Excluded Volume Effect for Large and Small Solutes in Water

M. V. Basilevsky,[†] F. V. Grigoriev,[‡,§] I. V. Leontyev,[†] and V. B. Sulimov*[,‡,§]

Department of Quantum Chemistry, Algodign, LLC, Bolshaya Sadovaya 8, 123379 Moscow, Russia, and Karpov Institute of Physical Chemistry, ul. Vorontsovo Pole, 10, 105064 Moscow, Russia

Received: March 10, 2005; In Final Form: May 31, 2005

The cavitation effect, i.e., the process of the creation of a void of excluded volume in bulk solvent (a cavity), is considered. The cavitation free energy is treated in terms of the information theory (IT) approach [Hummer, G.; Garde, S.; Garcia, A. E.; Paulaitis, M. E.; Pratt, L. R. *J. Phys. Chem. B* **1998**, *102*, 10469]. The binomial cell model suggested earlier is applied as the IT default distribution p_m for the number m of solute (water) particles occupying a cavity of given size and shape. In the present work, this model is extended to cover the entire range of cavity size between small ordinary molecular solutes and bulky biomolecular structures. The resulting distribution consists of two binomial peaks responsible for producing the free energy contributions, which are proportional respectively to the volume and to the surface area of a cavity. The surface peak dominates in the large cavity limit, when the two peaks are well separated. The volume effects become decisive in the opposite limit of small cavities, when the two peaks reduce to a single-peak distribution as considered in our earlier work. With a proper interpolation procedure connecting these two regimes, the MC simulation results for model spherical solutes with radii increasing up to $R = 10$ Å [Huang, D. H.; Geissler, P. L.; Chandler, D. *J. Phys. Chem. B* **2001**, *105*, 6704] are well reproduced. The large cavity limit conforms to macroscopic properties of bulk water solvent, such as surface tension, isothermal compressibility and Tolman length. The computations are extended to include nonspherical solutes (hydrocarbons C_1-C_6).

1. Introduction

The concept of the cavitation effect is an important element of the recent theory of solvation. The "cavity" is defined as a void of excluded volume in the bulk solvent prepared in order to accommodate the solute particle. Constituting, along with the van der Waals interaction contribution, the nonelectrostatic solvation component, the free energy of cavity formation is now considered as the main origin of the hydrophobic effect.[1−4]

The corresponding free energy change is a consequence of reorganization of the medium structure around the cavity. This reorganization takes place throughout the total volume of bulk solvent and represents a complicated many-particle collective phenomenon mainly responsible for the entropy increase accompanying the solvation process. This is why it is not as tractable as other components of solvation free energy, which are well understood in terms of conventional van der Waals and electrostatic potentials.

The theory of the cavitation effect is an active area of recent research.[1,4−13] The objective is to construct a unified approach covering the whole range of cavity sizes: from voids including small and ordinary organic molecules (the radius of the sphere boundary around the methane molecule is approximately 3 Å) up to extremely large cavities associated, for example, with protein structures, which can contain as many as 10^3-10^4 solvent particles.

The fundamental idea in the statistical theory of cavitation is based on the notion that formation of a cavity with given size and shape in bulk solvent can be identified as a density fluctuation. Denoting the probability of this fluctuation as P_0, the cavitation free energy is expressed as[14,15]

$$\Delta G_{cav} = -k_B T \ln P_0 \tag{1}$$

Earlier work was aimed at modeling this probability.[3,16−19] Later computer simulations[5,6,10,20−22] prompted a more appropriate formulation. It was suggested[1,5−7] to consider P_0 as a first member ($m = 0$) of the probability distribution P_m for observing exactly m solvent particles in a given cavity. The following prescription for modeling P_m was devised. As a first step, a primary (default) distribution p_m must be introduced as an initial guess based on some intuitive physical idea. The information theory (IT) then provides a method for systematically improving p_m by imposing additional constraints. These constraints are introduced in terms of exact values of the first two distribution moments $\langle m \rangle$ and $\langle m^2 \rangle$ which are available either from a simulation or from experimentation and which must be fitted under the condition of perturbing to a minimum extent the default distribution. In this way, the algorithm to express the resulting P_m as a modified default p_m reduces to a solution of IT equations in which exact values of $\langle m \rangle$ and $\langle m^2 \rangle$ are inserted as input data.

In the main body of original work,[5−7,23−25] no physical assumption was introduced in the default distribution. The p_m were considered as m-independent quantities, which provided a Gaussian law for p_m after performing the IT refinement procedure. This approach proved to be successful only for small cavities ($R < 3-4$ Å). In our previous work, we considered a more sophisticated binomial distribution

$$p_m = \binom{n}{m} y^m (1-y)^{n-m} \tag{2}$$

* To whom correspondence should be addressed. E-mail: vladimir.sulimov@gmail.com.
[†] Karpov Institute of Physical Chemistry.
[‡] Algodign, LLC.
[§] Current address: Research Computing Center, Moscow State University, Vorobjovy Gory, 119992 Moscow, Russia.

10.1021/jp051246z CCC: $30.25
Published on Web 07/19/2005

This model[13] finds its roots in the cell theory of dense fluids, which has a long history.[26−30] In the present context, the cell must be defined as a maximally sized element of space that can accommodate at most one solvent particle. Its shape is not explicitly specified; rather, one treats the distribution parameter n in (2) as the number of cells in a given cavity, thus absorbing implicitly any reference to the cell shape. The second parameter y in (2) means the probability of finding a single solvent particle in a given cell.

Having the IT refinement implemented, the binomial model proved to work well for spherical cavities with radii denoted as R up to $R = 6.4$ Å (capable of including 30−40 water molecules). It also successfully treated nonspherical real solutes,[13] namely, C_1-C_6 hydrocarbons, including linear, branched, and cyclic structures.[12] The prediction that ΔG_{cav} is proportional to the cavity volume v within the considered range of cavity sizes is in accord with recent conclusions of other authors.[9,10,31]

The binomial model, however, failed in attempts to treat large cavities ($R > 10$ Å). It can be shown that the default distribution (2) predicts the corresponding asymptotic ($R \rightarrow \infty$) result that $\Delta G_{cav} \propto v$, whereas it is the different law $\Delta G_{cav} \propto S$ (with S the surface area of the cavity boundary) that follows from the macroscopic thermodynamical consideration.[14,15,32] Creation of very large cavities obeys a macroscopic law consistent with the assumption that the thermodynamics of cavitation can be treated in terms of a continuum hydrostatic model of the water fluid.[4,32] This phenomenon originates from strong attractive forces between water molecules and results in a specific behavior of density fluctuations in the vicinity of the cavity boundary.[3,4,8,9] It can be also formulated as a tendency for those water molecules confined inside the cavity volume to stick to the cavity surface. The recent theoretical treatment of these effects is based on the phenomenological density functional methodology which is interpolated between large and small cavity extremes.[4,8,9]

The present work investigates an alternative route to a unified cavitation theory. We suggest a modification of the binomial default model which allows for properly describing the large cavity limit. The connection of the two extreme cases is performed by interpolation between the two types of the default distributions for large and small cavities, in analogy with the interpolation between the two extremal density functionals, as elaborated by Chandler, Weeks, Sun, and their co-workers.[4,8,9]

2. Two-Peak Default Model

The binomial distribution (2) was derived by counting different possibilities for arrangement of holes (i.e., empty cells) in the fluid considered as a lattice of cells. Being essentially combinatorial, this result represents a purely entropic effect and ignores the interaction between the particles contained in different cells. The interaction is implicitly introduced next, in the framework of IT equations, where the first and second moments of the particle number serve as a source of additional information on the interaction in the system. By this means, the binomial approach expresses the probability of the cavity formation as

$$P_0 = e^{-\lambda_0}(1-y)^n \tag{3}$$

where λ_0, the quantity extracted by solving the IT equations, is mainly responsible for the enthalpy component of the free energy. Therefore, the interaction is treated uniformly, without any additional assumptions of physical nature. Let us now introduce such a supplementary assumption about the peculiarity of interaction in large cavities. Consider m solvent particles in the cavity with n cells ($m \le n$). Provided m/n is of the order of unity, the default p_m is given in terms of the volume-defined binomial distribution (2). The essential change is invoked only at the final step of withdrawing particles from the large cavity, i.e., when $m/n \ll 1$. We assume then that the remaining m solvent particles stick to the cavity surface. Such a hypothesis, already mentioned above in section 1, conforms to general ideas underlying the theory of hydrophobicity.[1−4] At this moment, one begins a new combinatorial count of probabilities that includes only the interfacial layer with the number of cells $\nu \ll n$. This allows one to consider the case $n \propto S$ opposite to the case $n \propto v$ for the volume distribution (2). As will be shown below, this idea, for both surface and volume effects, is actually realized as a two-peak default distribution. Because for small cavities the single-peak binomial distribution must remain unchanged, we need to construct the interpolation procedure connecting the small and large cavity limits.

The new default model is initially formulated in terms of two auxiliary binomial distributions for the surface (p_m^s) and volume (p_m^v)

$$p_m^s = \binom{\nu}{m} y^m (1-y)^{\nu-m} \quad (m \le \nu)$$
$$p_m^s = 0 \quad (m > \nu)$$
$$p_m^v = \binom{n}{m} y^m (1-y)^{n-m} \quad (m \le n)$$
$$p_m^v = 0 \quad (m > n) \tag{4}$$

Similar to (2), y is the degree of cell occupation ($0 < y < 1$) and ν and n are the numbers of surface and volume cells, respectively. As usual, they can be considered as noninteger numbers (the binomial coefficients to be expressed in terms of Γ functions). It is known[13] that n is roughly proportional to v, and we define ν to be proportional to S

$$\nu = \frac{S}{\omega} \tag{5}$$

where ω denotes the area of a surface cell, treated as an adjustable parameter. Therefore, for large cavities

$$\nu/n \ll 1 \tag{6}$$

To assemble the desired combination of surface and volume peaks, we define, as a first step, the unnormalized distribution $\tilde{p}_m$

$$\tilde{p}_m = p_m^v (m > \bar{m})$$
$$\tilde{p}_m = p_m^s (m \le \bar{m}) \tag{7}$$

where $\bar{m}$ is the threshold number of solvent particles in the cavity switching the type of distribution.

The individual distributions (4) produce the averages

$$\langle m \rangle_v = ny; \quad \langle m \rangle_s = \nu y \tag{8}$$

and the large cavity limit ensures the following set of inequalities

$$n > \langle m \rangle_v \gg \bar{m} > \nu > \langle m \rangle_s$$
$$(\nu \rightarrow \infty, n \rightarrow \infty) \tag{9}$$

With $\bar{m}$ defined in this way, the distribution (7) represents the idea qualitatively formulated above. This distribution consists

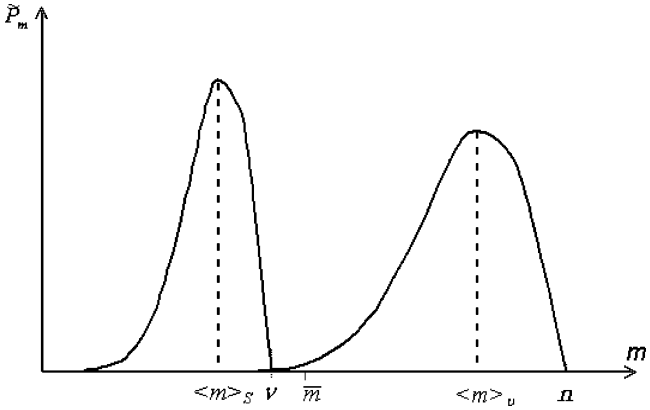

Figure 1. Two-peak default model for the large cavity limit, eq 9.

of a pair of peaks that practically do not overlap (see Figure 1). It can then be substituted by a simpler expression with similar properties

$$\tilde{p}_m = p_m^{\mathrm{s}} + p_m^{\mathrm{v}} \tag{10}$$

Equation 10 will be accepted as the desired default distribution (unnormalized). The presence of two well separated peaks in the large cavity limit (9) is its most important feature. In the limit of a small cavity, the peaks must merge, converting (10) into an ordinary binomial distribution. This is readily achieved in terms of the interpolation procedure modifying the large cavity definition (5) as

$$\omega = q\omega_\infty\left\{1 + \frac{1-q}{2q}\left[1 + \tanh\left[\left(\frac{n}{n_0}\right)^s - 1\right]\right]\right\} \tag{11}$$

Here the area of the surface cell is considered as a variable that decreases when n increases and reaches its minimum value $\omega = \omega_\infty$ in the large cavity limit (9); q, n_0, and s are the interpolation parameters. We need to assume additionally $q \geq 1$ and $\nu = n$ if $S/\omega > n$, which ensures physically relevant behavior of quantity n as defined by (5) and (11). Parameter n_0 serves for switching between the two asymptotic regimes, whereas s controls the width of the turnover region.

3. Information Theory Correction

The final distribution P_m corrected by means of the IT reads[5,6]

$$P_m = \tilde{p}_m \exp(-\lambda_0 - \lambda_1 m - \lambda_2 m^2) \tag{12}$$

Here $\mathrm{e}^{-\lambda_0}$ is a normalization factor whereas the Lagrangian multipliers λ_1 and λ_2 are selected so as to exactly reproduce in terms of (12) the first $\langle m\rangle$ and the second $\langle m^2\rangle$ distribution moments introduced as input data. For the present case, it is expedient to reformulate (12) as

$$P_m = (p_m^{\mathrm{s}} + p_m^{\mathrm{v}}) \exp(-\lambda_0 - \lambda_1 m) \tag{13}$$

by considering instead of λ_2 the parameter n (already present in p_m^{v}) as the second Lagrangian factor. The resulting set of IT equations defining λ_0, λ_1, and n is

$$1 = \sum_{m=0}^{n} P_m$$

$$\langle m\rangle = \sum_{m=0}^{n} m P_m$$

$$\langle m^2\rangle = \sum_{m=0}^{n} m^2 P_m \tag{14}$$

The probability of a density fluctuation creating the cavity is then

$$P_0 = \mathrm{e}^{-\lambda_0}[(1-y)^\nu + (1-y)^n] \tag{15}$$

When $n \gg \nu$, only the first term in brackets survives. In this way, the proportionality of the cavitation free energy to the cavity surface S is achieved for large cavities. It is implicitly assumed here that the dependence of λ_0 on the cavity size is weak; this assumption will be verified below in section 6.

The importance of the IT correction becomes apparent by the notion that the simple model (10) is unacceptable as a distribution describing a real solvent. Indeed, (10) with equal weights for the two peaks as shown in Figure 1 would result in pathologic macroscopic statistics. The correction factor $\mathrm{e}^{-\lambda_1 m}$ in (13) (or $\mathrm{e}^{-\lambda_1 m - \lambda_2 m^2}$ in (12)) strongly reduces the relative height of the surface peak provided λ_1 is negative, and this property indeed holds for the solution of eq 14 in the large cavity limit. As the calculation in section 6 shows, the heights of the surface and volume peaks form the ratio which decreases as a power of $1/n$ (either as n^{-2} or as $n^{-4/3}$).

By this means, due to the IT correction, the surface peak becomes negligible when macroscopic thermodynamic properties of the fluid are the main interest. It, however, keeps the dominating role in the expression for P_0 as is seen from (15).

Note that, according to (13) and (14), the number n of volume cells must be determined individually for every solute cavity, together with λ_0 and λ_1, as a solution to eq 14. For the further application, it is convenient to define

$$n = \vartheta\langle m\rangle \tag{16}$$

where ϑ is considered as a second variable satisfying (14). This approach appears different from that accepted earlier.[5,6,13] This could be circumvented by using the canonical formulation (12) where λ_2 is a regular Lagrangian multiplier. In essence, the two approaches prove to be equivalent (see the Appendix).

4. Computational Scheme

It is convenient to separate the normalization factor in (13)

$$P_m = \mathrm{e}^{-\lambda_0}\tilde{P}_m$$

$$\tilde{P}_m = \tilde{P}_m^{\mathrm{s}} + \tilde{P}_m^{\mathrm{v}} = (p_m^{\mathrm{s}} + p_m^{\mathrm{v}})\mathrm{e}^{-\lambda_1 m}$$

$$\tilde{P}_m^{\mathrm{s}} = p_m^{\mathrm{s}}\mathrm{e}^{-\lambda_1 m}; \quad \tilde{P}_m^{\mathrm{v}} = p_m^{\mathrm{v}}\mathrm{e}^{-\lambda_1 m} \tag{17}$$

The right-hand parts of (14) are expressed in terms of the sums

$$S_0(t) = \sum_{m=0}^{n} \tilde{P}_m(t) = S_0^{\mathrm{s}}(t) + S_0^{\mathrm{v}}(t)$$

$$S_1(t) = \sum_{m=0}^{n} m\tilde{P}_m(t) = S_1^{\mathrm{s}}(t) + S_1^{\mathrm{v}}(t)$$

$$S_2(t) = \sum_{m=0}^{n} m^2\tilde{P}_m(t) = S_2^{\mathrm{s}}(t) + S_2^{\mathrm{v}}(t) \tag{18}$$

We introduced here the notation

$$\exp(-\lambda_1) = t \tag{19}$$

The sums S_k^s and S_k^v, $k = 0, 1, 2$, are performed with surface and volume binomial distributions (4) modified by the factor $e^{-\lambda_1 m}$. For such a type of distribution, called "binomial-exponential"(BE), these sums can be treated analytically.[13] As a result we obtain (18) in the closed form

$$S_0(t) = (1 - y + yt)^\nu + (1 - y + yt)^n$$

$$S_1(t) = \nu yt(1 - y + yt)^{\nu-1} + nyt(1 - y + yt)^{n-1}$$

$$S_2(t) = S_1(t) + (yt)^2[\nu(\nu - 1)(1 - y + yt)^{\nu-2} + n(n - 1)(1 - y + yt)^{n-2}] \quad (20)$$

Finally the IT equations (14) are rewritten as follows:

$$e^{\lambda_0} = S_0(t) \quad (21)$$

$$\langle m\rangle = \frac{S_1(t)}{S_0(t)}$$

$$\sigma^2 = \langle m^2\rangle - \langle m\rangle^2 = \frac{S_2(t)}{S_0(t)} - \left(\frac{S_1(t)}{S_0(t)}\right)^2 \quad (22)$$

The pair of equations (22) define the values of two variables: ϑ (16) and t (19) for a fixed value of parameter y. The so determined root for t gives λ_0 according to (21), whereas n and ν are computed in terms of ϑ by means of (16) and (5), (11). The final result appears from the expression (15).

5. Asymptotic Analysis: The IT Equations

The expression for the cavitation free energy that follows from (15), namely

$$\Delta G_{cav} = -k_B T\{\ln[(1 - y)^\nu + (1 - y)^n] - \lambda_0\} \quad (23)$$

must now be examined in order to establish its asymptotic behavior. The logarithmic term produces the desired proportionality to S, but we must become convinced that the dependence of λ_0 on the cavity size is weaker than this S dependence. Therefore, we have to solve (21) and (22) in the asymptotic limit

$$\langle m\rangle \to \infty; \quad n \to \infty; \quad \nu \to \infty; \quad \nu/n \to 0 \quad (24)$$

In this limit the following relations prove to be true

$$t = 1 + \epsilon, \quad (\epsilon \to +0); \quad e^{ny\epsilon} = \gamma n \quad (25)$$

where γ is a quantity yet indeterminate. Equation 25 is a guess about the explicit form of the solution to the IT equations (22) when only their two leading terms, corresponding to the limit (24), are retained. By substitution of (25) into these truncated equations, we expect to obtain identities and, additionally, to establish the pertaining value of quantity γ.

A remarkable consequence of the first equation in (25) is $t = e^{-\lambda_1} > 1$. This implies $\lambda_1 < 0$, opposite to the ordinary case of small-cavity and single-peak solution, when λ_1 is always positive.[13] Due to this unusual peculiarity of the large-cavity solution, the relative heights of the surface and volume peaks in the distribution (13) renormalize as required, producing the ratio

$$\frac{\text{weight of the surface peak}}{\text{weight of the volume peak}} = \frac{1}{\gamma n} \quad (26)$$

In performing the expansion of the IT equations for the extreme case (24), the following asymptotic relations, a consequence of (25), are helpful:

$$(1 - y + yt)^l = [(1 + \epsilon y)^{1/\epsilon y}]^{ly\epsilon} = e^{ly\epsilon}; \quad l = \nu, n \quad (27)$$

$$1 - y + yt = 1 + \epsilon y = 1 + \frac{\ln(\gamma n)}{n}$$

After some manipulations, we obtain the asymptotic expressions for the sums in (22)

$$S_0 = 1 + \gamma n$$

$$S_1 = \gamma n^2 yD\left(1 + \frac{\nu}{\gamma n^2}\right)$$

$$S_2 = \gamma n^2 yD + \gamma n^2(n - 1)(yD)^2 \quad (28)$$

where $D = (1 + \ln(\gamma n)/(ny))/1 + \ln(\gamma n)/n$ asymptotically reduces to 1. In S_0 the first term comes from the surface component and the second term has the volume origin. This verifies (26). In S_1 and S_2, all terms represent volume contributions, except the second one in brackets for S_1, which proves to be negligible at the end of the calculations. Finally, by substituting (28) in (22), we obtain the leading terms of the IT equations

$$\langle m\rangle = ny$$

$$\sigma^2 = ny(1 - y + y/\gamma) \quad (29)$$

They define the value of γ introduced according to (25): by solving (29) for γ the assumption (25) is verified.

6. Asymptotic Analysis: The Basic Consequences

In the asymptotic limit the variance of the normalized number distribution P_m is proportional to its mean value

$$\sigma^2 = k\langle m\rangle \quad (30)$$

where k is a universal constant. The leading terms of the expansion for the parameters of (29) are

$$\vartheta = \frac{1}{y}; \quad \frac{1}{\gamma} = \frac{k + y - 1}{y} \quad (31)$$

Additionally, the leading terms for λ_0 and λ_1 follow from (21) and (27)

$$\lambda_0 = \ln S_0 = \ln(\gamma n)$$

$$\lambda_1 = -\epsilon = -\ln(\gamma n)/(yn) \quad (32)$$

An unexpected limitation arises due to the second equation in (31). It is seen from (32) that γ must necessarily be positive, which strongly restricts the choice of y by the condition $k + y > 1$. Because $k \ll 1$ for water (see section 7), this requires $1 - k < y < 1$, a quite undesirable constraint, which is unacceptable for describing small and moderately large cavities ($R < 10 \div 12$ Å). To circumvent this limitation, we recall that only leading members of the expansion are present in (31), and we assume here that the correction has the order $1/n$. This notion allows writing the second equation in (31) in the relaxed form

$$\frac{1}{\gamma} = \frac{k + y - 1 + c_1/n}{y} \quad (33)$$

where c_1 is an unknown constant; its value could be extracted if the expansion in terms of $1/n$ were extended for a smaller order of magnitude. We only expect that $c_1 > 0$; in this case y can be expressed as

$$y = 1 - k - \frac{c}{n}; \quad c > 0 \tag{34}$$

This ensures that $\gamma > 0$ provided $c_1 - c > 0$. Hoping this is indeed so, we therefore consider (34) as a definition of y in terms of another parameter c, when n is large. The consistency of such reasoning would be verified by a practical calculation provided it were shown that the numerical solution to full (not expanded) eq 22 does really exist for any (not necessarily large) value of n with y defined according to (34). Computations in section 8 confirm this guess.

We now infer the asymptotic free energy value from (5), (23), and (32)

$$\Delta G_{\text{cav}} = k_{\text{B}}T\left(-\frac{S}{\omega_\infty}\ln k + 2\ln n + \ln\frac{1-k}{c-c_1}\right) \tag{35}$$

Because the leading term must be equal to ΓS where Γ is the surface tension, the value of ω_∞ is fixed as

$$\omega_\infty = -\frac{k_{\text{B}}T}{\Gamma}\ln k \tag{36}$$

The constant term in (35) can be determined by examining the asymptotic free energy behavior found from a numerical solution.

A final comment is addressed to a specification of a higher order correction in (33) and (34) taken rather arbitrary as $O(1/n)$. A more relevant choice would seem to be $O(1/R)$; we can modify it as

$$\frac{1}{\gamma} = \frac{k + y - 1 + b_1/n^{1/3}}{y}; \quad b_1 > 0$$

$$y = 1 - k - b/n^{1/3}; \quad b > 0 \tag{37}$$

The pertaining counterpart of asymptotic eq 35 then becomes

$$\Delta G_{\text{cav}} = k_{\text{B}}T\left[-\frac{S}{\omega_\infty}\ln\left(k + \frac{b}{n^{1/3}}\right) + \frac{4}{3}\ln n + \text{const}\right] \cong \Gamma S\left(1 + \frac{b}{n^{1/3}k\ln k}\right) \tag{38}$$

This limit represents long-range asymptotic behavior compatible with a macroscopic thermodynamic treatment.[32,33] Short-range, (33) and (34), and long-range, (38), limits are discussed below in more detail.

7. Computational Results

Let us summarize the strategy of the suggested computational scheme. It deals with parameters of two levels. The first level parameters are ϑ (16) and t (19). Being determined as solutions to (22), they actually represent two Lagrangian factors inherent to the present model. The second level quantities are ordinary parameters inserted as input data. They include either c or b, defining y in terms of either (34) or (37), and also the parameters q, n_0, and s of the interpolation function (11). Quantity q allows for fitting the small cavity limit, whereas n_0 and s are responsible for the proper description of the turnover range of cavity sizes.

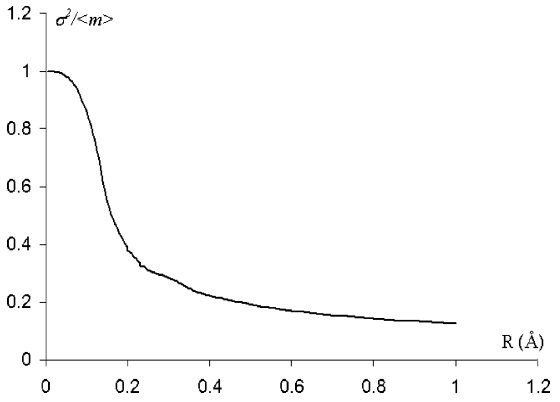

Figure 2. Illustration of the size dependence for the ratio $\sigma^2/\langle m\rangle$.

The total procedure will be called "two-peak BE scheme" in the following text. It computes the cavitation free energy in terms of eqs 1 and 15. The IT equations for ϑ ant t were solved numerically, producing λ_0 and λ_1, ϑ, n, and ν for different radii R. The MC simulation results for $\Delta G_{\text{cav}}(R)$[10] were fitted by adjusting second level parameters. These data are available for $R < 10$ Å. With this ultimate parametrization, the plot was extended for larger spheres by means of the present method. The moments $\langle m\rangle$ and σ^2 were extracted from MD simulation based on the TIP4P[34] trajectory for bulk water solvent. The GROMACS package[35] was used and Coulomb interactions were treated in terms of the PME technique[36] as discussed earlier [ref 13].The second moment was computed from the pair distribution function up to $R = 10$ Å. For larger values of R, an interpolation procedure was used. The resulting plot of $\sigma^2/\langle m\rangle$ is shown in Figure 2. It reaches the value $\sigma^2/\langle m\rangle = 0.102$ at $R = 15$ Å. Its asymptotic limit k can be extracted from the given value of isothermal compressibility χ_{T},[37] and we tried two options: $k = 0.08$ (the result of TIP4P simulation of χ_{T}[38]) and $k = 0.06$ (from the experimental values $\chi_{\text{T}} = 4.5 \times 10^{-4}$ MPa^{-1}, $\Gamma = 0.1036$ kcal/mol/Å^2 for $T = 298$ K[39]). The difference proved to be negligibly small and the value $k = 0.06$ was used in all following calculations.

The final results are illustrated by Figure 3a,b where two different values of parameter s in the interpolation function (11) were used. The value $s = 1$ scales the argument of the hyperbolic tangent as R^3 whereas $s = 1/3$ corresponds to its scaling as R (because n is roughly proportional to the cavity volume). The second case extends significantly the turnover region. Other second level parameters (c, q, and n_0) were optimized in order to gain the best fit to the simulation results for 3.2 Å $< R <$ 10.0 Å (Figure 3a).

Figure 4 illustrates the importance of the asymptotic analysis which suggests expression (34) for the basic parameter y. An obvious alternative option would be a simple variation of y as a second level parameter, independent of R. The failure of this approach, clearly demonstrated by curve (2), shows how sensitive is the computation in the turnover region (around $R \approx 10$ Å) to the details of a default model and its parametrization. An attempt to use eq 37 (and optimizing parameter b), with y depending on R more weakly than in eq 34, also proved to be unsuccessful. The result, similar to curve 2 in Figure 4, revealed large fluctuations in the turnover region. We therefore tried a more elaborate long-range model

$$y = 1 - k - \frac{c}{n} - \frac{b}{n^{1/3}} \tag{39}$$

The same asymptotic free energy expression (38) appears as

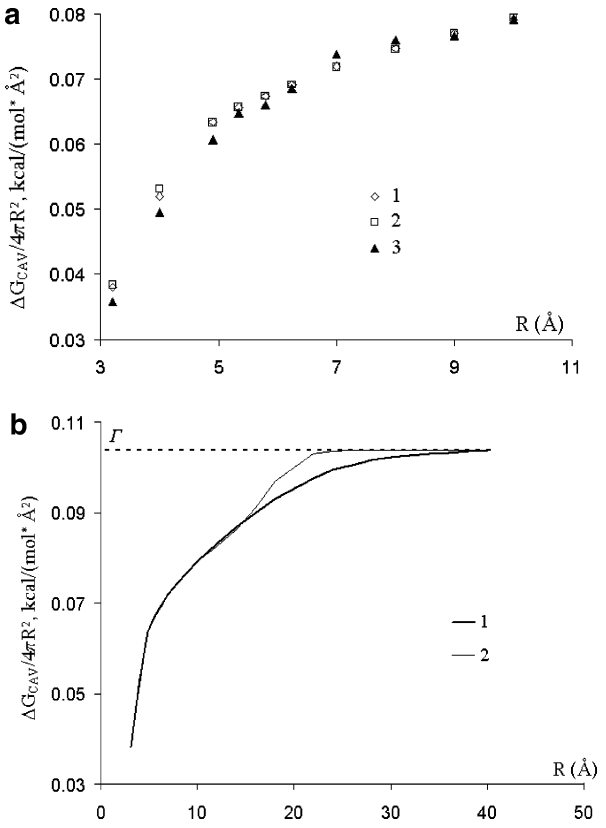

Figure 3. (a) Free energy profiles based on eq 34 with short-range asymptotic behavior. The fitting region 3.2 Å $\leq R <$ 10 Å. (1) $s = {}^1/_3$, $q = 1.36$, $n_0 = 745$, $c = 3.28$. (2) $s = 1$, $q = 1.31$, $n_0 = 526$, $c = 3.39$. (3) The simulation data [ref 10]. (b) Free energy profiles based on eq 34 with short-range asymptotic behavior. Overview for 10 Å $\leq R <$ 40 Å. Curves 1 and 2 correspond to $s = {}^1/_3$ and $s = 1$ with other parameters being the same as in Figure 3a.

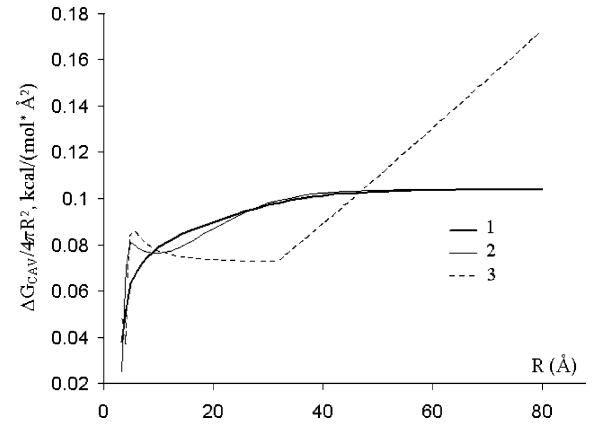

Figure 4. Free energy profiles for different models of y. (1) eq 34, the same curve as (1) in Figure 3b. (2) Constant $y = 0.94$, $s = {}^1/_3$ with other parameters optimized. (3) Constant $y = 0.92$, $s = {}^1/_3$ with other parameters optimized.

its result. The R dependencies of ΔG_{cav} within the extended range of R are shown in Figure 5 for several realizations of the definition (39); they are compared with the best short-range curve (eq 34 with $s = {}^1/_3$). If both constants c and b in (39) are optimized (Figure 5, curve 2), a negative value of b results, which seems to be incompatible with the conventional point of view.[10,32,35] Both cases $s = 1$ and $s = {}^1/_3$ were again considered, as in Figure 3, showing the preference of the latter option, only this choice being illustrated. We finally tested in eq 39 the value $b = 0.078$ based on the MD simulation of large water clusters.[40] Only parameter c was optimized in this case. The corresponding

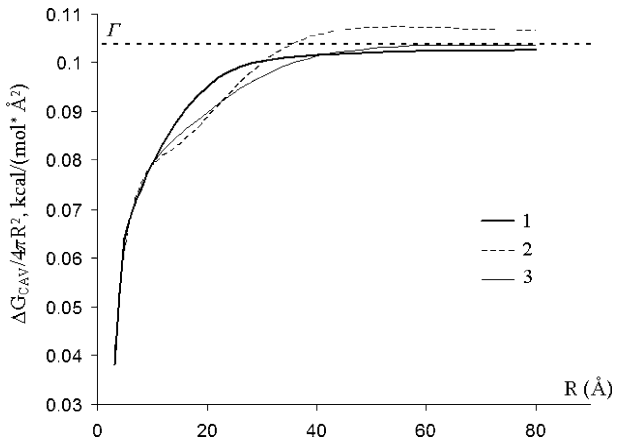

Figure 5. Free energy profiles based on eq 39 with long-range asymptotic behavior; $s = {}^1/_3$. (1) $b = 0.078$, $q = 1.40$, $n_0 = 149$, $c = 2.50$. The b value is extracted from the simulation data [ref 40]; other parameters optimized. (2) $b = -0.179$, $q = 1.84$, $n_0 = 575$, $c = 2.44$. All parameters, including b, are optimized. (3) $b = 0$, eq 34, the same profile as in Figure 3b (1).

curve 1 in Figure 5 describes well the turnover region and also shows reasonable long-range behavior. It is accepted below as a benchmark result.

8. Discussion

It is seen from Figures 3 and 5 how the two-peak BE model accomplishes a connection between the small and large cavity limits, making ΔG_{cav} proportional either to the volume or to the surface area of a given cavity. The ultimate result is qualitatively the same as that obtained with different versions of the density functional model.[4,8,9] By adjusting the second level parameters, i.e., c in (34) and (39) and the parameters of the interpolation function (11), the simulation results for 3.2 Å $< R <$ 10.0 Å available from the literature[10] can be adjusted with good accuracy (for the free energy rmsd = 0.7 and 0.8 kcal/mol with eqs 34 and 39, respectively; see profiles (1) in Figures 3a and 5). The lower boundary of the radius range accepted here corresponds to a molecule as small as methane. Considering smaller cavities makes no sense due to the obvious limitation $n \gg 1$ of the binomial model[13] (actually, $n \cong 6$ for $R = 3.2$ Å). The character of R dependence of the cavitation free energy for larger cavities is controlled by the power s in the argument of the hyperbolic tangent in the interpolation function (11). As seen from Figure 3b, a more natural looking smooth plot is obtained with $s = {}^1/_3$, which corresponds to the interpolation scale proportional to R. This conclusion was confirmed by the same test with eq 39. The short-range asymptotic limit with eq 34 is reached after $R = 40$ Å (Figure 3b). Additional input information (currently unavailable) is needed in order to fix definitely the profile of $\Delta G_{cav}(R)$ in the turnover region and thus establish the corresponding value of s. This could be obtained either by the extension to larger radii of the range of simulation data or by refining the theoretical model by treating explicitly the mechanism of density fluctuations in the turnover region.

The basic parameter y of binomial distributions deserves special comment. It represents the probability of cell occupation, being accepted to be the same for both volume and surface cells. Most remarkable is its dependence on the cavity size according to either (34) or (39). The second level parameter c, performing a connection of the two extremes of the free energy profile, is considered on equal grounds with interpolation parameters of (11). Both eqs 34 and 39 can formally be applied down to $y > 0$; a more realistic boundary condition is $n = \nu$. This unphysical

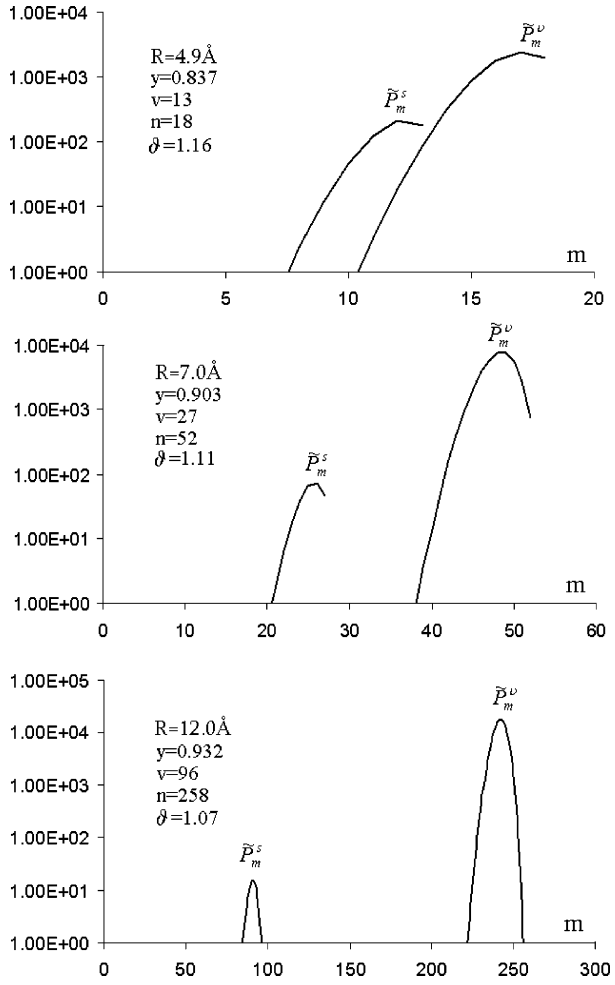

Figure 6. Influence of the sphere size on the shapes of two-peak distributions. The unnormalized volume ($\tilde{P}^{v}_{m}$) and surface ($\tilde{P}^{s}_{m}$) peaks are defined in eq 17. Logarithmic scale is used along the ordinate axis. The computations are based on the short-range y model (34) with parameters corresponding to Figure 3a(1). The peaks for small radii (R = 4.9 and 7.0 Å) are truncated stepwisely from the side of larger m.

bound is never reached in practical computations (with $R >$ 3.2 Å), and we did not try to adjust (34) and (39) for the case $n \rightarrow 0$. In practice, the y value reaches 0.5 for the small cavity limit, which is in accord with the earlier result obtained within the single-peak binomial scheme.[13]

The n dependence of y follows from the asymptotic analysis of section 6. Its importance is demonstrated by Figure 4. If y were considered, instead of c, as a simple second level parameter common for all spheres, the smooth matching of large and small cavity regimes would become impossible. The curve 2, with y fixed at constant value $1 - k = 0.94$ shows how the instability of solutions to the IT equations produces large fluctuations in the transition region, which cannot be eliminated by simple fitting of second level parameters with fixed y. Computations in this region are extremely sensitive to the input values of the first and second distribution moments which, therefore, must be determined very carefully. For curve 2, however, y represents its correct asymptotic limit. When this value is shifted (curve 3) even the asymptotic limit of the free energy is perturbed, although we have made the obligatory change in the definition (36) of ω_∞ ($\ln k$ must be substituted by $\ln(1-y)$). This is a signal that no correct solution with desired properties is available. As pointed out above in section 7, the model (37), where y varied too slowly, also displayed poor

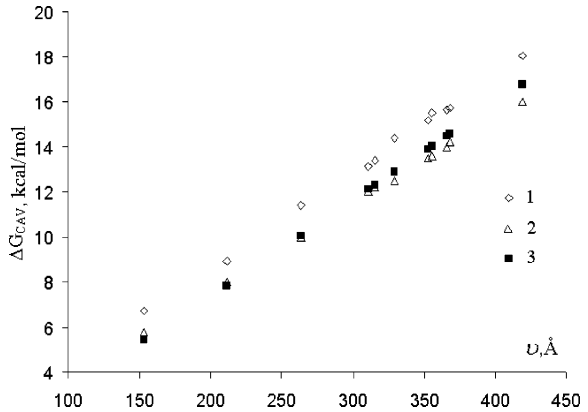

Figure 7. Cavitation free energies for alkanes and cycloalkanes (υ is the cavity volume). (1) MC simulation data,[12] eq 18a of ref 13. (2) MC simulation data,[12] eq 18b of ref 13. (3) The two-peak BE model based on eq 39. The parameters are specified in Figure 5 (1).

results, similar to curve 2 in Figure 4. Altogether, one can consider (34) or (39) as necessary ingredients of the two-peak default model.

The two-peak distributions are portrayed in Figure 6 for several values of R. Their important characteristics, such as y, n, ν, and ϑ, are also listed. The general trends expected based on the formulation of the two-peak model in sections 2 and 3 are clearly reproduced. The increase of the cavity size is indeed accompanied by the increase of the peak separation, and also the relative height of the volume peak greatly increases as compared to the surface one. The asymmetry of the peaks in the turnover region is remarkable.

The two-peak BE model is not limited to a spherical shape of solutes, and we tested it also for the case of nonspherical cavities. Figure 7 shows the results of the two-peak computation for a set of 11 hydrocarbons for which MC simulations of ΔG_{cav} have been reported.[12] This set has been studied earlier with several modifications of a single-peak binomial model.[13] All details about the systems and their computations can be found in refs 12 and 13. In the present test, the parameters of the two-peak computations were borrowed without any change from the computation of spheres as reported above. The results appear to be satisfactory, and we have only to comment on the underlying simulation data. Being obtained in the framework of the thermodynamic perturbation method,[12] they cannot represent quite precisely the hard-core cavitation effect. The true value of the "cavitation energy" is contained within a strip between the two soft-cavity bounds with a misfit of approximately 1−2 kcal/mol. This problem is discussed in more detail in ref 13, and Figure 7 shows both these bounds (eqs 8a and 18b of ref 13). Note that the upper bound also represents the interpretation of the cavitation free energy as suggested in the original simulation work.[12] The earlier reported single-peak BE computation was located closer to the lower bound very similarly to the present two-peak result.

9. Conclusion

By considering the two-peak distribution P_m for the number m of solvent particles occupying a given cavity in the bulk water solvent, it is possible to treat consistently the cavitation free energy in the entire range of the cavity size, beginning from the sphere with radius R = 3.2 Å as a smallest solute. The procedure based on the IT approach[5−7] inserts the first two distribution moments $\langle m \rangle$ and σ^2 as input data and applies the binomial cell model[13] for the individual peaks constituting P_m. The primary default distribution, originating from the cell theory

of dense fluids, takes into account only entropic effects accompanying cavity formation. It is systematically improved by solving IT equations for given $\langle m \rangle$ and σ^2; the interaction effects are introduced implicitly at this stage.

The parametrization for the asymptotic limit of large cavities fits two macroscopic properties of water, namely, surface tension Γ and isothermal compressibility χ_T. A more detailed long-range asymptotic treatment is available by adding Tolman length δ[33,40] as an extra input information. The large cavity (two-peak) and small cavity (single-peak) distributions are connected by an interpolation procedure which prescribes how the basic parameters of the model (the degree of cell occupation y and the area of surface cells ω) change during the variation of the cavity size between the two limits. The interpolation for y includes long-range (eq 39) and short-range (eq 34) options. The first one gives a correct long-range limit, but the second is simpler and describes better the smaller range beyond this limit.

The MC simulation data for small cavities[10] are well reproduced by fitting the interpolation parameters. The parametrization is, however, less definite in the region of the intermediate cavity size because of the lack of benchmark input information.

The spherical case is simple because one has to imitate a very simple function of a single variable R. A standard expansion in powers of $1/R$ seems to be sufficient.[10] Therefore, a test on a predictive power of any cavitation model must include nonspherical objects. In this respect, the computation for hydrocarbons (Figure 7) is a first step in the required direction. Further studies may be addressed largely to extending the range of sizes and shapes of test nonspherical solutes. Numerical simulations of cavitation energies for such systems should be the background for such studies.

Acknowledgment. The authors are indebted to Dr. S. Lushchekina for a computation of the TIP4P trajectory underlying Figure 2. M.V.B. and I.V.L. thank the Russian Foundation of Fundamental Research (Projects RFBR 02-03-33049 and 05-03-33015) for financial support.

Appendix: The Two-Peak BIT Technique

We have also considered a canonical IT formulation in terms of eq 12. Here a new Lagrangian factor λ_2 substitutes for the parameter n of the BE scheme (section 3) as a first level parameter involved in a solution of IT eq 22. Therefore, the definition of n must now be additionally specified. This more sophisticated scheme (the binomial information theory or BIT scheme) has been studied earlier at the single-peak level.[13]

Two-peak BIT computations for large cavities are complicated because no analytical expressions exist for binomial sums (18), as in the BE scheme. Straightforward calculations were possible up to $R = 20$ Å. For larger cavities, we used the Gaussian approximation for the binomial peaks[42] with the same mean values and variances

$$\langle m \rangle_s = \nu y; \quad \sigma_s^{\,2} = \nu y(1-y)$$

$$\langle m \rangle_v = n y; \quad \sigma_v^{\,2} = n y(1-y)$$

This is legitimate when sums (18) are considered because the tail effects are then negligible. However, truncation of these Gaussian peaks from the side of large m (them $m = \nu$ or n) is necessary, as can be seen from Figure 6. The sums were computed as integrals within the limits $(-\infty,\nu)$ and $(-\infty,n)$. Having determined in this way values λ_0, λ_1, and λ_2 by solving eqs 22, we could return to the true binomial expression (15) for the probability of cavity formation.

The interpolation procedures (eqs 11 and 34) were the same as in the main text. The asymptotic analysis performed with the Gaussian peak model repeated the results of sections 6 and 7.

These BIT computations produced no essential new results as compared to the much simpler BE scheme, and no extra illustrations are therefore required.

References and Notes

(1) Pratt, L. R. *Annu. Rev. Phys. Chem.* **2002**, *52*, 1.
(2) Smith, D. E.; Haymet, A. D. J. *Rev. Comput. Chem.* **2003**, *19*, 43.
(3) Stillinger, F. H. *J. Solut. Chem.* **1973**, *2*, 141.
(4) Lum, K.; Chandler, D.; Weeks, J. D. *J. Phys. Chem. B* **1999**, *103*, 4570.
(5) Hummer, G.; Garde, S.; Garcia, A. E.; Pohorille, A.; Pratt, L. R. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 8951.
(6) Hummer, G.; Garde, S.; Garcia, A. E.; Paulactis, M. E.; Pratt, L. R. *J. Phys. Chem. B* **1998**, *102*, 10469.
(7) Hummer, G.; Garde, S.; Garcia, A. E.; Pratt, L. R. *Chem. Phys. B* **2000**, *258*, 349.
(8) Sun, S. X. *Phys. Rev. E* **2001**, *64*, 021512.
(9) Rein ten Wolde, P.; Sun, S. X.; Chandler, D. *Phys. Rev. E* **2001**, *65*, 011201.
(10) Huang, D. M.; Geissler, P. L.; Chandler, D. *J. Phys. Chem. B* **2001**, *105*, 6704.
(11) Huang, D. M.; Chandler, D. *J. Phys. Chem. B* **2002**, *106*, 2047.
(12) Gallicchio, E.; Kubo, M. M.; Levy, R. *J. Phys. Chem. B* **2000**, *104*, 6271.
(13) Alexandrovsky, V. V.; Basilevsky, M. V.; Leontyev, I. V.; Mazo, M. A.; Sulimov, V. B. *J. Phys. Chem. B* **2004**, *108*, 15830.
(14) Reiss, H. *Adv. Chem. Phys*. **1965**, *9*, 1.
(15) Reiss, H.; Frish, H. L.; Helfand, F.; Lebowitz, J. L. *J. Chem. Phys.* **1960**, *32*, 119.
(16) Pierotti, R. A. *Chem. Rev.* **1976**, *76*, 717.
(17) Wilhelm, E.; Battino, R. *J. Chem. Phys*. **1971**, *55*, 4012; **1972**, *56*, 563.
(18) Irisa, M.; Nagayama, K.; Hirata, F. *Chem. Phys. Lett.* **1993**, *207*, 430.
(19) Tomas-Oliveira, I.; Wodak, S. J. *J. Chem. Phys*. **1999**, *111*, 8576.
(20) Pohorille, A. A.; Pratt, L. R. *J. Am. Chem. Soc*. **1990**, *112*, 5066.
(21) Pratt, L. R.; Pohorille, A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 2995.
(22) Floris, F. M.; Selmi, M.; Tani, A.; Tomasi, J. *J. Chem. Phys*. **1997**, *107*, 6353.
(23) Hummer, G.; Garde, S.; Paulaitis, A. E. *Proc. Natl. Acad. Sci. U.S.A.*. **1998**, *95*, 21.
(24) Garde, S.; Garcia, A. E.; Pratt, L. R.; Hummer, G. *Biophys. Chem.* **1999**, *78*, 21.
(25) Ausbangh, H. S.; Garde, S.; Hummer, G.; Kaler, E. W.; Paulaitis, M. E. *Biophys. J*. **1999**, *77*, 645.
(26) Cernuschi, F.; Eyring, H. *J. Chem. Phys.* **1939**, *7*, 547.
(27) Lennard-Jones, J. E.; Devonshire, A.F. *Proc. R. Soc. London, Ser. A* **1937**, *163A*, 53; **1938**, 165A, 1.
(28) Rowlinson, J. S.; Curtiss, C. F. *J. Chem. Phys.* **1951**, *19*, 1519.
(29) Hirshfelder, J. O.; Bird, R. B.; Curtiss, C. F. *Molecular Theory of Gases and Liquids*; Wiley: New York, 1954.
(30) Frenkel, Ja. I. *Kinetic Theory of Liquids*; Clarendon Press: Oxford, U.K., 1950.
(31) Hummer, G. *J. Am. Chem. Soc.* **1999**, *121*, 6299.
(32) Rowlinson, J. S.; Widom, B. *Molecular Theory of capillarity*; Dover: New York, 2002; Chapter 5.
(33) Tolman, R. C. *J. Chem. Phys*. **1949**, *17*, 333.
(34) Jorgensen, W.; Chandrasekhar, J.; Madura, J.; Impey, R.; Klein, M. *J. Chem. Phys.* **1983**, *79*, 926.
(35) Lindahl, E.; Hess, B.; van der Spoel, D. *J. Mol. Model.* **2001**, *7*, 306.
(36) Essmann, U.; Perera, L.; Berkowitz, M. L.; Darden, T.; Lee, H.; Pedersen, L. G. *J. Chem. Phys*. **1995**, *103*, 8577.
(37) Reichl, L. E. *A Modern Course in Statistical Physics*; Wiley: New York, 1998.
(38) Jorgensen, W. L.; Jenson, C. *J. Comput. Chem*. **1998**, *19*, 1179.
(39) *CRC Handbook of Chemistry and Physics*, 78th ed.; CRC Press: Boca Raton, FL, 1997.
(40) Zhukhovitskii, D. J. *J. Chem. Phys.* **1994**, *101*, 5076.
(41) The surface tension Γ_R of water clusters was treated as size-depended in this work, $\Gamma_R = \Gamma(1-2\delta/R)$. The estimate obtained for the Tolman length δ is $\delta = -0.232 r_l$ where $r_l = (3/(4\rho r))^{1/3}$ and r is the solvent density. By a comparison with eq 38 of the present work we found $b = 2\delta/r_l\,[(k \ln k)/(1-k)^{1/3}]$. The sign of δ is changed as we consider the case of a bubble rather than a drop.
(42) Feller, W. *An Introduction to Probability Theory and Its Application*; Wiley: New York, 1970; Vol. 1.