electric dipoles, quadrupoles, and point polarizabilities,[62] and models with TIPS[24] water molecules and charged soft-sphere ions.[63] These integral equation results have been paralleled in some cases by simulations[64,65] which give remarkably consistent results.

The picture (Fig. 8) emerging from these studies and the quite consistent results derived for the Patey model[62] is that the solvent-averaged solute–solute pair potentials derived from the various BO-level Hamiltonians show far more effect of the structure of the solvent than the Gurney-type models which have been adjusted to fit various solution properties.

Taken together, all of these aqueous solution results present some confusing inconsistencies, but they also show clearly that a new level of definitive and detailed structural interpretation of solution properties can be reached even by the continued application of known experimental and theoretical techniques, and it seems likely that there will be further development of the techniques themselves. The study of solution structures, although an old field, is far from mature.

Acknowledgments

I am grateful to the National Science Foundation for the support of this work, and to Professors P. J. Rossky and A. McCammon and Dr. M. Pettit for permission to present Fig. 8, which shows some of their unpublished research results.

[62] P. G. Kusalik and G. N. Patey, *J. Chem. Phys.* **79,** 4468 (1983).
[63] P. J. Rossky, M. Pettit, and A. McCammon, to be published.
[64] C. Pangali, M. Rao, and B. J. Berne, *J. Chem. Phys.* **71,** 2975, 2982 (1979).
[65] M. Berkowitz, O. A. Karim, A. McCammon, and P. J. Rossky, *Chem. Phys. Lett.* **105,** 577 (1984).

[2] Structural Chemistry of Biomolecular Hydration via Computer Simulation: The Proximity Criterion

By MIHALY MEZEI and DAVID L. BEVERIDGE

Introduction

The computer simulation of systems of biological molecules in aqueous solution, including other components of the environment such as counterions, is a challenging problem in theoretical biochemistry for the supercomputer age. Three aspects of biomolecular simulations require

serious attention as the field now emerges from infancy and establishes a broad-based credibility: (a) the development of accurate intermolecular functions, (b) the improvement of simulation methodology within the context of the Monte Carlo and molecular dynamics procedures, and (c) the analysis of results in a form accessible to a larger community of structural biochemists, molecular pharmacologists, and others requiring information from computer models to apply to their research studies. This chapter deals with the analysis issue and describes our effort to formulate a structural chemistry of hydration and environmental effects in general from the results of molecular simulation.

The first requirement of a structural chemistry of environmental effects is that we extend the idea of structure from that of individual molecules and complexes to the "statistical state" of the system, defined by the manifold possible complexions of a molecular assembly and their corresponding Boltzmann weighting factors. Also, we extend the idea of structure to composition, which includes both molecular geometry (conventional definition of structure) and also the energetic indices. In liquid state theory, the composition of a fluid follows from a knowledge of the molecular distribution function 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 theoretical approach to this problem was mapped out several years ago for pure fluids by Ben-Naim[1] 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, QCDFs 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. 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 devel-

[1] A. Ben-Naim, "Water and Aqueous Solutions," Plenum, New York (1974).

oped 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 biomolecules 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, 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 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 arbitrary partitioning of configuration space and to develop structural characteristics of the fluid environment within that region. 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 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.[2] 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 (GMDF). 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 attempt herein to collect our formal analysis of the problem in terms of the proximity criterion along with representative examples of applications to the study of the hydration of biomolecules and prototypes thereof. We feel one of the most potentially useful results emerging from this work is a well-defined idea of the "hydration complex" of a dissolved molecule, i.e., the solute and first shell of hydration extracted from a simulation of higher dimensionality. There are a number of projects now under way using hydration complexes defined from simulation on large assemblies in more rigorous electronic structure calculations at the level of quantum chemistry. Examples are the calculation of solvent effects on nuclear magnetic resonance (NMR) shielding constants,[3] electronic spectra,[4] optical rotatory strengths,[5] and vibrational spectra.[6] Also, the presentation of simulation results using computer graphics can be carried forth in terms of hydration complex theory, i.e., stereographic displays with solvent atoms color coded based on a proximity criterion analysis of their mode of hydration: ionic, hydrophilic, or hydrophobic.

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 structural and energetic characteristics of the system. The

[2] P. K. Mehrotra and D. L. Beveridge, *J. Am. Chem. Soc.* **102,** 4287 (1980).
[3] C. Giessner-Prettre and A. Pullman, *Chem. Phys. Lett.* **114,** 258 (1985).
[4] P. R. Callis, personal communication.
[5] G. A. Segal, personal communication.
[6] U. Gunnia, M. Diem, S. Cahill, M. S. Broido, G. Ravishanker, and D. L. Beveridge, *Conversat. Biomol. Stereodyn., 4th, Albany* (1985).

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 delineate 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.

We briefly review the formulation of these quantities for homogeneous 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 $\mathbf{X}^N$:

$$\mathbf{X}^N = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N\} \tag{1}$$

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

For any molecular property Q that is a function of the configurational coordinates $\mathbf{X}^N$ either directly or indirectly one can define a counting function $C_i^Q(\mathbf{X}^N, q)$:

$$C_i^Q(\mathbf{X}^N, q) = \delta[q - \mathrm{Q}_i(\mathbf{X}^N)] \tag{2}$$

where $Q_i(\mathbf{X}^N)$ gives the value of the property Q for molecule i, and $\delta[\ ]$ is the Dirac delta. The QCDF [1] $x_Q(q)$ is defined as

$$x_Q(q) = \int \ldots \int P(\mathbf{X}^N) \sum_{i=1}^{N} C_i^Q(\mathbf{X}^N, q)\, D_i^Q(\mathbf{X}^N)\, d\mathbf{X}^N \Big/ \int \ldots \int P(\mathbf{X}^N) \sum_{i=1}^{N} D_i^Q(\mathbf{X}^N)\, d\mathbf{X}^N \tag{3}$$

where $D_i^Q(\mathbf{X}^N)$ is a selector function whose value is either 0 or 1, which determines if molecule i will contribute to the particular $x_Q(q)$ or not. In applications to pure liquids, the selector function is usually taken as unity. For discrete properties, $x_Q(q)$ describes the configurational average of the fraction of molecules selected by $D_i^Q(\mathbf{X}^N)$ for which the value of the property Q is exactly q. For continuous properties, $x_Q(q)dq$ gives the configurational averaged fraction of molecules with property Q in the interval $[q, q + dq]$. For properties that are functions of a pair of molecules, the

corresponding QCDF is obtained as

$$X_Q(q) = \int \cdots \int P(\mathbf{X}^N) \sum_{i<j}^{N} C_{ij}^{Q}(\mathbf{X}^N, q)\, D_{ij}^{Q}(\mathbf{X}^N)\, d\mathbf{X}^N \Big/ \int \cdots \int P(\mathbf{X}^N) \sum_{i<j}^{N} D_{ij}^{Q}(\mathbf{X}^N)\, d\mathbf{X}^N \quad (4)$$

In general, by specifying $Q_i(\mathbf{X}^N)$, $D_i^Q(\mathbf{X}^N)$, or $Q_{ij}(\mathbf{X}^N)$, $D_{ij}^Q(\mathbf{X}^N)$, the corresponding QCDF $x_Q(q)$ is fully defined via Eqs. (2, 3, or 4). The configurational average of the property Q can be obtained from the corresponding QCDF as

$$\bar{Q} = \begin{cases} \int_{-\infty}^{\infty} x_Q(q)\, dq & \text{for continuous property Q} \\ \sum_{\{q\}} x_Q(q)\, q & \text{for discrete property Q} \end{cases} \quad (5)$$

In studying structural parameters in a statistical mechanical context, it should be noted that there are both probabilistic and energetic factors to consider and that the most favorable parameter value energetically may not be the most probable, particularly when it is associated with a relatively small region of configuration space. This circumstance is expressed quantitatively by a comparison of the QCDF $x_Q(q)$ and the corresponding quasi-component correlation function (QCCF) $q_Q(q)=x_Q(q)/v_Q(q)$, the latter quantity being normalized by the volume element of the configuration space with respect to the parameter q.

Figures 1 and 2 show several examples for the QCDFs and QCCFs computed for liquid water with the MCY model[7] in this laboratory.[8] Figure 1 gives the radial distribution function $g(R)$, the QCDF of the coordination number $x_C(K)$, binding energy $x_B(v)$, near-neighbor pair energy $x_P(\varepsilon)$, and near-neighbor dipole angle $x_D(\theta)$ and Fig. 2 defines the four hydrogen-bonding parameters R_{OO}, θ_H, θ_{LP}, and δ_D and gives their QCDFs and QCCFs.

The radial distribution function $g(R)$ can be defined as the QCCF of "distance." If we take the property $Q_{ij}(\mathbf{X}^N)$ to be the distance R_{ij} between molecules i and j and use $D_{ij}^R(\mathbf{X}^N) \equiv 1$, we obtain the QCDF $x_R(r)$. $x_R(r)\, dr$ gives the fraction of pairs whose distance R_{ij} falls into the interval $[r, r + dr]$. It is easy to see that

$$g_R(r) \equiv g(r) = x_R(r)/[4\pi^2 N/(N-1)V] \quad (6)$$

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

[8] D. L. Beveridge, M. Mezei, P. K. Mehrotra, F. T. Marchese, G. Ravi-Shanker, T. R. Vasu, and S. Swaminathan, *Adv. Chem. Ser.* **204,** 297 (1983).

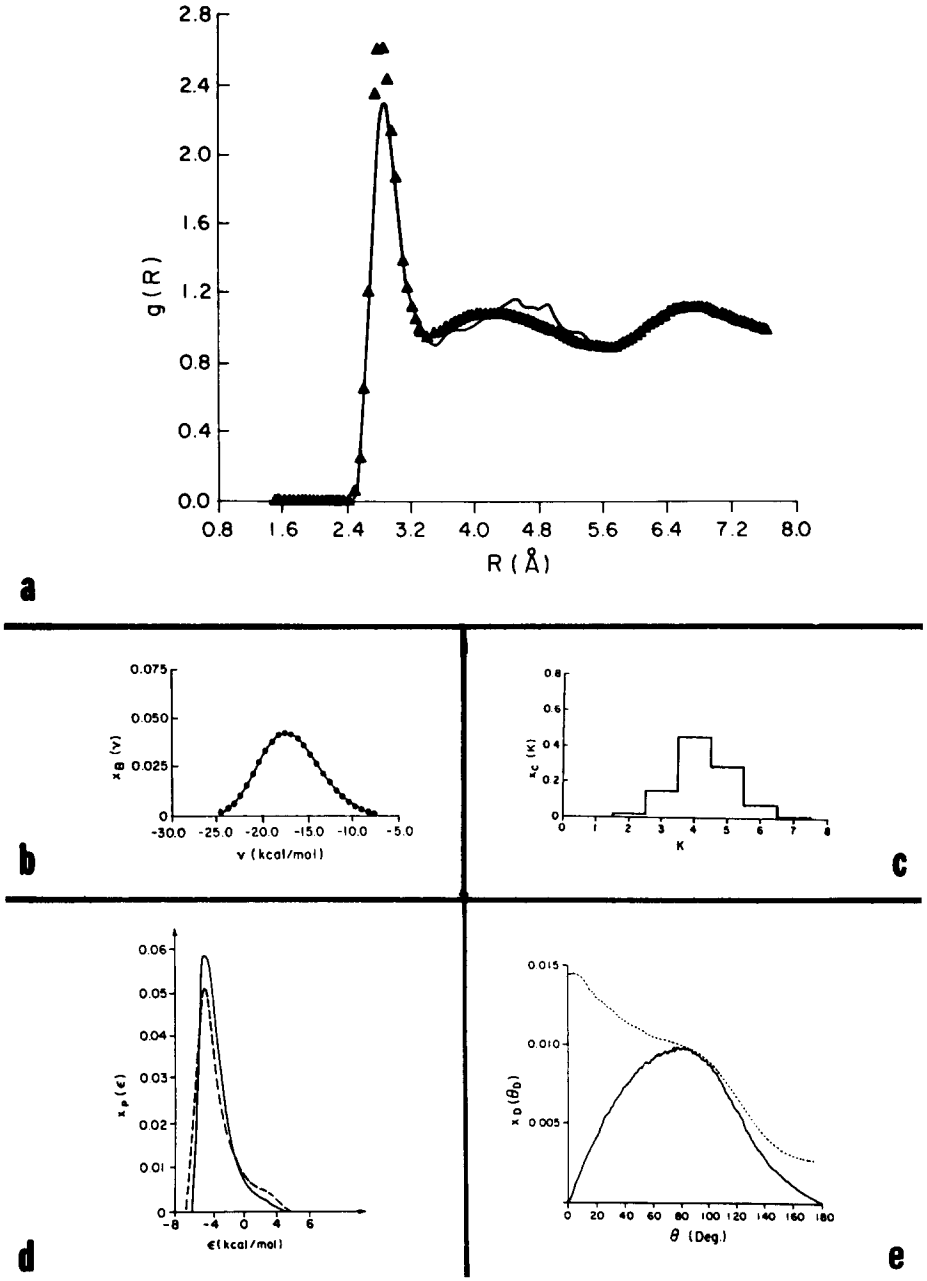


FIG. 1. Liquid water distribution functions computed from the MCY model at 25°. (a) $g(R)$ calculated: ▲; experimental[25]: full line; (b) $x_B(v)$; (c) $x_C(K)$; (d) $x_\phi(\varepsilon)$, ——, MCY model, and ---, ST2 model; and (e) ——, $x_D(\theta)$, and ····, $g_D(\theta)$ with $v(\theta) = \sin(\theta)$.

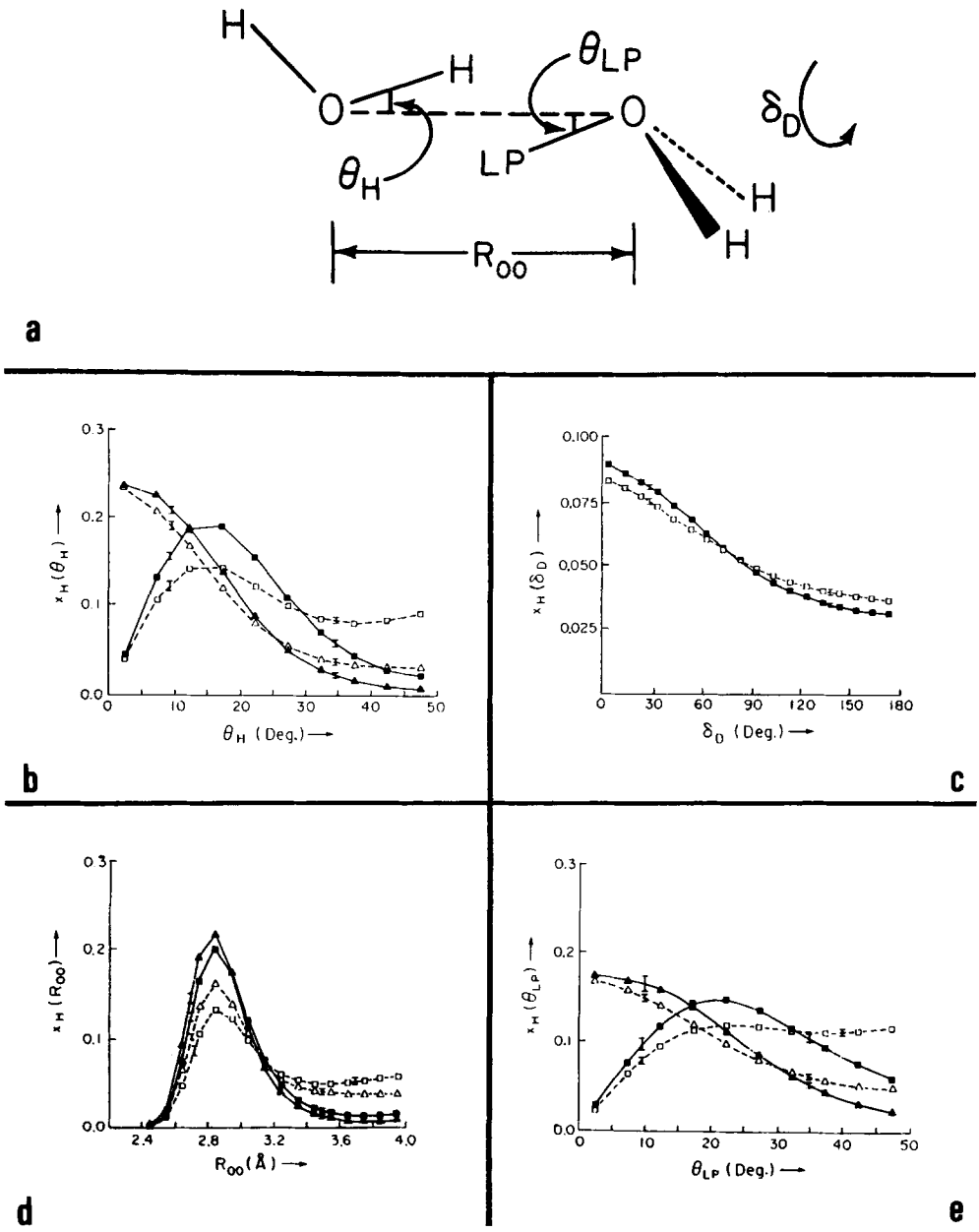


FIG. 2. QCDFs and QCCDs of the hydrogen-bonding parameters computed from the MCY model at 25° using two different hydrogen-bond definitions. The QCDFs are represented by squares and the QCCFs by triangles. Open symbols refer to the weak hydrogen bond and filled symbols to the strong hydrogen bond.[10] (a) The definition of the hydrogen-bond parameters; (b) $x_H(\theta_H)$ and $g_H(\theta_H)$; (c) $x_H(\delta_D)$; (d) $x_H(R_{OO})$ and $g_H(R_{OO})$; (e) $x_H(\theta_{LP})$ and $g_H(\theta_{LP})$.

that is, $v_R(r) = 4\pi r^2 N/(N-1)V$ defines the QCCF $g_R(r)$ as the commonly known radial distribution function $g(r)$.

The property "coordination number," $C_i(\mathbf{X}^N)$, for particle i is defined as

$$C_i(\mathbf{X}^N) = \sum_{j=1}^{N} h(R_{ij} - R_c) \tag{7}$$

where $h(R_{ij} - R_c)$ is a unit step function equal to unity if the interparticle separation between molecules i and j, 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 $C_i(\mathbf{X}^N)$ thus gives the number of other molecules that fall within the first coordination sphere of particle i in configuration $\mathbf{X}^N$. The quantity called "running coordination number" is simply the average coordination number $\bar{K}$ as a function of the cutoff radius R_c.

The binding energy of particle i in configuration $\mathbf{X}^N$ is defined as

$$B_i(\mathbf{X}^N) = E(\mathbf{X}_1, \ldots, \mathbf{X}_N) - E(\mathbf{X}_1, \ldots, \mathbf{X}_{i-1}, \mathbf{X}_{i+1}, \ldots, \mathbf{X}_N) \tag{8}$$

where E is the configurational energy of the system. $B_i(\mathbf{X}^N)$ is the negative of the vertical dissociation energy of the ith molecule. The thermodynamic configurational internal energy U is related to the average binding energy $\bar{\nu}$ by the expression

$$U = N\bar{\nu}/2 \tag{9}$$

Next we consider the distribution of the pair energies. This is a well-defined quantity only for pairwise additive potentials, although for certain types of cooperative potentials we can develop a meaningful partitioning of the total energy into contributions of pairs. This quantity is usually presented for all pairs of particles in the system. It has the disadvantage, however, that the large peak around zero, corresponding to the distant pairs, tends to dominate the curve. The quantity of principal interest, the peak corresponding to the optimum near-neighbor distance, may only appear as a shoulder in this distribution. However, by including only near-neighbor water pairs (i.e., waters whose distance is R_c or less), the dominant feature of the curve will be the peak corresponding to the optimum near-neighbor distance,[9] and the interpretation is then straightforward. The pair energy between two molecules i and j, $P_{ij}(\mathbf{X}^N)$, simply gives the interaction energy computed between the two molecules. If all pairs are to be taken into account, $D_{ij}^{P}(\mathbf{X}^N)$ is identically zero. If only near-

[9] M. Mezei and D. L. Beveridge, *J. Chem. Phys.* **76,** 593 (1982).

neighbor pairs are allowed to contribute, then

$$D_{ij}^{p}(\mathbf{X}^N) = h(R_{ij} - R_c) \tag{10}$$

The distribution of pair energies is usually a more sensitive indicator than the distribution of binding energies, since the latter is obtained from the averaging of the former.

The orientational correlations in the liquid can be described in several ways. Comparison of the atom–atom radial distribution functions usually gives an immediate clue for the preferred relative orientation. A more definitive answer can be obtained by examining QCDFs of various intermolecular angles. The QCDF of the dipole angle between pairs of molecules, $x_D(\theta_D)$, has been defined also for near-neighbor pairs. The spread in this distribution helps characterize the importance of orientational correlations, and the location of its maximum specifies the preferred orientation. For solute–solvent interaction, the QCDF of the angle between the solvent dipole and the solvent–solute direction has been evaluated. Here not only the QCDF $x_D(\theta)$ has been computed, but also the configurational average of θ as a function of distance:

$$\langle\theta(\mathrm{R})\rangle = \sum_{i=j}^{N} \delta[R - R_{ij}(\mathbf{X}^N)]\ \theta(\mathbf{X}^N) \Big/ \sum_{i-1}^{N} \delta[R - R_{ij}(\mathbf{X}^N)] \tag{11}$$

The function $\langle\theta(R)\rangle$ describes the variation in orientation as a function of solute–solvent distance. It can be used to determine the distance beyond which the orientational correlation between solute and solvent is negligible. $\langle\theta(R)\rangle = 90°$ corresponds to zero correlation. It is of particular interest, since orientational correlation can exist even when the radial density correlation is negligible, and vice versa.

For liquid water there exists a concept defined with respect to a pair of molecules, the hydrogen bond. The four internal coordinates of the water dimer that are relevant for the description of hydrogen bonding[10] are defined in Fig. 2a. Here R_{OO} is the interoxygen separation, the angle θ_H is the angle between the H–O and O–O bonds, and θ_{LP} is the angle between the LP–O and O–O bonds. The angle δ_D is the dihedral angle between the planes H–O–O and LP–O–O. In these definitions, LP is a suitably located "pseudoatom" on the water molecule, corresponding to the qualitative idea of tetrahedrally oriented lone-pair (LP) orbitals. For ST2 water, the LP pseudoatoms were chosen to coincide with the negative charges on the model water structure, while for the MCY water, they were placed in such a way that the LP–O–LP triangle is of the same dimensions as the H–O–H triangle and oriented perpendicular to it. Note that the LP posi-

[10] M. Mezei and D. L. Beveridge, *J. Chem. Phys.* **74,** 622 (1981).

tions for the analysis are not related to any terms in the analytical MCY potential function. For each water, the atom/pseudoatom participating in a hydrogen bond with another water was taken as the atom on the donor water closest to the oxygen atom of the acceptor water. A quantitative geometric definition of the H bond further requires the specification of cutoff values for each of these parameters. The strength assumed for the H bond can be modified by varying the cutoff values. Qualitative notions on the H bond place an upper bound on θ_H and θ_{LP}, since it is natural to require that the atoms on one molecule proximal to the oxygen of the other molecule should be an H and an LP, respectively. The tetrahedral character of the interaction leads to a "minimal" definition of the H bond as

$$\begin{aligned} R_{OO} &\leq R^{max} \\ \theta_H &\leq 70.53^\circ \\ \theta_{LP} &\leq 70.53^\circ \\ \delta_D &\leq 180.0^\circ \end{aligned} \tag{12}$$

A natural choice for R^{max} is the cutoff value R_c for the previously determined coordination number distribution function, 3.3 Å.

The four parameters described above determine four H-bonding QCDFs. The cutoff values chosen for the definition of the hydrogen bond determine the selector function to be used in the definition of all four hydrogen-bonding QCDFs: Its value is one only when all four hydrogen-bonding parameters fall below the preestablished threshold value:

$$D_{ij}^{H}(\mathbf{X}^N) = h(R_{OO}(\mathbf{X}^N) - R^{max})\, h(\theta_H(\mathbf{X}^N) - \theta^{max})\, h(\theta_{LP}(\mathbf{X}^N) - \theta^{max})\, h(\delta_D(\mathbf{X}^N) - \delta^{max}) \tag{13}$$

An alternative choice of selector function could consider the pair energy $P_{ij}(\mathbf{X}^N)$ and select pairs where $P_{ij}(\mathbf{X}^N)$ falls below a preestablished threshold value:

$$D_{ij}^{H}(\mathbf{X}^N) = h(P_{ij}(\mathbf{X}^N) - E^{max}) \tag{14}$$

This choice has been called the energetic definition of hydrogen bond.[11]

QCCFs have also been defined for the hydrogen-bonding QCDFs, with the volume functions chosen as

$$\begin{aligned} v_H(R_{OO}) &= 4\pi R^2 \\ v_H(\theta_H) &= \sin(\theta_H) \\ v_H(\theta_{LP}) &= \sin(\theta_{LP}) \\ v_H(\delta_D) &= 1 \end{aligned} \tag{15}$$

reflecting the change in the configurational space volume as a function of the hydrogen-bonding parameters. It is particularly important to consider

[11] A. Geiger, A. Rahman, and F. H. Stillinger, *J. Chem. Phys.* **70**, 263 (1979).

the QCCF of θ_H and θ_{LP}, since they look qualitatively different from the corresponding QCDF: Both QCCFs peak at 0°, while neither of the QCDF does, showing that the prevalence of bent hydrogen bonds is due to geometric factors. Note that an alterative way to derive the information contained in these two QCCFs is to consider $x_H(\cos\theta_H)$ and $x_H(\cos\theta_{LP})$ along with $x_H(\theta_H)$ and $x_H(\theta_{LP})$.

One may proceed along analogous lines to define other GMDFs. More detailed analyses of the statistical state can be obtained by developing GMDF for combined properties such as coordination number and binding energy together, using the combined counting function

$$C_i^{B,C}(\mathbf{X}^N, \nu, K) = C_i^B(\mathbf{X}^N, \nu) \cdot C_i^C(\mathbf{C}^N, K) \tag{16}$$

to give the joint QCDF $x_{B,C}(\nu, K)$ of binding energy as function of coordination number, as examined earlier.[12] We also computed the running coordination number as a function of pair energy, $\bar{K}(\varepsilon)$, which can be defined through another joint QCDF:

$$\bar{K}(\varepsilon) = \sum_{K=0}^{\infty} \int_{-\infty} K\, x_{P,C}(\varepsilon', K)\, d\varepsilon' \tag{17}$$

where $x_{P,C}(\varepsilon, K)$ is the joint QCDF of pair energy and coordination number, generated by the counting function

$$C_i^{P,C}(\mathbf{X}^N, \varepsilon, K) = C_i^P(\mathbf{X}^N, \varepsilon) \cdot C_i^C(\mathbf{X}^N, K) \tag{18}$$

Clearly, the limit of $\bar{K}(\varepsilon)$ at large ε is $\bar{K}$.

The graphics capabilities of most present-day computer systems encouraged the development of further analysis techniques. One such technique, the statistical state solvation site analysis,[13] displays the three-dimensional image of the envelope enclosing areas around a molecule where the density is above a threshold value. An alternative route displays the sequence of configurations generated in the simulation in rapid succession, thereby creating an animation. However, this requires a real-time vector graphics unit and is not amenable to easy reproduction and documentation in journals. Selected configurations, however, can be displayed as stereo images.

[12] S. Swaminathan and D. L. Beveridge, *J. Am. Chem. Soc.* **99,** 8392 (1977).
[13] P. K. Mehrotra, F. T. Marchese, and D. L. Beveridge, *J. Am. Chem. Soc.* **103,** 672 (1981).

Theory and Methodology

The basis for a general compositional analysis of the statistical state of molecular fluids must 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 proposed earlier the proximity criterion, which uniquely identifies each solvent molecule with a well-defined solute entity in each configuration.[2] 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. In geometrical terms, this is equivalent to saying that molecules that belong to the 1° solvation fall into the Voronoi polyhedron of A, generated by the solute atoms and the boundary of the system.[14] 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. Figure 3 shows the 1° solvation regions for formaldehyde as an example.

There are two normalization conditions on $N_A^{(k)}$ that follow directly from the defintion:

$$\sum_A N_A^{(k)} = N \text{ for any } k \tag{19}$$

and

$$\sum_k N_A^{(k)} = N \text{ for any } A \tag{20}$$

Here $N_A^{(k)}$ 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

[14] D. L. Beveridge, M. Mezei, P. K. Mehrotra, F. T. Marchese, V. Thirumalai, and G. Ravi-Shanker, *Ann. N.Y. Acad. Sci.* **367,** 108 (1981).

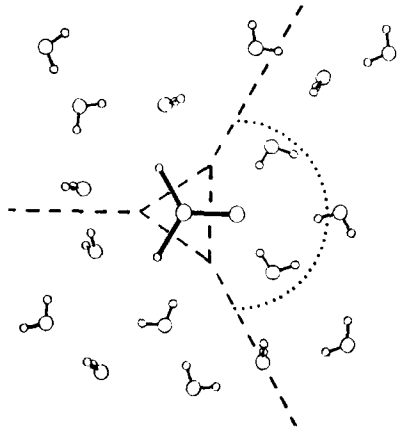


FIG. 3. Primary solvation regions of the formaldehyde molecule.

an N-particle configuration of the system $\mathbf{R}^N$,

$$\mathbf{R}^N = \{\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_N\} \tag{21}$$

Let us collect as a set the solute atoms listed in order of k. The members of this set are the "proximity indices" for solvent molecule i, $S^{(k)}(\mathbf{R}^N)$. Consider this set as the generalized property of the system in context of GMDF theory:

$$S_i(\mathbf{R}^N) = \{S_i^{(1^\circ)}(\mathbf{R}^N), S_i^{(2^\circ)}(\mathbf{R}^N), \ldots\} \tag{22}$$

where

$$S_i^{(1^\circ)}(\mathbf{R}^N) = \{A | R_{Ai} = \min_{M}\{R_{Mi}\}\} \tag{23}$$

i.e., the primary proximity index of solvent molecule is the solute atom A such that the distance R_{Ai} is the absolute minimum in the discrete set $\{R_{Mi}\}$ 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

$$S_i^{(2^\circ)}(\mathbf{R}^N) = \{A | R_{Ai} = \min_{M}\{R_{Mi}\}'\} \tag{24}$$

where the primed set, $\{R_{Mi}\}'$, is simply the set $\{R_{Mi}\}$ with the distance R_{Ai} corresponding to primary solvation deleted.

With the proximity indices thus defined for all solvent molecules, one may develop an analysis of the solvation of a solute molecule atom by atom. We are predominantly interested in the primary solvation of A, but we retain the superscript (k) notation for complete generality. For every

QCDF $X_Q(q)$, one can define the corresponding kth-order proximity QCDF $X_Q^{(k)}(q)$ by multiplying the selector function by $\delta[A - S_i^{(k)}(\mathbf{R}^N)]$.

The analysis described above can be readily extended for functional groups in a polyfunctional molecule. In this case, the selector function is simply

$$\sum_{A \in \{F\}} \delta[A - S_i^{(k)}(\mathbf{R}^N)] \tag{25}$$

where the set $\{F\}$ defines a functional group.

Two forms of radial distribution functions may be distinguished for each atom: the "total" atom–water radial distribution conventionally defined and denoted here by $g_{AW}^{tot}(R)$, where A refers to the solute atom; and an atom–water radial distribution function for those waters that are in the kth-order proximity region of the solute atom A, denoted by $g_{AW}^{(k)}(R)$. Figure 4 shows $g_{AW}^{tot}(R)$, $g_{AW}^{1°}(R)$, and $g_{AW}^{2°}(R)$, and $x_C^{i°}(K)$ for the atoms and functional groups of the formaldehyde molecule computed from Monte Carlo computer simulations.

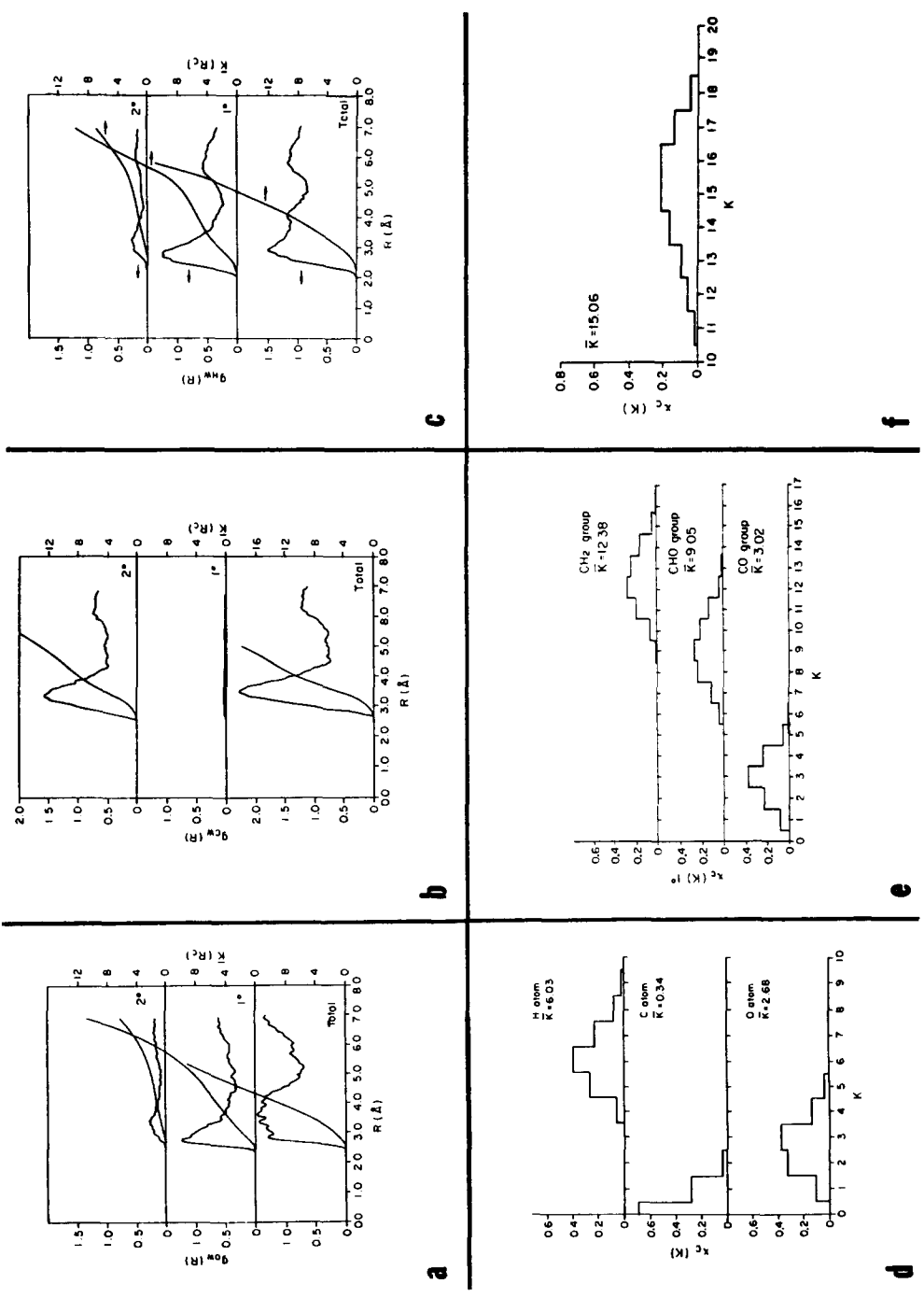
The generally irregular shape of the 1°, 2°, . . . regions around solute atoms brought up the question of normalization. The curves shown in Fig. 4 are normalized by the usual $4\pi R^2$, representing the volume element of a spherical shell. This has the advantage that the $g_{AW}^{(k)}(R)$'s satisfy a normalization condition similar to Eq. (20). However, the limit at $R \to \infty$ is not unity; therefore comparison of the $g_{AW}^{(k)}(R)$'s for atoms on different molecules is problematic. As an alternative, the volume element of the spherical shell in the Voronoi polyhedron associated with the primary region of the solute atom can be used instead of $4\pi R^2$:

$$g_{AW}^{(k)}(R) = N_A^{(k)}(R)/\rho V_A^{(k)}(R)\delta R \tag{26}$$

where $N_A^{(k)}(R)$ is the number of solvent molecules in the kth-order proximity region of solute atom A at a distance $r \in [R, R + \delta R]$ and $V_A^{(k)}(R)\delta R$ is the volume of the shell in the Voronoi polyhedra representing the kth-proximity region of solute atom A that is at a distance R from the solute atom A and is δR thick.[15] This normalization does account for the change in the shape of the Voronoi polyhedra as a function of distance and assures that at larger R, $g_{AW}^{(k)}(R)$ goes to unity. The present work computes $V_A^{(k)}(R)$ by a straightforward Monte Carlo procedure using $O(10^5)$ randomly generated points in the simulation cell.

The proximity criterion can also be used in conjunction with the statistical state solvation site or hydration shell analysis. In this case, density envelopes belonging to selected proximity regions are removed. This en-

[15] M. Mezei, P. K. Mehrotra, and D. L. Beveridge, *J. Biomol. Struct. Dyn.* **2**, 1 (1984).


a
b
c
d
e
f
Total
1°
2°
R (Å)
H atom
K̄ = 6.03
C atom
K̄ = 0.34
O atom
K̄ = 2.68
CH₂ group
K̄ = 12.38
CHO group
K̄ = 9.05
CO group
K̄ = 3.02
K̄ = 15.06
K

ables the user to separate the solvation of different atoms and provides a more comphrensible picture. The proximity criterion is also helpful in enhancing individually displayed configurations. Solvents solvating selected solute atoms or functional groups again can be removed or solvents can be color coded based on the proximity region to which they belong. The description of the solute effect on the solvent–solvent structure in the framework of the proximity criterion requires the separate study of the solvent–solvent interactions of waters in the various primary areas. First of all, solvent–solvent interactions are directly affected by the density fluctuations induced by the solute and characterized by the primary radial distribution functions. It is of interest, however, to see if there are any further effects. The quantities amenable to such study are the distribution of solvent–solvent pair properties (like the pair energy), the hydrogen-bonding angles θ_H and θ_{LP}, because these are independent of the density in the first order. Contributions from water pairs that lie in different primary areas contribute to both distributions; contributions from pairs that lie in the same primary area contribute with double weight. The pair properties studied previously in this laboratory are restricted to near-neighbor pairs in most of the cases. This restriction is not essential for the discussion above. It was introduced to eliminate the damping effect of the distant, noninteracting pairs. As a result, this study requires higher quality of convergence because the statistics are reduced when pair properties are examined for near-neighbor pairs only and due to the restriction of solvents to a given primary region. Also, the effects to be studied are quite small.

Results

The analysis of the solvation of a complex solute requires a decomposition of the calculated results. The proximity criterion allows the decomposition of the solvation environment of a solute into primary contributions from the different solute atom environments as well as the decomposition of the total environment into primary, secondary, etc. contributions for any solute atom. The normalization condition of Eq. (16) corresponds to the former and Eq. (20) to the latter. The results can then be used to (1) delineate the contributions of individual solute atoms and functional groups to the total picture (both structures and energetic), (2)

FIG. 4. Analysis of the Monte Carlo results on aqueous hydration of formaldehyde at 25°. (a) 1°, 2°, and total $g(R)$ around the O atom; (b) around the C atom; (c) around the H atom (averaged); (d) $x_C^{l^\circ}(K)$ for the O, C, and H atoms; (e) for the CH_2, CHO, and CO groups; (f) total $x_C(K)$.

aid us in obtaining a well-defined hydration shell without prior assumptions, and (3) extract characteristic features of atomic and functional group solvation environments that hold generally, i.e., independent of the molecule to which the atom or functional group is attached. In this section, examples will be provided for all the above.

The proximity criterion analysis has been applied in this laboratory to the analysis of Monte Carlo computer simulation of aqueous hydration of several prototype biomolecular solutes: formaldehyde,[2] glyoxal,[16] formamide,[17] benzene,[18] glycine zwitterion,[15] alanine dipeptide,[19,20] nucleic acid constituents,[21] and various amides.[22] In the following, we will draw examples from several of these with the understanding that full details of the analysis are available in the original papers. A review of some of the earlier results can be found in Ref. 14. Recent work by Rossky and co-workers[23] and by Jorgensen and co-workers[24] also employed the proximity criterion in the analysis of computer simulation results.

The decomposition of the total atomic radial distribution into primary and secondary atomic radial distribution for the formaldehyde molecule is shown on Fig. 4. The primary solvation of the atoms on the outside of the molecule show well-defined solvation structure, while the shielded carbon atom has essentially no primary solvation. The secondary solvation of the outside atoms is more diffuse, corresponding to the fact that the secondary solvation regions of those atoms are composed of several different contributions. The secondary solvation of the central carbon atom, however, is better defined.

The decomposition of the molecular solute–solvent radial distribution into primary atomic components is shown for the aqueous hydration of benzene. Figure 5 shows the solute–water center-of-mass radial distribution function and Fig. 6 the primary solute–atom–water radial distributions (averaged over equivalent atoms), using the individual volume–element normalization. A comparison of Figs. 5 and 6 shows that (1) the

[16] F. T. Marchese, P. K. Mehrotra, and D. L. Beveridge, *in* "Biophysics of Water" (F. Franks and S. Mathias, eds.). Wiley, New York, 1982.

[17] F. T. Marchese and D. L. Beveridge, *J. Phys. Chem.* **88,** 5692 (1984).

[18] G. Ravishanker, P. K. Mehrotra, M. Mezei, and D. L. Beveridge, *J. Am. Chem. Soc.* **106,** 4102 (1984).

[19] P. K. Mehrotra, M. Mezei, and D. L. Beveridge, *Int. J. Quantum Chem. Quantum Biol. Symp.* **11,** 301 (1984).

[20] M. Mezei, P. K. Mehrotra, and D. L. Beveridge, *J. Am. Chem. Soc.* **107,** 2239 (1985).

[21] D. L. Beveridge, P. V. Maye, B. Jayaram, G. Ravishanker, and M. Mezei, *J. Biomol. Struct. Dyn.* **2,** 261 (1984).

[22] G. Ravishanker, S. W. Harrison, R. Glacken, and D. L. Beveridge, to be published.

[23] R. A. Kuhapsky and P. J. Rossky, *J. Am. Chem. Soc.* **106,** 5786 (1984).

[24] C. J. Swenson and W. L. Jorgensen, *J. Am. Chem. Soc.* **107,** 569 (1985).

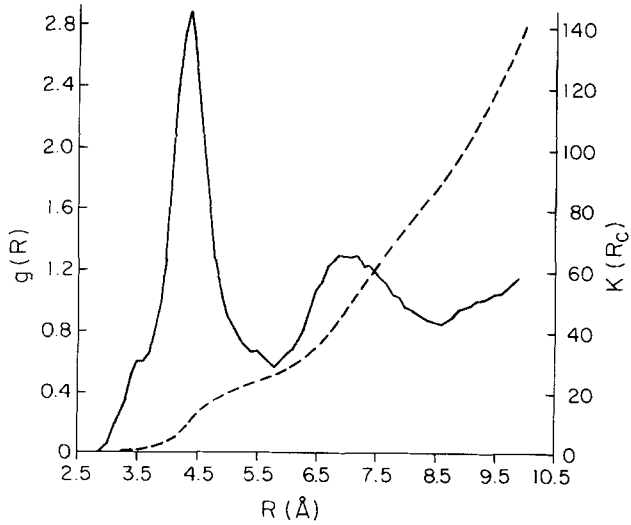


FIG. 5. Calculated center-of-mass solute–solvent $g(R)$ (full line) and the corresponding running coordination number (dashed line) for the aqueous hydration of benzene at 25°.

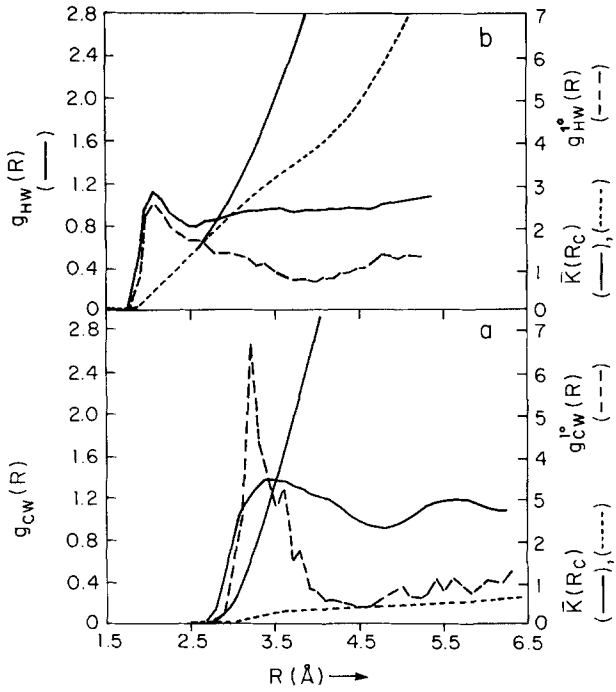


FIG. 6. Calculated total (——) and primary (----) solute–solvent radial distribution functions and the corresponding running coordination numbers for atoms in benzene, averaged over symmetry-equivalent atoms. (a) Carbon atom distribution; (b) hydrogen atom distribution.

sharp first peak in the molecular radial distribution function actually represents two peaks; (2) its sharpness, however, is due essentially to the hydrogen solvation; and (3) neither atomic primary radial distribution function shows a second peak, while the molecular RDF does. This latter implies that the first-solvation shell hydration complex presents itself as an essentially spherical entity.

Comparison of density and orientational correlations is shown for the methylene group of the glycine zwitterion. Figure 7 shows the primary atomic radial distribution functions on the hydrogen atoms and Fig. 8

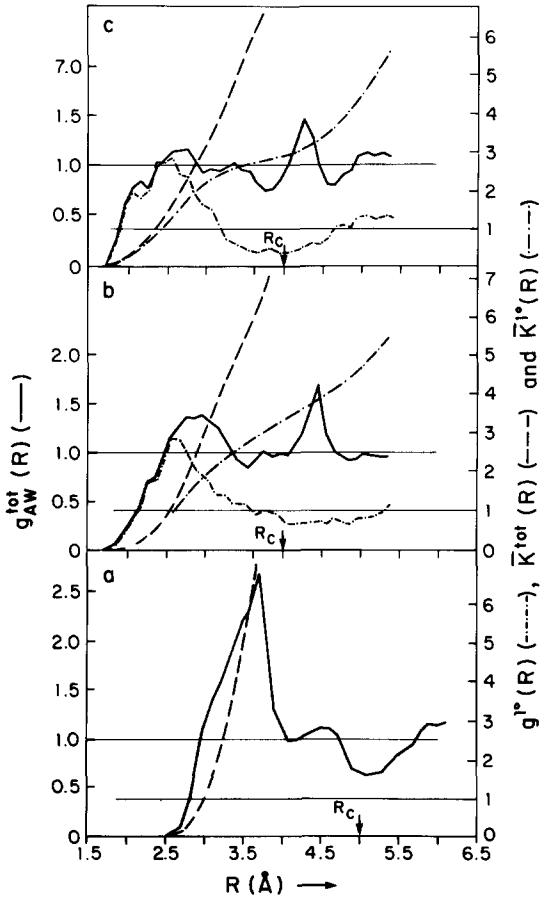


FIG. 7. Calculated atom–water radial distribution functions and running coordination numbers for the atoms in the CH_2 group of the glycine zwitterion. (a) $C(CH_2)$; (b) $H_2(CH_2)$; (c) $H_2(CH_2)$.

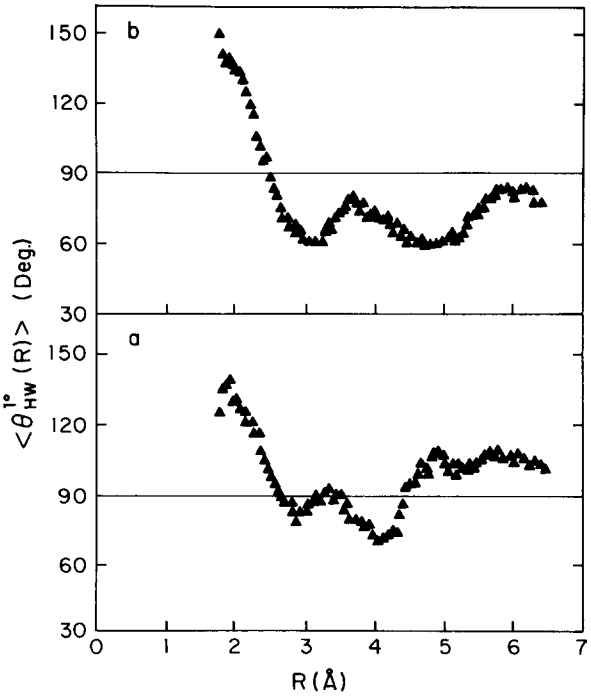


FIG. 8. Orientational correlation vs water–solute atom distance for the atoms in the CH_2 group in the glycine zwitterion. (a) $H(CH_2)$; (b) $H_2(CH_2)$.

shows the primary orientational correlation function, $\langle\theta^{1°}(R)\rangle$. It can be clearly seen that the orientational correlations have a longer range than the density correlations. This appears to be in accord with the commonly held notion that the aqueous hydration of a hydrophobic group is dominated by entropy effects.

The density envelopes of the statistical state solvation sites for glyoxal in aqueous solution are shown in Fig. 9a and its decomposition into primary contributions in Fig. 9b–d. It can immediately be seen that the proximity criterion allows us to see significant differences in the solvation of the different solute atoms that were not at all clear in the total molecular picture. It is particularly interesting to contrast the hydrogen and oxygen hydration showing a large degree of localized solvation for the oxygen atom and the opposite for the hydrogen atom.

Figures 10 and 11 show stereo views of a hydration complex from a Monte Carlo study of the aqueous hydration of the Ala dipeptide in the α_R conformation. In Fig. 10a, all waters are shown, while in Figs. 10b and

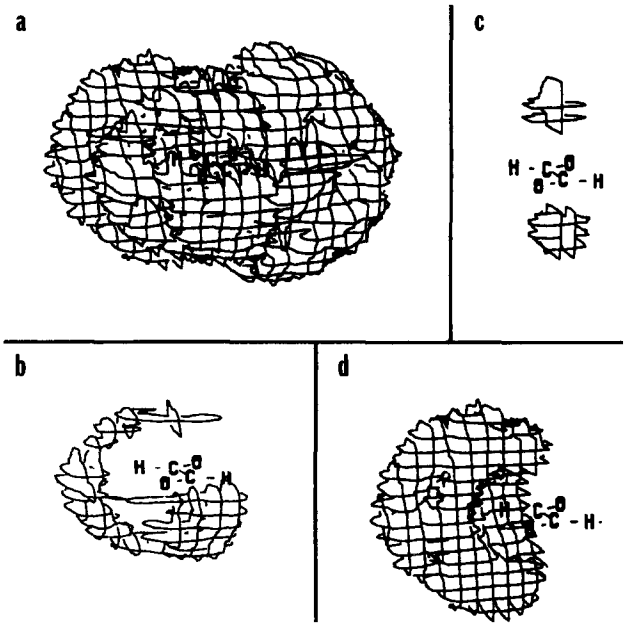


FIG. 9. Calculated solvent probability densities and statistical state solvation sites for *trans*-glyoxal in water. (a) Total density; (b) solvation of the oxygen atom; (c) solvation of the carbon atom; (d) solvation of the hydrogen atom.

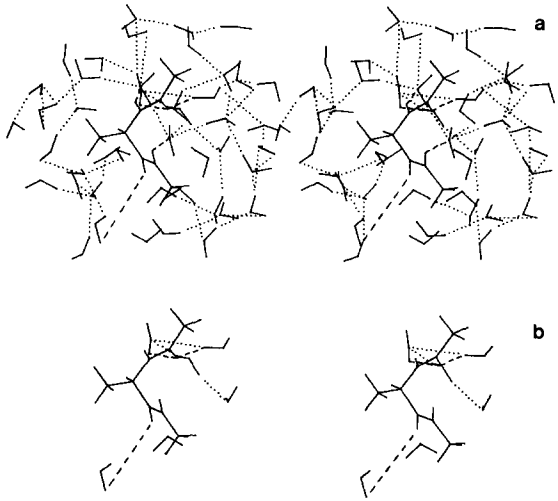


FIG. 10. Stereo view of a hydration complex around the Ala dipeptide in the α_R conformation. (a) All waters shown; (b) only waters proximal to the carbonyl groups are shown.

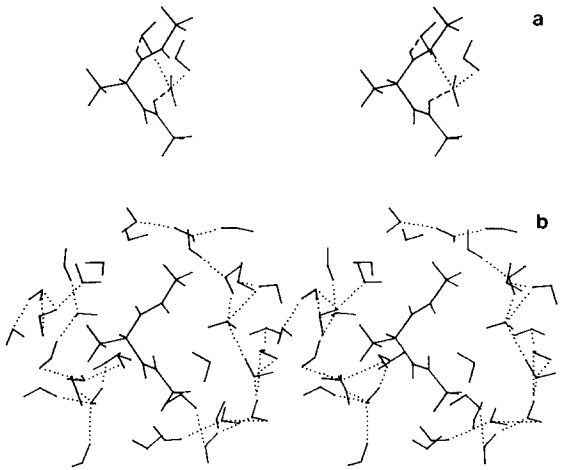


FIG. 11. Stereo view of a hydration complex of Fig. 10 showing only waters (a) proximal to the amine groups; (b) proximal to hydrophobic hydrogens.

11a and b only waters solvating primarily the methyl, amine, and carbonyl groups, respectively. The power of the proximity criterion to delineate the various solvation regions is thus demonstrated.

As mentioned earlier, the proximity criterion can also be used to define solvation shells around a solute without prior assumption of the solvation process itself. Using the primary radial distribution function for each solute atom, its minimum after the first peak gives a natural radius for the first solvation shell of that atom. Once this value is fixed, the coordination number QCDFs for each solute atom can be determined and average coordination numbers computed. This is illustrated of Fig. 11, containing also the total and primary QCDF of coordination numbers. It was one of the first illustrations of the contrasting behavior of the hydrophobic and hydrophilic groups, mentioned above about the glyoxal solvation. It was particularly interesting at that time, since earlier estimates based on solute–solvent energetics (solvation site models) did not describe the extensive first-shell solvation of the hydrophobic groups.

The transferability of the solute atom coordination numbers determined by the proximity criterion was recently examined on the nucleic acid constituents.[21] It was found that the atomic coordination numbers varied considerably both within atoms and within molecules. More consistent results were obtained when the functional group coordination numbers were considered. For the correct comparison, however, one has to consider the volume of the first coordination shell, since it varies from molecule to molecule. In Table I, we present the average functional group

TABLE I
COMPARISON OF FUNCTIONAL GROUP COORDINATION NUMBERS[a]

Functional group	$\langle K \rangle^b$	$\bar{K}_{min}$	$\bar{K}_{max}$	$\bar{K}_{max}/\bar{K}_{min}$	$\bar{K}'_{min}$	$\bar{K}'_{max}$	$\bar{K}'_{max}/\bar{K}'_{min}$
$-CH_3$	9.10	8.52	9.74	1.14	8.52	9.89	1.16
$-CH_2$	3.91	3.91	3.91	1.00	3.91	3.91	1.00
$\rangle CH$	1.65	0.97	1.95	2.01	1.56	1.86	1.19
$\rangle\!\!\rangle CH$	4.26	3.37	5.48	1.63	2.54	7.70	3.03
$-O-$	0.80	0.68	0.98	1.44	0.69	1.12	1.62
$-OH$	1.98	1.67	2.20	1.32	1.61	2.23	1.36
$>CO$	3.52	2.33	4.56	1.96	2.50	4.60	1.84
$-NH_2$	3.54	2.12	5.40	2.55	3.18	3.99	1.23
$>NH$	2.84	1.27	3.97	3.13	1.59	5.12	3.22
$=N-$	1.87	1.23	2.32	1.89	1.32	2.39	1.81

[a] The subscripts min and max refer to the smallest and largest value found for the functional group.
[b] $\langle\ \rangle$ represents average overall occurrences of the functional group in the systems studied.

coordination number K and the volume corrected $\bar{K}' = \bar{K}^*(V/\langle V \rangle)$ where V and $\langle V \rangle$ are the actual and average functional group first solvation shell volume. The results show good transferability for functional groups in the sp^3 hybridization state, but no transferability at all for the groups containing π bond; the data set, however, is quite small at this point.

There are various options in choosing the first solvation shell radius. The option used in the studies performed in our laboratory determined R_c for each solute atom from its respective primary radial distribution function. Therefore, comparisons between different molecules are complicated by the fact that the cutoffs used may be different. To eliminate this problem, Jorgensen suggested that cutoff values should be set for different solute atoms uniformly.[24] In our experience, however, the position of the first minimum in the primary radial distribution functions varies somewhat with the environment. Table II shows the range of values acceptable for first minimum position for the atoms in the Ala dipeptide in the C_5 conformation. It can be clearly seen that the terminal methyl groups have significantly different first solvation shell radii than the middle one. Furthermore, the volume correction introduced in Ref. 15 should also factor out the differences in the cutoff values.

[25] A. H. Narten and H. A. Levy, *J. Chem. Phys.* **55**, 2263 (1971).

TABLE II
COMPARISON OF THE PERMISSIBLE FIRST SOLVATION SHELL RADII FOR THE ATOMS IN THE Ala DIPEPTIDE IN THE C_5 CONFORMATION

Atom[a]	Range (Å)
H(C)	4.2–5.0
H(C)	3.7–4.6
H(C)	4.3–4.9
O(C)	3.5–3.9
H(N)	2.2–2.4
H(C)	3.7–4.1
H(C)	3.6–4.0
H(C)	3.5–4.1
H(C)	3.5–3.9
O(C)	3.1–3.3
H(N)	2.1–2.3
H(C)	4.0–4.7
H(C)	5.1–5.6
H(C)	4.2–5.0

[a] Atoms in parentheses indicate the functional group.

The distribution of the binding energies has been considered earlier as an important cue to decide the validity of mixture models or continuum models. Clearly, unimodal binding energy QCDFs were obtained for all water models studied.[8] This trend continued for the primary binding energy QCDFs obtained for different solutes.[16,18–22] Therefore, the consideration of the average primary binding energies is essentially adequate. The partition of the solute–solvent binding energy into contribution from solvent in the various primary solvation regions of the different functional groups is shown in Table III for the Ala dipeptide in four different conformations. It shows that the energetic differences between the various conformations are the consequence of the differences in the carbonyl group hydration.

The QCDF of the pair energy is generally a more sensitive indicator than the QCDF of binding energy. Figure 12 shows the QCDF of the solute–solvent pair energy for the glycine zwitterion, a distribution with several peaks. The decomposition of the distribution by the proximity criterion into primary functional group contributions, shown on Fig. 13, permits us to identify the functional groups contributing to the different peaks. It is interesting that the pair-energy QCDF for the methylene group

TABLE III
CALCULATED SOLUTE–WATER BINDING ENERGIES FOR THE C_7, C_5, α_R, AND P_{II} CONFORMATIONS OF AcAlaNHMe, RELATIVE TO C_7

Functional group	C_7	C_5	α_R	P_{II}
$-CH_3$	0	4.7	0.5	1.3
$>CO$	0	−1.1	−8.2	−5.9
$>NH$	0	−2.0	2.8	−2.2

is also bimodal, indicating that the primary solvation region of the methylene group is not homogeneous. The accompanying $\bar{K}(\varepsilon)$ curves give information about the number of solvent molecules involved with the different peaks in $X_P(\varepsilon)$.

The solute effect on the solvent structure has also been the subject of several studies. For example, we demonstrated earlier that the aqueous solvation environment of the methane is more structured than bulk water, while several simple ions were shown to decrease water–water structure.[9] Similar studies were also performed in the aqueous solvation study of the glycine zwitterion.[15] The average water–water pair energy was found to be −2.84, −2.98, and −2.87 kcal/mol for the NH_3^+, CH_2, and COO^- groups. Since the corresponding value in liquid water is −3.03 kcal/mol,

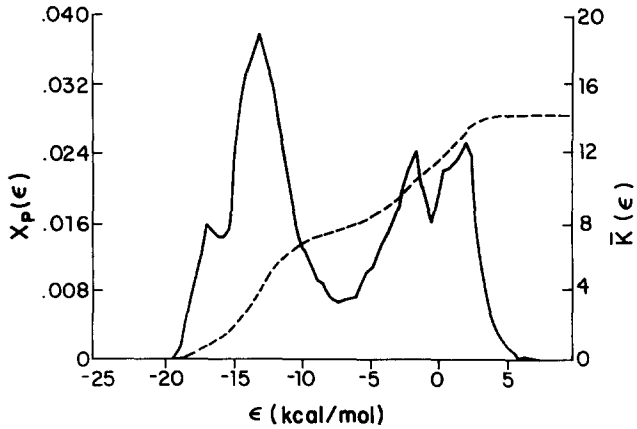


FIG. 12. Calculated $X_P(\varepsilon)$ (——) and $\bar{K}(\varepsilon)$ (---) for the glycine zwitterion. $\bar{\varepsilon}$ = −7.4 kcal/mol.

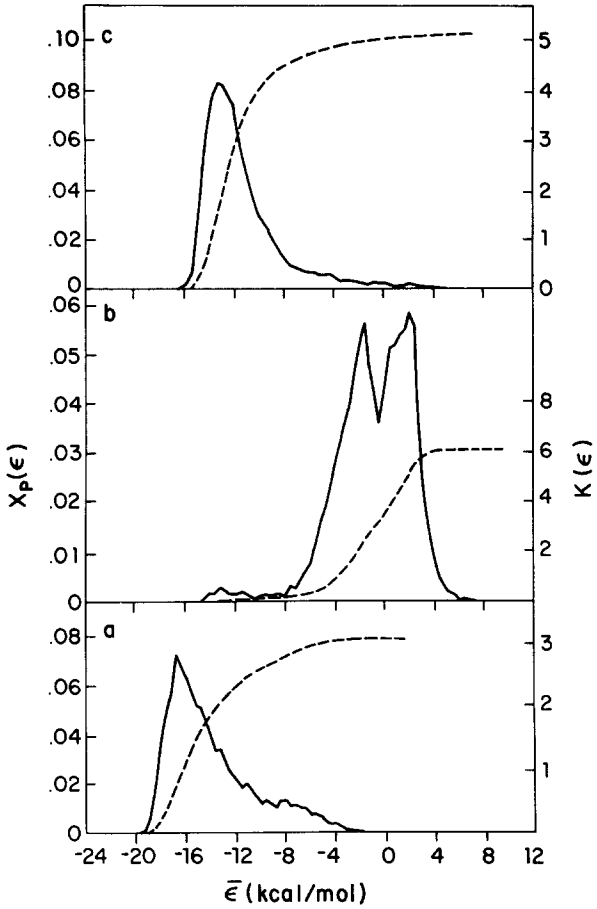


FIG. 13. Calculated $x_P(\varepsilon)$ (——) and $\bar{K}(\varepsilon)$ (---) for the functional groups of the glycine zwitterion. (a) NH_3^+ group, $\bar{\varepsilon} = -13.9$ kcal/mol; (b) CH_2 group, $\bar{\varepsilon} = -0.8$ kcal/mol; (c) COO^- group, $\bar{\varepsilon} = -11.3$ kcal/mol.

this clearly shows loss of structure for the ionic waters, but no loss for the methylene waters. Studies of the hydrogen-bond QCDFs showed smaller differences that did not exceed the stated error bounds, but were supportive of the above conclusion.

Acknowledgments

This research was supported by NIH grant GM 24914, NSF grant CHE-8203501, and a CUNY Faculty Research Award.