Grand Canonical Ensemble Monte Carlo Simulation of the dCpG/Proflavine Crystal Hydrate

H. Resat and M. Mezei
Department of Physiology and Biophysics, Mount Sinai School of Medicine, New York, New York 10029-6574 USA

ABSTRACT The grand canonical ensemble Monte Carlo molecular simulation method is used to investigate hydration patterns in the crystal hydrate structure of the dCpG/proflavine intercalated complex. The objective of this study is to show by example that the recently advocated grand canonical ensemble simulation is a computationally efficient method for determining the positions of the hydrating water molecules in protein and nucleic acid structures. A detailed molecular simulation convergence analysis and an analogous comparison of the theoretical results with experiments clearly show that the grand ensemble simulations can be far more advantageous than the comparable canonical ensemble simulations.

INTRODUCTION

Studies of the interactions of water molecules with biomolecules and of the role of water in biological activities have been an active area of research for numerous years. The internal and the first solvation shell waters constitute integral parts of proteins and nucleic acids, and it has been well established that solvation effects can have a significant influence on properties of biomolecules such as their structure, function, and dynamics. Internal waters are thought to be instrumental in structural stabilization and may play a role in the folding of proteins (Sekharudu and Sundaralingam, 1993), or they may be involved in the binding of ligands at the catalytic site (Dewar and Storch, 1985; Warshel et al., 1989a,b). It has also been proposed that internal waters may mediate electron transfer through the protein–water hydrogen bonds (Nar et al., 1991). Similarly, bridging water molecules may facilitate protein–protein (Janin and Chothia, 1990; Bhat et al., 1994; Ben-Naim, 1991) and protein–nucleic acid (Harrison and Aggarwal, 1990; Steitz, 1990) complex formation. A broad overview of the properties of water and of its role and function in biological systems can be found in the book edited by Westhof (1993).

In crystal structures the mobility of atoms can be estimated from temperature factors, also called *B* factors. Solvent–biomolecule interactions are generally weaker than the covalent intramolecular interactions of biomolecules. Therefore, in the absence of additional geometrical constraints, such weak interactions can put only a limited amount of restraint on the movement of the water molecules. This, combined with water's high mobility, even in the internally bound sites, results in temperature factors that are in general larger than those of the biomolecule atoms. Such large temperature factors complicate the detection of water molecules in x-ray or neutron diffraction studies (Karplus and Faerman, 1994). NMR spectroscopy seems to be more suitable for the study of properties of disordered waters if they are long lived (Otting et al., 1991). However, NMR techniques also have their deficiencies. Current methods have a sensitivity of ~50 ps and require waters to reside next to protons if they are to be detected. This limits the ability of these techniques to determine the positions of waters next to some carbonyl or carboxylate groups. It should be noted that diffraction and NMR techniques are in some ways complementary: they can be used in conjuction to detect better the strongly hydrogen bonded waters and the waters around the hydrophobic methyl groups (Levitt and Park, 1993; Karplus and Faerman, 1994).

For the above reasons, perhaps the most challenging part of the structure refinement process is the determination of the locations and the number of solvating water molecules. In this regard, computational studies can supplement experimental research. The major problem in theoretical approaches is the inefficient sampling of the studied quantities in a reasonable computation time. To overcome some of the statistical sampling deficiencies, we recently demonstrated the usefulness of the grand canonical ensemble in molecular simulations (Resat and Mezei, 1994). The present paper gives a more detailed account of the cavity-biased grand canonical Monte Carlo technique that is applied to the hydrated crystal structure of dCpG strand intercalated with the drug proflavine (Shieh et al., 1980). The 2:2 complex of dCpG/proflavine (Fig. 1) was chosen for this investigation because of its high-resolution crystal structure and the absence of potentially mobile counterions. The crystal structure shows a highly organized water network (Fig. 2); the minor groove waters form a flat polygonal disk, and the major groove waters are ordered as an array of edge-linked pentagonal disks (Neidle et al., 1980). Initial analysis predicted that there would be 100 waters in the crystal unit cell formed by four crystal-symmetry related asymmetric units (Shieh et al., 1980). Later, more-accurate density measure-

Received for publication 19 January 1996 and in final form 22 May 1996.

Address reprint requests to Dr. Mihaly Mezei, Department of Physiology and Biophysics, Mount Sinai School of Medicine, One Gustave L. Levy Place, New York, NY 10029. Tel.: 212-241-2186; Fax: 212-860-3369; E-mail: mezei@msvax.mssm.edu.

Dr. Resat's present address is Department of Physics, Koç University, Istinye-Istanbul 80860, Turkey.

0006-3495/96/09/1179/12 $2.00

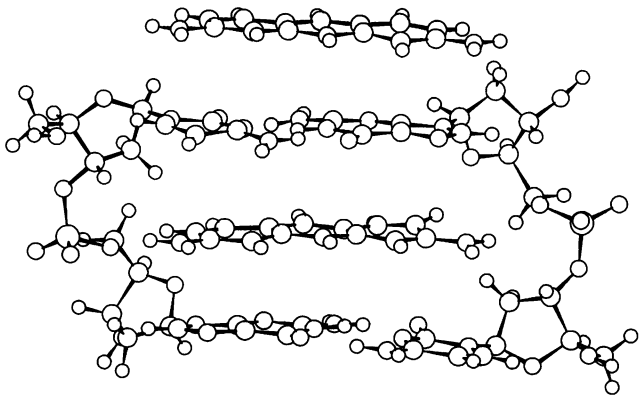

FIGURE 1 Ball-and-stick diagram of the dCpG/proflavine strand.

ments showed that the unit cell more likely contains 108 waters (Mezei et al., 1983). Moreover, subsequent diffraction studies at lower temperatures, −2 and −130°C, by Berman and co-workers indicated that the number of waters in the unit cell may be 120 or greater (Berman, 1994; Schneider et al., 1992). As stated by Schneider et al. (1992), the "new" waters observed in the lower-temperature structures are most likely present at room temperature but are disordered and cannot be detected on electron density maps.

The dCpG/proflavine complex has been the subject of several theoretical investigations by various groups. Mezei et al. (1983) reported a Monte Carlo (MC) investigation of the generic solvent site analysis (Mezei and Beveridge, 1984) in which almost two thirds of the experimental hydration sites were successfully reproduced. Kim et al. (1983) and Kim and Clementi (1985a,b) studied the same system again, using MC methods. To reduce the cutoff effects due to the small simulation cell size, they replicated the unit cell to form a simulation cell consisting of 12 unit cells. They also systematically increased the number of waters in the unit cell and predicted that the optimum number of waters in the crystal would be 122–132 per unit cell. The solvent density with their prediction is considerably larger than the early experimental density measurement but is closer to later predictions (Schneider et al., 1992). Swaminathan et al. (1990) and Herzyk et al. (1991) investigated the dynamical aspects of the dCpG/proflavine crystal structure. Both of these molecular dynamics (MD) studies used 108 waters per crystal cell. These studies showed that the overall rms fluctuations in the biomolecule atom positions are rather small, with sugar and phosphate groups having relatively larger mobilities as expected. Swaminathan et al. observed a systematic drift in sugar puckering. The water molecules not observed in experiments exhibited a much larger temperature factor relative to those experimentally observed, thus explaining the deficiencies in the diffraction study. Although the water network shows considerable flexibility, the pentagonal disk network and the polygon disk stayed intact during the simulation, with the latter exhibiting a bimodal behavior. Although the results of Herzyk et al. (1991) are similar to those of Swaminathan et al. (1990), Herzyk et al. observed distorted structures when the waters were not restrained close to their experimental positions. All these studies point out that molecular simulations are capable of reproducing the experimental results to a good degree. However, as discussed below, all these canonical ensemble studies might have been biased toward the initial configuration.

A close look at Fig. 2 reveals that the minor groove waters are "disconnected" from the major groove waters except via a link that is approximately one water molecule wide. Because of the high packing density of the system, the linkage waters, particularly OW8 (Fig. 3), would block the water exchange between the minor and major grooves. This effect was referred to as the "enclosed cavity effect" in our previous publication (Resat and Mezei, 1994). To eliminate such enclosed cavity effects, and to partially overcome the uncertainty in the density measurements, we advocated the simulation methods based on the grand canonical ensemble. Use of the grand canonical ensemble improves the statistical mechanical sampling rates in determining the solvation properties of crystal hydrates in two major ways: 1) The density measurements of the crystals are not accurate

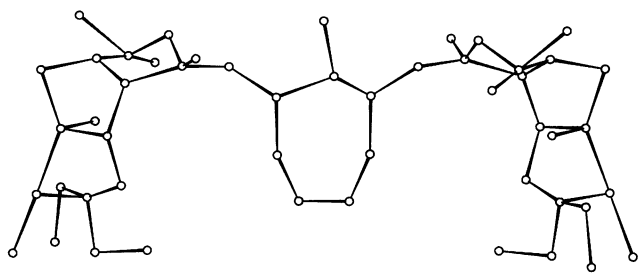

FIGURE 2 Water network in the dCpG/proflavine crystal hydrate as observed by Neidle et al. (1980). Note that the bonds are not covalent bonds and are drawn to show the network.

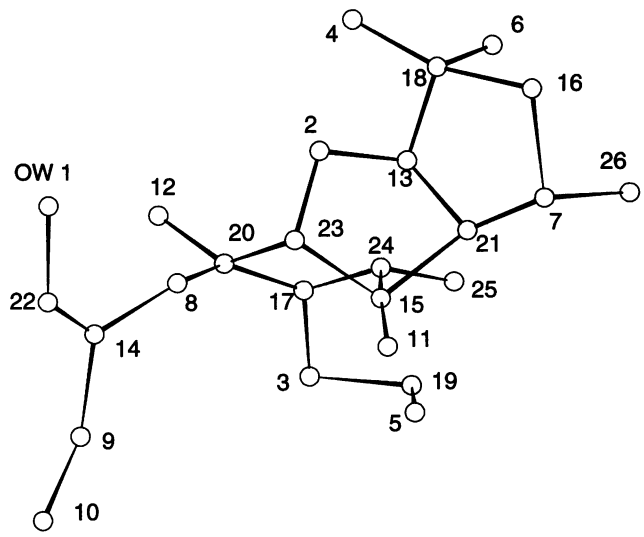

FIGURE 3 Water network numbering scheme for the experimentally determined water molecules (Shieh et al., 1980) as used in Tables 1 and 2.

enough to permit us to determine exactly the total number of water molecules inside the unit cell. The grand canonical ensemble approaches partially solve this problem by allowing the number of water molecules to fluctuate. The statistical distribution profile and information on the energetics and the location of added and removed solvent molecules can help in the determination of the probable number densities. 2) The water molecules of the crystal hydrates are generally enclosed in unconnected water pockets. An appropriate simulation technique has to allow for molecule exchange between the enclosed cavities. Conventional simulation methods based on canonical and microcanonical ensembles fix both the total number of water molecules and their distribution among the pockets from the start, and molecule exchange between the enclosed cavities is virtually impossible. This results in a strong bias toward the initial configuration and the number distribution among the cavities. In canonical ensemble MC methods the particle exchange problem might be solved by use of very large moves, but in such attempts the move acceptance rates become extremely low, which results in a very poor statistical sampling. Unless the biomolecule has "breathing modes," i.e., large fluctuating modes to allow for water exchange between different pockets, bias toward the initial configuration would be particularly strong in the canonical ensemble MD studies.

The advantages of the grand canonical ensemble simulation techniques were initially demonstrated by a study of the crystal hydrate polydisaccharide hyaluronic acid performed with the cavity-biased grand canonical Monte Carlo (CB-GCMC) method (Resat and Mezei, 1994). Here we apply the CB-GCMC method to investigate the hydration structure in the dCpG/proflavine system and compare our results with experiments and with the earlier canonical ensemble studies. Our results indicate that (see Results) there may be fewer waters in the minor groove and that they are quite delocalized. This raises some questions on the experimental and the previous theoretical studies cited above.

In the report that follows we first give a brief description of the CB-GCMC method, describe the molecular simulation setup, and then present the results. The last section summarizes our findings and discusses future directions.

THEORY

In statistical mechanical treatments, such as molecular simulations, the aim is to calculate the ensemble averages $\{\langle\mathcal{O}\rangle\}$ of the desired quantities $\{\mathcal{O}\}$. For example, in the canonical ensemble such ensemble averages for an N-particle system interacting with potential U_N are given (with $d\Gamma = d^{3N}r$ and $\beta = 1/kT$) as

$$\langle\mathcal{O}\rangle = \frac{\int d\Gamma\, \mathcal{O}e^{-\beta U_N}}{\int d\Gamma\, e^{-\beta U_N}} \equiv \frac{1}{Z_N}\int d\Gamma\, \mathcal{O}e^{-\beta U_N}, \qquad (1)$$

where Z_N is the configuration integral. In contrast, the grand canonical ensemble allows for number fluctuations and the ensemble averages are given (for a single-component system) as

$$\langle\mathcal{O}\rangle = \frac{1}{\Xi}\sum_N^\infty \frac{z^N}{N!}Z_N\langle\mathcal{O}\rangle_N. \qquad (2a)$$

In Eq. 2a the grand ensemble partition function Ξ is given as

$$\Xi(\mu,V,T) = \sum_N^\infty \frac{z^N}{N!}Z_N, \qquad (2b)$$

where, with Λ denoting the thermal de Broglie wavelength, $z = e^{\beta\mu}/\Lambda^3$ is the fugacity (activity) function. As Eqs. 2 show, in essence a grand canonical simulation is equivalent to a set of appropriately weighted canonical ensemble simulations. Because of this similarity, the grand canonical ensemble and bicanonical ensemble simulation methods were developed mainly by generalizing the existing canonical simulation methods (Adams, 1974, 1975; Cagin and Pettitt, 1991; Panagiotopoulos, 1992; Beutler and van Gunsteren, 1994; Swope and Anderson, 1995; Resat et al., 1996). A formal derivation of the above equations can be found in the book by Friedman (1985).

In this study, we follow Adams's approach (1974, 1975) to grand canonical ensemble Monte Carlo (GCMC) simulations. Recasting the fugacity in terms of the chemical potential of an ideal gas of particles of the same mass and the same average number of molecules $\bar{N}$, volume V, and temperature, we can express the grand ensemble partition function at constant volume and temperature as

$$\Xi = \sum_N^\infty \frac{1}{N!}e^{NB}\int V^{-N}\, d\Gamma e^{-\beta U_N}. \qquad (3)$$

The B parameter is defined as

$$B = \beta_\mu + \ln(V/\Lambda^3), \qquad (4a)$$

with

$$\beta\mu_e = B - \ln\bar{N}, \qquad (4b)$$

where μ_e is the excess chemical potential [over $\beta^{-1}\ln(n\Lambda^3)$, the chemical potential of an ideal gas with number density $n = \bar{N}/V$]. Noting the similarity to the canonical ensemble simulations, Adams developed a GCMC simulation scheme in which the move attempts used in canonical ensemble simulations to generate a Markov chain to sample the phase space are replaced with two types of move: 1) regular displacement moves as in the canonical ensemble and 2) insertion–deletion moves to allow for fluctuations in the number of molecules. There is no rigorous rule for combining these two move attempts, and in this study we use a 1:1 ratio; i.e., every regular move is followed by an insertion–deletion attempt.

As Eq. 4a shows, the B parameter and the chemical potential differ by a constant, and therefore a constant μ

ensemble is equivalent to using a constant B parameter in GCMC simulations. In implementing the GCMC scheme (Resat et al., 1996), the B parameter is adjusted at the beginning until the targeted average number of molecules is approximately achieved. After fine tuning, the B parameter is kept constant during the data acquisition and the average number of molecules is calculated in the same simulation as well. Then the chemical potential can be calculated at the end by use of the relation among the excess chemical potential, the B parameter, and the average number of molecules (Eq. 4b). In generalizing the above equations to multispecies cases, one defines a species-dependent chemical potential (or the corresponding B parameter). The number of every species can fluctuate during the simulations. However, in our simulation we keep the dCpG/proflavine complex fixed in the unit cell; i.e., the nucleic acid or the proflavine is not allowed to be added or deleted. Since some of the components are kept at the same number density, the utilized ensemble is not grand canonical in the purely theoretical sense. However, the label grand canonical ensemble is still appropriate for the following reasons: Because of the lack of free space, addition of a second biomolecule is not possible. This leaves the "zero biomolecule state" (i.e., the deletion of the existing molecule) as the only other possible state. Owing to the very favorable solvation effects, thermochemical partitioning between these two states tells us that the statistical sampling rate of the zero biomolecule state should be approximately vanishing. Therefore, the neglect of the zero biomolecule state from the statistical sampling would introduce only a very small and unimportant error into the calculations.

CALCULATIONS

The object of this study is to determine what are the likely locations for the solvating waters in the dCpG/proflavine crystal hydrate by using the CB-GCMC method. The cavity-biased formulation of GCMC was used because, at high densities, it considerably improves the statistical sampling efficiency (Mezei, 1980, 1987). It was observed in the earlier MD studies that the sugar rings and phosphate groups may undergo conformational changes (Swaminathan et al., 1990; Herzyk et al., 1991). Such conformational changes may be the result of favorable solvation effects, or they may be due to the artifacts of the utilized interaction potential parameters. Our aim here is to determine the water molecule locations and compare them with experimental results. A direct comparison with experiments will be more meaningful if the solute molecule is kept in its crystal conformational state. Therefore, to eliminate the effects of the solute movements, we use a rigid solute molecule in which the solute atoms are fixed in their experimental crystal structures throughout the simulation run and only the water molecules are allowed to move. This approach is in line with the earlier MC simulation studies mentioned in the previous section and with the potential of mean force expansion approach of Hummer et al. (1995).

The dCpG/proflavine crystal has a $P2_12_12_1$ symmetry, and the rectangular unit cell with dimensions 32.991 Å × 21.995 Å × 13.509 Å is formed by four symmetry-related asymmetric subunits. The simulation cell was set equal to the crystal unit cell; thus it consisted of four 2:2 dCpG/proflavine complexes plus the waters. Periodic boundary conditions were applied. Interaction parameters of nucleic acid bases, sugar, and phosphate groups were represented by the AMBER force field (Weiner et al., 1984). The short-range parameters for proflavine were taken from the AMBER force field, and proflavine site charges were derived by fitting to the electrostatic potential and were kindly provided to us by P. A. Kollman (personal communication). Proflavine site charges used in our simulations are shown in Fig. 4. The water model chosen was the TIP3P model (Jorgensen et al., 1983), and solute–solvent interaction parameters were calculated by use of the geometric mean rule for both σ and ϵ. The temperature was 300 K. Solute–water interactions were treated with the minimum image boundary condition, and the water–water interactions were truncated with a spherical cutoff at 6.75 Å. After sufficient equilibration the simulation was run for 28×10^6 steps, with a 1:1 ratio of displacement to insertion–deletion attempts; i.e., there were 28 million attempts each of displacements and insertions–deletions. Solvent molecule displacement sizes were chosen to yield an ~50% acceptance rate. The cavity bias technique (Mezei, 1980, 1987) enabled us to obtain an acceptance rate of 1.3×10^{-3} for the insertion–deletion attempts. The configurations at every 2000 MC steps were

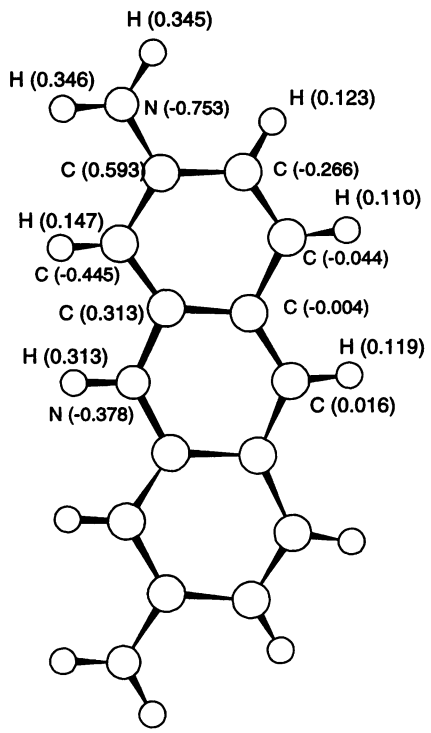

FIGURE 4 Proflavine molecule and its site charges used in the computations.

recorded to be used in the hydration analysis; a total of 14,000 configurations were used in the results analysis. The water chemical potential was adjusted to be −8.2 kcal/mol, so the unit cell contained 108 waters on average. Solvation chemical potential of the TIP3P model bulk water was calculated by Beglov and Roux (1994) as −6.4 ± 0.5 kcal/mol. Comparison shows, as expected, that because of the favorable solute–solvent interactions the water solvation free energy in the dCpG/proflavine crystal hydrate is considerably lower than the bulk water chemical potential.

ANALYSIS

We analyzed simulation results by using three complementary methods to determine the hydration properties of the crystal waters. Different ways of analysis gave supporting results. First, the water density was calculated on a Cartesian grid. In this approach one determines the singlet water density function on a uniform grid, and then the grid points with densities larger than a certain cutoff are reported as the most likely locations for the hydrating waters. The second approach utilized the generic solvent site (GSS) idea (Mezei and Beveridge, 1984). In the GSS approach one overlaps the subsequent configurations to determine the GSSs for the likely water locations. In the GSS approach the waters do not carry labels; therefore the molecule exchanges between GSSs are allowed during the molecule assignment to the sites. We determined GSSs by assigning the molecules by the graph theoretical Hungarian method (Berge, 1962), which is an efficient way of solving the optimal minimization problem.

The third analysis method was a hybrid approach between the connected-cluster hydration method of Lounnas and Pettitt (1994a) and the Hungarian method. In a series of papers, Lounnas and Pettitt and their co-workers (Lounnas et al., 1992, 1994; Lounnas and Pettitt 1994a,b) developed an elegant analysis method to study protein–solvent interfaces. In their method the local densities initially calculated on a grid are iteratively density-weight averaged with the nearby sites. The averaging radius is set to a small value at the beginning and then increased in small increments until convergence is obtained (Lounnas and Pettitt, 1994a). Then the most probable hydration sites are given by the local maxima of the density distribution. Lounnas and co-workers applied the method to study the hydration patterns around myoglobin in solution phase and showed that the hydration pattern is much less organized than what is seen in crystallography experiments. For our case, however, when the averaging radius was increased to values of more than 1.2 Å, some of the "well-" converged sites had an occupancy of more than 1. Such sites were in the major groove, where some of the waters are well defined and the density distributions for the other waters are somewhat diffuse. Thus, during the density-weight averaging process some of the density near a diffuse site would also be incorporated into the well-defined nearby site that had an occupancy of 1. Therefore, our connected-cluster analysis had to be terminated at a somewhat small weighted density averaging radius, 1.2 Å, to avoid having sites with occupancies larger than unity. The sites found by the Lounnas-Pettitt method, a total of distinct 250 sites, were used as the starting points in a subsequent GSS-type calculation. The Hungarian method was used to assign the site occupancies as in a GSS calculation. The sites with very low occupancies were eliminated in several steps until a total of 116 sites were left (corresponding to 29 distinct waters in each asymmetric unit cell). In this regard, the second and the third analysis methods are closely related, and indeed the results were almost identical.

We discuss the results communicated in the next section by separating the distinct crystal waters into three categories: 1) The waters of the minor groove, which form a heptagonal ring. With the labeling of Shieh et al. (1980), water OW22 occupies the bottom corner (Figs. 2 and 3) and the symmetry-related edges of the heptagon are formed by waters OW9, 10, and 14. Also, the half-occupancy water OW1 resides next to OW22, forming a tail. Note that OW22 is also a half-occupancy water and may reside in one of the two possible locations. 2) The waters that link the minor and major groove waters, OW8, 12, and 20. 3) The major groove waters, which form a pentagonal water network.

RESULTS

Minor groove hydration

Fig. 5 shows the results of the grid calculation for the minor groove hydration pattern. In the figure the yellow spheres represent the grid points where the water density is larger than the uniform solvent density. The radii of the spheres that mark the grid points are proportional to the magnitude of the calculated water densities: A larger radius corresponds to a higher water density. As Fig. 5 reveals, the GCMC simulation reproduces the experimental minor groove water locations very well. Notice that the edges are actually formed by continuous high-density regions. Similar continuous high-density regions were observed by Hummer et al. (1995), and they were interpreted as an indication of structural flexibility in the water network. Flexibility of the heptagonal ring was actually observed in MD simulations (Swaminathan et al., 1990). Our calculations show that the high-density region is quite tubular in shape, with a diameter approximately the size of that of a water molecule. Thus, because of this geometrical constraint, the movement of heptagonal network waters has to be a concerted motion.

Subsequent analysis using GSS and the connected-cluster hydration method (see Analysis) established that it would be possible to assign three to six sites as the edge of the ring (six is the experimental value). The most likely locations of water oxygens as determined by GSS analysis of the GCMC trajectories were overlapped with experimentally detected locations; the results are reported in Table 1. Agreement for the tail part of the polygon disk is very good, within 0.17 Å, and the sites of the polygon edges are reproduced within

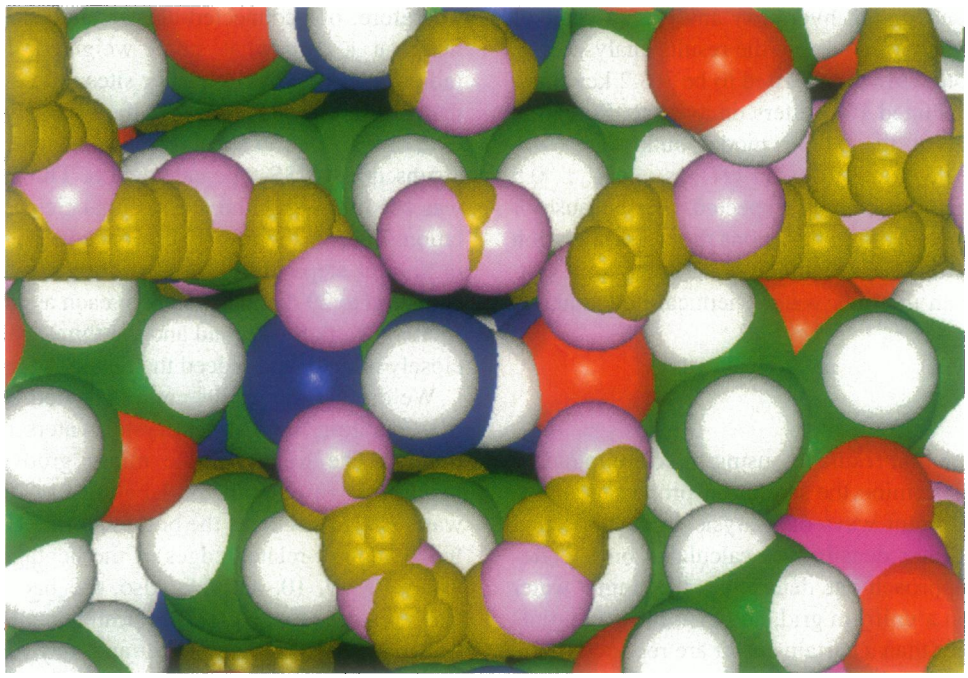

FIGURE 5 Grid density calculations for the minor groove and the link region (see Table 1). Green, blue, and white show the DNA and proflavine carbons, nitrogens, and hydrogens, respectively. Magenta and red, respectively, show the phosphate and oxygens of the phosphate group. Pink spheres represent the experimentally determined water locations. Two overlapping spheres at the bottom of the ring show the two possible locations for the half-occupancy water molecule OW22. Shown as yellow are the grid points where the water oxygen density is larger than the uniform solvent density. The radii of the spheres are proportional to the magnitude of the calculated water density. For clarity, the radii of spheres marking the grid points are kept small; they have a range of 0.40–0.55 times the van der Waals radius of oxygen. For the same reason, the experimental water spheres have a radius equal to 0.80 times the oxygen van der Waals radius.

1.22 Å. Although six sites would provide a better representation of the tubular high-density region, average density calculations showed that edges of the heptagon are actually formed by approximately three waters. To quote from Table 1, the average densities for waters OW9, 10, and 14 are 0.64, 0.32, and 0.58, respectively. These densities sum to 3.08 for the edges of the heptagon. The validity of this finding, however, can be verified by reanalysis of the experimental data. Although it may be an artifact of the utilized force field, the overall agreement with experiment makes us believe that placement of six waters (rather than three or four) have most probably been due to the facts that the solvent density distribution along the ring edges is continuous and that there is empty volume available for solvent to occupy. Notice that such features cannot be easily captured by the structure refinement algorithms. In this respect computer simulations can complement the experimental studies, thus permitting incorporation of the motions of the solvent at the molecular level.

Note that such effects were not observed in earlier theoretical studies either. However, all the earlier approaches used canonical ensemble methods, which might explain the differences. Because of the structure of the water network, the minor and major groove waters are connected through a linkage region that is quite narrow. In a canonical ensemble study that starts with all six waters for the polygon edge, the only way in which these waters could leave the minor groove would be through that linkage. However, the linkage region is almost always occupied (Table 1), and the density of the grooves at both ends of the linkage region is high. Therefore, actually to observe the escape of a water molecule from the minor groove would require extremely lengthy canonical ensemble simulations. Such limitations on the movement of water molecules have also been commented on by Kim et al. (1983). In contrast to canonical ensemble methods, such transfer mechanisms are intrinsic and are automatically taken care of in the grand canonical ensemble simulations. To show that such advantages of the grand canonical ensemble simulations can be utilized as powerful tools to study certain types of biological problems is the main objective of this study. Observation of such "disagreements" with other theoretical studies when more-conventional approaches are used in fact further illustrates the possible advantages of the grand canonical ensemble methods.

Linkage region

Fig. 5 also shows the grid calculation results for the linkage region. As can be seen, the linkage region waters have a large mobility and occupy a large region. Average density calculations with GSS analysis (Table 1) show that this region is occupied almost fully: Only one of the three

TABLE 1 Analysis of the distinct water sites

OW	ΔR(Expt-GCMC)	σ_{asym}	ρ_{occ}
Minor groove			
1	0.13	0.04	0.94
22	0.17	0.06	0.89
9	1.22	0.07	0.64
14	1.00	0.71	0.58
10	0.64	0.08	0.32
Link region			
12	0.41	0.03	1.00
20	2.15	0.79	0.99
8	1.63	0.47	0.79
Major groove			
2	0.30	0.67	1.00
4	0.68	0.35	1.00
5	0.93	0.80	1.00
7	0.70	0.40	1.00
11	0.16	0.05	1.00
15	0.56	0.67	1.00
3	1.20	1.78	0.99
13	0.91	1.22	0.99
16	0.35	0.79	0.99
17	0.24	0.53	0.99
26	0.56	0.59	0.99
19	0.64	0.71	0.98
23	0.46	0.43	0.98
18	1.04	0.88	0.97
24	1.17	0.62	0.97
21	2.72	0.88	0.96
25	1.26	0.49	0.96
6	0.66	0.46	0.95
Unassigned sites			
*u*1	–	0.97	1.00
*u*2	–	1.19	0.96
*u*3	–	8.18	0.96
*u*4	–	1.23	0.87

$\sigma^2_{asym} \equiv \Sigma_i (r_i - r_{ave})^2/4$, where r_{ave} is the mean position of a distinct water and i is summed over the four symmetry related sites. ΔR and σ_{asym} are in angstroms. ρ_{occ} is the site occupation density as determined with GSS analysis.

sites has a density lower than the full occupancy, ~0.8. High-density regions are somewhat shifted away from the experimental positions though, with deviations from the experimental positions of 1.62, 0.41, and 2.15 Å for the three sites of this hydration region. However, these shifted locations do not change the hydration pattern (Fig. 6). The observed shifts in the locations of the linkage waters might be due to the small shift in the position of the lower edge corner of the minor groove polygon disk, OW14. Given that there is a unique site-to-site correspondence, and that the sites are fully occupied, it can be safely stated that the overall comparison of the GCMC results with experiments is quite satisfactory.

Major groove

Most of the hydrating waters reside in the major groove. The chemical potential in the simulations was adjusted such that there were 108 waters on average in the simulation cell. Because the unit cell consists of four symmetry-related

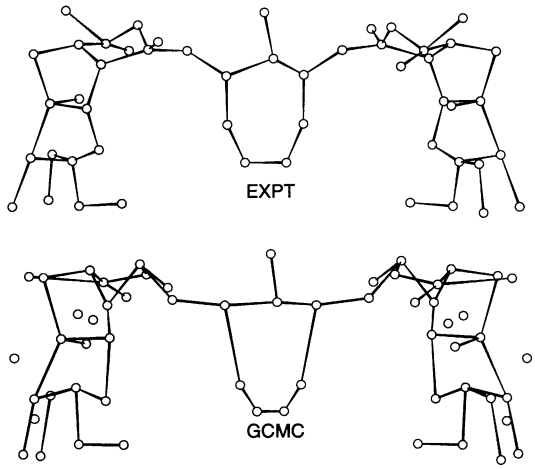

FIGURE 6 Comparison of the water network calculated in GCMC simulations with experiments. Unbonded spheres show the sites that have no equivalent experimental locations (*u*1–*u*4 in Tables 1 and 2).

subunits, 108 waters corresponds to 27 distinct molecules. Twenty-seven distinct waters rather than the twenty-five of the experimental study (Shieh et al., 1980) were included because, as detailed in the Introduction, later experimental studies predicted more waters in the crystal cell. As discussed in the first subsection of Results, some of the minor groove sites were found to be half occupied. Thus, 1.5 of 3 distinct waters of the minor groove "migrate" to a major groove. To accommodate this, we performed a GSS hydration analysis, using a total of 29 distinct sites (i.e., a total of 116 sites as cited in Analysis). All the additional sites, i.e., the sites that do not appear in the experimental structure, do of course appear in the major groove, thus crowding the region.

Fig. 7 reports the results of the grid density calculation for the major groove. Notice that the simulation results predict the experimental waters quite reliably. The deviations between the most likely locations as determined from the simulation trajectories and experimental positions are reported in Table 1. Except for the large disagreement for one of the sites (2.72 Å), the predicted sites deviate from their experimentally detected positions by less than 1.26 Å, with an average difference of 0.70 Å. When the site with large disagreement is included, the average deviation becomes only 0.81 Å, which is still smaller than the experimental resolution of 0.83 Å. As was predicted in another earlier study (Schneider et al., 1992), additional water molecules appear inside the "empty core" of the major groove (Fig. 6) and complicate the pentagonal water network. Kim and Clementi (1985a,b) extensively investigated the structural and energetic properties of these additional waters. In their MD study, Swaminathan et al. (1990) observed that one of the extra waters had a large *B* factor. Schneider et al. (1992) supported this finding that these additional waters, which are observed at lower-temperature studies, are not observed at room temperature because they are disordered. In fact, one of the unassigned waters in our study has a *B* factor that

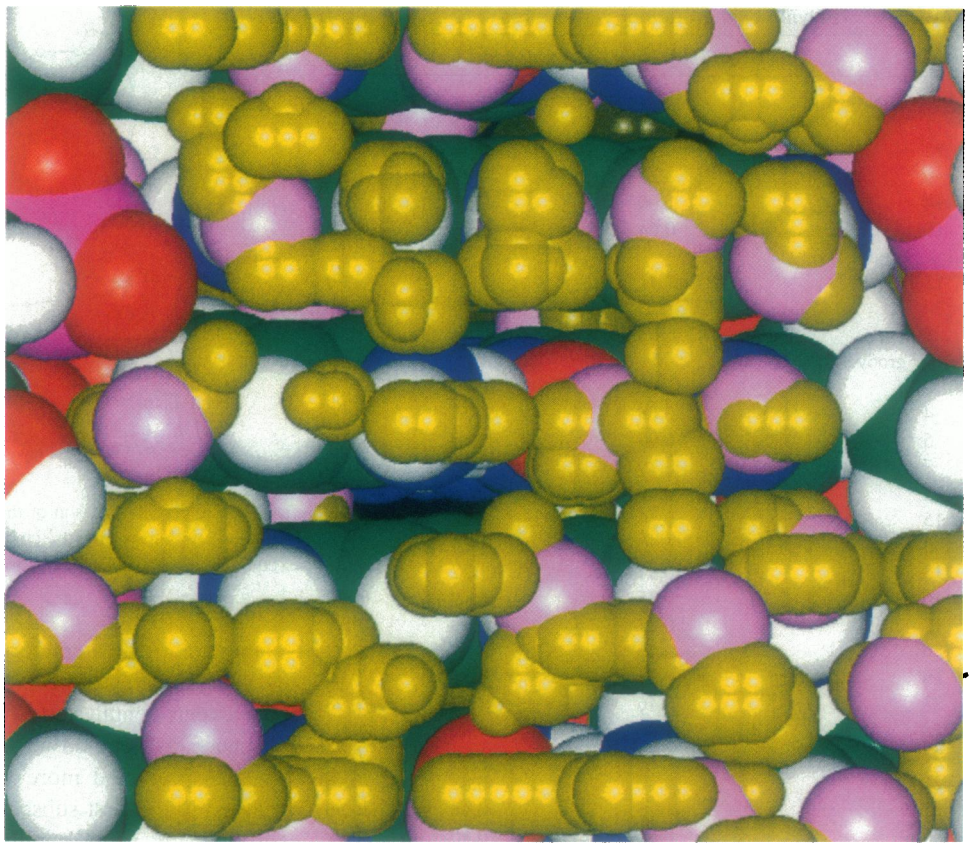

FIGURE 7 Grid density calculations for the major groove; details are as in Fig. 5.

is almost three times the average B factors of the other waters (Table 2). However, the other three unassigned waters have B factors very close to the B factors of the "detected" waters.

Symmetry analysis and convergence

The simulation unit cell consisted of four symmetry-related asymmetric subunits. We can use this fact to investigate the convergence characteristics of the simulations. For this we overlap the four symmetry-related sites, using the relevant transformations, and calculate a mean position for each distinct water. Inasmuch as water molecules are not constrained to obey the symmetry relations during the simulation, deviations about the mean position would be a reliable indicator of convergence; for a fully converged simulation the deviations should approximately vanish. As Table 1 shows, only two, OW3 and 13, of the twenty-six distinct waters have an asymmetry sizably larger than the experimental resolution. Note that these two waters belong to the part of the major groove that is "empty" in terms of experimentally detected waters. Fig. 8 compares the amount of asymmetry for each individual water and the disagreement with its experimentally determined position; in the figure the x and y axes, respectively, stand for the distance between the experimental site to the position determined in the simulation and the amount of asymmetry between the four symmetry-related sites of a distinct water molecule, as reported in Table 1. Almost all the points are either about the $y = x$ line or below it, which shows that the disagreements between the experimental locations and those determined in the simulations are genuine, meaning that the fluctuations in the simulations cannot account for the differences.

Energy analysis

Kim et al. (1983) and Kim and Clementi (1985a,b) extensively analyzed the energy and hydration pattern in the dCpG/proflavine crystal by varying the number of waters in the simulation cell. The availability of number fluctuations in the GCMC simulations allows us to do a comparable analysis from just one GCE simulation. Although the force fields are different in these two studies, as will be shown below, the conclusions are in agreement to a very good degree. Fig. 9 shows the fluctuations in the number of waters during the GCMC simulation; the distribution is nicely peaked around 108 waters, and its shape resembles a normal distribution with a half-width of four molecules.

Because the nucleic acid and proflavine molecules are kept fixed during the simulation, the relevant total energy is the summation of two contributions: solute–solvent and solvent–solvent terms. Figure 10 shows the total and the solute–solvent energies as a function of the number of waters. Compare these figures with Fig. 4 of Kim and Clementi (1985a). Except for the magnitude of the energies, which are very sensitive to the employed interaction poten-

TABLE 2 *B* factors

OW	B_{ave}	B_1	B_2	B_3	B_4
Minor groove					
1	4.33	2.29	6.38	–	–
22	20.84	15.51	26.16	–	–
9	13.63	12.52	13.40	15.01	13.58
14	44.90	79.41	16.21	40.17	43.82
10	8.27	4.39	4.44	19.94	4.31
Link region					
12	1.69	1.75	1.86	1.61	1.54
20	14.39	13.80	3.76	11.16	28.83
8	24.55	9.71	25.81	46.32	16.36
Major groove					
2	8.36	2.53	2.53	25.24	3.14
4	16.11	2.75	4.12	2.65	54.92
5	4.32	1.21	13.55	1.07	1.47
7	13.00	4.93	11.40	13.55	22.11
11	1.70	2.37	1.65	1.56	1.22
15	8.91	10.67	8.85	10.28	5.85
3	5.77	2.24	2.79	4.18	13.86
13	20.19	21.23	18.07	11.07	30.39
16	12.68	13.38	4.71	10.94	21.68
17	13.59	26.64	13.17	6.26	8.30
26	23.63	14.09	11.95	39.46	29.02
19	11.32	9.42	7.14	19.21	9.50
23	11.29	5.65	13.58	13.68	12.26
18	13.67	5.31	9.04	23.51	16.81
24	16.73	14.11	25.75	20.13	6.93
21	17.60	8.43	9.87	29.34	22.77
25	26.87	8.99	11.70	52.75	34.05
6	16.83	11.43	16.52	16.73	22.65
Unassigned sites					
*u*1	5.78	3.72	5.91	8.00	5.47
*u*2	25.53	8.79	10.05	52.11	31.19
*u*3	19.95	9.23	19.82	12.94	37.80
*u*4	41.60	21.43	26.55	54.95	63.47

The B factor for each site, $B \equiv (8\pi^2/3N_f) \Sigma_f (r_f - \bar{r})^2$, is in square angstroms and f is summed over the number of molecular simulation configurations N_f in which the site is occupied. B_{ave} is the average B factor of the four symmetry-related sites. Note that OW1 and 22 are half-occupancy sites.

tials, the trends in the calculated energies are very similar (however, the range in the studied number of molecules is much smaller in our case). As expected, because of the increased density effects the solute–solvent energy per water molecule (Fig. 11 *a*) and the average water–water pair interaction energy (Fig. 11 *d*) increase with increasing number of waters in the unit cell. The short-range interactions do not seem to vary with number of waters (Fig. 11 *c*), and the changes in the solute–solvent interaction energy are due solely to the electrostatic effects (Fig. 11 *b*).

DISCUSSION

We have demonstrated the computational efficiency of the grand canonical Monte Carlo simulations in studying the hydration pattern of biological macromolecules. As discussed, in the absence of "clean" experimental data the enclosed cavity effects cause the conventional canonical or microcanonical ensemble-based methodologies to be unsuitable for use in determining the solvation patterns in the cavities of the biological systems. The above statement is particularly true when the water pockets are of arbitrary shape and are disconnected from one another. Such shape irregularities make the prediction of how many waters should be inside each cavity almost impossible. Such problems are automatically eliminated in GCMC ensemble simulations.

We demonstrated the success of the GCMC simulations by studying the hydration pattern in the dCpG/proflavine crystal hydrate. Most of the experimentally detected waters were successfully predicted. Observed disagreements with experiments (which may be due as well to the force fields) for the minor groove can actually be utilized as a test of the predictive power of the computer simulations. A reanalysis of the experimental data incorporating our prediction that there may be fewer waters in the minor groove could be very fruitful, as it would reveal whether the grand canonical ensemble simulations overcome some of the intrinsic shortcomings of the canonical ensemble–based simulation methodologies.

It should be pointed out that the enclosed cavity effects are not only limited to crystal hydrates. Another possible application area of the grand canonical ensemble methodology would be to use it as a "soaking" algorithm. Generally

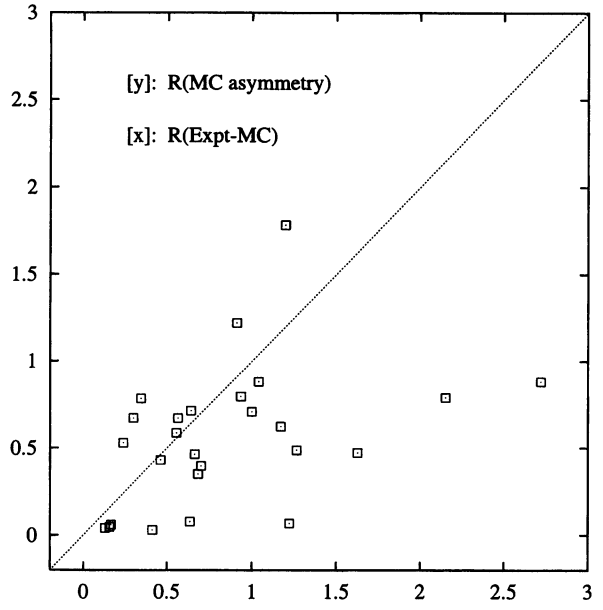

FIGURE 8 Comparison of the asymmetry of the symmetry-related waters calculated in GCMC simulations (*y axis*) and the disagreement with the experimental positions (*x axis*) for each distinct solvent molecule. The dotted line represents $y = x$ and is only a guide for the eye.

only a limited percentage of waters in the crystal or the solution structures of biomolecular systems can be detected in experiments. However, a complete study of biological systems requires that the waters, especially the hydrogen-bonded ones, be known. By placing the molecules at certain locations, a rough grand canonical simulation study would efficiently pinpoint structurally or energetically important waters or both. Of course, a longer and fine-tuned simulation study, as in this report, can be used as a predictive tool to determine accurately the most likely locations and the energy profile of the solvating water molecules.

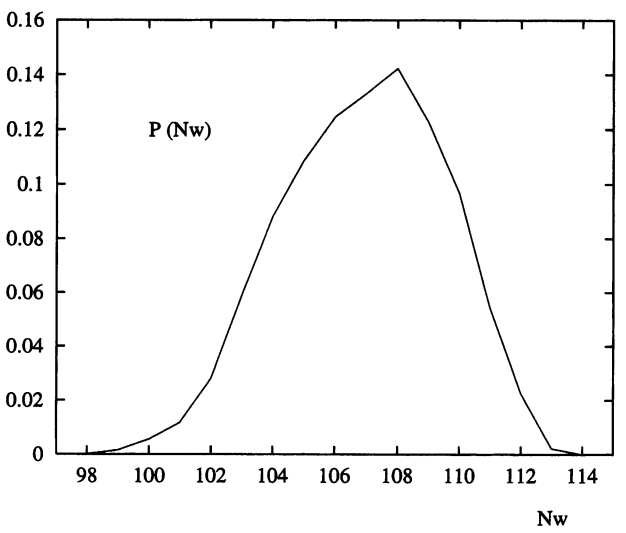

FIGURE 9 Distribution of the number of water molecules in the GCMC simulations.

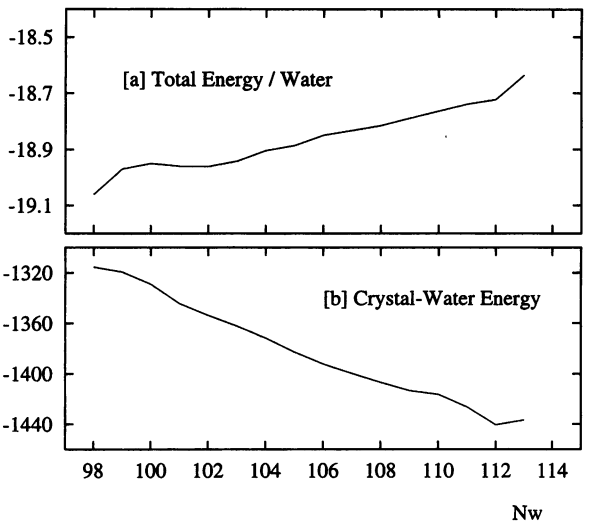

FIGURE 10 Energy profile as a function of number of waters: (*a*) total energy per water molecule, (*b*) solute–solvent interaction energy. Energies are in kcal/mol.

In many reactions between biological molecules, the reaction mechanism produces an enclosed region at the active site, and the entrapped waters may play an essential role (see the Introduction). For such systems the role of involved waters can be easily investigated by the approach advocated in this study. Another example case would be the biological association reactions in which the water does not play a direct role. For example, many biological reactions occur

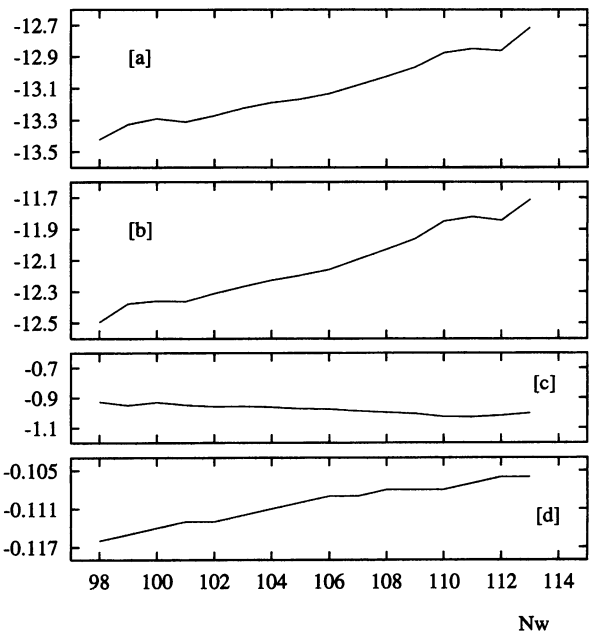

FIGURE 11 Solvent energy profile as a function of number of waters: (*a*) solute–solvent energy per water, (*b*) solute–solvent electrostatic interaction energy per water, (*c*) solute–solvent Lennard–Jones interaction energy per water. The average solvent–solvent interaction energy per water pair is shown in (*d*). Energies are in kcal/mol.

between molecules with shape complementarity. In such cases a water pocket gets created when the reacting molecules start to form a complex. Theoretical approaches to studying such reactions would require those waters to be emptied out of the formed pocket so that the reaction can progress. Emptying out those trapped waters can prove to be cumbersome with the canonical ensemble simulation methodologies; however, it can be achieved rather trivially through deletion steps in the grand canonical ensemble simulations (Resat et al., 1996). For example, we recently investigated the association reaction of the enzyme, trypsin, with its inhibitor, benzamidine (H. Resat, T. Marrone, and J. A. McCammon, in preparation), for which the GCMC method was very successful in providing good statistics in simulations.

It should also be noted that the additional computational expense of using grand canonical rather than canonical ensemble Monte Carlo simulations is only a few percent. This small increase in the expense, which permits the achievement of better statistical sampling, makes the grand canonical ensemble simulation methods a powerful approach in studying the solvation properties of biological systems and in the crystallographic data analysis.

This study was supported by National Institutes of Health Grant No. R55-GM43500.

REFERENCES

Adams, D. J. 1974. Chemical potential of hard-sphere fluids by Monte Carlo methods. *Mol. Phys.* 28:1241–1252.

Adams, D. J. 1975. Grand canonical ensemble Monte Carlo for a Lennard–Jones fluid. *Mol. Phys.* 29:307–311.

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

Ben-Naim, A. 1991. Strong forces between hydrophilic macromolecules: Implications in biological systems. *J. Chem. Phys. 95:* 8186–8210.

Berge, C. 1962. The Theory of Graphs and Its Applications. John Wiley and Sons, New York.

Berman, H. M. 1994. Hydration of DNA: take 2. *Curr. Opin. Struct. Biol.* 4:345–350.

Beutler, T. C., and W. F. van Gunsteren. 1994. Molecular dynamics simulations with first order coupling to a bath of constant chemical potential. *Mol. Simul.* 14:21–34.

Bhat, T. N., G. A. Bentley, G. Boulot, M. I. Greene, D. Tello, W. D. Acqua, H. Souchon, F. P. Schwarz, R. A. Mariuzza, and R. J. Poljak. 1994. Bound water molecules and conformational stabilization help mediate an antigen-antibody association. *Proc. Natl. Acad. Sci. USA.* 91: 1089–1093.

Cagin T., and B. M. Pettitt. 1991. Grand molecular dynamics: a method for open systems. *Mol. Simul.* 6:5–26.

Dewar, M. J. S., and D. M. Storch. 1985. Alternative view of enzyme reactions. *Proc. Natl. Acad. Sci. USA.* 82:2225–2229.

Friedman, H. L. 1985. A Course in Statistical Mechanics. Prentice Hall, Englewood Cliffs, NJ.

Harrison, S. C., and A. K. Aggarwal. 1990. DNA recognition by proteins with the helix-turn-helix motif. *Annu. Rev. Biochem.* 59:933–969.

Herzyk, P., J. M. Goodfellow, and S. Neidle. 1991. Molecular dynamics simulations of dinucleoside and dinucleoside-drug crystal hydrates. *J. Biomol. Struct. Dyn.* 9:363–386.

Hummer, G., A. E. Garcia, and D. M. Soumpasis. 1995. Hydration of nucleic acid fragments: comparison of theory and experiment for high resolution crystal structures of RNA, DNA, and DNA-drug complexes. *Biophys. J.* 68:1639–1652.

Janin, J., and C. Chothia. 1990. The structure of protein-protein recognition sites. *J. Mol. Biol.* 265:16027–16030.

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

Karplus, P. A., and C. Faerman. 1994. Ordered water in macromolecular structure. *Curr. Opin. Struct. Biol.* 4:770–776.

Kim, K. S., and E. Clementi. 1985a. Energetics and pattern analysis of crystals of proflavine deoxydinucleosite phosphate. *J. Am. Chem. Soc.* 107:227–234.

Kim, K. S., and E. Clementi. 1985b. Hydration analysis of the intercalated complex of deoxydinucleosite phosphate and proflavin: computer simulations. *J. Phys. Chem.* 89:3655–3663.

Kim, K. S., G. Corongiu, and E. Clementi. 1983. Networks of water molecules in a proflavine deoxydinucleosite phosphate complex. *J. Biomol. Struct. Dyn.* 1:263–285.

Levitt, M., and B. H. Park. 1993. Water: now you see it, now you do not. *Structure.* 1:223–226.

Lounnas, V., and B. M. Pettitt. 1994a. A connected-cluster of hydration around myoglobin: correlation between molecular dynamics simulations and experiment. *Proteins Struct. Funct. Genet.* 18:133–147.

Lounnas, V., and B. M. Pettitt. 1994b. Distribution function implied dynamics versus residence times and correlations: solvation shells of myoglobin. *Proteins Struct. Funct. Genet.* 18:148–160.

Lounnas, V., B. M. Pettitt, L. Findsen, and S. Subramaniam. 1992. A microscopic view of protein solvation. *J. Phys. Chem.* 96:7157–7159.

Lounnas, V., B. M. Pettitt, and G. N. Phillips, Jr. 1994. A global model of the solvent-protein interface. *Biophys. J.* 66:601–604.

Mezei, M. 1980. A cavity biased (T, V, μ) Monte Carlo method for the computer simulation of fluids. *Mol. Phys.* 40:901–906.

Mezei, M. 1987. Grand canonical ensemble Monte Carlo study of dense liquid Lennard-Jones, soft spheres and water. *Mol. Phys.* 61:565–582; erratum: 67: 1207–1208 (1989).

Mezei, M., and D. L. Beveridge. 1984. Generic solvent sites in a crystal. *J. Comp. Chem.* 5:523–527.

Mezei, M., D. L. Beveridge, H. M. Berman, J. M. Goodfellow, J. L. Finney, and S. Neidle. 1983. Monte Carlo studies on water in the dCpG/proflavin crystal hydrate. *J. Biomol. Struct. Dyn.* 1:287–297.

Nar, H., A. Messerschmidt, R. Huber, M. van de Kamp, and G. W. Canters. 1991. X-ray crystal structure of the 2 site-specific mutants his35gln and his35leu of azurin from *pseudomonas-aeruginosa. J. Mol. Biol.* 218: 427–447.

Neidle, S., H. M. Berman, and S. H. Shieh. 1980. Highly structured water network in crystals of a deoxydinucleoside-drug complex. *Nature (Lond.)* 288:129–133.

Otting, G., E. Liepinsh, and K. Wuthrich. 1991. Protein hydration in aqueous solution. *Science.* 254:974–980.

Panagiotopoulos, A. Z. 1992. Direct determination of fluid phase equilibria by simulation in the Gibbs ensemble: a review. *Molec. Simul.* 9:1–23.

Resat, H., and M. Mezei. 1994. Grand canonical Monte Carlo simulation of water positions in crystal hydrates. *J. Am. Chem. Soc.* 116: 7451–7452.

Resat, H., M. Mezei, and J. A. McCammon. 1996. Use of the grand canonical ensemble in potential of mean force calculations. *J. Phys. Chem.* 100:1426–1433.

Schneider, B., S. L. Ginell, and H. M. Berman. 1992. Low temperature structures of dCpG-proflavine. Conformational and hydration effects. *Biophys. J.* 63:1572–1578.

Sekharudu, C. Y., and M. Sundaralingam. 1993. Hydration of protein secondary structures: the role in protein folding. *In* Water and Biological Macromolecules. E. Westhof, editor. CRC Press, Boca Raton, FL. 148–162.

Shieh, H. S., H. M. Berman, M. Dabrow, and S. Neidle. 1980. The structure of drug-deoxydinucleoside phosphate complex: generalized

conformational behavior of intercalation complexes with RNA and DNA fragments. *Nucleic Acids Res.* 8:85–97.

Steitz, T. A. 1990. Structural studies of protein-nucleic acid interaction: the sources of sequence-specific binding. *Q. Rev. Biophys.* 23:205–280.

Swaminathan, S., D. L. Beveridge, and H. M. Berman. 1990. Molecular dynamics simulation of a deoxydinucleoside-drug intercalation complex: dCpG/proflavin. *J. Phys. Chem.* 94:4660–4665.

Swope, W. C., and H. C. Anderson. 1995. A computer simulation method for the calculation of chemical potentials of liquids and solids using the bicanonical ensemble. *J. Chem. Phys.* 102:2851–2863.

Warshel, A., J. Aqvist, and S. Creighton. 1989a. Enzymes work by solvation substitution rather than by desolvation. *Proc. Natl. Acad. Sci. USA.* 86:5820–5824.

Warshel, A., G. Naray-Szabo, F. Sussmann, and J.-K. Hwang. 1989b. How do serine proteases really work? *Biochemistry.* 28: 3629–3637.

Weiner, S. J., P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta, Jr., and P. Weiner. 1984. A new force field for molecular mechanical simulation of nucleic acids and proteins. *J. Am. Chem. Soc.* 106:765–784.

Westhof, E. 1993. Water and Biological Macromolecules. CRC Press, Boca Raton, FL.