62	14.26
Myoglobin	19.59	18.73

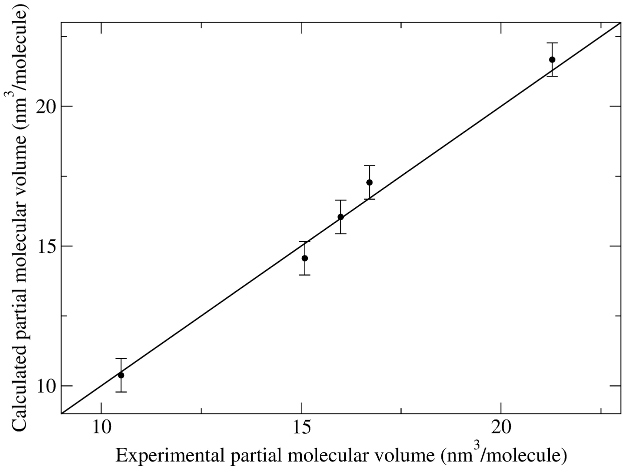

Fig. 5 Correlation between the calculated protein partial molecular volumes, as obtained by means of equation 7, and the experimental values. Error bars correspond to plus/minus two standard errors of our calculated values and the solid line is the plane bisector.

modynamic properties such as the protein partial molecular volume. In fact, from the definition of the partial molecular volume of a solute $v = (\partial V/\partial N)_{p,T,N_{solvent}}$ (i.e. the partial derivative of the system volume V with respect to the number of solute molecules N at constant pressure p, temperature T and number of solvent molecules $N_{solvent}$), it follows

$$v = V_{box} - \frac{N_{SPC}}{\rho_{bulk}} \qquad (7)$$

where V_{box} is the volume of the SPC-protein simulation box with pressure identical, within the noise, to the pressure of the pure SPC reference box, N_{SPC} is the number of SPC molecules within the SPC-protein simulation box and ρ_{bulk} is the density of the pure SPC reference box (i.e. the bulk SPC density at 33.321 molecules/nm^3). In figure 5 we show the correlation between our calculated protein partial molecular volumes and the corresponding experimental values[33–35], together with the bisector of the plane. Remarkably, the computational values we obtained are, within the noise, indistinguishable from the experimental corresponding values thus indicating not only, once again, that the computational procedure used is reliable and accurate but also that it can be utilized to obtain a proper evaluation of protein partial molecular volumes not experimentally characterized yet.

The method described above provides a direct way to compute protein partial molecular volumes from MD simulations, mimicking the experimental procedure of measure. In fact, we can make use of the definition of the hydration shell to rationalize and dissect the protein partial molecular volume in terms of excluded volume and hydration shell density. By using eq. 6 (i.e.

Table 6 Comparison of the protein partial molecular volumes (v) as obtained either *via* equation 7 or equation 9. For all the values reported in the first column the standard error is about 0.3 nm^3/molecule and for the values reported in the second column the standard error is about 0.03 nm^3/molecule.

	v (eq. 7) (nm^3/molecule)	v (eq. 9) (nm^3/molecule)
Ubiquitin	10.30	9.90
Lysozyme	17.27	17.15
Ribonuclease A	16.04	16.03
Barnase	14.54	14.53
Myoglobin	21.67	21.47

$\langle V_{shell} \rangle - \langle V_{ex} \rangle = \langle n_{shell} \rangle / \rho_{shell}$) and the SPC bulk density ρ_{bulk}, we can express the protein partial molecular volume v as

$$v = \langle V_{ex} \rangle + \langle n_{shell} \rangle \left(\frac{1}{\rho_{shell}} - \frac{1}{\rho_{bulk}} \right) \tag{8}$$

where obviously $\langle n_{shell} \rangle = \rho_{shell} (\langle V_{shell} \rangle - \langle V_{ex} \rangle)$ and then

$$v = \langle V_{ex} \rangle - \eta \left(\langle V_{shell} \rangle - \langle V_{ex} \rangle \right) \tag{9}$$

with $\eta = (\rho_{shell} - \rho_{bulk}) / \rho_{bulk}$ corresponding to the relative density increment of the SPC molecules in the hydration shell accessible volume with respect to the bulk density. When using eq. 9 to obtain the partial molecular volumes on the basis of the values of $\langle V_{ex} \rangle$, $\langle V_{shell} \rangle$ and η as provided by the MD simulations, we obtain indistinguishable partial molecular volumes, within the noise, from the ones provided by eq. 7, thus showing that the estimate of 1 nm for the hydration shell thickness from the protein ellipsoid surface we used for evaluating V_{shell} and η is consistent and reliable (see table 6). Moreover, the partial molecular volumes as obtained by eq. 9, reproducing the experimental values with relative deviations always well below 5 %, are characterized by a much smaller noise than the corresponding partial molecular volumes as provided by eq. 7. Therefore, in order to reduce the statistical noise of our estimates, the use of eq. 9 can be preferred and even necessary when evaluating small partial molecular volume variations like the unfolding partial molecular volume change.

Finally, eq. 9 can also furnish a simple expression to roughly evaluate the protein partial molecular volume on the basis of only the crystal structure, without using any MD simulation data (note that although protein crystal structures may have excluded volumes different from the corresponding solution mean excluded volumes, see table 4, the protein crystal structure ellipsoid semiaxes are always very close to the ones obtained by MD simulations). In fact, when using the protein crystal structure ellipsoid to estimate the protein mean excluded volume and hydration shell volume (i.e. by adding to the crystal structure semi-axes for the former 0.2 nm and for the latter 1 nm, see the subsection Protein hydration shell) and considering a fixed η value corresponding to the mean value over the five proteins investigated in this study ($\bar{\eta}$=0.101), by means of equation 9 we can obtain a general expression providing a rough estimate of the partial molecular volume of any protein with known crystal structure. Note that the effective mean thickness of the ellipsoid surface we used (0.06 nm) was obtained by tuning its value in this rough estimate of the partial molecular volumes in order to best reproduce the experimental partial molecular volumes over a sample of 15 globular proteins[33] including the 5 proteins investigated in this paper (i.e. the thickness value such that the linear regression of the calculated versus experimental points coincided with the plane bisector). Results are shown in figure 6 which clearly indicates that for most of the 15 proteins considered the use of such a simple and approximated procedure provides a rather reasonable estimate of the partial molecular volume, thus confirming that η is about 0.1 in general for different proteins (the four proteins significantly deviating from the model behaviour are, in partial molar volume ascending order, Insulin, Pancreatic Trypsin Inhibitor, Adenylate kinase and the Bence-Jones protein REI).

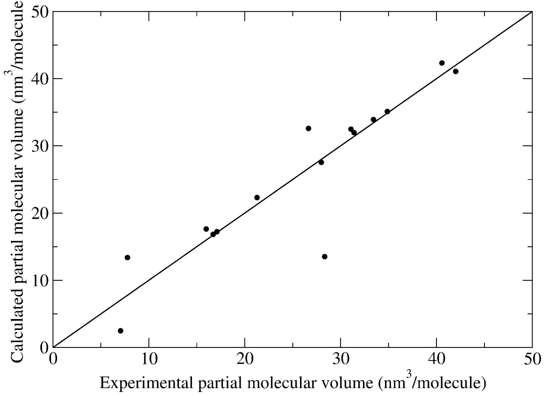

Fig. 6 Correlation between the calculated protein partial molecular volumes, as obtained by means of equation 9 and using the crystal structure protein ellipsoid, and the experimental values for 15 different proteins. The solid line is the plane bisector.

4 Conclusion

In this paper we showed that proper atomistic NVT MD simulations can furnish a powerful tool to quantitatively characterize the hydration shell and obtain a reliable estimate of the partial molecular volume of solvated proteins. The computational procedure described, based on mimicking the isobaric insertion at room temperature of a protein molecule in a pure solvent system, provided an accurate reproduction of the experimental (native state) partial molecular volumes of the five globular proteins we investigated in this work. These results illustrate the advantage of using NVT simulations instead of NpT simulations in order to suppress the large noise due to the simulation box volume fluctuations, hence allowing a reliable direct estimate of the protein partial molecular volume and its related properties. Moreover, our data show the importance of using the experimental bulk solvent density rather than the pressure to set up the MD simulation box to be used. In fact, the simulation isobar identified by the density (33.321 molecules/nm^3) of the reference pure SPC box (i.e. the experimental liquid water density at about 300 K) we utilized to mimic the typical liquid water conditions for inserting the protein molecule into the solvent, corresponds within our simulation conditions to $\approx$ 560 bar instead of the experimental $\approx$ 1 bar. It follows that the use of protein-SPC simulations at $\approx$ 1 bar of pressure would result in a significantly lower solvent bulk density leading to rather different and unreliable protein partial molecular volumes and hydration shells. Our results also indicate that, based on the solvent density analysis, the protein hydration shell can be conceived as an ellipsoid with surface at about 1 nm from the ellipsoid approximating the protein shape and characterized by an inner solvent density within the accessible volume, about 10% higher than the bulk density, in excellent agreement with the few experimental data available.

All the data discussed in this paper show that when including the whole hydration shell and a significant number of bulk solvent molecules in the simulation box, a fully consistent behaviour with hence reasonably converged estimates of the protein thermodynamic properties can be achieved. Therefore, the computational procedure presented is very promising and suited for systematically evaluating solutes partial molecular volumes and investigating the hydration effects on protein properties including, in particular, the observables involved in the unfolding transitions.

Acknowledgement

This work was supported by the Ministero dell' Istruzione, Università e Ricerca (R.Levi-Montalcini fellowship to M.D.). The authors acknowledge Prof. Ruggero Caminiti (Chemistry Dept., Sapienza University of Rome) and CINECA for providing computing time on HPC facilities.

References

1 D. I. Svergun, S. Richard, M. H. J. Koch, Z. Sayers, S. Kuprin and G. Zaccai, *Proc. Natl. Acad. Sci.*, 1998, **95**, 2267–2272.

2 F. Merzel and J. C. Smith, *Proc. Natl. Acad. Sci.*, 2002, **99**, 5378–5383.

3 A. C. Fogarty and D. Laage, *J. Phys. Chem B*, 2014, **118**, 7715–7729.

4 M. Heyden and M. Havenith, *Methods*, 2010, **52**, 74–83.

5 P. Rani and P. Biswas, *J. Phys. Chem. B*, 2015, **DOI:10.1021/jp511691c**,.

6 V. C. Nibali, G. D'Angelo, A. Paciaroni, D. J. Tobias and M. Tarek, *J. Phys. Chem. Lett.*, 2014, **5**, 1181–1186.

7 O. Sushko, R. Dubrovka and R. S. Donnan, *J. Chem. Phys.*, 2015, **142**, 055101.

8 U. Heugen, G. Schwaab, E. Bründermann, M. Heyden, X. Yu, D. M. Leither and M. Havenith, *Proc. Natl. Acad. Sci.*, 2006, **103**, 12301–12306.

9 S. Ebbinghaus, S. J. Kim, M. Heyden, X. Yu, U. Heugen, M. Gruebele, D. M. Leitner and M. Havenith, *Proc. Natl. Acad. Sci.*, 2007, **104**, 20749–20752.

10 V. A. Makarov, B. K. Andrews and B. M. Pettitt, *Biopolymers*, 1998, **45**, 469–478.

11 J. J. Virtanen, L. Makowski, T. R. Sosnick and K. F. Freed, *Biophysical Journal*, 2010, **99**, 1–9.

12 W. R. Taylor, J. M. Thornton and W. G. Turnell, *Journal of Molecular Graphics*, 1983, **1**, 30–38.

13 S. Harding, J. C. Horton and H. Cölfen, *Eur. Biophys. J.*, 1997, **25**, 347–359.

14 Y. E. Ryabov, C. Geraghy, A. Varshney and D. Fushman, *J. Am. Chem. Soc.*, 2006, **128**, 15432–15444.

15 T. Yamazaki, T. Imai, F. Hirata and A. Kovalenko, *J. Phys. Chem. B*, 2007, **111**, 1206–1212.

16 V. P. Voloshin, N. N. Medvedev, N. Smolin, A. Geiger and R. Winter, *Phys. Chem. Chem. Phys.*, 2015, **17**, 8499–8508.

17 A. V. Kim, N. N. Medvedev and A. Geiger, *Journal of Molecular Liquids*, 2014, **189**, 74–80.

18 I. Brovchenko, M. Andrews and A. Olienikova, *Phys. Chem. Chem. Phys.*, 2010, **12**, 1233..4238.

19 N. Patel, D. N. Dubins, R. Pomés and T. V. Chalikian, *J. Phys. Chem. B*, 2011, **115**, 4856–4862.

20 J. G. kirkwood and F. P. Buff, *J. Chem. Phys.*, 1951, **19**, 774.

21 Y. Harano, T. Imai, A. Kovalenko and M. Kinoshita, *J. Chem. Phys.*, 2001, **114**, 9506.

22 H. J. C. Berendsen, D. van der Spoel and R. van Drunen, *Comput. Phys. Commun.*, 1995, **91**, 43–56.

23 V. Hornak, R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, *Proteins*, 2006, **65**, 712–725.

24 C. Oostenbrink, A. Villa, A. E. Mark and W. F. van Gunsteren, *J. Comput. Chem.*, 2004, **25**, 1656–1676.

25 H. J. C. Berendsen, J. R. Grigera and T. P. Straatsma, *J. Chem. Phys.*, 1993, **98**, 10089–10092.

26 D. J. Evans and G. P. Morriss, *Statistical mechanics of non equilibrium liquids*, Accademic Press, London, 1990.

27 B. Hess, H. Bekker, H. J. C. Berendsen and J. Fraaije, *J. Comput. Chem.*, 1997, **18**, 1463–1472.

28 T. Darden, D. Tork and L. Pedersen, *J. Comput. Chem.*, 1997, **18**, 1463–1472.

29 F. Eisenhaber, P. Lijnzaad, P. Argos, C. Sander and M. Scharf,

J. Comput. Chem., 1995, **16**, 273–284.

30 P. Marracino, F. Apollonio, M. Liberti, G. D'Inzeo and A. Amadei, *J. Phys. Chem. B*, 2013, **117**, 2273–2279.

31 M. D'Alessandro, A. Amadei, M. Steiner and M. Aschi, *J. Comput. Chem.*, 2015, **36**, 399–407.

32 T. Matsuo, T. Arata, T. Oda and S. Fujiwara, *Biophysics*, 2013, **9**, 99–106.

33 L. R. Murphy, N. Matubayasi, V. A. Payne and R. M. Levy, *Folding & Design*, 1998, **3**, 105–118.

34 Y. K. Griko, G. I. Makhatadze, P. L. Privalov and R. W. Hartley, *Protein Science*, 1994, **3**, 669–676.

35 T. Imai, S. Ohyama, A. Kovalenko and F. Hirata, *Protein Science*, 2007, **16**, 1927–1933.