Generic Solvent Sites in a Crystal

Mihaly Mezei and David L. Beveridge

Chemistry Department, Hunter College of the City University of New York, 695 Park Avenue, New York, New York 10021

Received 28 March 1984; accepted 25 June 1984

A numerical procedure is described and tested for the determination of solvent sites in a crystal hydrate from computer simulation results. The method does not require the computation of density distributions.

I. INTRODUCTION

Liquid state computer simulation, using both Monte Carlo (MC) and molecular dynamics (MD) procedures, are currently being used to study the organization of water in crystals in comparison with results on ordered water positions obtained from x-ray and neutron diffraction experiments. A useful set of indices for comparison in the calculated results is the mean water positions obtained as a configurational average in the simulation. However, the calculation is not straightforward, since solvent molecules can in principle diffuse during the simulation, and thus a given mean position may have fractional contributions from several specific water molecules in the course of the simulation. What is required is a set of generic mean positions for solvent molecules, defined in such a way that diffusional interchange is taken into account.

In the evaluation of computer simulation the determination of solvent sites thus presents a problem. In principle, the computer simulation can be used to determine the three-dimensional density distribution of solvents, the peaks of which can be identified with solvent sites. However, the representation of a three-dimensional density distribution requires data points the number of which is inversely proportional to the cube of the gridsize. Accurate estimate of the density at all grid points requires long runs. The length of the runs referred to in this article is adequate to obtain the density accurately on a grid of about 0.4 Å. The accurate representation of the density on a 0.1 Å grid would require $4^3 = 64$ times longer runs. The large number of grid points also presents a storage problem that can be overcome, however, by obtaining the density in slices. Previous computer simulation studies[1-4] dealt with this problem by determining two-dimensional density maps in $O(1$ Å) thick planes. This procedure, however, inherently limits the accuracy of the result.

Note also that the Fourier coefficients of the density also can be obtained from a simulation without obtaining the density distribution itself. In general, the product of the probability density $\rho(\mathbf{R})$ and a function $F(\mathbf{R})$ can be obtained from

$$\int F(\mathbf{R})\rho(\mathbf{R})\, d\mathbf{R} = \sum_i F(\mathbf{R}_i)/n \qquad (1)$$

where the sum is taken over all positions $\mathbf{R}_i$ that were sampled in the simulation by any of the solvents, since the simulation methods are constructed to ensure that the positions $\mathbf{R}_i$ are sampled with probability $\rho(\mathbf{R})$.

The purpose of this article is to present an alternative procedure for the determination of solvent position in a crystal that does not require the determination of the density distribution entirely. A somewhat similar approach was applied recently in this Laboratory to the analysis of ionic hydration.[5] Section II presents the procedure and Section III describes its performance on a dinucleotide/proflavin crystal hydrate, where experimental water positions have been determined,[6,7] and MC computer simulations have been performed.[8,9]

II. THEORY AND ALGORITHMS

For a system of one solvent molecule, it can be shown by straightforward differentiation that if $\mathbf{R}^k$ is the position of the molecule at the kth

CCC 0192-8651/84/060523-05$04.00

observation/step, the mean position

$$\mathbf{S}_i = \left(\sum_{k=1}^{n} \mathbf{R}^k \right) / n \tag{2}$$

minimizes the expression

$$\sum_{k=1}^{n} |\mathbf{R}^k - \mathbf{S}_i|^2 \tag{3}$$

The average in eq. (2) is a specific average. (It happens to coincide with a generic average, since there is only one solvent in the system.) Thus for a system of N solvent molecules one should seek a set of points $\mathbf{S}_i$, $i = 1, \ldots, N$, where the sum of N expressions of the type (3) is minimized. However, to account for the diffusion of the solvent molecules, the choice of $\mathbf{R}_j^k$'s to be used (the subscript j refers to the molecule the position of which is $\mathbf{R}_j^k$) also have to be subjected to optimization. In other words, one is looking for the generic average and not the specific averages over positions of the individual solvents. These requirements are satisfied if such a set $\{\mathbf{S}_i,\ i = 1, \ldots, N\}$ is sought that minimizes

$$\text{MSD}(\{\mathbf{S}_i\}) = \sum_{i=1}^{N} \left(\sum_{k=1}^{n} \sum_{j=1}^{N'} |\mathbf{R}_j^k - \mathbf{S}_i|^2 O_{i,j}^k \right) / n \tag{4}$$

where N is the number of sites, N' is the number of molecules, and $\mathbf{O}^k$ is a matrix such that

$$O_{i,j}^k = 0 \quad \text{or} \quad 1 \tag{5}$$

$$\sum_{j=1}^{N'} O_{i,j}^k \begin{cases} \leq 1 & \text{if } N > N' \\ = 1 & \text{if } N \leq N' \end{cases} \quad \text{for } i = 1, \ldots, N \tag{6}$$

$$\sum_{i=1}^{N} O_{i,j}^k \begin{cases} \leq 1 & \text{if } N < N' \\ = 1 & \text{if } N \geq N' \end{cases} \quad \text{for } j = 1, \ldots, N' \tag{7}$$

$$\sum_{i=1}^{N} \sum_{j=1}^{N'} O_{i,j}^k |\mathbf{R}_j^k - \mathbf{S}_i|^2 \quad \text{is minimum} \tag{8}$$

Notice that we allow for $N \neq N'$. If $N < N'$ then there are solvents that are uniformly distributed (delocalized solvents), while the case of $N < N'$ represents a system where some of the sites are only fractionally occupied. The positions of the nonzero elements in $\mathbf{O}^k$ associate one $\mathbf{R}_j^k$ with each $\mathbf{S}_i$ for $N \leq N'$ and one $\mathbf{S}_i$ with each $\mathbf{R}_j^k$ for $N \geq N'$. The conditions (5)–(7) on the matrix $\mathbf{O}^k$ are necessary to ensure that each $\mathbf{R}_j^k$ and $\mathbf{S}_i$ are selected at most once and the maximum number of assignments are made. The expression

$$\sum_{j=1}^{N'} |\mathbf{R}_j^k - \mathbf{S}_i|^2 O_{i,j}^k \tag{9}$$

simply provides a single term to the MSD sum.

The problem defined by eqs. (5)–(8) is known as the assignment problem. It can be solved exactly, although the solution is time consuming since it is an integer programming problem. The best method available is the Hungarian method.[10,11] However, the special nature of the matrix associated with this problem allows for a simple but approximate algorithm that appears to work reasonably well for this type of problem. The effort involved in the approximate algorithm proposed here is equivalent to the effort of obtaining the initial solution in the exact algorithm, where an undetermined (but limited) number of iterative refinements follow. It is to be stressed that the procedure described here works with any method that solves the assignment problem, and our choice here was dictated by economy.

II A. An Iterative Algorithm for Minimizing the MSD

A practical approach and an iterative algorithm for the solution for the peak position consists of the following steps:

(1) Select a set of representative configurations. For MC computer simulations, the configurations should be chosen infrequently enough so that all solvent molecules move between the selected configurations.

(2) Obtain an estimate for $\{\mathbf{S}_i\}$. Either it can come from an outside source (e.g., experiment), or else the program has to generate it from the data. In Section II A two algorithms will be described for this purpose.

(3) For each k, solve eqs. (5)–(8) for the matrix $\mathbf{O}^k$. An approximate solution can be obtained as follows:

(3a) Set all elements of $\mathbf{O}^k$ to zero.

(3b) $N \leq N'$:

For every $\mathbf{S}_i$ find an index $j(i)$ such that

$$|\mathbf{R}_{j(i)}^k - \mathbf{S}_i|^2 = \min_{j'} |\mathbf{R}_{j'}^k - \mathbf{S}_i|^2 \tag{10a}$$

$$j' \in \left\{ j \,\middle|\, \sum_{i=1}^{N} O_{i,j}^k = 0 \right\}$$

and set $O^k_{i,j(i)} = 1$. In words, assign to the peak estimate $\mathbf{S}_i$ the solvent that is the nearest to it among the unassigned solvents.

(3c) $N \geq N'$:

For every $\mathbf{R}^k_j$ find an index $i(j)$ such that

$$|\mathbf{R}^k_j - \mathbf{S}_{i(j)}|^2 = \min|\mathbf{R}^k_j - \mathbf{S}_{i'}|^2 \qquad (10b)$$

$$i' \in \left\{ i \middle| \sum_{j=1}^{N'} O^k_{i,j} = 0 \right\}$$

and set $O^k_{i(j),j} = 1$. In words, assign to the solvent j to the peak estimate that is the nearest to it among the unassigned ones.

The complexity of step (3b) or (3c) is proportional to N^2. At the end of step (3b) or (3c), the $\mathbf{O}^k$ matrix has been filled for all k selected. Clearly, for $N = N'$ either (3b) or (3c) can be used.

(4) Compute the new estimate of site positions as

$$\mathbf{S}'_i = \left(\sum_{k=1}^{n} \sum_{j=1}^{N'} O^k_{i,j} \mathbf{R}^k_j \right) \Big/ n \qquad (11)$$

for each peak $\mathbf{S}_i$. Equation (11) represents the mean position of solvents assigned to peak $\mathbf{S}_i$. Notice, that the matrix multiplication in eq. (11) and in similar expressions in subsequently described algorithms is only a notational convenience to select the appropriate $\mathbf{R}^k_j$ to be used. In practice, the arrays $i(j)$ and/or $j(i)$ are stored in the computer thus the complexity of this step is proportional only to N.

(5) Substitute $\{\mathbf{S}'_i\}$ into eq. (4) to obtain the new MSD'. If MSD' < MSD, repeat from step (3).

(6) If the new estimate is not better than the old one [in fact, it may even be worse, due to the approximate minimization in step (3)], stop and accept the previous estimate as the solution.

II B. Generation of the Initial Estimate

The initial estimates for the solvation sites can be obtained from external sources or generated from the data itself. We first describe two algorithms that generate initial estimates for $N = N'$. Both perform well.

The first algorithm, called the "sequential algorithm" consists of the following steps:

(1) Take the first configuration as the first approximation to the estimate.

(2) In general, from the kth approximation $\{\mathbf{S}^k_i\}$, obtain the $(k+1)$th by solving eqs. (5)–(8) for $\mathbf{O}^{k+1}$ using $\{\mathbf{S}^k_i\}$ for $\{\mathbf{S}_i\}$. Then

$$\mathbf{S}^{k+1}_i = [k/(k+1)]\mathbf{S}^k_i + 1/(k+1) \sum_{j=1}^{N'} \mathbf{R}^{k+1}_j O^{k+1}_{i,j} \qquad (12)$$

(3) If step (2) is repeated $n-1$ times, the configurations will be exhausted and $\{\mathbf{S}^n_i\}$ will be the initial estimate to be used in step (2) of the iterative process.

The second algorithm, called "pairing algorithm" proceeds as follows:

(1) Divide the configurations into successive pairs.

(2) For each pair, solve eqs. (5)–(8) using the first element of the pair as $\{\mathbf{S}_i\}$ and the second element as $\{\mathbf{R}^k_j\}$. Prepare a composite using the solution obtained:

$$0.5\mathbf{R}^k_i + 0.5 \sum_{j=1}^{N'} \mathbf{R}^{k+1}_j O^{k+1}_{i,j} \qquad (13)$$

(3) Replace each pair of configurations with their composite according to eq. (13).

(4) If there is more than one pair, repeat from step (1).

Note that this algorithm requires that the number of configurations used is a two-power. [To relax this restriction, we should have introduced in eq. (13) some cumbersome weighting factors.] This restriction, however, is not a serious one since by adjusting the frequency of selecting representative configurations, it can be easily satisfied.

The two algorithms require about the same computational effort: The solvent assignment problem has to be solved $n-1$ times.

If $N < N'$ is required, one has to drop $N' - N$ sites from $\{\mathbf{S}_i\}$. It is reasonable to drop those that have the largest MSD.

For $N > N'$, additional sites have to be generated. One alternative could be a search for cavities that are left after having placed the solvent molecules at the initial N' sites. This has the drawback that the actual crystal atoms have to be introduced into the calculation of the initial estimate.

Another possibility is to generate all the sites related by the space group symmetry of the crystal to the first N' estimates and choose from these those that are the farthest from the already existing sites. This procedure would "help"

to introduce into the result the required symmetry.

Once N' is chosen the procedure described here works automatically. The choice of N', however, is not necessarily simple, since it depends on the existence of disordered waters or partially occupied sites. This information is either available from outside sources or could be derived from a calculation of the type proposed here. In particular, sites with large MSD can be considered to be in a region of disordered solvent and be dropped. Partially occupied sites are indicated when the density of the crystal shows that there is some empty space in it. The volume of this empty space can give an estimate of the possible extra sites. If a calculation is performed, the occupancy of the resulting sites will either confirm the N' chosen from this estimate or show that some sites have too small occupancies to be considered and thus have to be dropped.

II C. Upper Bound on the Error of the Solution of the Assignment Problem

It is possible to obtain a rigorous upper bound on the error in our solution of the solvent assignment problem as follows.

Let us define $j(i)$ as such j that $O^k_{i,j(i)} = 1$, $i(j)$ as such i that $O^k_{i(j),j} = 1$ [as in eq. (10)], and $i'(j)$ as such i for which $|\mathbf{R}^k_j - \mathbf{S}_i|^2$ is minimum. Furthermore, if for a j no such i exists that $O^k_{i(j),j} = 1$, let $i(j) = 0$ by definition. It is easy to see that the possible improvement of the MSD by reassigning the solvents is limited by

$$\sum_{\{j|i(j)\neq 0\}} |\mathbf{R}_j - \mathbf{S}_{i(j)}|^2 - |\mathbf{R}_j - \mathbf{S}_{i'(j)}|^2 \quad (14)$$

It follows from this argument that if $O^k_{i'(j),j} = 1$ for all j examined, then the solution obtained is exact. In words, if the solvent configuration is such that for all peaks the solvent that is closest to one of the peak estimate has no other peak estimate closer to it, then the algorithm provides the exact solution.

Note that this procedure did not require consideration of unassigned solvent molecules since an unassigned solvent is always farther from any site than the solvent assigned by our procedure to that site. Thus, if the exact solution assigns to a site a solvent unassigned by our procedure, the contribution of that site to the MSD will necessarily increase.

II D. Possible Improvements of the Solution of the Assignment Problem

It follows from the argument in the previous section that the result will depend on the order of scanning $\{\mathbf{S}_i\}$, since if one solvent is the nearest to more than one site then it will be assigned to the site that is looked at first. This comment gives a trivial way of obtaining different and possibly better approximations to the assignment problem: Repeat the procedure from step (3), but change the order of scanning the peaks. By performing a certain number of retries one may obtain reasonable assurance that the solution obtained is close to the exact one.

A somewhat more systematic approach, that is about as costly as a retry, can proceed as follows. For each solvent such that $i(j) > 0$, if

$$|\mathbf{S}_{i(j)} - \mathbf{R}_j|^2 + |\mathbf{S}_{i'(j)} - \mathbf{R}_{j(i'(j))}|^2 > |\mathbf{S}_{i'(j)} - \mathbf{R}_j|^2 - |\mathbf{S}_{i(j)} - \mathbf{R}_{j(i'(j))}|^2 \quad (15)$$

then set

$$O^k_{i(j),j} = 0 \qquad O^k_{i'(j),j} = 1 \quad (16)$$

In words, check the effect of the interchange in the assignment on the MSD and if it decreases the sum, make the interchange. Since the interchange does not affect any other term in the MSD sum, the approximation has clearly been improved. Note, that this procedure is the same for $N > N'$ and $N \leq N'$. Naturally, for $N < N'$, $i(j)$ is never zero.

III. RESULTS AND DISCUSSION

The dCpG/proflavin crystal hydrate was studied by x-ray diffraction[6,7] and by MC computer simulation.[8,9] Three simulations have been performed on the unit cell in this Laboratory, differing from each other in the starting configuration used and the number of waters simulated as specified in Table I. Simulation I used the experimentally obtained oxygen positions for the initial configuration, and for simulation III, eight more waters were placed in cavities found. For simulation II, the initial configuration was obtained by placing 100 waters on a grid spanning the region of the crystal that is not occupied by the solute. The resulting mean positions are compared with the experimental data in Ref. 7. The average deviations between the three simu-

Table I. Results of the site determination in the *d*CpG/Proflavin crystal hydrate from MC simulations.

	Simulation		
	I	II	III
Number of waters	100	100	108
Length of the simulation	2600 K	1600 K	1600 K
Number of configurations used	8192	4096	4096
Sampling frequency	317	395	397
Number of iterations	6	6	6
MSD/water (in Å^2)	0.73	0.51	0.27
Error bound on MSD (in Å^2)	0.24	0.16	0.02
Refinement effect (in Å^2)	0.20–0.12	0.01–0.09	0.08–0.001
Number of singly occupied sites	71	71	69

lations are significantly larger than the error bounds on the MSDs obtained, thus the differences between the three simulation results are mainly due to the difference in the initial assumptions. The analysis results are also collected in Table I for the three simulations. The error bounds obtained show that the exact solution of the assignment problem could have decreased the MSD by 30% or less. The refinement procedure, described in Section II D, was useful, but not crucial since it improved the MSD only by 16% or less. We also tried the changing of the scanning order, but it only affected the MSD by a few percent and was subsequently dropped. The results also showed that 70–80% of the sites were occupied by the same water molecule during the entire length of the simulation; the rest of the sites were "visited" by more than one water. The sites obtained were also compared with two-dimensional density maps drawn in 1 Å thick slices of the unit cell. Each generated site was found to be at or near the center of a density peak, showing that the approximate solution of the assignment problem did not introduce artifacts in the results.

This research was supported by NIH grant 5-R01-GM-24914 and a CUNY Faculty Research Award. Useful discussions with Dr. Helen Berman and Dr. Julia Goodfellow are greatly appreciated.

References

1. A. T. Hagler and J. Moult, *Nature*, **272**, 222 (1978).
2. A. T. Hagler, J. Moult, and D. Osguthorpe, *Biopolymers*, **19**, 395 (1980).
3. V. Madison, D. J. Osguthorpe, P. Dauber, and A. T. Hagler, *Biopolymers*, **22**, 27 (1983).
4. J. Hermans and M. Vacatello, in *Water in Polymers*, S. P. Rowland, Ed., ACS Symposium Series 127, American Chemical Society, Washington, D.C., 1980.
5. F. T. Marchese and D. L. Beveridge, *J. Am. Chem. Soc.*, **106**, 3713 (1984).
6. S. Neidle, H. M. Berman, and H. S. Shieh, *Nature*, **288**, 129 (1980).
7. H. S. Shieh, H. M. Berman, and S. Neidle, *Nucleic Acids Res.*, **8**, 85 (1980).
8. M. Mezei, D. L. Beveridge, H. M. Berman, J. Goodfellow, J. Finney, and S. Neidle, *J. Biomol. Struct. Dyn.*, **1**, 287 (1983).
9. K. S. Kim, G. Corongiu, and E. Clementi, *J. Biomol. Struct. Dyn.*, **1**, 263 (1983).
10. E. Egervary, *Mat. Fiz. Lapok*, **38**, 16 (1931), translated by H. W. Kuhn, Office of Naval Research Logistic Project Report, Dept. of Math., Princeton University.
11. H. W. Kuhn, *Naval Res. Logist. Quart.*, **2**, 83, (1955).