JOURNAL

OF THE AMERICAN CHEMICAL SOCIETY

Registered in U.S. Patent Office. © *Copyright, 1980, by the American Chemical Society*

VOLUME 102, NUMBER 13 JUNE 18, 1980

Structural Analysis of Molecular Solutions Based on Quasi-Component Distribution Functions. Application to $[H_2CO]_{aq}$ at 25 °C

P. K. Mehrotra and David L. Beveridge*

Contribution from the Chemistry Department, Hunter College of the City University of New York, New York, New York 10021. *Received October 17, 1979*

Abstract: A general, uniquely defined method for the analysis of structure and energetics in the statistical state of a molecular solution is proposed. The key element in the analysis is the proximity criterion, whereby solvent molecules in a given many-particle configuration of the system are classified on the basis of the nearest solute atom. The proximity criterion is cast analytically in the form of a property of the system. A compositional analysis is then developed in terms of joint quasi-component distribution functions involving proximity indexes and other properties of the system such as coordination number, binding energy, etc. The solvation of a general molecular solute can then be formally described in terms of atoms, functional groups, or subunits. The procedure is fully illustrated by an analysis of the results of a Monte Carlo computer simulation on the dilute aqueous solution of formaldehyde at 25 °C.

I. Introduction

The motional degrees of freedom of molecules in liquids and solutions at laboratory temperatures mandate theoretical studies in this area to be problems in statistical mechanics and dynamics.[1] The composition, both structural and energetic, of such systems must be defined on the statistical state of the system, and compositional indexes must be defined in terms of statistically weighted structural alternatives rather than any single supermolecular structure. In principle, the composition of a fluid follows from a knowledge of the molecular distribution functions (MDF) for the system. The various atom-atom pair correlation or radial distribution functions (RDF), $g(R)$, can in principle be deduced from diffraction experiments as well as theoretical calculations and are thus the most important of this class of functions. The analysis of the composition of a molecular fluid thus requires an interpretation of the statistical distribution functions in structural and energetic terms.

A general theoretical approach to this problem was mapped out several years ago by Ben-Naim[2] based on generalized molecular distribution functions and the closely related quasi-component distribution functions (QCDF), and involves developing the distribution of particles with certain well-defined values of a compositional characteristic on the statistical state of the system. In particular, QCDF with respect to coordination number and binding energy have been used extensively in conjunction with Monte Carlo computer simulation methodology in a series of recent research studies on molecular liquids and solutions reported from this laboratory.[3-6] Ben-Naim's approach has proved to be a very graphic and effective means of dealing with compositional problems in fluids.

The use of QCDFs to interpret RDFs and composition in fluids has up to this point been focused on systems in which the local environment of the particles is simple and isotropic enough that structure can be developed in terms of relatively simple orientationally averaged distribution functions. Here the various atom-atom RDFs display a well-developed shell structure, and along with the calculated RDF between interparticle centers of mass can be used to formally and uniquely define a useful structural property such as coordination number. Furthermore, the various energetic environments represented in binding-energy distributions can be determined without serious ambiguities.

In extending this approach to solutions of molecules with low symmetry and considerable structural anisotropy, orientationally averaged distribution functions and related quantities are not adequate to elucidate the complexity of structural detail in the system. This is clearly due to the fact that simple extension of the orientationally averaged quantities results in quantities which reflect a composite of contributions from the environments of different substructures (i.e., atoms, functional groups, or subunits) of the solute molecule. The solute-solvent atom-atom RDFs are correspondingly more complicated in appearance and the definitions of properties such as coordination number for use in QCDF are no longer straightforward. Furthermore, simply stepping back a level in the reduction of the distribution function, i.e., eliminating all the orientational averaging, leads to an analysis with too much dimensionality to interpret in accessible descriptive terms.

The research studies having to contend with this point to date are relatively few. The approach of choice to date has been to discuss the structure of the local solution environment of different substructures of a polyatomic solute in aqueous solution by means of a physically sensible but necessarily arbitrary partitioning of configuration space, and developing structural characteristics of the fluid environment within that region.[7] While the calculations based on this approach have provided accurate data and useful insight on the structure of individual systems, we have come to question this idea of partitioning configuration space as a general procedure. Problems arise in uniquely defining such a partitioning for the same functional groups in different molecules, and the consequent limitations in the transferability of results. Also, when the local solution environments of two proximal functional groups on a solute encroach upon one another, there is no simple and systematic way to pursue the analysis.

We have considered the analysis of solutions in the context of the problems outlined above with particular cognizance of the facts that (a) the contributions from the local environment of the various substructures of the system must be resolved without ambiguity, and (b) orientational averaging must be involved to some extent in order to simplify the results. The ensuing analysis is developed on the basis of a unique definition of the total solvation of a solute substructure, be it atom, functional group, or subunit, in terms of the "proximity criterion", whereby solvent molecules in a given many-particle configuration of the system are classified on the basis of the nearest solute substructure. This classification can be formally cast in the form of an abstract property of the system. Analysis of structure can then be developed in terms of generalized molecular distribution functions. With this in place one can proceed to discuss theoretically the solvation of a solute molecule atom by atom, functional group by functional group, or subunit by subunit as desired, and solvent effects on structure and process in solution can be developed in similar, formally defined terms. Furthermore, the solvation state of a given type of functional group in different molecular environments can be quantitatively compared.

We present herein the formalism for the analysis of statistical state of solutions based on the proximity criterion, and illustrate the procedure with an analysis of the local environment of formaldehyde in infinitely dilute aqueous solution. This system demonstrates all essential features of the analysis in prototype. The background for this project is reviewed in section II. The basic idea and the related formalism are presented in section III. The calculations on $[H_2CO]_{aq}$ are described in section IV and the analysis of results based on the proximity criterion is given in detail. The results are discussed in section V followed by summary and conclusions.

II. Background

Generalized molecular distributions were developed by abstracting the procedure involved in formulating ordinary molecular distribution functions for positional correlations in a fluid, and extending the procedure to encompass other structural and energetic characteristics of the system.[2] The basic idea is to select a well-defined property of the particles of the system, and impose a condition on that property. A counting function is formulated to quantitatively determine the number of particles for which the condition is satisfied in a given N-particle configuration of the system. The average number of particles satisfying the condition on the property is obtained by configurational averaging. A definition of the composition of the system in terms of this property is obtained by determining the distribution of particles for all possible values of the condition in the statistical state of the system.

The leading examples of QCDFs for homogeneous isotropic fluids are those for coordination number K and binding energy ν. We briefly review the formulation of these quantities for homogenous systems in order to introduce certain notation and terminology relevant to the analysis of solutions introduced in the following section. Consider a system of N identical molecules. The supermolecular geometry of a given N-particle configuration of the system is fully specified by the configurational coordinate $\mathbf{X}^N$:

$$\mathbf{X}^N = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\} \qquad (1)$$

where the configurational coordinates $\mathbf{X}_i$ of each particle i are the product of positional and orientational coordinates $\mathbf{R}_i$ and $\mathbf{\Omega}_i$, respectively.

Certain properties of the system such as coordination number and also aspects of the analysis introduced herein depend only on the positional coordinates of the particles $\mathbf{R}^N$:

$$\mathbf{R}^N = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N\} \qquad (2)$$

For particle i in a given N-particle configuration of the system, the property "coordination number", C_i, is defined as

$$C_i(\mathbf{R}^N) = \sum_{j \neq i}^{N} h(R_{ij} - R_C) \qquad (3)$$

where $h(R_{ij} - R_C)$ is a unit step function, equal to unity if the interparticle separation R_{ij} is less than the radius of the coordination sphere R_C. To be as consistent as possible with conventional chemical connotations of coordination number, R_C is chosen as the distance corresponding to the first minimum in the intermolecular center of mass $g(R)$. The quantity C_i thus gives the number of other molecules that fall within the first coordination sphere of particle i in configuration $\mathbf{R}^N$. The counting function for this property

$$N_C(\mathbf{R}^N,K) = \sum_{i=1}^{N} \delta[C_i(\mathbf{R}^N) - K] \qquad (4)$$

where the Kronecker δ is unity whenever the property $C_i(\mathbf{R}^N)$ achieves condition K, and is identically zero otherwise. This sum counts the total number of particles whose coordination number is K in configuration $\mathbf{R}^N$. The average number of such particles in the statistical state of the system is

$$N_C(K) = \int \ldots \int P(\mathbf{R}^N) N_C(\mathbf{R}^N,K)\, d\mathbf{R}^N = \langle N_C(\mathbf{R}^N,K)\rangle \qquad (5)$$

where the bracket notation denotes integration over the configurational coordinates of the system. The quantity $N_C(K)$ is a singlet generalized molecular distribution function for coordination number. The mole fraction of particles $x_C(K)$ for which the coordination number is identically K is simply

$$x_C(K) = N_C(K)/N \qquad (6)$$

The quantity $x_C(K)$ can be viewed as the component of a vector

$$\mathbf{x}_C = \{x_C(0), x_C(1), \ldots\} \qquad (7)$$

which defines the composition of the system with respect to the classification according to coordination numbers. The average coordination number is

$$\overline{K} = \sum_{K=0}^{\infty} K x_C(K) = \int_0^{R_C} g(R) 4\pi R^2\, dR \qquad (8)$$

where $g(R)$ is the interparticle center-of-mass RDF. The binding energy of particle i in configuration $\mathbf{X}^N$ is defined as

$$B_i(\mathbf{X}^N) = E(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_i, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_N) - E(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_N) \qquad (9)$$

where E is the configurational energy of the system. The counting function for binding energy is

$$N_B(\mathbf{X}^N, \nu) = d\nu \sum_{i=1}^{N} \delta[B_i(\mathbf{X}^N) - \nu] \qquad (10)$$

which is the number of particles having a binding energy between ν and $\nu + d\nu$ for the specified configuration $\mathbf{R}^N$; note that δ here is a Dirac δ function. The average number of particles having a binding energy between ν and $d\nu$ is

$$N_B(\nu)\, d\nu = d\nu \langle N_B(\mathbf{X}^N,\nu)\rangle \qquad (11)$$

and the corresponding mole fraction is

$$x_{\mathrm{B}}(\nu)\,\mathrm{d}\nu = N_{\mathrm{B}}(\nu)\,\mathrm{d}\nu/N \tag{12}$$

The quantity $x_{\mathrm{B}}(\nu)$ can be viewed as the component of a compositional vector

$$\mathbf{x}_{\mathrm{B}} = \{x_{\mathrm{B}}(\nu)\}, \qquad \nu = -\infty, +\infty \tag{13}$$

which defines the composition of the system with respect to binding energy. The average binding energy is given by

$$\bar{\nu} = \int_{-\infty}^{+\infty} \nu x_{\mathrm{B}}(\nu)\,\mathrm{d}\nu = \langle E(\mathbf{X}^N)\rangle/N \tag{14}$$

such that the thermodynamic configurational internal energy is related to binding energy by the expression

$$U = {}^1\!/_2 N\bar{\nu} \tag{15}$$

One may proceed along analogous lines to define other GMDFs; a characteristic ϕ related to partial molar volume is also developed in ref 1. More detailed analyses of the statistical state can be obtained by developing GMDF for combined properties such as coordination number and binding energy together, giving

$$x_{\mathrm{B,C}}(\nu,K) = N_{\mathrm{B,C}}(\nu,K)/N \tag{16}$$

such that the distribution of binding energy as function of coordination number may be examined. Numerical examples of all the GMDF formally defined in this section have been determined for model liquids,[2] water,[3] the dilute aqueous solution of methane,[4] and dilute aqueous solutions of monatomic cations and anions.[5,6]

III. Theory

The basis for a general compositional analysis of the statistical state of molecular fluids must, as stressed in section I, be a unique definition of the local solution environment of each identifiable substructure—atom, functional group, or subunit—of the solute. To accomplish this we propose the "proximity criterion", which uniquely identifies each solvent molecule with a well-defined solute entity in each configuration. In this section we show how the proximity criterion, formally defined, leads directly and systematically to a general structural analysis of the system based on generalized molecular distribution functions.

Consider an infinitely dilute solution consisting of one solute molecule with a volume V together with N solvent molecules. The analysis as presented can be developed in terms of the coordinates of the N solvent molecules defined relative to the solute center of mass with no loss of generality. In any given configuration of the system, each of the N solvent molecules is classified on the basis of the nearest solute atom, A. The set of solvent molecules closer to A than to any other solute atom are henceforth referred to as the total 1° solvation of A. Higher orders of total solvation may also be defined; the set of molecules for which A is the second nearest solute atom gives the total 2° solvation of A, and so on for 3°, 4°, etc. The first normalization conditions follow directly:

$$\sum_k N_{\mathrm{A}}^{(k)} = N \text{ for any A} \tag{17}$$

and

$$\sum_{\mathrm{A}} N_{\mathrm{A}}^{(k)} = N \text{ for any } k \tag{18}$$

Here $N_{\mathrm{A}}^{(k)}$ is the total solvation number of A at order k.

We now proceed to cast the proximity criterion into the language of GMDF and to analyze the composition of the various orders of total solvation of solute atoms on this basis. For a given solvent molecule i in an N-particle configuration of the system $\mathbf{R}_N$, let us collect as a set the solute atoms listed in order of k. The members of this set are the "proximity indexes" for solvent molecule i, $S_i^{(k)}(\mathbf{R}^N)$. Consider this set as a generalized property of the system in context of GMDF theory:

$$\mathbf{S}_i(\mathbf{R}^N) = \{S_i^{(1°)}(\mathbf{R}^N), S_i^{(2°)}(\mathbf{R}^N), \ldots\} \tag{19}$$

where

$$S_i^{(1°)}(\mathbf{R}^N) = (\mathrm{A}|R_{\mathrm{A}i} = \min\{R_{\mathrm{M}i}\}) \tag{20}$$

i.e., the primary proximity index of solvent molecule is the solute atom A such that the distance $R_{\mathrm{A}i}$ is the absolute minimum in the discrete set $\{R_{\mathrm{M}i}\}$ of all distances between the M solute atoms and the center of mass of the ith solvent molecule. Higher orders of solvation are defined, for example, as

$$S_i^{(2°)}(\mathbf{R}^N) = (\mathrm{B}|R_{\mathrm{B}i} = \min\{\mathbf{R}_{\mathrm{M}i}\}') \tag{21}$$

where the primed set $\{R_{\mathrm{M}i}\}'$ is simply the set $\{R_{\mathrm{M}i}\}$ with the distance $R_{\mathrm{A}i}$ corresponding to primary solvation deleted.

The counting function for this property is

$$N_{\mathrm{S}}(\mathbf{R}^N,\mathrm{A}^{(k)}) = \sum_{i=1}^{N} \delta(S_i^{(k)}(\mathbf{R}^N) - \mathrm{A}) \tag{22}$$

where the δ function is unity when the proximity index $S_i^{(k)}(\mathbf{R}^N)$ is (logically) equal to substructure A and is zero otherwise. The quantity $N_{\mathrm{S}}(\mathbf{R}^N,\mathrm{A}^{(k)})$ is then equal to the number of solvent molecules associated with atom A at solvation order k. We are predominantly interested in the primary solvation of A, but we retain the superscript (k) notation for complete generality. The average number of solvent molecules assigned to A in the statistical state of the system is

$$N_{\mathrm{S}}(\mathrm{A}^{(k)}) = \langle N_{\mathrm{S}}(\mathbf{R}^N,\mathrm{A}^{(k)})\rangle \tag{23}$$

where $N_{\mathrm{S}}(\mathrm{A}^{(k)})$ may be considered a singlet GMDF for solvation, S. The mole fraction of the solvent molecules in the system identified with A at order k is

$$x_{\mathrm{S}}(\mathrm{A}^{(k)}) = N_{\mathrm{S}}(\mathrm{A}^{(k)})/N \tag{24}$$

Collecting the $x_{\mathrm{S}}(\mathrm{A}^{(k)})$ for all solute atoms A, B, . . . , leads to the quasi-component compositional vector

$$\mathbf{x}_{\mathrm{S}}^{(k)} = \{x_{\mathrm{S}}(\mathrm{A}^{(k)}), x_{\mathrm{S}}(\mathrm{B}^{(k)}), \ldots\} \tag{25}$$

the distribution of solvent molecules with respect to solute atoms according to the proximity criterion at order k. The mole fractions are defined such that

$$\sum_{\mathrm{A}} x_{\mathrm{S}}(\mathrm{A}^{(k)}) = 1 \qquad \text{for any } k \tag{26}$$

and

$$\sum_{k} x_{\mathrm{S}}(\mathrm{A}^{(k)}) = 1 \qquad \text{for any A} \tag{27}$$

With the proximity indexes thus defined for all solvent molecules, one may develop an analysis of the solvation of a solute molecule atom by atom. The radial distribution function for the kth order solvation of substructure A is

$$g_{\mathrm{AW}}^{(k)}(R) = \rho^{-2}\int \ldots \int P(\mathbf{R}^N) \sum_i [\delta(L_i(\mathbf{R}^N) - R)] \times [\delta(S_i^{(k)}(\mathbf{R}^N) - \mathrm{A})]\,\mathrm{d}R^N \tag{28}$$

The $g_{\mathrm{AW}}^{(k)}(R)$ are related to the $g_{\mathrm{AW}}(R)$ via the expression

$$g_{\mathrm{AW}}(R) = \sum_k g_{\mathrm{AW}}^{(k)}(\mathbf{R}) \tag{29}$$

resolving a composite quantity into contributions from the various orders of solvation. If the analysis scheme is successful, these contributions individually should be much easier to interpret than the full radial distribution function. Particularly, we expect $g_{\mathrm{AW}}^{(1°)}(R)$ in many cases to have a well-developed shell structure and thus permit a unique choice of R_{C} and a definition for coordination number of A. Analogously, the solute–water contribution to the internal energy of the system U_{SW} can be resolved into contributions developed in terms of the proximity criterion:

$$U_{\mathrm{SW}} = \sum_{\mathrm{A}} \sum_{k} U_{\mathrm{AW}}^{(k)} \tag{30}$$

where

$$U_{\mathrm{AW}}^{(k)} = \langle E_{\mathrm{AW}}^{(k)}(X^N) \rangle \tag{31}$$

and

$$E_{\mathrm{AW}}^{(k)}(X^N) = \sum_{i} E_{\mathrm{SW}}(X_{\mathrm{S}},X_i^{\mathrm{W}})\delta(S_i^{(k)} - \mathrm{A}) \tag{32}$$

assuming pairwise additivity of intermolecular interactions.

A further structural analysis of the local solution environment of solute atoms follows in terms of the combination properties $x_{\mathrm{C,S}}(K,\mathrm{A}^{(k)})$ and $x_{\mathrm{B,S}}(\nu,\mathrm{A}^{(k)})$, allowing the distribution of coordination numbers and binding energies for a substructure to be examined. Note here the definition of the average quantities

$$\overline{K}_{\mathrm{A}}^{(1°)} = \sum_{K} K x_{\mathrm{C,S}}(K,\mathrm{A}^{(1°)}) = \int_0^{R_{\mathrm{C}}} g_{\mathrm{AW}}^{(1°)}(R) 4\pi R^2 \,\mathrm{d}R \tag{33}$$

and

$$\bar{\nu}_{\mathrm{A}}^{(1°)} = \int_{-\infty}^{+\infty} \nu x_{\mathrm{B,S}}(\nu,\mathrm{A}^{(1°)}) \,\mathrm{d}\nu = U_{\mathrm{AW}}^{(1°)} \tag{34}$$

and their relationship to radial distribution functions and mean energies.

The analysis in terms of solute atoms described above can be readily extended to encompass compositional analyses of the local solution environment of functional groups in a polyfunctional molecule or of subunits, residues, or other well-defined components of molecules and macromolecules. Let us define the counting function for the kth order solvation of functional group F as

$$N_{\mathrm{S}}(\mathbf{R}^N,\mathrm{F}^{(k)}) = \sum_{\mathrm{A}\in\mathrm{F}} N_{\mathrm{S}}(R^N,\mathrm{A}^{(k)}) \tag{35}$$

where $N_{\mathrm{S}}(R^N,\mathrm{A}^{(k)})$ is defined in eq 22. The average number of solvent molecules associated with F to order (k) is

$$N_{\mathrm{S}}(\mathrm{F}^{(k)}) = \langle N_{\mathrm{S}}(\mathbf{R}^N,\mathrm{F}^{(k)}) \rangle \tag{36}$$

and the corresponding mole fraction quantity

$$x_{\mathrm{S}}(\mathrm{F}^{(k)}) = N_{\mathrm{S}}(\mathrm{F}^{(k)})/N \tag{37}$$

The compositional analysis of the local solution environment of functional groups can be pursued further in terms of coordination numbers and binding energies by means of the joint distribution functions $x_{\mathrm{C,S}}(K,\mathrm{F}^{(k)})$ and $x_{\mathrm{B,S}}(\dot{\nu},\mathrm{F}^{(k)})$, defined analogously to the corresponding quantities in eq 33 and 34. The average coordination number and binding energy for the functional groups are simply

$$\overline{K}_{\mathrm{F}}^{(k)} = \sum_{\mathrm{A}\in\mathrm{F}} \overline{K}_{\mathrm{A}}^{(k)} \tag{38}$$

and

$$\bar{\nu}_{\mathrm{F}}^{(k)} = \sum_{\mathrm{A}\in\mathrm{F}} \bar{\nu}_{\mathrm{F}}^{(k)} \tag{39}$$

Analogous considerations follow for subunit or residue analysis. Note that normalization considerations do not follow straightforwardly on these quantities since the same atom may appear in more than one F.

IV. Calculations

Calculations on the dilute aqueous solution of formaldehyde, $[\mathrm{H_2CO}]_{\mathrm{aq}}$, were organized to demonstrate the analysis of a molecular solution based on the proximity criterion as described above. Formaldehyde is a molecular solute small enough to be treated fairly rigorously in computations, yet anisotropic enough to exhibit most of the aforementioned analysis problems particular to molecular solutions. The formaldehyde structure can be partitioned into three different chemically relevant fragments (the carbonyl group, C=O, the methylene group, $\mathrm{CH_2}$, and the aldehyde moiety, HCO), and thus provides ample opportunity to display various alternative ways in which the $[\mathrm{H_2CO}]_{\mathrm{aq}}$ analysis may be organized. Also, suitable intermolecular potential functions are already available for this system.[8] We note in passing the well-known tendency of formaldehyde as well as certain other organic aldehydes and ketones to react with solvent water to form diols. The study of this aspect of the aqueous hydration of the carbonyl group would require explicit consideration of the diol–water interaction and is not dealt with herein. Thus the formaldehyde solute, constrained here to be planar, should be considered as a means to represent the carbonyl group in prototype rather than a realistic treatment of the aqueous hydration of formaldehyde. Statistical thermodynamic (T, V, N) ensemble Monte Carlo calculations on $[\mathrm{H_2CO}]_{\mathrm{aq}}$ were carried out using a modified Metropolis method on one formaldehyde molecule and 124 water molecules at $T = 25$ °C and a density of 1 g/cm^3. The condensed phase environment of the system was modeled by simple cubic periodic boundary conditions. Convergence characteristics and error bounds on each of the calculated quantites were determined using control functions.[3,4]

The configurational energy of the system was developed under the assumption of pairwise additivity of intermolecular interactions using potential functions representative of ab initio quantum mechanical calculations of the water-water and formaldehyde-water interaction energies. For the water-water interaction energy, we continue to use the potential function developed by Matsuoka, Clementi, and Yoshimine[8] (MCY) based on moderately large configuration interaction calculations on the water dimer and used in previous studies. For the formaldehyde-water interaction, we have recently reported an analytical potential function representative of ab initio 6-31G molecular orbital calculations, and this is used without modification here.[9]

The use of MCY-CI water-water potential function is justified on the basis of the very good agreement between calculated and the observed radial distribution function for liquid water. No such direct test on the quality of the formaldehyde-water potential is available. A slice of the potential-energy hypersurface for the formaldehyde-water interaction based on this function is shown in Figure 1. The basic features expected from structural chemical considerations are all in place, and the calculated C=O···H hydrogen bond energy of −5.4 kcal/mol is reasonable. We consider this function at least of sufficient accuracy to adequately demonstrate the analysis.

The remainder of this section deals with details of the computer simulation. All potential functions in the simulation were truncated at a spherical cutoff of 7.45 Å. The initial configuration in the Monte Carlo calculation was an equilibrated geometry taken from work in progress on $[\mathrm{Na^+}]_{\mathrm{aq}}$ by Mezei et al.[5] and replacing $\mathrm{Na^+}$ with $\mathrm{H_2CO}$. The standard Metropolis sampling procedure was modified to include preferential sampling within a radius of 6.35 Å of the solute in the manner suggested by Owicki and Scheraga.[10]

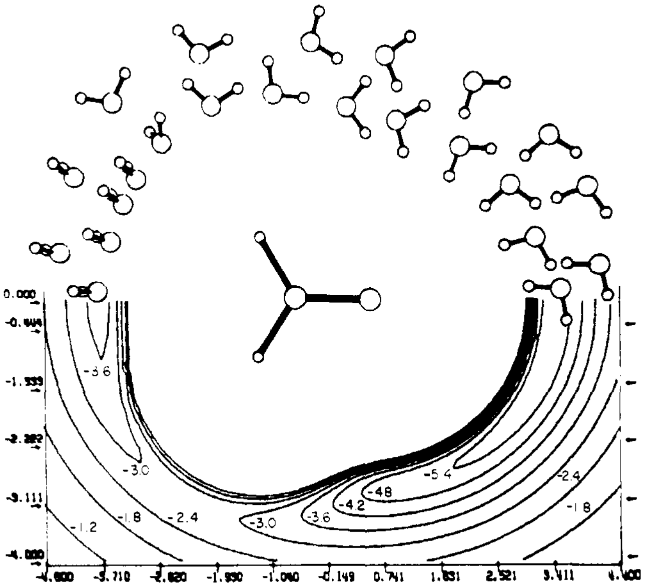

Figure 1. Isoenergy (kcal/mol) contour map of orientationally optimized formaldehyde-water interaction energies calculated from the solute-solvent intermolecular potential function used herein. The distance coordinates refer to the separation between the center of mass of formaldehyde and water. The molecular geometry corresponding to a given energy is depicted in mirror-image position in the top half of the figure. The relative size of the formaldehyde and water molecular structures are scaled down to improve legibility of the figure.

Table I. Calculated Internal Energy and Related Quantities for $[H_2CO]_{aq}$ at 25 °C (kcal/mol)

U_{SW} (N_W = 124, N_S = 1)	−1089.30 ± 10.23
$U_{W'}$ (N_W = 124)	−1067.35 ± 10.36
$\underline{U}_W$ (N_W = 124)	−1072.6 ± 4.96
$\overline{U}_{S'}$	−21.95 ± 2.0
$\overline{U}_{rel}$	5.25 ± 11.49
$\overline{U}_S$	−16.7 ± 11.67

The complete simulation involved a total of 1200K configurations. The initial 600K configurations were used to equilibrate the system, and the ensemble average properties were determined over the remaining 600K. The convergence characteristics of the calculation are displayed in Figure 2. The calculated internal energy and related quantities for $[H_2CO]_{aq}$ are collected in Table I. The quantities entered here are the mean energy U_{SW} of the solution of one solute, N_S = 1, and 124 solvent molecules, N_W = 124; the energy U_W for 124 water molecules in $[H_2O]_l$; U_W', the corresponding energy of solvent water in $[H_2CO]_{aq}$; $\overline{U}_S$, the calculated partial molar internal energy of transfer of H_2CO into water; and $\overline{U}_S'$ and $\overline{U}_{rel}$, the solute-solvent and solvent relaxation contributions to $\overline{U}_S$. Each of these quantities is formally defined in eq 1-12 and Figure 12 of ref 4. The comparison of calculated with observed values for the solvation energy of formaldehyde is not relevant here owing to the aforementioned problem with diol formation.

V. Results

We present in this section an analysis of the computer simulation results on $[H_2CO]_{aq}$ at 25 °C based on the proximity criterion. The analysis is first presented on a solute atom by atom basis, then developed in terms of functional groups. Finally, the integrated results for the entire solute molecule are given. The results and their implications are discussed in the following section.

We begin with an analysis of the local solution environment of the oxygen atom in H_2CO. The calculated oxygen-water radial distribution function is given in Figure 3. (All solute-water radial distribution functions presented herein are referred to the center of mass of solvent water molecules.) The total $g_{OW}(R)$ for oxygen and the corresponding running coordination number $\overline{K}(R_C)$ are given in Figure 3a. The broadness of the first main peak (2.4-5.2 Å) and the lack of well-defined structure are a consequence of the composite nature of this quantity, which involves contributions from the solvent in the vicinity of both the carbonyl and methylene regions of the solute molecule.

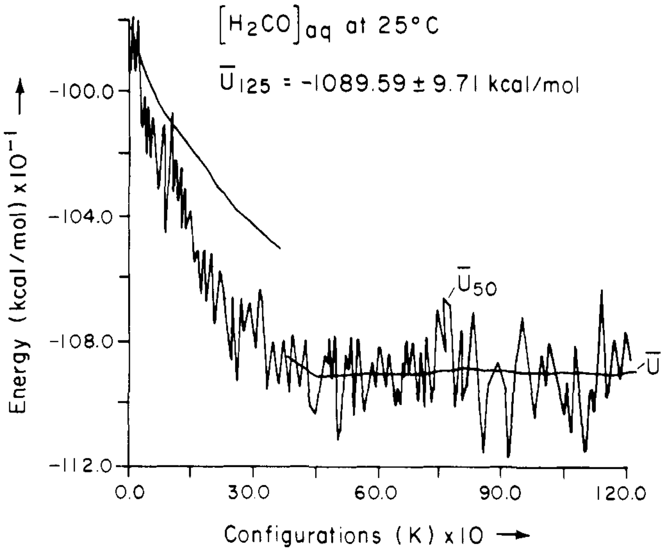

Figure 2. Convergence profile for the Monte Carlo realization on $[H_2CO]_{aq}$ at T = 25 °C. The solid line labeled $\overline{U}$ is the cumulative mean energy and the $\overline{U}_{50}$ are the Wood control functions taken over 50K intervals in the realization.

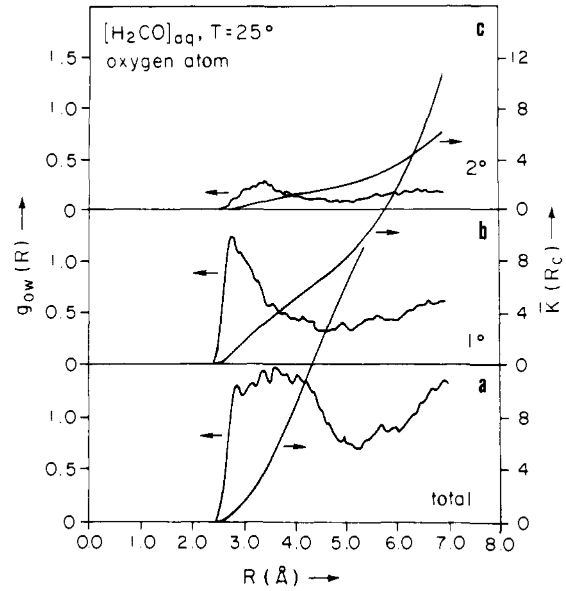

Figure 3. Calculated solute oxygen-solvent water distribution functions for $[H_2CO]_{aq}$.

Application of the proximity criterion permits the solvent molecules 1° and 2° to the oxygen atom as well as higher order contributions to be identified. The corresponding quantities $g_{OW}^{(1°)}(R)$ and $g_{OW}^{(2°)}(R)$ are displayed in Figure 3 along with the total $g_{OW}(R)$. The primary contribution is clearly dominant and does exhibit a considerably simpler appearance than the total $g_{OW}(R)$. The first peak in $g_{OW}^{(1°)}(R)$ is still quite broad and shows a minimum at 4.5 Å, thus spanning the region associated with the first two hydration shells in the corresponding $[H_2O]_l$. This possibility is developed further

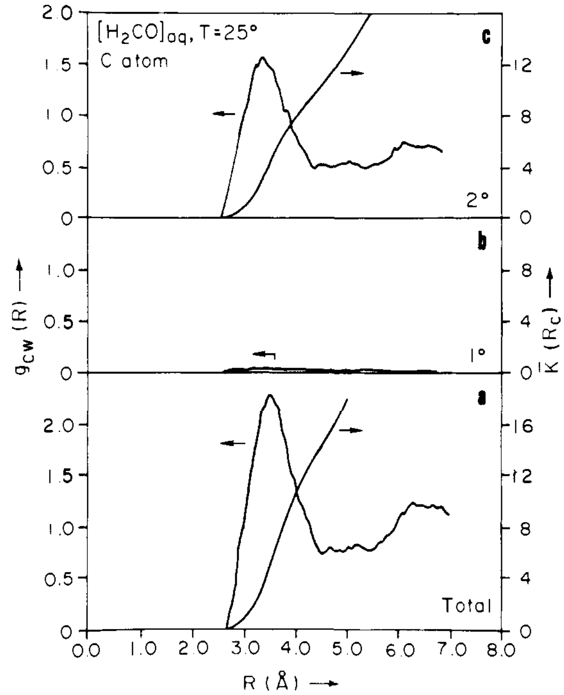

Figure 4. Calculated solute carbon-solvent water distribution functions for $[H_2CO]_{aq}$.

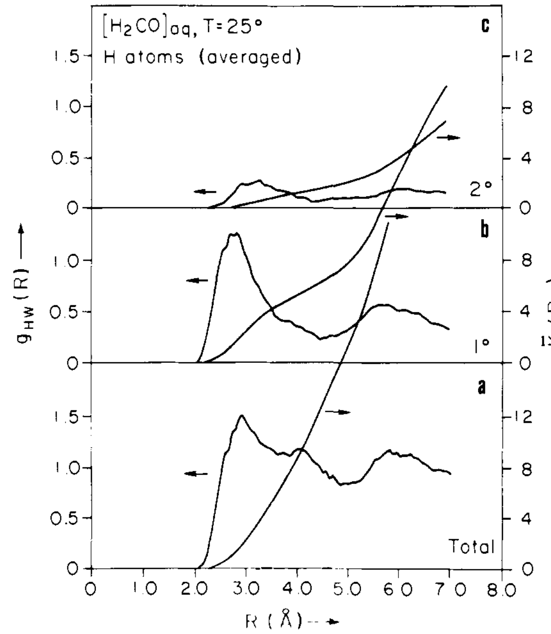

Figure 5. Calculated solute hydrogen-solvent water distribution functions for $[H_2CO]_{aq}$.

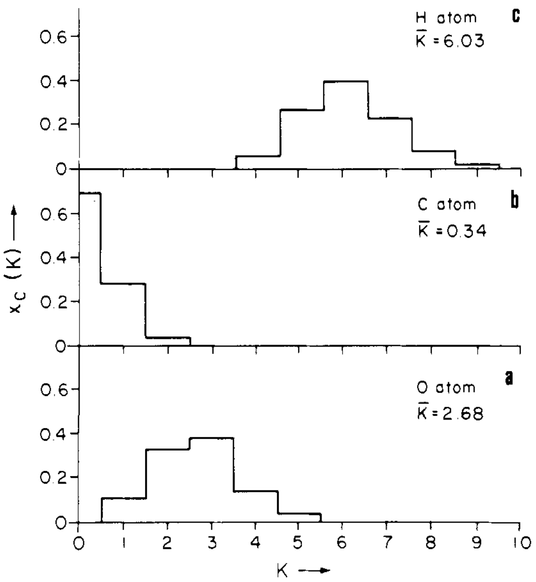

Figure 6. Calculated QCDF for primary solute-solvent coordination number on an atom by atom basis in $[H_2CO]_{aq}$.

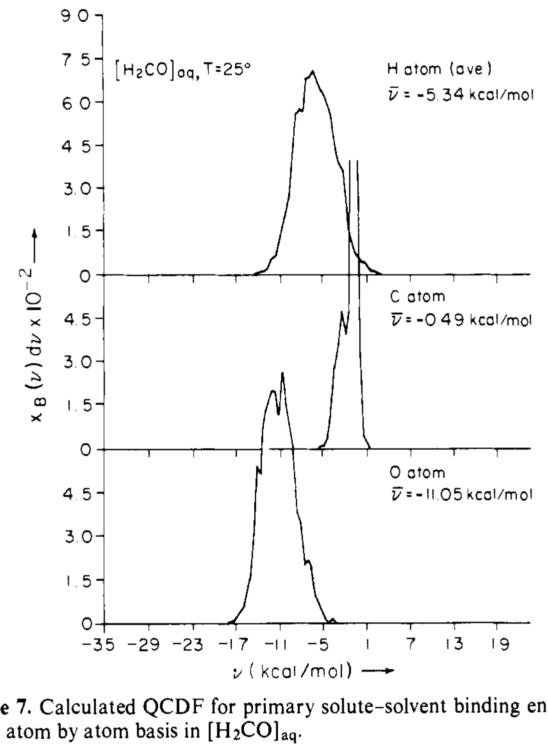

Figure 7. Calculated QCDF for primary solute-solvent binding energy on an atom by atom basis in $[H_2CO]_{aq}$.

below. An R_C value of 3.2 Å carried over from $[H_2O]_l$ serves to resolve these two types of water in the definition of coordination number. The total area of the first peak in $g_{OW}^{(1°)}(R)$ up to R_C corresponds to $\bar{K}$ = 2.68. The net contribution of $g_{OW}^{(2°)}(R)$ is relatively small. Note that the sum of the 1° and 2° contributions will not account for the entire $g_{OW}(R)$, indicating that there are also clearly significant higher order contributions to $g_{OW}(R)$. Examination of these results shows the $g_{OW}^{(3°)}(R)$ to be the dominant remaining contribution. In this and subsequent cases herein we have chosen not to present all the higher order results (>2°) since they are generally of limited interest in the structural chemistry of the solution.

The analysis of the calculated local solution environment for the carbon atom is shown in Figure 4. The total $g_{CW}(R)$ and running coordination number are shown in Figure 4a. The resolution of $g_{CW}(R)$ into contributions based on the proximity criterion leads to the results for $g^{(1°)}(R)$ and $g^{(2°)}(R)$ shown in Figures 4b and 4c, respectively. Here we see that the $g_{CW}(R)$ is dominated by 2° rather than 1° contributions, a consequence of the limited solvent accessibility of the carbon atom. The corresponding results for the hydrogen atoms in H_2CO are shown in Figure 5. Here the individual results for the two symmetry-equivalent hydrogen atoms have been averaged together to obtain improved statistics. The 1° contribution is clearly dominant, and $g_{HW}^{(1°)}(R)$ exhibits a clear shell structure, whereas the total $g_{HW}(R)$ does not.

We turn now to an atom-by-atom analysis of the 1° composition of the local solution environment of H_2CO in terms

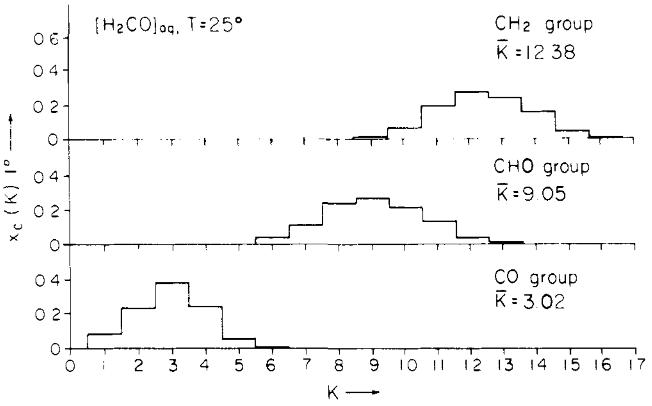

Figure 8. Calculated QCDF for primary solute-solvent coordination number on a functional group basis in $[H_2CO]_{aq}$.

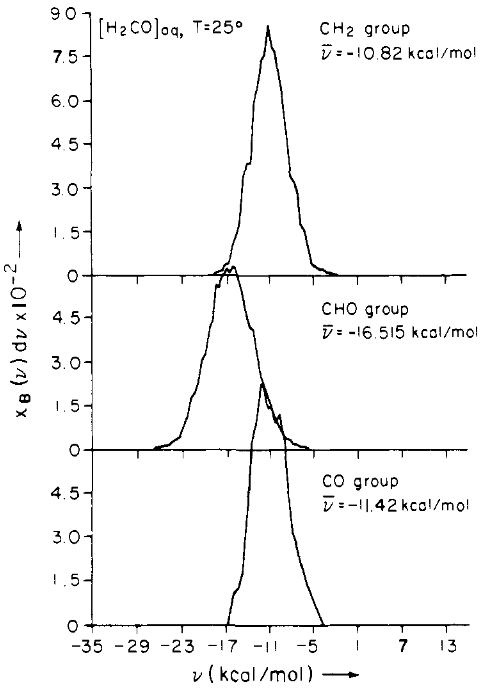

Figure 9. Calculated QCDF for primary solute-solvent binding energy on a functional group basis in $[H_2CO]_{aq}$.

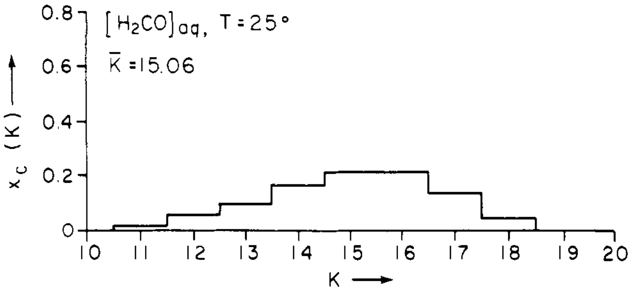

Figure 10. Calculated QCDF for primary solute-solvent coordination number on a molecular basis in $[H_2CO]_{aq}$.

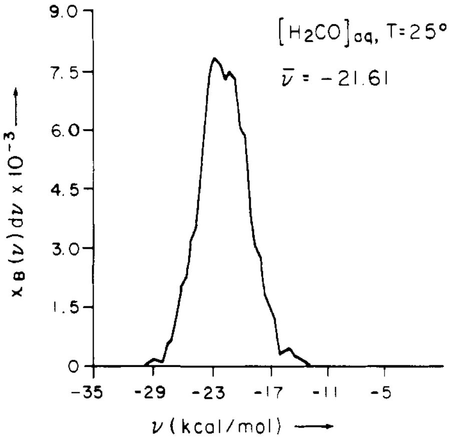

Figure 11. Calculated QCDF for primary solute-solvent binding energy on a molecular basis in $[H_2CO]_{aq}$.

of quasi-component distributions functions. The 1° radial distribution functions were used along with R_c values of 3.2, 5.0, and 4.5 Å for oxygen, carbon, and hydrogen, respectively. Since $g_{CW}^{(1°)}(r)$ is not helpful for determining R_C for carbon the value of 5.0 was inferred from $g_{CW}(R)$ in $[CH_4]_{aq}$ determined in previous work. The hydrogen R_c value was determined from the first minimum in $g_{HW}^{(1°)}(R)$.

The distribution of 1° coordination number is shown for each atom in Figure 6. For the oxygen atom, Figure 6a, the distribution ranges from $K = 1$ to $K = 5$, with the dominant contribution coming from $K = 2$ and $K = 3$. The average primary coordination number for oxygen is 2.68. The primary coordination numbers for carbon range from 0 to 2, with zero dominant and $\bar{K} = 0.34$. For hydrogen, the distribution ranges from $K = 4$ to $K = 8$ with $K = 4$, 5, and 6 all important. The average primary coordination number for hydrogen is 6.03. The corresponding analysis of 1° composition in terms of binding energy is shown in Figure 7. The average primary binding energies for oxygen, carbon, and hydrogen are −11.05, −0.49, and −5.34 kcal/mol, respectively.

We turn next to the analysis of the composition of the local solution environment of $[H_2CO]_{aq}$ in terms of the carbonyl, aldehyde, and methylene functional groups. This analysis is developed for each functional group in terms of primary compositional indexes of each of the contributing atoms as discussed at the end of section III. The results on the distribution of functional group coordination numbers and functional group binding energies are given in Figures 8 and 9, respectively. Average coordination and binding energies for each functional group are noted on each of the figures.

The integrated results on the primary solvation of the entire formaldehyde molecule are given in Figures 10 and 11. The distributions of primary coordination numbers and binding energies are given in Figures 10 and 11, respectively. The average solvent coordination number for formaldehyde in $[H_2CO]_{aq}$ is 15.06 and the average binding energy is −21.61 kcal/mol.

In concluding this section, we present in Figures 12 and 13 two stereographic views of formaldehyde and first hydration shell extracted from significant contributions to the realization. Of course many thousands of configurations go into the ensemble average, and these can only be provisionally considered representative. However, a large number of the configurations have very similar structural characteristics, and these views considered in perspective provide a useful alternative view of the system.

Considering the contributing structures together with the result on $g_{OW}^{(1°)}(R)$ in Figure 3 shows there to be two or three solvent waters at <3.2 Å distinctly hydrogen bonded to the carbonyl oxygen with one of them making an especially good linear hydrogen bond. The $\bar{\nu}$ figure of −11.05 for oxygen is also consistent with the existence of two or three O···H_2O hydrogen bonds. The bent hydrogen bonds seen in Figures 12 and 13 are similar in appearance to those seen previously in $[H_2O]_l$ and $[CH_4]_{aq}$. The solvent waters in the first peak >3.2 Å are primarily interacting with the water molecules proximal to the

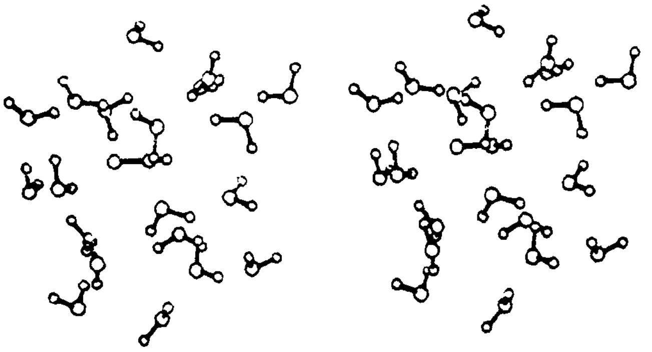

Figure 12. Stereographic view of a significant molecular structure contributing to the statistical state of $[H_2CO]_{aq}$.

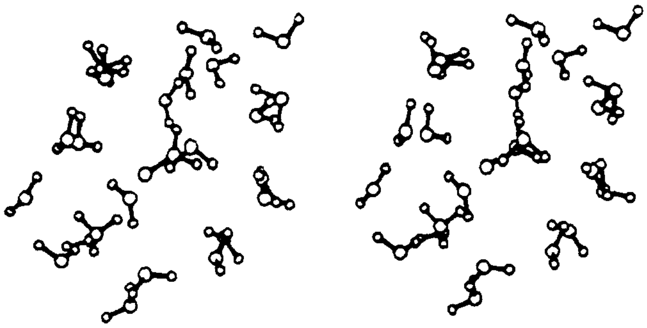

Figure 13. Stereographic view of another significant molecular structure contributing to the statistical state of $[H_2CO]_{aq}$.

carbonyl group rather than with the carbonyl group directly. This collectively is the justification for choosing $R_c = 3.2$ for oxygen.

VI. Discussion

The results presented in the previous section demonstrate the extent to which total solute-solvent radial distribution functions can be resolved into physically significant contributions based on the proximity criterion. The procedure was successful in establishing a reasonable means for the definition of coordination number of solute atoms. The distributions of coordination-number values obtained for each of the atoms of H_2CO in aqueous solution are reasonable in light of the relative electronegativity and solvent exposure of each of the atoms. The values of coordination number for the individual functional groups and for the entire molecule compounded from the atomic coordination numbers are generally in line with expectations of structural chemistry; however, the total first shell solvent coordination number of 15.06 is somewhat higher than is customarily assumed in solvation models.

A preliminary look at the transferability of solvation-number values obtained for functional groups under the proximity criterion can be developed in terms of the methylene group. The 1° coordination number of the methylene functional group in $[H_2CO]_{aq}$ is ≈ 11. Referring back to previous calculations from this laboratory on $[CH_4]_{aq}$ the average methane coordination number was found to be 19.35, or approximately 10 per methylene unit. Thus the 1° solvent coordination of CH_2 in $[H_2CO]_{aq}$ of $K = 12$ is observed to be approximately the same as for CH_2 in $[CH_4]_{aq}$. Thus the transferability of 1° coordination numbers for atoms and functional groups in aqueous solution appears quite promising.

In conclusion, we note in structural biochemistry an interesting and powerful means of studying environmental effects on macromolecular structure and function in terms of the topological "solvent accessibility" of the various subunits of the structure. It appears to us that reasonable quantitative estimates of the solvation energy and solvent coordination numbers of a biomacromolecule could be made from a combination of 1° solvation analyses of the residues of the structure used in conjunction with solvent accessibility data. Subsequent studies will be organized on this point.

VII. Summary and Conclusions

In the preceding sections, we introduced the idea of a structural analysis of the statistical state of solutions based on the proximity criterion, developed the requisite formalism, and demonstrated the procedure in the analysis of a statistical thermodynamic Monte Carlo computer simulation on $[H_2CO]_{aq}$ at $T = 25$ °C. The results provide a detailed interpretation of the local solution environment of the formaldehyde molecule in water. The solute atom-water radial distribution functions were resolved into contributions from successive orders of solvation based on the proximity criterion, and the results for 1° coordination permit a meaningful definition of solute atom coordination number. An analysis of the local solution environment of the carbonyl, methylene, and aldehyde functional groups was developed in terms of atomic contributions and analyzed based on quasi-component distribution functions for coordination number and binding energy. A preliminary consideration of the transferability of local solution environments for solute atoms and functional groups was developed by comparing the 1° solvent coordination number of methylene in $[H_2CO]_{aq}$ with corresponding values for methylene in $[CH_4]_{aq}$. Distinct correspondence was observed for both the distribution function and coordination number results. If the results for $[H_2CO]_{aq}$ are indicative of the method, the proximity criterion may well be the key to the development of a simply understood yet formally general structural chemistry of the statistical state of molecular solutions, and form the basis for a rigorously based descriptive chemistry of solution structure and processes.

Acknowledgment. This research was supported by Grant 5-R01-GM-24914-5 and a City University of New York Faculty Research Award. Discussions with Drs. M. Mezei and S. Swaminathan are also gratefully acknowledged.

References and Notes

[1] J. A. Barker and D. H. Henderson, *Rev. Mod. Phys.*, **48**, 587 (1976).

[2] A. Ben-Naim, "Water and Aqueous Solutions", Plenum Press, New York, 1974.

[3] S. Swaminathan and D. L. Beveridge, *J. Am. Chem. Soc.*, **99**, 8392 (1977).

[4] S. W. Harrison, S. Swaminathan, and D. L. Beveridge, *J. Am. Chem. Soc.*, **100**, 5705 (1978).

[5] M. Mezei and D. L. Beveridge, submitted for publication.

[6] D. L. Beveridge, M. Mezei, S. Swaminathan, and S. W. Harrison in "Computer Modeling of Matter", P. G. Lykos, Ed., American Chemical Society, Washington, D.C., 1978.

[7] See, for example, P. Rossky and M. Karplus, *J. Am. Chem. Soc.*, **101**, 1913 (1979).

[8] O. Matsuoka, E. Clementi, and M. Yoshimine, *J. Chem. Phys.*, **64**, 1351 (1976).

[9] S. Swaminathan, R. J. Whitehead, E. Guth, and D. L. Beveridge, *J. Am. Chem. Soc.*, **99**, 7817 (1977).

[10] J. C. Owicki and H. A. Scheraga, *Chem. Phys. Lett.*, **47**, 600 (1977).