Evaluation of Functional Group Contributions to Excess Volumetric Properties of Solvated Molecules

Daren M. Lockwood and Peter J. Rossky*

Department of Chemistry and Biochemistry, University of Texas at Austin, Austin, Texas 78712-1167

Received: October 30, 1998; In Final Form: December 31, 1998

We develop extensions to a previous methodology for evaluating the excess compressibility of solvation. These extensions make it possible to analyze solution compressibilities in terms of the hydration shell model of solvation. The methodology is applied to three model solutes—water, methane, and methanol—in water. We find that the compressibility is accounted for by localized effects of the solute on the solvent. In addition, for the case of methanol, we find that the localized effects of one solute functional group are independent of the effects of the other. This creates the opportunity for estimation of group-additive contributions to the compressibility, and we illustrate a technique for the extraction of such contributions from molecular dynamics simulation data. The technique is easily generalized for examination of large solutes, including biologically important macromolecules.

1. Introduction

The complex and diverse behaviors exhibited by proteins cannot be fully characterized without a solid understanding of protein solvation effects. Experimental measures of volumetric properties of protein solutions, such as solution compressibility, have been found to be particularly sensitive to solvation effects.[1−4] Such experimental measures are increasingly precise and accessible,[1,3,4] allowing investigators to probe such phenomena as protein conformational changes[2] and interior fluctuations.[3] However, interpretation of experimental data is dependent upon the models used to isolate the solvation effects from other, so-called "intrinsic," contributions to measured properties. One of the most important models used for this purpose is the hydration-shell model of solvation.[5] This model attributes excess thermodynamic properties to a perturbed solvent layer near the solute surface and assumes that solvation effects fall off rapidly with distance from the solute surface.

Our conception of the hydration shell model has undergone significant recent evolution, in part due to advances in molecular dynamics calculations, which provide detailed information about the molecular arrangement of solvent molecules. Results from such calculations have helped to supplant early "iceberg" models[6] with a more rich and dynamic picture and have provided evidence that the influence of individual solute functional groups is very localized,[7] with solvent perturbation by the solute limited primarily to the first hydration shell. However, the extent to which, and the circumstances under which, the hydration shell model is quantitatively valid are not entirely resolved.[5]

Evidence implying that the hydration shell model provides a valid description of volumetric properties can be found in studies of molecules composed of repeating subunits, such as series based on alternative alkyl components. The volumetric properties of these molecules are remarkably linear functions of the number of subunits.[4,8−10] Such linearity suggests to investigators that the effects of the subunits are nearly independent,[4] and therefore that each subunit induces effects that are highly localized. Indeed, if the subunits had long-ranged effects on the solvent, and so acted upon common regions of solvent, additivity would not be expected. On the basis of the assumption that volumetric properties can be explained in terms of independent and localized effects, several investigators have attempted to extract such "group-additive" contributions to volumetric properties, and to employ this information in studies of protein behavior.[2,3,9] These investigators have taken particular interest in determining group-additive contributions to the compressibility. This property has unique advantages as a probe of protein behavior, including its potential role in the isolation of hydration features, the extreme sensitivity of the measure, and its relatively straightforward implementation as a probe of protein states.[1−4] In light of the success and activity of investigators in this field, recent theoretical work by Matubayasi and Levy[5] that cast doubt upon the adequacy of the hydration shell model in accounting for the excess compressibility was both surprising and important.

The present study was undertaken with two goals in mind. First, we were interested in determining how the theoretical results of Matubayasi and Levy could be reconciled with the apparent successes of additive models which imply that solute functional groups have localized and independent effects on the solvent. Second, we were interested in developing more practical methods for the extraction of potential group-additive contributions from molecular dynamics calculations. To this end, we have developed extensions to the previous methodology. The development is done primarily using a model system of pure water. This model system has the advantage that each water molecule may be treated as a solute, and so one can average over the trajectories of each water "solute," providing benchmark results for which the statistical error is negligible.

The remainder of the paper is organized as follows. In section 2, we present a statistical mechanical formulation of excess volumetric properties. We note that such properties may be expressed in terms of alternative solvent distribution functions, and we outline a procedure for selecting partial distribution functions which are particularly well suited to calculation of

* Author to whom correspondence should be addressed.

10.1021/jp984297n CCC: $18.00
Published on Web 03/02/1999

contributions to volumetric properties. Particular attention is paid to the asymptotic behavior of distribution functions with distance from the solute, and a method is discussed for analyzing simulated distribution functions with which there is associated significant statistical error. In section 3, we describe procedures for evaluation of volumetric properties by way of molecular dynamics simulations. In section 4, we discuss results obtained by application of the methodology to specific chemical systems. We examine methanol solvation in water; normal alcohols are ideal model solutes[11] for understanding solution compressibilities, due to their simplicity, the availability of experimental data, and the nearly linear relationship between experimental compressibility data for normal alcohols and the number of $-CH_2-$ subunits. The techniques we develop are shown to be easily generalized to larger and more complex solutes, and the promise of these techniques is corroborated by analysis of methanol solvation in terms of solute functional groups.

2. Theory

The following statistical mechanical formulation follows closely the thoughtful treatment of Matubayasi and Levy.[5] Several significant extensions are introduced. In the first of four sections, we discuss the excess volume of solvation. The previous work is extended to take directly into account the fact that, in the canonical ensemble, the solvent density approaches the average solvent density only asymptotically with system volume. In the second section, we discuss the latitude that one has in choosing a solvent distribution function for study of solution volumetric properties, and explain the advantages that one solvent distribution function may have over another for elucidating these volumetric properties. We outline a procedure for the selection of a particularly useful solvent distribution function. In the third of the four sections, we consider additive contributions. We express the excess compressibility in terms of the excess volume, and show how the excess compressibility may be analyzed in terms of a set of partial distribution functions. Finally, in the fourth section, we discuss the way in which the maximum entropy method may be used to analyze simulation data with which there is associated significant statistical error. These four enhancements will later be shown to permit analysis of volumetric properties in terms of the hydration shell model, and also in terms of group-additive contributions to volumetric properties of solution.

2.1. Excess Volume. In this section, we introduce the excess volume for a single solute in pure solvent, and express this quantity in terms of the solute−solvent particle distribution function. In addition, we introduce a new formula for the excess volume evaluated via the canonical ensemble, which conveniently takes into account the solvent distribution in the limit of large distance from the solute. This excess volume[12] of solvation is defined to be the difference between the partial molar volume, V_m, in solution and that for an ideal-gas particle

$$\Delta V \equiv V_m - V_m^{\text{i.g.}} \tag{1}$$

where V_m is the volume change associated with insertion of the solute at constant pressure, and $V_m^{\text{i.g.}}$ may be expressed in terms of the isothermal compressibility of the pure solvent, κ_o, as

$$V_m^{\text{i.g.}} = kT\kappa_o \tag{2}$$

As demonstrated by Ben-Naim[13] and others, this definition implies that the excess volume is given by the volume change due to insertion of the solute into the solvent at a fixed location at constant pressure. We choose to evaluate the required elements of ΔV in the canonical ensemble with cubic periodic boundary conditions, in which case the fixed solute of interest and some number, N, of solvent molecules are taken to be located in a cubic unit cell of volume V, and these conditions give rise to an average pressure, P, for the system.[14] The excess volume corresponding to this pressure is then given in the thermodynamic limit by

$$\Delta V = V - V_o \tag{3}$$

where V_o is the volume, per N solvent molecules, that the system would occupy in the absence of the solute, if the average pressure exhibited was the same. It has been shown by Matubayasi and Levy[5] that in the thermodynamic limit one has

$$\Delta V = V - N/\rho(\infty) \tag{4}$$

where $\rho(\infty)$ is the solvent density in the limit of infinite distance from the solute. This follows from the fact that, in the thermodynamic limit, the asymptotic value of the density at a given pressure is unchanged by solute insertion.

Equation 4 for the excess volume may be expressed as an integral over the volume of the unit cell[5]

$$\begin{aligned}\Delta V &= V - N/\rho(\infty)\\ &= \int_V d\vec{r} - \left(\int_V d\vec{r}\,\rho(\vec{r})\right)/\rho(\infty)\\ &= \int_V d\vec{r}\,(1 - \rho(\vec{r})/\rho(\infty))\end{aligned} \tag{5}$$

where $\rho(\vec{r})$ represents the solute−solvent particle distribution function. Furthermore, if the hydration shell model were strictly valid, then one could define a cutoff boundary $|\vec{r}| = \lambda$ around the solute beyond which the solvent properties would equal their asymptotic limit. In that case, the λ-dependent value given by

$$\Delta V(\lambda) = \int_{|\vec{r}|<\lambda} d\vec{r}\,(1 - \rho(\vec{r})/\rho(\infty)) \tag{6}$$

would yield the correct value of ΔV since $\rho(\vec{r}) \approx \rho(\infty)$ for $|\vec{r}| \geq \lambda$. As the cell volume V is increased, the asymptotic value of the density N/V_o also approaches the average density of solvent molecules in the box, $\rho_o = N/V$. However, the approximation to eq 6 in which these two values are taken to be equal (see ref 5) was not found to be valid for the unit cell lengths and values of λ used in this study. We therefore depart here from the work of Matubayasi and Levy[5] and instead determine $\Delta V(\lambda)$ self-consistently. We note first that the asymptotic value $\rho(\infty)$ can be estimated alternatively using $V_o = V - \Delta V$ by the expression

$$\rho_\lambda(\infty) = N/(V - \Delta V(\lambda)) \tag{7}$$

This suggests the following definition of $\Delta V(\lambda)$, which also has the correct limiting behavior as λ is increased:

$$\begin{aligned}\Delta V(\lambda) &\equiv V - N/\rho_\lambda(\infty)\\ &= V - N/[N/(V - \Delta V(\lambda))]\\ &= \int_{r<\lambda} d\vec{r}\,[1 - \rho(\vec{r})\,(V - \Delta V(\lambda))/N]\end{aligned} \tag{8}$$

It is straightforward to show that, for λ less than half the unit cell length, this rearranges to give

$$\Delta V(\lambda) \equiv \frac{\int_{r<\lambda} \mathrm{d}\vec{r}\,(1 - \rho(\vec{r})/\rho_o)}{1 - N^{-1}\int_{r<\lambda} \mathrm{d}\vec{r}\,\rho(\vec{r})} \tag{9}$$

For λ significantly less than half the simulation unit cell length, the denominator is approximately equal to unity, and one obtains the expression used by Matubayasi and Levy.[5] If a value of λ exists that is sufficiently far from the solute to yield bulk solvent behavior and sufficiently small to be short compared to the sample size, all expressions will yield the same result. However, if one is interested in larger λ values (or smaller unit cells), the expression given here is to be preferred. Indeed, as λ approaches half the unit cell length, assuming that the number of solvent molecules inside a spherical integral cutoff is proportional to the volume therein, the ratio of $\Delta V(\lambda)$ from eq 9 to the approximation where the denominator is set to unity is approximately $(6 - \pi)/6$, yielding an error close to 100%.

We next turn to the issue of how one may select an advantageous solvent distribution function with which to evaluate volumetric properties via eq 9.

2.2. The Solvent Distribution Function. In eq 5, we introduced the solute−solvent particle distribution function $\rho(\vec{r})$ into the formulation of the excess volume. We note now that this particle distribution function is not a uniquely defined function, as it allows an arbitrary choice of molecular center[12] for the solvent molecules. Any choice of molecular center leaves both the integral of the particle distribution function over the unit cell and the asymptotic value of the density $\rho(\infty)$ unchanged, and so the value of the integral given in eq 6 is in fact unique for an appropriate value of λ. The chosen molecular center may be any point on the solvent molecules, including so-called "auxiliary sites" which do not contribute to the intermolecular potential.[12]

In fact, we note that since any chosen molecular center—labeled α—will give the correct value of ΔV, ΔV can be expressed as an average of n alternative expressions for ΔV, each a function of a different solvent distribution function $\rho_\alpha(\vec{r})$ of a different solvent molecular center, α. It easy to demonstrate that, correspondingly, the excess volume may be calculated by means of a solvent distribution function which is itself an average over individual alternative solvent particle distribution functions, namely

$$\rho(\vec{r}) = n^{-1}\sum_{\alpha=1}^{n} \rho_\alpha(\vec{r}) \tag{10}$$

In the case where the distribution of one or more auxiliary sites about an arbitrarily selected molecular center, β, is described by a probability distribution function $P(\vec{r} - \vec{r}')$ of the displacement from β, one attains for the new solvent distribution function

$$\rho(\vec{r}) = \int \mathrm{d}\vec{r}\,'\rho_\beta(\vec{r}\,')\,P(\vec{r} - \vec{r}\,') \tag{11}$$

In the completely general case, one may also allow that P is itself a function of $\vec{r}'$; however, in this study we will be concerned only with probability distribution functions P that are dependent only on the relative intramolecular displacement.

Using the convolution theorem, eq 11 can be reexpressed in a manner that facilitates calculation of the new solvent distribution function[15]

$$\rho(\vec{r}) = (2\pi)^{-3}\int \mathrm{d}\vec{k}\; e^{i\vec{k}\cdot\vec{r}}\,\hat{\rho}_\beta(\vec{k})\,\hat{P}(\vec{k}) \tag{12a}$$

where the Fourier transforms employed on the right-hand side of the equation are given by

$$\hat{P}(\vec{k}) = \int \mathrm{d}\vec{r}\; e^{-i\vec{k}\cdot\vec{r}}\,P(\vec{r}) \tag{12b}$$

$$\hat{\rho}_\beta(\vec{k}) = \int \mathrm{d}\vec{r}\; e^{-i\vec{k}\cdot\vec{r}}\,\rho_\beta(\vec{r}) \tag{12c}$$

If, for example, we consider the simple case where the function P is uniform within a certain distance R of the molecular center β, and zero otherwise, the Fourier transform is given by

$$\begin{aligned}\hat{P}(\vec{k}) &= \int_{|\vec{r}|<R} \mathrm{d}\vec{r}\; e^{-i\vec{k}\cdot\vec{r}}\,(3\pi^{-1}R^{-3}/4) \\ &= \int_0^R 4\pi r^2\,\mathrm{d}r\left(\frac{\sin kr}{kr}\right)(3\pi^{-1}R^{-3}/4) \\ &= 3k^{-2}R^{-2}[k^{-1}R^{-1}\sin kR - \cos kR]\end{aligned} \tag{13}$$

In this paper, we will frequently have cause to refer to solvent distribution functions that are determined by the convolution of a physical solvent−solute particle distribution function with this particular simple uniform intramolecular probability distribution function. As such, we will want to have a descriptive name for this new kind of solvent distribution function. We note that these solvent distribution functions are themselves a function of a choice of radius R, which defines the spherical distribution of auxiliary sites about a molecular center on the solvent molecules, and that as the radius R goes to zero ($P(\vec{r} - \vec{r}') \rightarrow \delta(\vec{r} - \vec{r}')$), one attains a conventional particle distribution function. We therefore choose to refer illustratively to the new solvent distribution functions as "sphere distribution functions," and adopt this term throughout the rest of this work.

We are now in a position to make advantageous selections of the solvent distribution function. We first note that, due to packing forces, solute−solvent distribution functions are typically oscillatory, and such oscillations are responsive to pressure both with respect to amplitude and location. In particular, Matubayasi and Levy[5] showed that a solute-water oxygen distribution function (β = oxygen; $R \rightarrow 0$ in eqs 12 and 13) manifests significant long-ranged oscillations, and so does not attain its asymptotic value within a computationally readily accessible range of the solute. Since integral truncation like that employed in eq 6 is necessary in practical molecular dynamics calculations, it is important that the distribution function employed fulfill the previously described condition, namely that for a computationally accessible displacement λ from the solute molecule, $\rho(\vec{r}) \approx \rho(\infty)$ for $|\vec{r}| \geq \lambda$. Beyond this computational reason, such behavior is also important to allow definitive analysis of potential group-additive models for the property in question.

While the choice $R \rightarrow 0$ is not always advantageous, as R increases toward the length of half the unit cell, $L/2$, one ensures that the solute influence on the solvent distribution function will extend to the integral cutoff distance $\lambda < L/2$, and so again $\rho(\lambda) \approx \rho(\infty)$ is not attained. Intermediate values, however, remain as possible choices of R that might yield more rapid convergence to the asymptotic value. We will demonstrate that a choice of R of molecular scale does, in fact, lead to good convergence and that, as a result, it is possible to understand volumetric properties in terms of both the hydration shell model of solvation and group-additive contributions.

Having outlined the ways in which the excess volume can be addressed, we turn now to the formulation of the excess

compressibility in terms of the excess volume, and consider additive contributions.

2.3. Excess Compressibility. As suggested by Matubayasi and Levy, we estimate the excess compressibility by finite difference, namely,

$$\Delta\kappa(\lambda) \equiv -\rho_0(P)\,\partial\Delta V(\lambda,P)/\partial P \approx -\rho_0(P)\,(\Delta P)^{-1}[\Delta V(\lambda,P+\Delta P/2) - \Delta V(\lambda,P-\Delta P/2)] \quad (14)$$

Here ΔP represents a finite change in the pressure, $\rho_0(P)$ is the average solvent density in the unit cell at the pressure P, and we have noted explicitly that ΔV is a function of the pressure P. Errors due to the finite difference approximation scale as the square of the pressure differential, ΔP, and may be neglected for a suitably small differential.[16] Since the excess compressibility takes the form of a difference between excess volumes, the results of the preceding sections may be applied directly to calculation of the excess compressibility.

Substitution of the excess volume expression in eq 6 into eq 14 gives

$$\Delta\kappa(\lambda) = \rho_0(P)\,(\Delta P)^{-1}\int_{|\vec{r}|<\lambda} d\vec{r}\left[\frac{\rho(\vec{r},P+\Delta P/2)}{\rho(\infty,P+\Delta P/2)} - \frac{\rho(\vec{r},P-\Delta P/2)}{\rho(\infty,P-\Delta P/2)}\right] \quad (15)$$

where the solvent distribution functions are now also explicitly represented as functions of pressure. We find for the systems in this study that the above expression may be replaced by the approximate form[5]

$$\Delta\kappa(\lambda) = (\rho_0/\Delta P)\int_{|\vec{r}|<\lambda} d\vec{r}\,[g(\vec{r},P+\Delta P/2) - g(\vec{r},P-\Delta P/2)] \quad (16)$$

where $g(\vec{r},P+\Delta P)$ is the solvent distribution function at the pressure $P+\Delta P$, normalized so that its integral over the cell volume is equal to the number of solvent molecules in the unit cell, N.

To permit analysis in terms of separate functional groups on the solute, we note that eq 16 takes the form of an integral over the full space $|\vec{r}| < \lambda$. Thus, it may be easily reexpressed as a sum of integrals over separate regions of that space. For example, if a plane bisects the solute, the points within the space $|\vec{r}| < \lambda$ that are on one side of the plane constitute one region of the space, while the points on the other side of the plane form a second region of the space. Denoting these regions as R_1 and R_2, and noting that the combination of these two regions gives the complete space, $|\vec{r}| < \lambda$, we can represent eq 16 as

$$\Delta\kappa(\lambda) = (\rho_o/\Delta P)\{\int_{R_1} d\vec{r}\,[g(\vec{r},P+\Delta P/2) - g(\vec{r},P-\Delta P/2)] + \int_{R_2} d\vec{r}\,[g(\vec{r},P+\Delta P/2) - g(\vec{r},P-\Delta P/2)]\} \quad (17)$$

Generalization to further subdivision of space is straightforward. Associated with a region R_j, it is convenient to define a partial distribution function $g_j(\vec{r},P+\Delta P)$, which is equal to $g(\vec{r},P+\Delta P)$ within R_j, and 0 elsewhere. Then, for example, eq 17 takes the form

$$\Delta\kappa(\lambda) = (\rho_o/\Delta P)\{\int_{|\vec{r}|<\lambda} d\vec{r}\,[g_1(\vec{r},P+\Delta P/2) - g_1(\vec{r},P-\Delta P/2) + g_2(\vec{r},P+\Delta P/2) - g_2(\vec{r},P-\Delta P/2)]\}$$
$$= (\rho_o/\Delta P)\{\int_{r<\lambda} 4\pi r^2 dr\,[g_1(r,P+\Delta P/2) - g_1(r,P-\Delta P/2) + g_2(r,P+\Delta P/2) - g_2(r,P-\Delta P/2)]\} \quad (18)$$

In the last equality, $g_j(r,P+\Delta P)$ represents the angle-averaged radial partial distribution function. These radial distribution functions are more conveniently used in eq 12.

Selection of the regions $\{R_j\}$ on which to base the distribution functions $\{g_j(r,P+\Delta P)\}$ may be carried out in a number of ways. In the case where each region R_j is taken to be composed of those points closer to the origin of $g_j(r,P+\Delta P)$ than to the origins of the remaining partial distribution functions, the partial distribution functions are equivalent to the "1°" distribution functions recommended and described by Mehrotra and Beveridge in their work on quasi-component distribution functions.[17] However, we do not restrict ourselves to such distribution functions in the present work.

2.4. The Maximum Entropy Method. In the previous three subsections, we have outlined new methodology which will be seen to permit analysis of the excess compressibility in terms of separate localized effects of the solute on the solvent. However, there are significant statistical errors associated with the estimation of such contributions from computationally convenient simulation times. Therefore, we now describe the way in which the maximum entropy method may be used to make improved estimates from noisy distribution function data. A number of good references regarding the maximum entropy method are available.[15,18,19]

The maximum entropy method is applicable to the general problem of calculating a discretized probability density function ("pdf") $\{p_i\}$ from a set of data points $\{d_i\}$ with which there are associated statistical errors $\{\sigma_i\}$. In the case where these statistical errors are appreciable, a number of significantly different pdf's may be considered to be consistent with the data set $\{(d_i,\sigma_i)\}$. Consistency of the pdf with the data set may be measured in a number of ways; one well-established measure is used in the so-called "historic" method,[15] in which consistency with the data set is imposed by the constraint $\chi^2 \leq 1$, where χ^2 is the reduced χ^2 function:

$$\chi^2 = \sum_i [(p_i - d_i)^2/\sigma_i^2] \div \sum_i [1] \quad (19)$$

According to the maximum entropy principle, one should choose that pdf $\{p_i\}$ consistent with the dataset $\{(d_i,\sigma_i)\}$ that minimizes the function

$$-S = \sum_i m_i - \sum_i p_i + \sum_i p_i \ln(p_i/m_i) \quad (20)$$

where $\{m_i\}$ is the so-called "default model" for $\{p_i\}$ and is chosen on the basis of the "principle of indifference"[19] to express total ignorance of the dataset. The function $-S$ is a measure of the information content of the pdf, and is referred to as the cross entropy, or as the directed divergence of $\{p_i\}$ from $\{m_i\}$; alternatively, the function S is referred to as the entropy of the pdf. Since the entropy-optimized pdf has the minimum information content, it has the advantage that any features present are justified by the data. On the other hand, one must understand that features can be lost, and that inferences made via entropy optimization are by nature conservative. In addition, we note

that in many cases, including the examples in this study, the above expression for the cross entropy may be replaced by the more straightforward leading term in an expansion of $\{p_i\}$ about $\{m_i\}$, without significantly affecting the results. One finds

$$-S \approx \sum_i |p_i - m_i|^2/(2m_i) \tag{21}$$

In order to maximize the entropy subject to consistency with the data set, one generally maximizes the function

$$Q = \alpha S - (1/2)\chi^2 \tag{22}$$

Here, the constant α acts as a regulation parameter, balancing maximization of a χ^2 fit to the data against minimization of divergence from the default pdf. In the following, we select α so that the previously-described historic condition $\chi^2 \leq 1$ is satisfied at maximum Q.

The excess compressibility (eq 16) is proportional to the integral

$$F(\lambda) = V^{-1}(\lambda)\int_{V(\lambda)} d\vec{r}\,[g(\vec{r},P + \Delta P/2) - g(\vec{r},P - \Delta P/2)] \tag{23}$$

where $V(\lambda)$ is the volume of the space defined by $|\vec{r}| < \lambda$. This integral is estimated as a discrete sum

$$F(\lambda) \approx V^{-1}(\lambda) \sum_{i<i_{max}} [V_i g_i(P + \Delta P/2) - V_i g_i(P - \Delta P/2)] \tag{24}$$

where the set of incremental finite volumes V_i constitute the total volume of the space $V(\lambda)$, and $g_i(P + \Delta P)$ represents the distribution function mean value within the small volume V_i. Summation now takes place over the difference between two discretized pdf's. According to the principle of indifference, the default model for each term in brackets in eq 24 corresponds to uniform density

$$m_i = V_i g_i^{(o)} = V_i(1) = V_i \tag{25}$$

Sivia[19] suggests that for a function $\{a_i\}$ which is the difference between two pdf's $\{p_i\}$ and $\{p_i'\}$, each with the same default model, one should maximize the sum of the two entropy functions for $\{p_i\}$ and $\{p_i'\}$, subject to a χ^2 constraint on the difference between the pdf's. The corresponding entropy measure, expressed in terms of the mean of the two pdf's $\{\bar{p}_i\}$, is then

$$-S(\{a_i\};\{\bar{p}_i\}) = \sum_{i<i_{max}} \{2m_i - 2\bar{p}_i + (\bar{p}_i + a_i/2)\ln[(\bar{p}_i + a_i/2)/m_i] + (\bar{p}_i - a_i/2)\ln[(\bar{p}_i - a_i/2)/m_i]\} \tag{26}$$

Since the χ^2 constraint is exclusively on the function $\{a_i\}$, it is possible to select the function $\{\bar{p}_i\}$ so as to maximize the entropy in the absence of a corresponding constraint. Setting the derivative of $-S$ (eq 26) with respect to $\bar{p}_i$ equal to zero, one finds

$$\bar{p}_i = (m_i^2 + a_i^2/4)^{1/2} \tag{27}$$

Substituting this into eqs 26 and 22, one obtains the following function of $\{a_i\}$ to optimize

$$Q = \alpha \sum_{i<i_{max}} \{-2m_i + (a_i^2 + 4m_i^2)^{1/2} - a_i \ln[(\sqrt{a_i^2 + 4m_i^2} + a_i)/(2m_i)]\} - (1/2)\sum_{i<i_{max}} [(a_i - a_{exp,i})^2/\sigma_i^2]$$
$$= -\alpha \sum_{i<i_{max}} (|a_i|^2/4m_i) - (1/2)\sum_{i<i_{max}} [(a_i - a_{exp,i})^2/\sigma_i^2] + \vartheta(\sum_{i<i_{max}} |a_i|^3) \tag{28}$$

In the present case, the functions of interest are $\{a_i/V_i = g_i(P + \Delta P/2) - g_i(P - \Delta P/2)\}$ and $\{m_i/V_i = 1\}$. Substituting this into eq 28, one finds

$$Q = -\alpha \sum_{i<i_{max}} (V_i|\Delta g_i|^2/4) - (1/2)\sum_{i<i_{max}} [(\Delta g_i - \Delta g_{exp,i})^2/\sigma_{\Delta gi}{}^2] + \vartheta(\sum_{i<i_{max}} |\Delta g_i|^3) \tag{29}$$

where Δg_i denotes the difference between the two distribution functions.

We estimate the errors $\sigma_{\Delta gi}$ used in the χ^2 function (eq 29) by dividing the simulated trajectory into five equal parts, and calculating a standard deviation over the values corresponding to each segment. It should be noted that a more sophisticated representation of uncertainty in a data set has been employed by Gallicchio and Berne,[18] who consider cross-correlations between data points. It is possible that such a representation could yield even better estimates of $\Delta\kappa(\lambda)$ than those calculated according to the more straightforward constraints on entropy optimization discussed herein.

The formulation presented so far is not ideal for the present purposes. We find that, in the present context, the function $\{a_i = V_i\Delta g_i\}$ corresponding to the maximum of Q shows undesirable dependence on the choice i_{max}. This is because the entropy function places a high penalty on the large values of a_i obtained within the first hydration shell, and the χ^2 function allows for greater and greater reduction of this contribution as i_{max} is increased. This is undesirable not only because it is not physical, but also because the Bayesian likelihood of the function of interest (that function being an integral over a_i as a function of integral cutoff distance) is more heavily dependent upon lower-r values. Hence we require additional accountability of the pdf's to the data, namely that they also satisfy a weighted χ^2 fit to the data, $\chi_w^2 \leq 1$, where

$$\chi_w{}^2 \equiv \left(\sum_i \Delta g_{exp,i} \frac{(\Delta g_i - \Delta g_{exp,i})^2}{\sigma_{\Delta gi}{}^2}\right) \div \sum_i \Delta g_{exp,i} \tag{30}$$

Finally, we note that in some cases the entropy-optimized pdf suggests that further refinement of the default model is a reasonable direction for improvement.[19] In the present case, such refinement takes the form of information regarding deviation of p_i and p_i' from one another. To derive the information measure, one again optimizes the sum of the entropy functions for the two pdf's, but now the default models deviate from one another, and we represent these default models as $\{m_i + \Delta m_i/2\}$ and $\{m_i - \Delta m_i/2\}$. The leading term in the entropy is then

$$S = \sum_i \{|a_i - \Delta m_i|^2 \div [m_i(4 - \Delta m_i^2/m_i^2)]\} \tag{31}$$

3. Simulation Procedures

Molecular dynamics simulations[14] were carried out in the canonical ensemble, at a temperature of 25 °C, using the periodic boundary conditions convention with cubic unit cells. Cell edge lengths of 26.86 and 26.00 Å were used to define the pair of densities, and the number of solvent water molecules in every case was 647. In each simulation performed, the cell contained a fixed solute molecule in addition to the solvent molecules; solutes considered were methane, methanol, and water itself. Methane and methanol molecules were described by the OPLS[20] parameter set, while the SPC[21−23] potential was used for water molecules. Velocities were sampled according to the Boltzmann distribution every 500 time steps throughout the entire simulation;[14] the time step was 2 fs. After 50 ps of equilibration, coordinates were recorded every 25 time steps in a 4 ns trajectory, and used to determine solute−oxygen particle distribution functions.[14] (In the case of methane, only 2 ns were recorded.)

Equations 12 and 13 were employed in the conversion of these solute−oxygen particle distribution functions into the defined sphere distribution functions. Sphere radii were selected, for reasons outlined previously, to be commensurate with the dimensions of a water molecule. In the next section, we concentrate on the specific choice $R = 2.3$ Å, which corresponds to the distance over which the pure-water, oxygen−oxygen particle distribution function is 0. We emphasize, however, that the specific value of this radius is not crucial.

The sphere distribution functions were in turn used in eqs 9, 16, and 18 to calculate the thermodynamic quantities of interest. In eqs 16 and 18, we used the experimental value[24] of $\Delta P = 2791$ atm, which is the pressure difference between pure water systems with densities of 1.0 and 1.1 g/cm^3. While the presence of a small solute does causes a change in this pressure difference, the error is on the order of 1 part in 647 (the number of solute molecules per solvent molecule)[12,14] and so may be neglected for the purposes of this study.

The calculated excess compressibilities can be converted to experimentally accessible partial molar compressibilites of dilute aqueous solutions by means of the expression

$$\begin{aligned}\kappa_2^o &\equiv -\partial(V_m)/\partial P \\ &= -\partial(\Delta V + kT\kappa_o)/\partial P \\ &= -\partial(\Delta V)/\partial P - \partial(kT\kappa_o)/\partial P \\ &= \rho_o^{-1}\Delta\kappa - kT(\partial\kappa_o/\partial P)\end{aligned} \tag{32}$$

where $\Delta\kappa$ is the value obtained by simulation, and $(\partial\kappa_o/\partial P)$ is taken from experiment.

4. Results and Discussion

We begin by considering the two pure water systems described in the Procedures section. In this case, the solute−oxygen particle distribution function was taken to be the oxygen−oxygen particle distribution function. Corresponding 2.3-Å sphere distribution functions were generated. Excess volumes were calculated on the basis of each alternative distribution function, as a function of integral cutoff. Figure 1a shows the excess volume function based directly upon the oxygen−oxygen particle distribution function, while Figure 1b demonstrates the result based on use of the sphere distribution function. As is clearly evident, the use of the sphere distribution function leads to much more rapid convergence of the excess volume to its asymptotic value, as a function of the integral cutoff distance. The appearance of Figure 1b is much more in line with what one would anticipate on the basis of the hydration shell model, and the result suggests that the effect of at least some solutes on solvent can indeed be regarded as local to the solute. The excess-volume values obtained, at half the unit cell length, by this method are approximately $\Delta V_{1.0} = 28$ Å^3 and $\Delta V_{1.1} = 26$ Å^3, which is in excellent agreement with the values of 28.1 and 26.0 Å^3 calculated by means of eqs 1 and 2 and the experimental data in reference 24.

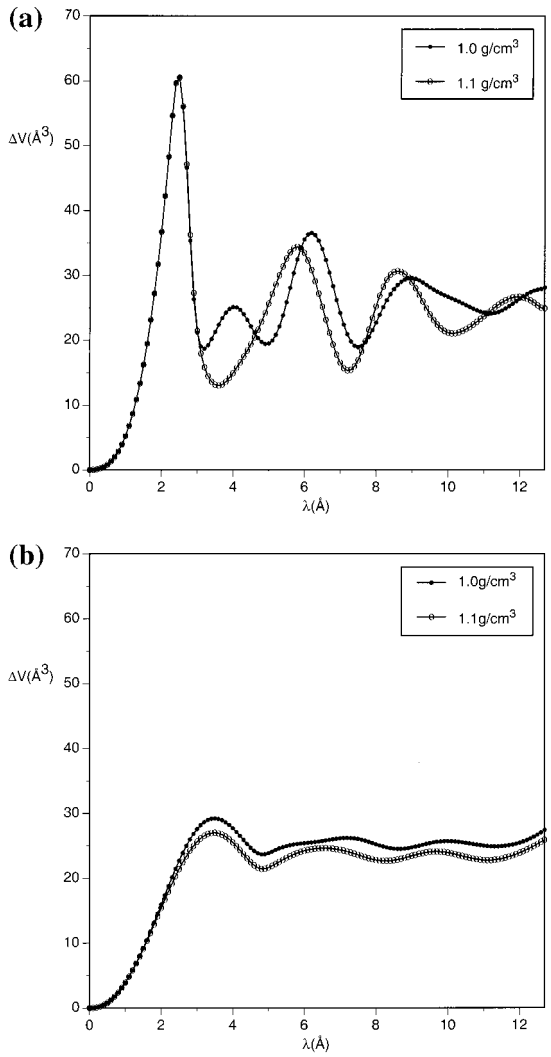

Figure 1. (a) Excess volume for a water "solute" in water solvent, as a function of integral cutoff distance, at two densities (1.0 and 1.1 g/cm^3), using the oxygen−oxygen particle distribution functions. (b) As in part (a), but the oxygen−oxygen based 2.3 Å-sphere distribution functions are used rather than the oxygen−oxygen particle distribution functions.

It is important to note that the introduction of the sphere distribution functions does not introduce new physical content, but rather they elucidate it. As is evident comparing Figure 1a to Figure 1b, the asymptotic converged value is the same in both cases and is essentially achieved in both cases by the end of the first hydration shell. However, using the sphere distributions, one can demonstrate this convergence and, more importantly, make clear estimates of the converged values at shorter distances. The sphere distribution, in effect, carries out a coarse grained averaging over a radial region. The purely oscillatory contributions associated with packing effects are then averaged out, and net contributions from solvent regions are emphasized.

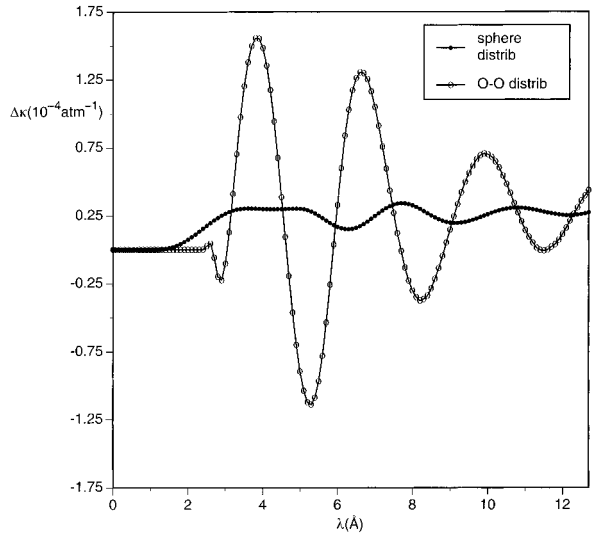

Figure 2. Excess compressibility for a water "solute" in water solvent, as a function of integral cutoff distance, using either the oxygen−oxygen particle distribution functions or the corresponding 2.3 Å-sphere distribution functions.

The $\Delta V(\lambda)$ functions of Figure 1 for the pure water system were used to calculate excess compressibilities as a function of integral cutoff distance, via eq 16. The results are shown in Figure 2. Use of the sphere distribution function results in dramatically improved convergence with distance, and evidently greater consistency with the hydration shell model. The resultant excess-compressibility value at half the unit cell length is 0.28 $\times$ 10^{-4}/atm; by comparison, the experimental value[24] attained at a water density of 1.05 g/cm^3 is 0.34 $\times$ 10^{-4}/atm. The difference is consistent with the incomplete convergence of the plot (the same plot can be seen in larger scale in Figure 3a, where it is marked by solid symbols), but it is equally likely to be associated with the water model.

The use of a pure water system is advantageous because each water molecule could be treated as a solute, allowing us to average over the trajectories of each water "solute," and so gain results for which the statistical error was greatly reduced. Before moving on to other single-solute systems, we first made use of the pure water system to gain an expectation for the statistical errors that would typically be found. In addition to the previous calculation of the "exact" excess compressibility function for water, in which we averaged over all possible water solutes, we also calculated functions which were an average over smaller numbers of water solutes. In Figure 3a, we show four representative plots, generated by averaging over 1, 2, 10, and 646 solutes respectively, over a span of 3 ns. Assuming that the extent of ensemble sampling—as a function of the number of independent configurations—is the same, whether it is time or the number of solutes that is increased, then the present four cases may be considered to correspond to "effective" single-solute run times of 3, 6, 30, and 1938 ns, where we have multiplied the number of solutes by the simulation time. The plots clearly demonstrate that for run times typically employed in modern molecular dynamics simulations—say, on the order of 3 ns—a high level of statistical error is to be expected for single-solute systems. And to determine compressibilities with negligible error via the route shown, the required computational expense would be on the order of 1 μs!

To address this issue of statistical error, we reexamined the data using entropy optimization. The entropy-optimized plots are shown in Figure 3b; they differ from those in Figure 3a in that features not justified by the data are removed. For example,

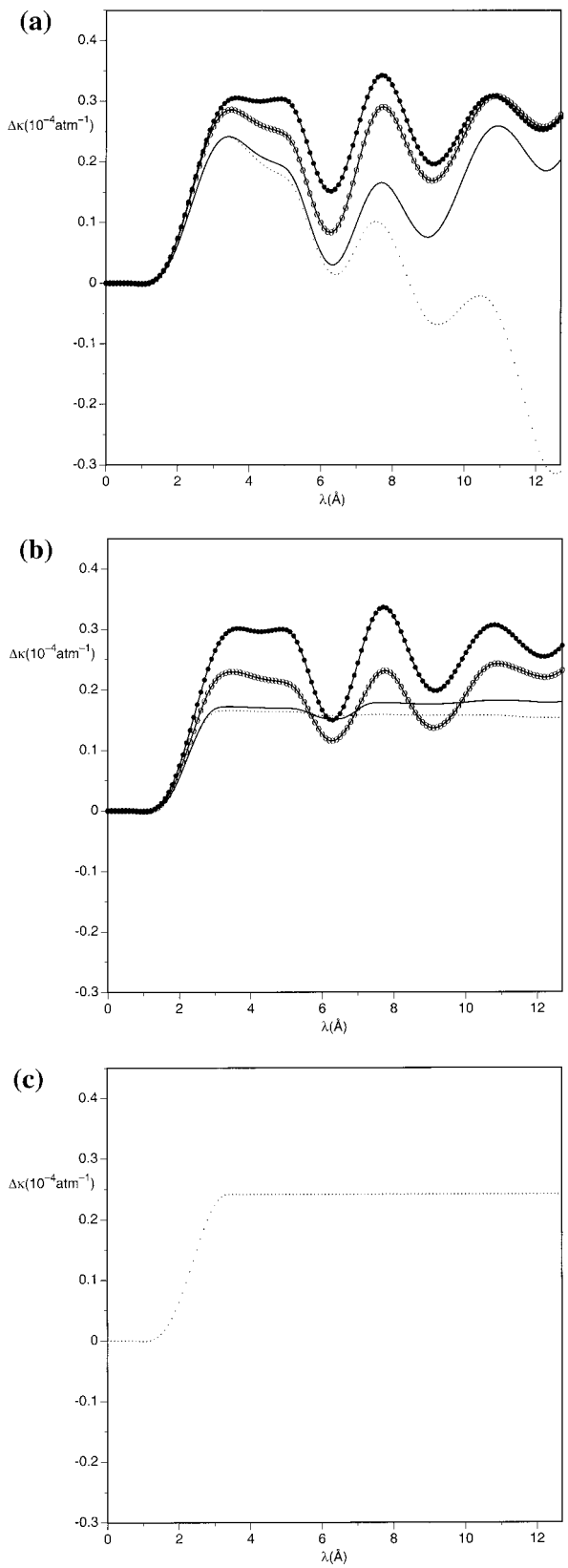

Figure 3. (a) Excess compressibility for a water "solute" in water solvent (using 2.3 Å-sphere distribution functions) as a function of integral cutoff distance. The simulation time is 3 ns, and distribution functions are averaged over a variable number of water "solute" molecules: 1 (broken line), 2 (plain), 10 (open symbols), 646 (solid symbols). This is suggestive of convergence rate of the excess compressibility with simulation time. (b) As in part (a), but using entropy optimization to regulate the integrands. (c) As in part (b), but for one water "solute" only, and using modified default model, incorporating raw lower λ data (λ < 3.2 Å) into the default model prior to entropy optimization.

the entropy-optimized excess compressibility, as calculated from the "one-solute" data, is found to be nearly independent of integral cutoff distance, for cutoffs above about 3.2 Å. This means that in this case, solvent farther than 3.2 Å from the solute is taken to have no effect on the compressibility, due to the fact that statistical errors in this region are too high to justify any inferences regarding the compressibility. The advantage of such a conservative treatment of the data is that more reliable inferences from the data can be made, allowing one to attain reasonable estimates of excess compressibilities from more computationally accessible run times. Nevertheless, it is clear that there is a reduction in numerical accuracy for the less averaged data.

While the one-solute plot shown in Figure 3b is nearly independent of the integral cutoff distance for distances above about 3.2 Å, features below 3.2 Å are determined to be "real." In other words, in the lower-r region, entropy optimization does not entirely remove features in Δg. However, entropy optimization does reduce the magnitude of Δg features, to the extent permitted by the historic constraint on the entropy optimization. Given that the most likely values of Δg in the region $r < 3.2$ Å are the computed average values, and in light of both the relatively high accuracy and the importance of these values to our estimate, we suggest that a reasonable refinement of the entropy-optimized plot can be obtained by including these average values of Δg in the default model (see eq 31). The result obtained using the modified default model is shown in Figure 3c; very similar results are found for different choices of individual water solute. The magnitude of the excess compressibility that is obtained by this method is much improved and in reasonable agreement with the results from greater sampling. Thus we conclude that it is possible to obtain good estimates of the excess compressibility from significantly shorter runs than those required to determine all the features of the value's dependence on integral cutoff.

Using what we have learned from the pure water system, we proceed to analyze solvation of methanol in an aqueous environment. Our first interest is in the examination of potentially sensible additive contributions to the excess compressibility for this system. To this end, we consider a plane passing through the Lennard−Jones center of the carbon atom on the methanol molecule, and perpendicular to the C−O bond. This creates two distinct regions of space, each composed of points on a given "side" of the plane. We refer here to the exterior half-space not incorporating the oxygen as the "C-end" of methanol and the other half space as the remainder. Then, following the procedure outlined in the Theory section, we compute the two corresponding additive contributions to the excess compressibility.

In Figure 4a, we show the contribution to the excess compressibility of the C-end of methanol, as a function of integral cutoff distance, utilizing directly the carbon−oxygen particle distribution function. For comparison, we also show one-half of the excess compressibility for methane as a function of cutoff distance, results for which are also published elsewhere.[5] Standard deviations are shown for the two plots, and one sees that, within the statistical error of the simulations, the two plots are indistinguishable. Thus one finds that the excess compressibility contribution around the C-end of the methanol solute is indeed largely independent of the neighboring hydroxyl group. The result is consistent with the success of group-additive models for alcohol compressibility.[11]

To obtain a better estimate of the excess compressibility contribution of this region, we follow the development used

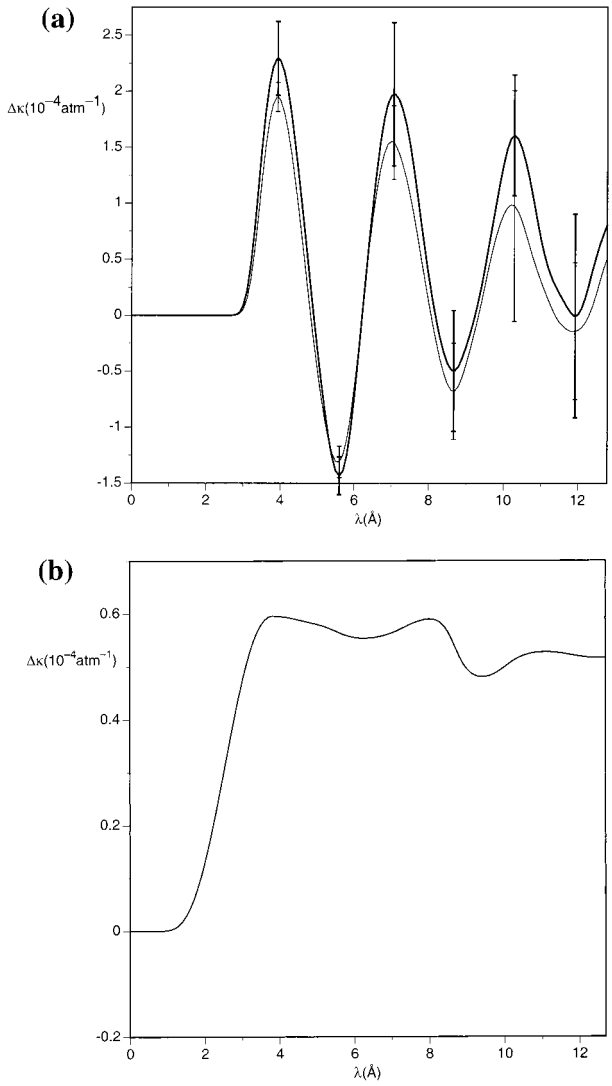

Figure 4. (a) Estimated excess compressibility contribution, as a function of integral cutoff distance, for two related solute functional groups: half of a methane molecule (thin solid line) and "C-end" of methanol (bold solid line). Calculations are based on carbon−oxygen particle distribution functions. Error bars represent standard deviation over five equal subdivisions of the simulation time. (b) Estimated excess compressibility contribution, as a function of integral cutoff distance, for "C-end" of methanol. Carbon−oxygen particle distribution functions have been converted to 2.3-Å sphere distribution functions. Entropy optimization is used to regulate the integrands, and the default model is modified as for the single water "solute."

for water and convert the carbon−oxygen partial particle distribution function into a partial 2.3 Å-sphere distribution function. The excess compressibility is recalculated, employing the entropy-optimization techniques developed on the pure water system, including modification of the lower-r region of default model as discussed just above. The result is shown in Figure 4b. Again, the improvement compared to Figure 4a is dramatic, and we estimate an excess compressibility contribution of 0.5 $\times 10^{-4}$/atm. This is consistent within statistical uncertainty with the asymptote of the previously reported $\Delta\kappa(\lambda)$ for methanol,[5] but use of the sphere distribution function gives a much more definitive estimate of this quantity.

Finally, we turn to the remainder region of the methanol molecule. Since the choice of molecular center for distribution functions is arbitrary, we consider two reasonable methanol−water partial distribution functions: carbon−oxygen and oxygen−oxygen. Each of these is converted to a 2.3 Å-sphere distribution

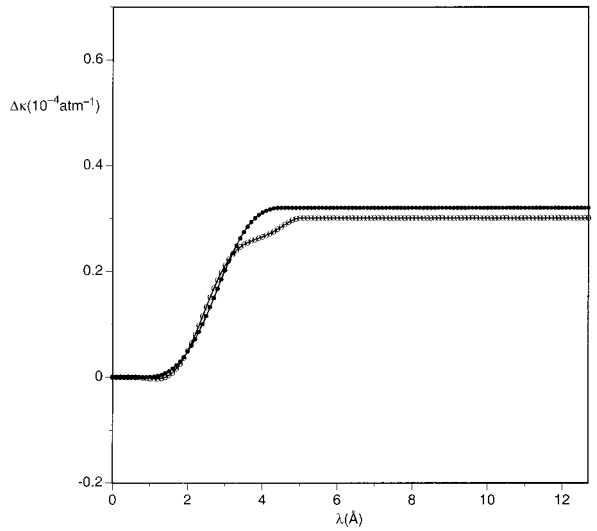

Figure 5. Estimated excess compressibility contribution, as a function of integral cutoff distance, for the remainder of the methanol solute. Either carbon−oxygen (open symbols) or oxygen−oxygen (closed symbols) particle distribution functions have been converted to 2.3-Å sphere distribution functions. Entropy optimization is used to regulate the integrands, and the default model is modified as for the single water "solute."

function, which is in turn used to calculate the excess compressibility contribution as a function of integral cutoff distance, again using entropy-optimization techniques developed on the pure water system, including modification of the lower-r region of the default model. The results are shown in Figure 5. We estimate an excess compressibility contribution of 0.3×10^{-4}/atm from either plot, yielding a total excess compressibility for methanol in water of 0.8×10^{-4} /atm. By way of comparison, experimental excess compressibilities for methanol solvation at atmospheric pressure can be calculated by means of eq 32 and data in reference 11; the result at 25 °C is 0.67×10^{-4}/atm. The deviation from experiment is readily accounted for by statistical error, uncertainties due to the classical potential functions employed, and perhaps most importantly the fact that the simulation corresponds to significantly higher-than-atmospheric pressure; we are not aware of experimental data available for this higher pressure region.

Conclusion

We have developed a number of extensions to the theoretical development of Matubayasi and Levy[5] focused on obtaining partial molar volumetric quantities from computer simulations of solutions. The extensions allow both the reduction of statistical errors, using entropy optimization, and the elucidation of the spatial range of contributions around solutes, using alternative solvent distribution functions. The results derived for water (as solute) and for methanol are close to comparable experimental values. Further, results for the partial contributions of the methanol methyl group have been shown to be equivalent to those from methane within statistical error. Thus, we find that the methods developed in this paper can be of use in interpreting experimental values for volumetric properties of solution in terms of independent, additive contributions of solute functional groups. Having estimated group-additive contributions for methanol, extension to larger solutes is straightforward. The techniques developed in this paper are currently being applied, in this laboratory, to other molecular systems of chemical and biological interest.

Acknowledgment. The authors are grateful for the support of this research through grants from the NIH (GM49204) and the R. A. Welch Foundation (F-0761). The authors also thank Ronald M. Levy, Nobuyuki Matubayasi, and Emilio Gallicchio for stimulating discussions.

References and Notes

(1) Chalikian, T. V.; Totrov, M.; Abagyan, R.; Breslauer, K. J. *J. Mol. Biol.* **1996**, *260*, 588.
(2) Chalikian, T. V.; Breslauer, K. J. *Proc. Natl. Acad. Sci. U.S.A.* **1996**, *93*, 1012.
(3) Kharakoz, D. P.; Sarvazyan, A. P. *Biopolymers* **1993**, *33*, 11.
(4) Chalikian, T. V.; Sarvazyan, A. P.; Breslauer, K. J. *Biophys. Chem.* **1994**, *51*, 89.
(5) Matubayasi, N.; Levy, R. M. *J. Phys. Chem.* **1996**, *100*, 2681.
(6) Frank, H. S.; Evans, M. W. *J. Chem. Phys.* **1945**, *13*, 507.
(7) Rossky, P. J.; Karplus, M. *J. Am. Chem. Soc.* **1979**, *101* (1), 1913, and references therein.
(8) Iqbal, M.; Verral, R. E. *J. Phys. Chem.* **1987**, *91*, 967.
(9) Kharakoz, D. P. *J. Phys. Chem.* **1991**, *95* (5), 5634.
(10) Chalikian, T. V.; Sarvazyan, A. P.; Funck, T.; Breslauer, K. J. *Biopolymers* **1994**, *34*, 541.
(11) *Water Science Reviews 1*; Franks, F., Ed.; Cambridge University Press: New York, 1985, and references therein.
(12) Hansen, J. P.; McDonald, I. R. *Theory of Simple Liquids*; Academic Press: New York, 1986. Excess quantities defined in the way used in this paper are discussed in Chapter 2.
(13) Ben-Naim, A. *Solvation Thermodynamics*; Plenum Press: New York, 1987.
(14) Allen, M. P.; Tildesley, D. J. *Computer Simulation of Liquids*; Oxford University Press: Oxford, 1987.
(15) *Numerical Recipes*; Press, W. H., Teukolsky, S. A., Vetterling, W. T., Flannerey, B. P., Eds.; Cambridge University Press: New York, 1992.
(16) Kincaid, D.; Cheney, W. *Numerical Analysis*; Brooks/Cole Publishing Company: Pacific Grove, CA, 1991.
(17) Mehrotra, P. K.; Beveridge, D. L. *J. Am. Chem. Soc.* **1980**, *102* (2), 4287.
(18) Gallicchio, E.; Berne, B. J. *J. Chem. Phys.* **1996**, *105*, 7064.
(19) Sivia, D. S. *Data Analysis: A Bayesian Tutorial*; Oxford University Press: New York, 1996.
(20) Jorgensen, W. L. *J. Phys. Chem.* **1986**, *90*, 1276.
(21) Jorgensen, W. L.; Chandrasekhar, J.; Madura, J. D.; Impey, R. W.; Klein, M. L. *J. Chem. Phys.* **1983**, *79*, 926.
(22) Berendsen, H. J. C.; Postma, J. P. M.; Van Gunsteren, W. F.; Hermans, J. In *Intermolecular Forces*; Pullman, G., Ed.; Reidel: Dordrecht, The Netherlands, 1981.
(23) Andersen, H. C. *J. Computat. Phys.* **1983**, *52*, 24.
(24) Gibson, R. E.; Loeffler, O. H. *J. Am. Chem. Soc.* **1941**, *63* (3), 898.