Determination of the Interfacial Water Content in Protein-Protein Complexes from Free Energy Simulations

Peter Monecke, Thorsten Borosch, Jürgen Brickmann, and Stefan M. Kast
Eduard Zintl-Institut für Anorganische und Physikalische Chemie, Technische Universität Darmstadt, 64287 Darmstadt, Germany

ABSTRACT The question as to how many tightly or weakly bound water molecules are located in interfaces between protein-protein complex constituents is addressed from a phase equilibrium point of view by developing a theory in the canonical ensemble. A fast method based on free energy simulations is described for computing the number of water molecules in the interface regions. Results are given for 211 interfacial cavities of 26 antigen-antibody complexes for which experimentally determined structures are found in the Protein Data Bank. The accuracy of the method is assessed and the computational water content is compared with experimental data, revealing the amount of water molecules not resolved by experimental approaches.

INTRODUCTION

The stability of protein-protein complexes is determined on one hand by direct interactions between the proteins like hydrogen bonds, salt bridges, hydrophobic effects, and van der Waals contacts. An additional important factor for the stability of protein conformations and of noncovalent protein complexes is, on the other hand, the water content of the interface between the proteins. The water molecules are localized in interfacial cavities formed due to the non-complementarity of the protein surfaces. They stabilize the complexes by acting as hydrogen bond bridges between donors and acceptors of the proteins (1). Experimental investigations of the interfaces identified for protein-protein complex crystal structures revealed an average of 18–20 water molecules, i.e., one molecule per 100 $Å^2$ (2,3) although this number does not reflect the large fluctuations observed. Janin (4) introduced the terminology of dry interfaces if water molecules populate only the perimeter around central water-free binding regions, and wet interfaces if the binding region contains a number of water-filled cavities. For protein-protein complexes and for the particular case of antigen-antibody complexes, interfacial water molecules contribute considerably to the shape complementarity of the binding region and to the enthalpy-driven complex stability (5–9). The true water content in the binding regions is, however, most likely larger than found in the experimental structures collected in the Protein Data Bank (PDB) (10). This is far more likely for larger cavities, since a certain degree of immobility and crystalline local order is necessary for identifying water positions in x-ray experiments (11). Typical exchange times for buried water as determined from NMR spectroscopy vary strongly, from 10^{-8} to 10^{-2} s (11), yet are small enough to ensure a state of thermodynamic equilibrium on experimentally relevant timescales.

Due to the relevance of water molecules for complex stability, it is important to include their presence in empirical scoring functions for predicting binding modes and stability constants. At least for protein-ligand complexes such a strategy is widely accepted (12–15), whereas for modeling protein-protein interactions an efficient method is still lacking. This is mainly due to the fact that no easily applicable computational procedure for filling large, apparently empty cavities is presently available in the context of well-established algorithms for identifying cavities in protein structures (16–21). Based on free energy criteria, Zhang and Herman's DOWSER approach (22) is capable of positioning water molecules in single layers on internal cavity and pocket surfaces. As an alternative, molecular dynamics simulations are an established tool for characterizing hydration features of protein moieties since the late eighties of the past century (23,24). Free energy simulations represent an important methodology to determine the thermodynamic stability of water sites in protein environments (22,25–28). Sequential hydration of monomer cavities has been studied using a free energy model within a continuum electrostatics approach quite recently (29). The question, however, as to how many water molecules, including experimentally invisible ones, occupy a given cavity from a thermodynamic equilibrium point of view has only recently been addressed for the KcsA potassium channel by Roux and co-workers (30). In the latter work, a theory is developed in the grand canonical ensemble within a Monte Carlo simulation framework.

In this work, we present a fast and reliable method for the determination of the total number of water molecules that populate interfaces in protein-protein complexes from standard canonical free energy simulations. We adopt a phase equilibrium point of view between internal water and the bulk phase by directly equating the excess chemical potential, $\Delta\mu$, derived from the simulations of a finite water

Submitted April 29, 2005, and accepted for publication October 20, 2005.

Address reprint requests to Stefan M. Kast, E-mail: kast@pc.chemie.tu-darmstadt.de.

Peter Monecke's present address is Aventis Pharma Deutschland GmbH, Industriepark Höchst G838, 65926 Frankfurt, Germany.

0006-3495/06/02/841/10 $2.00

doi: 10.1529/biophysj.105.065524

amount in the cavities with that of pure water. In this way, full cooperativity between water molecules in both phases as well as the atomic detail of the protein environment is accounted for. Standard state corrections for the finite interfacial cavity volume play only a negligible role.

In the following, we describe the theory and practical aspects of the methodology in detail. The approach is validated and the accuracy is assessed by application to a specific antigen-antibody complex. Simulation parameters were varied to find the computationally most efficient set. Finally, the method was applied to 211 cavities in 26 antigen-antibody complexes. Results are compared with experimentally found water numbers, and future directions of research are discussed.

THEORY

Following Kirkwood (31), we start with the finite-system chemical potential for N particles (assumed to be structureless for simplicity) as the difference between Helmholtz free energies A for systems at constant volume V and constant temperature T that differ by a single molecule,

$$\mu(N) = A(N,V,T) - A(N-1,V,T), \tag{1}$$

and noting that the classical Helmholtz free energy is expressed as

$$-\beta A = \ln Z_{\mathrm{N}} - \ln N! - 3N \ln \Lambda, \tag{2}$$

with $\beta = 1/kT$ (k is the Boltzmann constant), the configuration integral Z_{N}, and the thermal De Broglie wavelength Λ, we get

$$-\beta\mu = \ln\frac{Z_{\mathrm{N}}}{Z_{\mathrm{N-1}}} - \ln N - \ln \Lambda^3. \tag{3}$$

Upon introduction of coupling parameter integration for the ratio of configuration integrals (the coupling parameter λ switches on the interaction between the particle singled out and the rest), we have

$$\ln\frac{Z_{\mathrm{N}}}{Z_{\mathrm{N-1}}} = \ln\frac{Z_{\mathrm{N}}(\lambda=1)}{Z_{\mathrm{N}}(\lambda=0)} + \ln V. \tag{4}$$

The first term on the right-hand side of the latter expression, when multiplied by $-kT$, is identical with the excess chemical potential $\Delta\mu$ such that we finally obtain with the number density $\rho = N/V$

$$\mu = \beta^{-1} \ln \rho\Lambda^3 + \Delta\mu. \tag{5}$$

Assuming equilibrium for a species distributed between the cavity of volume V_{cav} and chemical potential μ_{cav} and the surrounding bulk at a density ρ_{bulk}, chemical potential μ_{bulk}, and the same temperature, we end up with the canonical ensemble expression

$$\Delta\mu_{\mathrm{bulk}} = \Delta\mu_{\mathrm{cav}}(N) + \beta^{-1} \ln \frac{N}{V_{\mathrm{cav}}\rho_{\mathrm{bulk}}}. \tag{6}$$

The physical picture behind the right-hand side of this relation can be interpreted as a three-stage process: First, a virtual volume is populated by a prescribed number N of molecules in an ideal gas state, its size adjusts to the specified bulk density. Second, the volume is compressed/expanded to the actual cavity size, giving rise to the second term. Third, the interactions are switched on, leading to the excess term. The phase equilibrium condition used in this work,

$$\Delta\mu_{\mathrm{bulk}} \approx \Delta\mu_{\mathrm{cav}}(N), \tag{7}$$

then follows from assuming the same ideal gas reference state concerning density and temperature for both subsystems. This assumption is quite reasonable since the volume correction term (the volume itself is actually hard to specify for soft potentials) is very small for our systems, as we shall demonstrate later.

Clearly, such an approach is reliable only if N is large enough to be representative of the average number of molecules in an open system equilibrium within a treatment based on the grand canonical ensemble. Besides the explicit grand canonical simulations for water in biological finite volumes (30), water in spherical cavities has also been treated within the Gibbs ensemble simulation framework by Geiger and co-workers (32) where knowledge of the bulk chemical potential can be avoided. It is clear, however, that for finite systems, and particularly for extremely small ones, the correspondence between first order quantities among different ensembles is no longer given. In this sense, if only a small number of water molecules fits into the cavity, it might well be that the bulk value of the chemical potential cannot be reached by adding or removing an integer number of particles, see González et al. (33) for a thorough discussion. We proceed by introducing an excess free energy interpolation for a continuous number N to be able to differentiate, yielding an equally continuous chemical potential that can be equated with the bulk quantity:

$$\Delta\mu_{\mathrm{cav}}(N) = \left(\frac{\partial \Delta A(N)}{\partial N}\right)_{\mathrm{V,T}}. \tag{8}$$

In practice, this leads to a very simple and cost-effective method to estimate the water content: a), The cavity under consideration (directly taken from the experimental crystal structure and assumed to be completely rigid for the calculations) is sealed with dummy atoms to prevent water from escaping the finite volume; b), the excess Helmholtz free energy is computed (34) for increasing numbers of water molecules using free energy perturbation (FEP) and thermodynamic integration (TI) molecular dynamics simulations, i.e., by gradually inflating/deflating the interaction potentials between water and the protein environment and among the water molecules using a coupling parameter; c), the resulting function of N is smoothed and continuously interpolated by a spline function that allows for differentiation; and d), the optimal water number, N_{opt}, is found by locating the intersection of the derivative with the bulk excess chemical potential of water.

We require the (fractional) expectation number for the occupancy from a grand canonical treatment, which represents an average over an ensemble of cavities, to correspond to the resulting (fractional) number from our method. Although lacking some rigor, we think that this represents a pragmatic way of resolving the finite-size difficulty, particularly if the inherent (and neglected) cavity flexibility is taken into account as an additional effect that perturbs an average integer content. As we will demonstrate, the reliability is stressed in our view by the correct prediction of single water cavities representing the extremal case for our method on one hand and the experimentally best characterized situation on the other. A deeper mathematical treatment will be left open to future work and discussions.

METHODS

Preparation of the simulation systems

The size and shape of the interfacial cavities have to be defined before the free energy calculations to restrict water movement to a closed volume. Water can escape particularly in the case of weak interactions with the protein environment for small values of the coupling parameter connecting initial and final Hamiltonian at the beginning/endpoints of the free energy calculations. To overcome this problem, different approaches were developed. Wade et al. (25) assigned a very large mass to the water oxygen inside the cavity. Such an approach is suitable only for small cavities with rather limited range of water movement. Furthermore, due to the slower dynamics, adequate sampling becomes difficult. Another possibility to confine the water molecules inside the cavity is the application of a harmonic restraining potential to the test particle to keep it close to a reference point. Roux et al. (27) used this method in their investigation of water in the bacteriorhodopsin proton channel. They applied a restraint to the oxygen atom of the water molecules. The reference positions and force constants for the restraints were obtained from evaluation of the water positions and fluctuations in preliminary simulations. A similar approach was used by Helms and Wade (26). They applied a flat-bottomed harmonic well potential to the test particle. This potential is zero if the molecule is within a given radius of a reference point and does, therefore, not artificially perturb the motion of the molecule. In the case of a larger distance from the reference point, the molecule is restrained to a spherical region.

Such restraining procedures need some degree of advance knowledge and are difficult to automate and to apply to nonspherical cavities. Therefore, an alternative grid-based algorithm (see below) was used that seals holes and channels connecting the cavities with the exterior by essentially repulsive dummy Lennard-Jones atoms ($\varepsilon = 0.001$ kcal mol^{-1}, $\sigma = 1$ Å). In this way, the simulations can be done quite efficiently without an embedding water box. For a rigid cavity, free energy simulations directly yield the Helmholtz free energy. A complete free energy simulation run of an average cavity with 10 water molecules takes only ~30 min on a single 1 GHz Pentium III processor.

The coordinates of the complexes were taken from the PDB, removing all experimentally identified water molecules. The program package CHARMM 24g2 (35) with an all atom force field was used for all subsequent modeling and simulation steps. Missing hydrogen atoms were generated using the HBUILD facility of CHARMM (36). After creation of the H atoms, all known heavy and hydrogen atoms of the proteins were kept fixed. The newly created hydrogen sites were optimized with the Newton-Raphson algorithm for 2000 steps or until the change in energy from a minimization step was below 10^{-8} kcal mol^{-1}. The resulting structures were manually split into moieties that form a complex with a single common interface.

The solvent-accessible surfaces of the protein moieties and the entire complex were created using a grid-based algorithm (37) equivalent to the Connolly method (38). A sphere radius of 1.2 Å was used for the protein moieties and 1.8 Å for the entire complex. The complex was subsequently placed into a regular grid with 0.3 Å spacing. Grid points representing the interior of the interfacial cavity were defined as those located outside the solvent accessible surface of the protein moieties and inside the solvent accessible surface of the entire complex. Every grid point matching this condition and with enough space to place a sphere with a radius of 1.2 Å on it without overlapping the van der Waals surface of the surrounding atoms was selected for representing the cavity (Fig. 1, *top*). On these points that define the cavity (shorthand term: cavity points) nonoverlapping spheres with a radius of 0.8 Å were placed to represent possible cavity water positions used as starting points (Fig. 1, *middle*). The radius of 0.8 Å was used to fill the cavity with more water molecules than possible in reality, representing an upper limit to the number of water molecules for the subsequent free energy simulations. The water positions were successively occupied starting with the cavity point with the largest distance from the geometric center of mass of all cavity points. The list of cavity points was subsequently searched for the next available position where a nonoverlapping sphere can be placed. This step was repeated until no further cavity point is available, yielding a strong overestimation of the water number (N_{max}). The required number N of water molecules was selected by deleting the first N_{max}-N entries from the list. After adding hydrogen atoms (HBUILD) to the remaining set of points (assumed to be oxygen sites) and geometry optimization, the resulting configuration formed the starting point for the free energy simulations. The entire protein including the cavity points was placed into a regular grid with 1.0 Å spacing. All grid points within a distance between 2.5 and 3.5 Å from the outer cavity points and with a larger distance than 1.4 Å from a protein atom were selected as dummy atoms (Fig. 1, *bottom*). Dummy atoms were placed separately for each cavity.

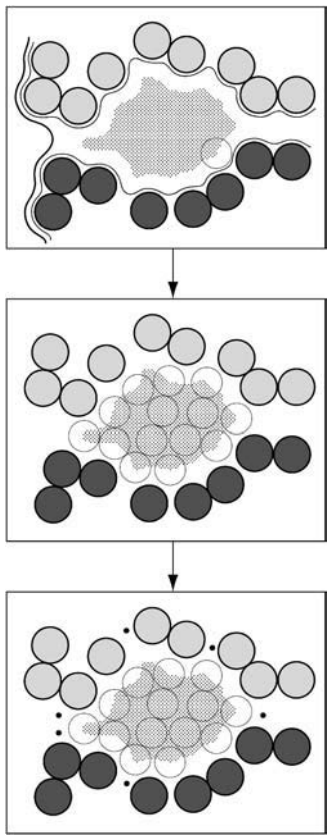

FIGURE 1 Cavity with protein moieties in light and dark gray, points defining the cavity (*dots*, *top*), possible cavity water positions (*middle*), and sealing black dummy atoms (*bottom*).

Free energy simulations

In the free energy simulations, each cavity was subsequently filled with water, and the corresponding Helmholtz free energies for insertion (forward direction, $\Delta A^{0\rightarrow N}$) and removal (backward direction, $\Delta A^{N\rightarrow 0}$) of the water molecules were calculated. All protein moieties, the corresponding dummy atoms, and possible water positions of the cavity were loaded into CHARMM. To reduce CPU time, all protein atoms with a distance larger than $r_c = 10$ Å from a possible water position of the cavity were deleted for production runs. In all calculations, a cutoff scheme was used for nonbonded interactions, with a heuristic update frequency of the nonbonded list. An atom-based switching function was used for the Lennard-Jones potential driving nonbonded interactions to zero between 10 Å and the cutoff distance 12 Å. For the Coulomb potential, an atom-based shifting function with a corresponding cutoff of 12 Å and a relative dielectric constant of unity was applied. All protein and dummy atoms were kept fixed in all calculations. The TIP3P model was used to represent water molecules (39). The cavity water positions were first optimized with the steepest descent method for 500 steps and further with the adopted-basis Newton-Raphson algorithm for 1000 steps or an energy change by $<10^{-8}$ kcal/mol. In all molecular dynamics simulations, the SHAKE algorithm (40) was used to constrain the bonds of the cavity water. A time step of 2 fs was applied with the Verlet algorithm.

The system was heated from 0 K to 298 K within 3000 steps (6 ps). After heating, the temperature was kept constant at 298 K using the Hosé-Hoover thermostat for simulation in the canonical ensemble (41). The free energy calculations were conducted using the TSM module of CHARMM for forward and backward simulations with TI and FEP. The coupling parameter values $\lambda = 0.05, 0.15, \ldots, 0.95$ were used for the simulations with a linear coupling between initial and final states. For the FEP simulations, double wide sampling was employed that also covers the end states while simultaneously avoiding the interaction potential singularity problem. For TI, the missing values of $\lambda = 0$ and $\lambda = 1$ were extrapolated by applying the perturbation expression on the basis of the $\lambda = 0.05$ and $\lambda = 0.95$ ensembles, respectively. Since water interactions do not vanish at any stage, all molecules are effectively restrained in the cavity by the sealing dummy atoms. Mutation simulations were performed in both the creation and annihilation directions, for the whole set of water molecules simultaneously and not using step by step insertion/deletion. Although the risk of a water phase transition exists when using such a procedure, we opted for the direct Helmholtz free energy evaluation since the resulting hysteresis errors from forward and backward runs were much smaller as compared to those from stepwise simulations.

For each of the 10 coupling parameter windows in the production simulations, the systems were equilibrated for 2 ps, starting from the last configuration of the previous window and continued by sampling runs of 10 ps length. Hence, the overall simulation time for one system was 126 ps for heating, equilibration, and sampling. A mean statistical inefficiency over all cavities of the test complex 1VFB used for calibration (see below) was determined to be ~100 time steps (200 fs) from a blocking method (42), indicating strong statistical correlation. The resulting correlation time was taken as the same throughout for all other complexes. The total error was calculated as the sum of hysteresis and the adjusted statistical errors of the creation and annihilation simulations. The free energy was computed with growing amounts of water in the cavity until the mean excess free energy $\Delta A(N)$ started to increase as shown exemplarily for our test complex in Fig. 2 (*top*).

Postprocessing of the free energy simulations

The pair of values $N = 0$, $\Delta A(N) = 0 \pm 0.01$ kcal mol^{-1} was added to the results as a starting point. For evaluating the chemical potential $\Delta\mu_{cav}$ by differentiation (Fig. 2, *bottom*), the mean excess free energy was interpolated by a weighted third order spline approximant (43). The reciprocal total error was used as the weighting factor and a smoothing parameter of 20 times the number of reference points (optimized in preliminary experiments with respect to maximizing smoothness without erasing important details). This ensures that the weight of a free energy value with a large error has less influence on the spline parameters than free energy values with small errors, resulting in smooth curves (Fig. 2, *top*). In the case of only three points ($N = 0, 1, 2$), a third order spline cannot be applied. Therefore a second order spline was used. In cavities with only one cavity point, and therefore also with only one geometrically placed water molecule inside, a direct comparison with the reference value $\Delta\mu_{bulk}$ for bulk TIP3P water was done to decide whether the cavity would be empty or filled: Only for a chemical potential less than or equal to the bulk value can the presence of a water molecule be expected. The reference value used in this work is −6.4 kcal mol^{-1} for pure TIP3P water as given by Beglov and Roux (44). This value corresponds with the most recent estimate for TIP3P water in the full potential limit of –5.8 kcal mol^{-1} (45), corrected with respect to the effect of the shifted-force/cutoff conditions as applied in this work. This correction has been determined to be –0.6 kcal mol^{-1} from an integral equation theory using a formalism recently developed by us (46). Bulk water and cavity water models are thus guaranteed to agree. From the analytical derivative (Fig. 2, *bottom*) the optimal water amount is determined by intersecting the computed curve with the bulk value.

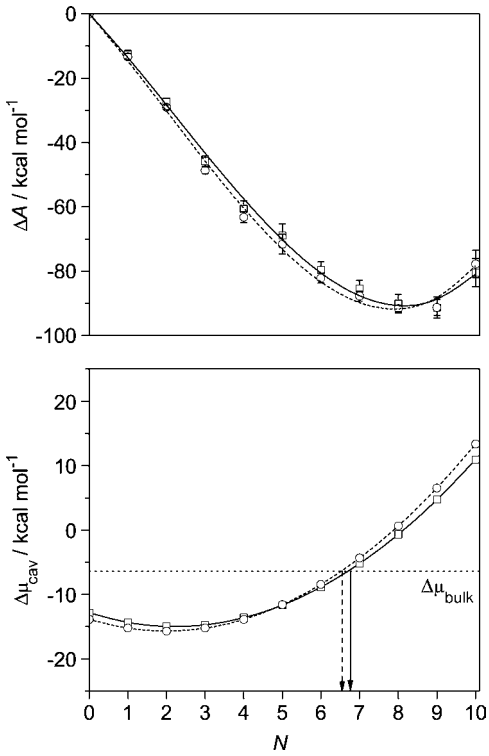

FIGURE 2 Total excess free energy (*top*) and associated chemical potential (*bottom*) as a function of the number of cavity water molecules N in cavity No. 2 of the complex 1VFB; results of TI (*solid line*) and FEP (*dashed line*) calculations. The optimal water amount in this cavity (as indicated by *arrows*) is 6.7 (TI) and 6.5 (FEP).

RESULTS

Calibration of the simulation parameters

We first tried to optimize simulation times and cutoff parameters for the example of the cavities of the Fv fragment of mouse monoclonal antibody D1.3 complexed with hen egg lysozyme (PDB entry 1VFB) (47). This complex was chosen

TABLE 1 Experimental and calculated water content of the cavities in the complex 1VFB from standard ordering of initial positions, varying simulation times and cutoff conditions

		N_{opt}(reference: 10 Å, 2/10 ps)		N_{opt}(6 Å, 2/10 ps)		N_{opt}(12 Å, 2/10 ps)		N_{opt}(10 Å, 4/20 ps)		N_{opt}(10 Å, 8/40 ps)	
Cav.	N_{exp}	TI	FEP	TI	FEP	TI	FEP	TI	FEP	TI	FEP
1	11	19.9	19.5	20.8	20.6	20.0	19.8	20.1	19.9	19.9	19.6
2	4	6.7	6.5	7.0	6.8	6.7	6.5	6.7	6.5	6.7	6.5
3	2	3.9	3.8	3.9	3.9	3.9	3.9	3.9	3.8	4.0	3.9
4	1	1.0	0.9	0.9	0.8	0.7	0.7	1.1	1.0	1.1	0.9
5	4	8.2	8.1	8.2	8.1	8.0	8.0	8.2	8.1	8.3	8.0

Reference simulation: cutoff $r_c = 10$ Å, 2 ps equilibration/10 ps sampling per λ-window (short notation: 10 Å, 2/10 ps); for $r_c = 6$ and 12 Å: 2/10 ps; for 4/20 ps and 8/40 ps: $r_c = 10$ Å.

because small, medium-sized, and large cavities are found in the interface representing the typical variety. Differences between FEP and TI results, furthermore, give some indication about the robustness of the presented approach.

In addition to the reference simulation using only protein atoms within the 10 Å cutoff spheres around the cavity points, simulations with cutoffs of 6 Å and 12 Å were carried out to investigate the influence of the cutoff distance on the results. Further calculations with a 10 Å cutoff around the cavity were performed with 4 ps equilibration/20 ps sampling time and 8 ps equilibration/40 ps sampling per coupling parameter window to determine whether the simulation time of the reference simulation (2 ps equilibration/10 ps simulation time per window) was long enough to sample the cavity sufficiently for converged results.

The results of the free energy calculations (N_{opt}) of the complex 1VFB in comparison with the experimental numbers (N_{exp}) of solvent molecules from the x-ray diffraction structure are given in Table 1. Only those experimental water molecules are counted for which the distance between the oxygen atom and any cavity point is not larger than 0.3 Å. As a criterion emphasizing the quality of the simulation approach, the single water molecule found in cavity 4 is also consistently found computationally with numbers close to one for any parameter setting, solely on the basis of excess chemical potentials excluding any volume term. The importance of continuous differentiation is, for instance, seen by the fractional occupation number of 6.5 for cavity 2 (reference simulation, FEP), meaning that the probability of finding six or seven water molecules is roughly equal if a symmetric distribution is assumed and as far as the interpolation approach is accepted to be equivalent to a grand canonical treatment.

The simulations with different cutoffs around the cavities reveal that the cavity water content from the 6 Å cutoff simulations differs from that of the reference simulation with a 10 Å cutoff particularly for the large cavities, giving a discrepancy of a maximum of one water molecule. The differences between the simulations with cutoffs of 10 and 12 Å are smaller and differ by no more than 0.3. The comparison of the results for different simulation times shows no significant changes of the optimal cavity water content determined from longer simulation times. This suggests that the reference simulation parameters (2 ps equilibration/10 ps sampling per coupling parameter window) are adequate for sufficient sampling of the accessible phase space within the cavities. The difference between FEP and TI results (see Fig. 2) is, furthermore, a measure for the robustness of the method: For small cavities this difference is almost negligible, for the largest cavity the discrepancy is ~0.5 water molecules. This number reflects the maximum error to be expected from this approach though the statistical uncertainty in the reference value for $\Delta\mu_{bulk}$ adds a small further amount.

To check other sources of error, we have done two further experiments: First, the influence of the initial water positions was estimated by simply inverting the initial position order, i.e., by taking the first instead of the last elements from the list of accessible locations. As expected (see Table 2), the

TABLE 2 Calculated water content of the cavities in the complex 1VFB from simulations under reference production conditions, varying initial conditions, including/excluding volume term, cavity volume V_{cav}, cavity density ρ_{cav}, and volume contribution $\Delta\mu_{id}$ to chemical potential

	N_{opt}(std. order, no vol. term)		N_{opt}(inv. order, no vol. term)		N_{opt}(std. order: with vol. term)		N_{opt}(inv. order: with vol. term)				
Cav.	TI	FEP	TI	FEP	TI	FEP	TI	FEP	V_{cav}/Å^3	ρ_{cav}/Å^{-3}	$\Delta\mu_{id}$/kcal mol^{-1}
1	19.9	19.5	20.0	20.0	19.8	19.5	19.9	20.0	510.2	0.0386	0.08
2	6.7	6.5	6.4	6.3	6.7	6.5	6.4	6.3	166.8	0.0396	0.10
3	3.9	3.8	4.0	3.9	3.9	3.8	3.9	3.8	72.3	0.0533	0.28
4	1.0	0.9	1.0	0.9	1.0	0.9	1.0	0.9	28.0	0.0339	0.01
5	8.2	8.1	8.6	8.4	8.1	8.0	8.5	8.3	183.6	0.0444	0.17

All simulations: reference conditions (cutoff $r_c = 10$ Å, 2 ps equilibration/10 ps sampling per λ-window); N_{opt}(std. order, no vol. term) corresponds to N_{opt}(reference) in Table 1, $\rho_{cav} = <N_{opt}(\text{std. order, no vol. term})>_{TI,FEP}/V_{cav}$, $\Delta\mu_{id} = \beta^{-1}\ln(\rho_{cav}/\rho_{bulk})$ with $\rho_{bulk} = 0.0334$ Å^{-3}.

influence of the starting positions is small; the data actually confirm our maximum error estimate of ~0.5 water molecules per cavity. Additionally, FEP simulation turns out to be slightly more robust as compared to TI in terms of consistency of the results under varying initial conditions. Due to the sufficiently strong structural perturbations during energy minimization and simulations, we have not attempted to select a randomized initial set to maintain reproducibility.

Second, the volume term in Eq. 6 has been evaluated for the average (over TI and FEP results) optimal water amount from the reference simulations with standard ordering by considering a cavity volume approximated by the Connolly surface of a cluster of spheres that can be filled into the cavity (using a radius of 1.2 Å for both, cluster creation, and surface generation). The results are also summarized in Table 2: The resulting volume work (for $\rho_{bulk} = 0.0334$ Å^{-3}) is on average 0.13 kcal mol^{-1} and reaches a maximum of 0.28 kcal mol^{-1} for the model complex 1VFB, much smaller than the excess chemical potential for the liquid water model and within range of the respective statistical error. It is therefore safe to neglect the volume term. To prove this assertion we have done additional calculations for 1VFB, this time with explicit consideration of the volume term and standard initial ordering (note that one should avoid $N = 0$ for the spline computation in this case; we used $N = 0.1$ and $\Delta A = 0$ as the lower boundary in this case): The maximum discrepancy in the predicted number of water molecules is 0.1, as also given in Table 2. One should note in particular that the approximation is clearly the better justified the larger the cavity is, i.e., in the most interesting cases where experiments typically resolve far too few water molecules.

Analysis of antigen-antibody complexes

Finally, with the simulation settings determined to be optimal in the preceding section (10 Å cutoff around the cavity and 2 ps equilibration/10 ps sampling time per window, standard initial ordering, neglect of volume term), the methodology was applied to 211 interfaces of 26 antigen-antibody complexes found in the PDB. With such a large database it is possible to estimate the tendencies for the amount of experimentally invisible buried water molecules.

In Table 3, the number of experimentally visible water molecules and the calculated optimal water contents from TI and FEP simulations is summarized. Again, experimental water is counted only for oxygen atom/cavity point distance not larger than 0.3 Å. The cavity water content from TI and FEP evaluation methods in Table 3 are very similar. A correlation plot of both results is shown in Fig. 3, being very close to the ideal diagonal. Pearson's correlation coefficient yields 0.99943, indicating a high degree of consistency. In Fig. 4, the calculated water numbers are plotted against the corresponding experimentally found water content according to Table 3. As expected, the calculated water amount is significantly larger (roughly by a factor of two) than the

TABLE 3 Experimental and calculated water content for 211 cavities in 26 antigen-antibody complexes

PDB code	Cav.	N_{exp}	N_{opt}(TI)	N_{opt}(FEP)
1A2Y	1	2	4.6	4.5
	2	5	7.0	6.8
	3	2	2.7	2.7
	4	1	1.2	1.1
	5	0	0.5*	0.5*
	6	2	4.5	4.4
1AR1	1	0	2.4	2.3
	2	0	3.5	3.4
	3	0	2.1	2.0
	4	0	0.9	0.9
	5	0	1.4	1.4
	6	0	0.5	0.5
	7	2	4.4	4.3
	8	0	0.5	0.5
	9	0	1.5	1.5
1BOG	1	1	2.3	2.3
	2	0	2.0	1.9
	3	0	1.2	1.1
	4	0	1†	1†
	5	0	1.1	1.0
	6	0	4.9	4.7
1BQl	1	3	8.5	8.4
	2	2	6.4	6.2
	3	0	0.9	0.9
	4	0	1.2	1.2
	5	1	6.9	7.1
	6	1	2.0	1.6
	7	1	1.2	1.2
	8	0	1.5	1.4
	9	0	1.5	1.4
	10	0	1.5	1.5
	11	0	3.0	3.0
1CIC	1	0	2.4	2.4
	2	4	7.5	7.4
	3	1	3.4	3.4
	4	1	0.7	0.7
	5	0	0.7	0.7
	6	1	1.8	1.7
	7	0	0.5	0.5
1CU4	1	0	2.3	2.2
	2	0	1.0	1.0
	3	1	6.3	6.2
1DQJ	1	1	3.6	3.6
	2	0	1†	1†
	3	0	0.5	0.5
	4	1	1.4	1.4
	5	1	1†	1†
	6	0	0.9	0.9
	7	0	1.2	1.2
	8	1	1†	1†
	9	0	1.6	1.6
	10	0	0.6	0.6
	11	1	1.5	1.4
	12	2	2.3	2.2
	13	0	2.9	2.9
1EJO	1	1	2.8	2.7
	2	0	1.8	1.8

(*Continued*)

Table 3 (*Continued*)

PDB code	Cav.	N_{exp}	N_{opt}(TI)	N_{opt}(FEP)
	3	1	2.4	2.4
	4	0	1.2	1.2
	5	1	1.6	1.6
1EO8	1	0	1.1	1.1
	2	0	2.4	2.4
	3	0	0.8	0.7
	4	2	3.0	2.9
	5	1	2.6	2.6
	6	0	1.6	1.6
	7	0	0.3‡	0.3‡
	8	0	0.7	0.6
	9	0	0.7	0.7
1FNS	1	1	1.1	1.1
	2	1	1.2	1.2
	3	2	4.4	4.3
	4	1	2.3	2.2
	5	1	2.8	2.7
1IGC	1	0	0.7	0.7
	2	1	4.2	4.0
	3	1	2.0	2.0
	4	0	1†	1†
	5	0	0.8	0.8
	6	0	0‡	0‡
	7	0	0‡	0‡
	8	0	1.9	2.0
	9	0	0.6	0.6
1JRH	1	0	3.5	3.4
	2	0	6.0	5.8
	3	0	1.8	1.8
	4	1	0.7*	0.7*
	5	0	1.9	1.9
	6	0	1.5	1.5
1KIP	1	8	22.3	21.9
	2	3	5.9	5.8
	3	2	2.7	2.7
	4	1	1.0	1.0
	5	0	3.4	3.4
1KIQ	1	1	2.3	2.2
	2	2	3.7	3.7
	3	4	6.0	5.9
	4	2	3.2	3.1
	5	0	0.7*	0.7*
	6	0	0†,‡	0†,‡
	7	1	0.7	0.7
	8	0	0.3*	0.3*
	9	0	1†	1†
	10	1	4.1	4.0
1KIR	1	9	18.6	18.3
	2	4	6.1	6.0
	3	2	2.6	2.6
	4	1	0.9	0.9
	5	1	1.5	1.5
	6	1	2.4	2.3
1MLC	1	0	0.5	0.5
interface 1	2	0	0.6*	0.5*
	3	0	2.2	2.0
	4	1	0.6*	0.6*

(*Continued*)

Table 3 (*Continued*)

PDB code	Cav.	N_{exp}	N_{opt}(TI)	N_{opt}(FEP)
	5	0	0.8	0.8
	6	2	2.1	2.0
	7	0	1†	1†
	8	0	1†	1†
	9	0	0.2*	0.2*
1MLC	1	0	0.7	0.6
Interface 2	2	2	2.1	2.1
	3	0	0.8*	0.7*
	4	1	1†	1†
	5	0	1.9	1.9
	6	0	0.6	0.5
	7	0	0.6	0.6
	8	0	0†,‡	0†,‡
1NCA	1	0	1.9	1.8
	2	0	1.4	1.4
	3	0	2.4	2.5
	4	0	1†	1†
	5	0	0†,‡	0†,‡
	6	0	1†	1†
	7	0	2.8	2.9
	8	0	0.6	0.6
	9	0	0*,‡	0*,‡
	10	0	0.1‡	0.6‡
1NCB	1	0	0.5	0.5
	2	0	1.7	1.7
	3	0	2.8	2.8
	4	0	1.9	1.8
	5	0	0.8	0.8
	6	0	0.7	0.7
	7	0	2.5	2.4
	8	0	3.5	3.4
	9	1	2.4	2.4
	10	0	1.1	1.1
	11	0	0.4*,‡	0.4*,‡
1NCD	1	0	0.6	0.5
	2	0	1.5	1.5
	3	0	1.3	1.2
	4	0	2.4	2.4
	5	0	0.6	0.6
	6	0	0†,‡	0†,‡
	7	0	1.6	1.5
	8	0	3.0	2.8
	9	0	0.4‡	0.4‡
	10	0	0.6‡	0.3‡
	11	0	1.3	1.3
	12	0	1†	1†
	13	0	1.0	1.0
1OAK	1	0	1†	1†
	2	0	0†,‡	0†,‡
	3	1	0.8	0.8
	4	2	3.6	3.6
	5	1	1.9	1.9
	6	1	2.1	2.1
1OSP	1	7	12.0	11.7
	2	0	1.1	1.1
	3	0	1†	1†
	4	0	0.7*	0.6*
	5	1	2.3	2.2

(*Continued*)

Table 3 (*Continued*)

PDB code	Cav.	N_{exp}	N_{opt}(TI)	N_{opt}(FEP)
	6	0	0†,‡	0†,‡
	7	0	2.1	1.9
	8	1	1.2	1.2
1QFU	1	0	0‡	0‡
	2	0	0.7	0.7
	3	0	1†	1†
	4	0	0.5	0.5
	5	0	5.5	5.4
	6	2	4.4	4.4
	7	0	1†	1†
	8	1	1.8	1.7
	9	1	1.9	1.9
	10	0	1.3	1.3
	11	1	2.4	2.5
1VFB	1	11	19.9	19.5
	2	4	6.7	6.5
	3	2	3.9	3.8
	4	1	1.0	0.9
	5	4	8.2	8.1
2IFF	1	0	8.8	8.3
	2	0	3.1	3.0
	3	0	0.4	0.4
	4	0	0.9*	0.8*
	5	2	7.3	7.1
	6	0	2.3	2.3
	7	1	1.0	1.0
	8	0	0.3	0.4
	9	0	1.6	1.6
	10	0	3.7	3.6
	11	0	1.5	1.5
3HFM	1	0	4.9	5.0
	2	0	4.1	4.0
	3	0	1.3	1.3
	4	0	1.4	1.4
	5	0	5.7	5.7
	6	0	0‡	0‡
	7	0	1.3	1.2
	8	0	1.4	1.3
	9	0	0.8	0.8

*Evaluation with second order spline.
†Only one cavity point, therefore no spline evaluation possible but only direct comparison with bulk value for chemical potential.
‡$\Delta\mu_{cav}(N) > \Delta\mu_{bulk}(N)$ for all numbers N of cavity water.

experimental number for medium-sized and large cavities, sometimes even dramatically. But also for the smaller cavities we typically find water where the experiment does not resolve a single molecule. A deeper look at the packing structure will be necessary for understanding the relation between cavity shape/surface properties and water content and mobility to understand the sources for the lack of experimental visibility.

CONCLUDING REMARKS

This work provides a computational methodology for determining the water content of interfaces of protein-protein complexes that complements unclear or missing structural information from experimental investigations due to water mobility and/or amorphous character. Our approach is based on the assumption that the cavity water is in a state of thermodynamic equilibrium with the surrounding bulk water phase, allowing for fractional, i.e., noninteger occupation numbers even though a canonical ensemble formalism is applied instead of the formally correct grand canonical treatment. The related cavity water chemical potential is derived from standard canonical free energy simulations as opposed to computationally more demanding grand canonical procedures. The computational approach makes use of a simple strategy to restrict water movement to the cavity, thereby allowing for a drastically reduced model of the protein. As a consequence, the water content can be computed reliably within minutes on a single CPU. We have characterized the accuracy and robustness of the method by determining the

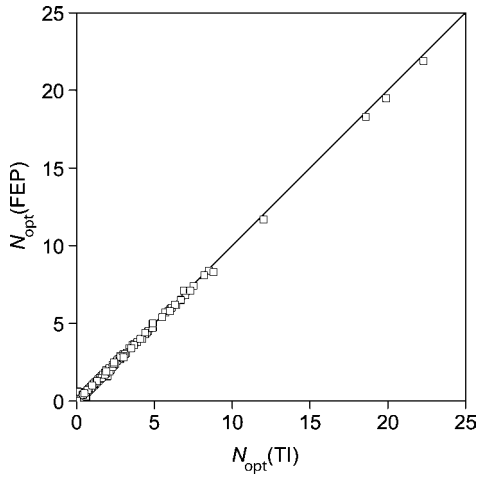

FIGURE 3 FEP versus TI values of optimal cavity water molecule amount in antigen-antibody complexes using data from Table 3.

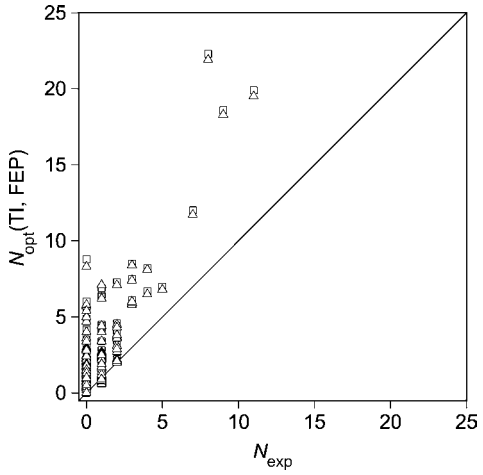

FIGURE 4 Calculated versus experimentally found cavity water numbers in antigen-antibody complexes using data from Table 3. (□) TI; (△) FEP results.

error as the difference between different simulation protocols and by comparison with experiments. The technique was ultimately applied to a large database of 211 cavities in antigen-antibody complexes, revealing the extent of missing water information.

A number of questions are of course still open and need further elaboration: The key assumption developed in this work, namely the correspondence between grand canonical ensemble expectation values for the occupancy and the result of continuous free energy interpolation for small numbers, needs further mathematical analysis. We have, furthermore, not yet taken a closer look at the true water mobility and packing structure as a function of cavity shape and surface properties to explain which water molecules should be visible experimentally. The solvent-free simulation conditions applied in this work should, however, allow for very large simulation times to facilitate deeper insight. Additionally, we did not investigate the generality of our findings for other types of complexes besides antigen-antibody structures where more experimental data might be available. Even more important for future research will be the development of empirical water packing models for even more rapid estimation of the water content. Such an approach will ultimately be necessary for the development of robust and accurate novel scoring functions for characterizing protein-protein complex stability. The results of this work will guide the parametrization of such a model.

Financial support by the Fonds der Chemischen Industrie, the Adolf-Messer-Stiftung, and the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

REFERENCES

1. Purkiss, A., S. Skoulakis, and J. M. Goodfellow. 2001. The protein-solvent interface: a big splash. *Philos. Trans. R. Soc. Lond. A.* 359: 1515–1527.
2. Conte, L. L., C. Chothia, and J. Janin. 1999. The atomic structure of protein-protein recognition sites. *J. Mol. Biol.* 285:2177–2198.
3. Xu, D., C. J. Tsai, and R. Nussinov. 1997. Hydrogen bonds and salt bridges across protein-protein interfaces. *Protein Eng.* 10:999–1012.
4. Janin, J. 1999. Wet and dry interfaces: the role of solvent in protein-protein and protein-DNA recognition. *Struct. Fold. Des.* 7:R277–R279.
5. Braden, B. C., B. A. Fields, and R. J. Poljak. 1995. Conservation of water molecules in an antibody-antigen interaction. *J. Mol. Recognit.* 8:317–325.
6. Bhat, T. N., G. A. Bentley, G. Boulot, M. I. Greene, D. Tello, W. Dall'Acqua, H. Souchon, F. P. Schwarz, R. A. Mariuzza, and R. J. Poljak. 1994. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc. Natl. Acad. Sci. USA.* 91:1089–1093.
7. Covell, D. G., and A. Wallqvist. 1997. Analysis of protein-protein interactions and the effects of amino acid mutations on their energetics. The importance of water molecules in the binding epitope. *J. Mol. Biol.* 269:281–297.
8. Davies, D. R., and G. H. Cohen. 1996. Interactions of protein antigens with antibodies. *Proc. Natl. Acad. Sci. USA.* 93:7–12.
9. Fields, B. A., F. A. Goldbaum, W. Dall'Acqua, E. L. Malchiodi, A. Cauerhff, F. P. Schwarz, X. Ysern, R. J. Poljak, and R. A. Mariuzza. 1996. Hydrogen bonding and solvent structure in an antigen-antibody interface. Crystal structures and thermodynamic characterization of three Fv mutants complexed with lysozyme. *Biochemistry.* 35:15494–15503.
10. Bernstein, F. C., T. F. Koetzle, G. J. Williams, E. E. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. 1977. The Protein Data Bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112:535–542.
11. Levitt, M., and B. H. Park. 1993. Water: now you see it, now you don't. *Structure.* 1:223–226.
12. Raymer, M. L., P. C. Sanschagrin, W. F. Punch, S. Venkataraman, E. D. Goodman, and L. A. Kuhn. 1997. Predicting conserved water-mediated and polar ligand interactions in proteins using a K-nearest-neighbors genetic algorithm. *J. Mol. Biol.* 265:445–464.
13. Garcia-Sosa, A. T., R. L. Mancera, and P. M. Dean. 2003. WaterScore: a novel method for distinguishing between bound and displaceable water molecules in the crystal structure of the binding site of protein-ligand complexes. *J. Mol. Model. (Online).* 9:172–182.
14. Rarey, M., B. Kramer, and T. Lengauer. 1999. The particle concept: placing discrete water molecules during protein-ligand docking predictions. *Proteins.* 34:17–28.
15. Pastor, M., G. Cruciani, and K. A. Watson. 1997. A strategy for the incorporation of water molecules present in a ligand binding site into a three-dimensional quantitative structure-activity relationship analysis. *J. Med. Chem.* 40:4089–4102.
16. Voorinholt, R., and W. G. Hol. 1989. A very fast program for visualizing protein surfaces, channels and cavities. *J. Mol. Graph.* 7: 243–245.
17. Delaney, J. S. 1992. Finding and filling protein cavities using cellular logic operations. *J. Mol. Graph.* 10:174–177.
18. Levitt, D. G., and L. J. Banaszak. 1992. POCKET: a computer graphics method for identifying and displaying protein cavities and their surrounding amino acids. *J. Mol. Graph.* 10:229–234.
19. Goodfellow, J. M., W. R. Pitt, O. S. Smart, and M. A. Williams. 1995. New methods for the analysis of the protein-solvent interface. *Comput. Phys. Commun.* 91:321–329.
20. Laskowski, R. A. 1995. SURFNET: a program for visualizing molecular surfaces, cavities, and intermolecular interactions. *J. Mol. Graph.* 13:323–330.
21. Exner, T., M. Keil, G. Moeckel, and J. Brickmann. 1998. Identification of substrate channels and protein cavities. *J. Mol. Model. (Online).* 4: 340–343.
22. Zhang, L., and J. Hermans. 1996. Hydrophilicity of cavities in proteins. *Proteins.* 24:433–438.
23. Levitt, M., and R. Sharon. 1988. Accurate simulation of protein dynamics in solution. *Proc. Natl. Acad. Sci. USA.* 85:7557–7561.
24. Ahlström, P., O. Teleman, and B. Jönsson. 1988. Molecular dynamics simulation of interfacial water structure and dynamics in parvalbumin solution. *J. Am. Chem. Soc.* 110:4198–4203.
25. Wade, R. C., M. H. Mazor, and J. A. McCammon. 1991. A molecular dynamics study of thermodynamic and structural aspects of the hydration of cavities in proteins. *Biopolymers.* 31:919–931.
26. Helms, V., and R. Wade. 1995. Thermodynamics of water mediating protein-ligand interactions in cytochrome P450cam: a molecular dynamics study. *Biophys. J.* 69:810–824.
27. Roux, B., M. Nina, R. Pomes, and J. C. Smith. 1996. Thermodynamic stability of water molecules in the bacteriorhodopsin proton channel: a molecular dynamics free energy perturbation study. *Biophys. J.* 71: 670–681.
28. Helms, V., and R. Wade. 1998. Hydration energy landscape of the active site cavity in cytochrome P450cam. *Proteins.* 32:381–396.
29. Borodich, A. I., and G. M. Ullmann. 2004. Internal hydration of protein cavities: studies on BPTI. *Phys. Chem. Chem. Phys.* 6:1906–1911.
30. Woo, H.-J., A. R. Dinner, and B. Roux. 2004. Grand canonical Monte Carlo simulations of water in protein environments. *J. Chem. Phys.* 121:6392–6400.

31. Kirkwood, J. G. 1935. Statistical mechanics of fluid mixtures. *J. Chem. Phys.* 3:300–313.

32. Brovchenko, I., D. Paschek, and A. Geiger. 2000. Gibbs ensemble simulation of water in spherical cavities. *J. Chem. Phys.* 113:5026–5036.

33. González, A., J. A. White, and F. L. Román. 1998. How the structure of a confined fluid depends on the ensemble: hard spheres in a spherical cavity. *J. Chem. Phys.* 109:3637–3650.

34. Kollman, P. A. 1993. Free energy calculations: applications to chemical and biochemical phenomena. *Chem. Rev.* 93:2395–2417.

35. Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. 1983. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *J. Comput. Chem.* 4:187–217.

36. Brunger, A. T., and M. Karplus. 1988. Polar hydrogen positions in proteins: empirical energy placement and neutron diffraction comparison. *Proteins.* 4:148–156.

37. Keil, M. 2002. Modellierung und Vorhersage von Strukturen biomolekularer Assoziate auf der Basis von statistischen Datenbankanalysen. PhD thesis. Technische Universität Darmstadt (D17), Darmstadt, Germany.

38. Connolly, M. L. 1983. Solvent-accessible surfaces of proteins and nucleic acids. *Science.* 221:709–713.

39. Jorgensen, L. W., J. Chandrasekhar, J. D. Madura, R. W. Impey, and M. L. Klein. 1983. Comparison of simple potential functions for simulating liquid water. *J. Chem. Phys.* 79:926–935.

40. Ryckaert, J. P., G. Ciccotti, and H. J. C. Berendsen. 1977. Numerical integration of the Cartesian equations of motion of a system with constraints: molecular dynamics of n-alkanes. *J. Comput. Phys.* 23: 327–341.

41. Evans, D. J., and B. L. Holian. 1985. The Nose-Hoover thermostat. *J. Chem. Phys.* 83:4069–4074.

42. Allen, M. P., and D. J. Tildesley. 1989. Computer Simulation of Liquids. Oxford Science Publications, Clarendon Press, Oxford.

43. Dierckx, P. 1975. An algorithm for smoothing, differentiation and integration of experimental data using spline functions. *J. Comput. Appl. Math.* 1:165–184.

44. Beglov, D., and B. Roux. 1994. Finite representation of an infinite bulk system: Solvent boundary potential for computer simulations. *J. Chem. Phys.* 100:9050–9063.

45. Deng, Y., and B. Roux. 2004. Hydration of amino acid side chains: nonpolar and electrostatic contributions calculated from staged molecular dynamics free energy simulations with explicit water molecules. *J. Phys. Chem. B.* 108:16567–16576.

46. Schilling, B., J. Brickmann, and S. M. Kast. 2006. Hybrid simulation/integral equation model for enhancing free energy computations. *Phys. Chem. Chem. Phys.* In press. DOI:10.1039/b514185k.

47. Bhat, T. N., G. A. Bentley, G. Boulot, M. I. Greene, D. Tello, W. Dall'Acqua, H. Souchon, F. P. Schwarz, R. A. Mariuzza, and R. J. Poljak. 1994. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc. Natl. Acad. Sci. USA.* 91:1089–1093.