Morphometric Approach to the Solvation Free Energy of Complex Molecules

Roland Roth,[1,2] Yuichi Harano,[3] and Masahiro Kinoshita[3]

[1]*Max-Planck-Institut für Metallforschung, Heisenbergstrasse 3, D-70569 Stuttgart, Germany*
[2]*ITAP, Universität Stuttgart, Pfaffenwaldring 57, D-70569 Stuttgart, Germany*
[3]*International Innovation Center, Kyoto University, Uji, Kyoto 611-0011, Japan*

(Received 21 April 2006; published 14 August 2006)

We show that the solvation free energy of a complex molecule such as a protein can be calculated using only four geometrical measures of the molecular structure and corresponding thermodynamical coefficients. We compare results from this morphometric approach to those obtained by an elaborate statistical-mechanical theory in liquid state physics for a large variety of different structures of protein *G* and find excellent agreement. Since the computational time is drastically reduced, the new approach provides a practical and efficient way for calculating the solvation free energy which can be employed when this quantity has to be calculated for a large number of structures, as in a simulation study of protein folding.

DOI: 10.1103/PhysRevLett.97.078101 PACS numbers: 87.14.Ee, 05.70.−a, 61.20.−p

The solvent has enormous effects on the structure and properties of flexible, complex polyatomic molecules immersed in it. Among a variety of such molecules, a protein is undoubtedly the most important object in physics, chemistry, and biology. It is quite large, and the candidate structures which could be stabilized in local free energy minima are just innumerable. Nevertheless, the protein folds into a unique, native structure. Currently, the analysis on the structural stability of proteins and the prediction of the native structure is one of the most enthusiastic subjects [1]. The protein itself aims for the structure that corresponds to the global minimum of its intramolecular free energy F_{protein}, whereas the solvent tries to force the protein to form the structure minimizing the solvation free energy $F_{\text{solvation}}$. The structural stability is determined by the complicated interplay of these two factors [2,3]. A key point is that the solvent and the protein must be modeled on the same level because the marginal balance of F_{protein} and $F_{\text{solvation}}$ is essential. If one employs a sophisticated model only for the protein and regards the solvent as a continuum, for example, the predicted results would turn out rather unreliable.

Despite the crucial importance of $F_{\text{solvation}}$ of the protein in a given structure, relatively little is known about effective ways of calculating this quantity. Calculations by molecular dynamics computer simulations are limited to small solute molecules and infeasible for large polyatomic molecules like proteins. A practicable means of calculation is the three-dimensional (3D) version [4–6] of the integral-equation theory, an elaborate statistical-mechanical theory in liquid state physics [7]. It is capable of treating the atomic details of the given protein structure immersed in an infinitely large number of solvent molecules and calculating the ensemble-averaged solvent configuration in equilibrium with the structure. However, it has the disadvantage of large computer storage demands. Further, the long computation time required becomes serious when one attempts to calculate $F_{\text{solvation}}$ for a huge number of different, candidate structures.

In this Letter, we demonstrate the remarkable power of our new morphometric method [8] when applied to the calculation of $F_{\text{solvation}}$ of a large, complex molecule such as a protein in a given structure, if the molecule-solvent and solvent-solvent interactions are specified. To this end, we study how the mesoscopic morphometric approach fares compared to the microscopic 3D integral-equation theory. We compare the solvation free energy obtained by these two very distinct approaches for 600 different structures of protein *G*, which is a protein with 56 residues and 855 atoms. Results from the 3D integral-equation theory serve as a benchmark for the morphometric approach. In the latter, $F_{\text{solvation}}$ is determined by only four geometrical measures of the protein structure and corresponding thermodynamical coefficients [8]. This separation of $F_{\text{solvation}}$ into geometrical and thermodynamic coefficients allows for an extremely fast calculation. Here we report that the morphometric approach predicts results which are almost indistinguishable from those of the 3D integral-equation theory in a computation time that is over 4 orders of magnitude shorter.

We start by recalling the basic aspects of the 3D integral-equation approach, because we wish to highlight the remarkable difference between this theory and the morphometric approach described later. One great advantage of the 3D integral-equation theory is that details of the polyatomic structure of a solute molecule can explicitly be taken into account. The theory is briefly described for the cases where the solvent is a simple fluid.

A solute molecule, denoted as *I*, in a prescribed structure is immersed into a solvent of small spheres. The bulk density of the solvent ρ_S is given. The solute *I* consists of a set of fused atoms. The basic equations are expressed in terms of the correlation functions, the radial-symmetric solvent-solvent (SS) and 3D solute-solvent (IS) correlation

functions. There are two principal equations. The first one is the IS Ornstein-Zernike equation [7] and is expressed in Fourier space as

$$W_{\mathrm{IS}}(k_x, k_y, k_z) = \rho_S C_{\mathrm{IS}}(k_x, k_y, k_z) H_{\mathrm{SS}}(k), \tag{1}$$

where the capital letters C, H, and W represent the Fourier transforms of c, h, and $w = h - c$, respectively. c is the direct and h is the total correlation function. $k^2 = k_x^2 + k_y^2 + k_z^2$. The second equation, which is the closure relation, is expressed in real space as

$$c_{\mathrm{IS}}(x, y, z) = \exp\{-\beta u_{\mathrm{IS}}(x, y, z)\} \times \exp\{w_{\mathrm{IS}}(x, y, z) + b_{\mathrm{IS}}(x, y, z)\} - w_{\mathrm{IS}}(x, y, z) - 1, \tag{2}$$

where u_{IS} denotes the solute-solvent interaction potential and $\beta = 1/(k_B T)$ with Boltzmann's constant k_B and the temperature T. Here we employ the HNC approximation where the bridge function b is set to zero [7]. The reliability of this approximation has been verified [9].

The two principal equations are numerically solved on a cubic grid. The center of mass of the protein molecule is chosen as the origin of the coordinate system. The numerical procedure is briefly summarized as follows. In the initialization, the solute-solvent interaction $u_{\mathrm{IS}}(x, y, z)$ is calculated at each grid point and $w_{\mathrm{IS}}(x, y, z)$ is set to zero. Then follows a loop in which the pair direct correlation function $c_{\mathrm{IS}}(x, y, z)$ is calculated from Eq. (2) and transformed into $C_{\mathrm{IS}}(k_x, k_y, k_z)$ using the 3D fast Fourier transform (3D-FFT). From this, with the help of Eq. (1), one obtains $W_{\mathrm{IS}}(k_x, k_y, k_z)$ and transforms it to $w_{\mathrm{IS}}(x, y, z)$ using the inverse 3D-FFT. This loop is iterated until the input and output functions become identical within convergence tolerance.

In principle, this microscopic approach is capable of including detailed chemical information about the protein-solvent interaction by specifying the interaction potentials between each atom of the protein with the solvent. However, in the present study where we focus mainly on the power of the morphometric approach and not so much on the results, we wish to keep the model as simple as possible, while treating the polyatomic structure and the solvent on equal footing. We model the solvent molecules as hard spheres and the solute molecule as a set of fused hard spheres. Because of the hard-core nature of the interaction $u_{\mathrm{IS}}(x, y, z)$ we consider here, the Boltzmann factor $\exp\{-\beta u_{\mathrm{IS}}(x, y, z)\}$ is zero on each grid point where a solvent particle and at least one of the atoms overlap. Otherwise, the Boltzmann factor is unity. For most calculations, the grid spacing (Δx, Δy, and Δz) is set at $0.2 d_S$ and the grid resolution ($N_x \times N_y \times N_z$) is chosen to be $256 \times 256 \times 256$. It has been verified that the spacing is sufficiently small and the box size ($N_x \Delta x$, $N_y \Delta y$, and $N_z \Delta z$) is large enough to get a good overview and to avoid numerical artifact due to finite-size effects. However, we will also discuss how the results are affected if the grid spacing is reduced.

The microstructure of the solvent near solute I is described by $g_{\mathrm{IS}}(x, y, z)$ where $g = h + 1$, or equivalently, by the density profile $\rho(x, y, z) = \rho_S g_{\mathrm{IS}}(x, y, z)$. Once the solvent structure is known, the solvation free energy (SFE) of I, which we denote by $F_{\mathrm{solvation}}^{\mathrm{3D\text{-}HNC}} = \Delta\mu_I$, is obtained from the 3D integration [5,6] expressed by

$$\beta\Delta\mu_I = \rho_S \iiint \left\{ \frac{h_{\mathrm{IS}}(x, y, z)^2}{2} - c_{\mathrm{IS}}(x, y, z) - h_{\mathrm{IS}}(x, y, z) \frac{c_{\mathrm{IS}}(x, y, z)}{2} \right\} dx dy dz. \tag{3}$$

The great advantage of this approach is that we have access to microscopic information of the solvent distribution around the solute and to the solvation free energy $F_{\mathrm{solvation}}$ within the same theoretical framework. However, this approach is computationally quite demanding because of the lack of spacial symmetry in the problem and the iterative numerical calculation using 3D grids.

We employ our 3D integral-equation theory in order to generate a set of benchmark data for a large variety of structures of protein G. To this end, we calculate the solvation free energy of 600 structures of protein G in a hard-sphere solvent. Those structures were taken from local-minimum states of the energy function found in a replica-exchange molecular dynamics simulation using all-atom potentials [10]. We used structures generated by computer simulations to avoid unrealistic overlaps of the polypeptide chain and energetically unreasonable structures. The 600 conformations cover a very wide range of different structures. The solvent diameter $\sigma_S = 2.8$ Å and density $\rho_S \sigma_S^3 = 0.7$ are set to mimic corresponding values of water [6]. The diameters of the atoms in the protein are the Lennard-Jones diameters of AMBER99.

In Fig. 1 we show the resulting solvation free energy $F_{\mathrm{solvation}}^{\mathrm{3D\text{-}HNC}}$ calculated by the 3D integral-equation approach, as a function of the radius of gyration for the set of 600 structures of protein G. It is the variation in the solvation free energy from structure to structure, rather than its absolute value, that is important. As a general trend, one can identify that a structure corresponding to a relatively low solvation free energy possesses a small radius of gyration. This indicates that compact structures are favored by the solvent. A fragmented protein structure with a large radius of gyration would be susceptible to external forces and could easily change its structure. This, however, would hinder the protein to fulfill its biological function, which requires a robust structure.

The connection between the solvation free energy and the radius of gyration is substantiated by the structures shown as an illustration in Fig. 1. The structure with the lowest solvation free energy is compact and has a small volume V and surface area A. Another structure, explicitly depicted in Fig. 1, is less compact and thereby has a larger volume and surface area and corresponds to a larger solvation free energy. In contrast, we also show a random coil structure of protein G, with a much larger volume and

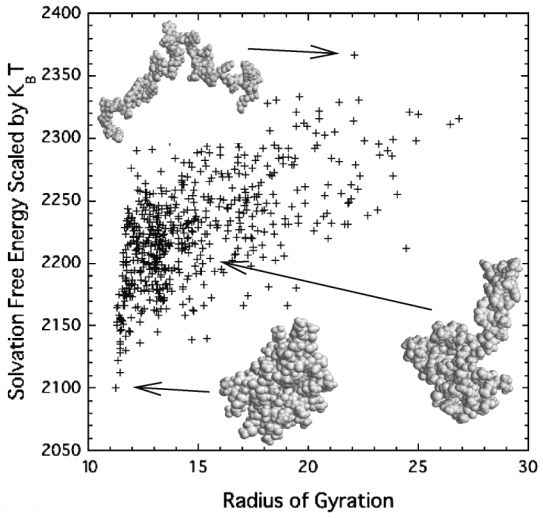

FIG. 1. Benchmark data $F_{\text{solvation}}^{\text{3D-HNC}}$ obtained from the 3D integral-equation theory for 600 different structures of protein G in a hard-sphere solvent as a function of the radius of gyration. We use a set of complex and quite distinct structures, as illustrated explicitly by three structures. For these calculations, we used a 3D grid with $\Delta x = \Delta y = \Delta z = 0.2 d_S$ and $N_x = N_y = N_z = 256$.

surface area, corresponding to a highly increased solvation free energy.

We now turn to our second approach, which is based on the morphometric form of thermodynamical potentials [8]. The idea of this approach is to predict the solvation free energy of the protein in a fixed structure based on only four geometrical measures and corresponding thermodynamical coefficients:

$$F_{\text{solvation}}^{\text{morph}} = pV + \sigma A + \kappa C + \bar{\kappa} X, \tag{4}$$

where V, A, C and X are the volume excluded by the protein, the surface area accessible to the solvent and the integrated mean and Gaussian curvatures of the accessible surface, respectively. The corresponding thermodynamic coefficients are the pressure p, the surface tension σ of the solvent at a planar wall, and two bending rigidities κ and $\bar{\kappa}$ which account for curvature effects [8]. The geometrical measures C and X are defined by

$$C \equiv \int_{\partial V} H dA, \quad \text{and} \quad X \equiv \int_{\partial V} K dA \tag{5}$$

as the integrated (over the surface area A) mean and Gaussian curvatures $H = (1/R_I + 1/R_{II})/2$ and $K = 1/(R_I R_{II})$, respectively. R_I and R_{II} are the two principal radii of curvature. In order to be able to uniquely calculate these geometrical measures we have to define the surface of the protein first. We employ the definition of the solvent accessible surface due to Lee and Richards [11] with the same hard-sphere diameters for the atoms of the protein and the solvent as employed in the integral-equation approach. The area A is then determined by the surface that is accessible to the centers of solvent spheres. All the geometrical measures and the thermodynamic coefficients are calculated with respect to this definition. V is the volume that is enclosed by the surface area. For a given structure of the protein, both A and V can be calculated *exactly* [12,13]. It is interesting to note that A and V are connected via a normal derivative

$$A = \partial_\varepsilon V = \lim_{\varepsilon \to 0} \frac{V_\varepsilon - V}{\varepsilon}, \tag{6}$$

where V_ε is the volume of the structure resulting by increasing the radius of each sphere by $\varepsilon \to 0$.

The calculation of the integrated curvatures is more involved. In addition to the surface contributions to C and X, there are contributions C_l and X_l from lines of intersecting spheres and point contributions X_p where three such lines meet [14]. The surface terms are readily calculated, as H and K are constant on the surface.

The line contributions C_l and X_l can be calculated by considering the curvature in a parallel surface displaced by an infinitesimal amount ε from the molecular surface. Following Ref. [14] we find

$$C_l = -\frac{\phi R_c}{2}(\theta_1 + \theta_2), \qquad X_l = -\phi(\sin\theta_1 + \sin\theta_2), \tag{7}$$

where ϕ is the angular length, ϕR_c the arc length of the intersection, and θ_1 and θ_2 the angles between the spheres and the plane of intersection as defined in Refs. [12,13]. The contribution X_p is the solid angle spanned by the surface normals at the intersection of three lines [14].

An important check for the correctness of the numerical calculation of the integrated curvatures is the property of X, the integrated Gaussian curvature, or Euler characteristics. $X = 4\pi N$, $N = 0, \pm 1, \pm 2, \ldots$ is a topological invariant and one finds $N = 1 - N_h + N_c$, where N_h counts the number of holes and N_c the number of cavities. To illustrate the meaning of these numbers, consider a torus, which has $N_h = 1$ and $N_c = 0$, so that its Euler characteristics vanish, and a hollow sphere, which has $N_h = 0$ and $N_c = 1$, so that $X = 8\pi$.

The morphometric form of the solvation free energy, Eq. (4), separates the geometry and the thermodynamical coefficients. This feature allows one to determine the values of p, σ, κ and $\bar{\kappa}$ in simple geometries. We determine these coefficients from calculations of the solvation free energy of spherical solutes with varying radius using a radial-symmetric integral-equation implementation. In principle, these coefficients can be determined using an arbitrary theoretical approach like integral-equation and density-functional theories or computer simulations. The nature of the solvent-solvent and protein-solvent interaction is reflected in the values of p, σ, κ and $\bar{\kappa}$.

For a hard-sphere solvent, we calculate the solvation free energy for the same structures of protein G that we studied using the 3D integral-equation approach. Figure 2 shows the deviation $D = 100(F_{\text{solvation}}^{\text{3D-HNC}} - F_{\text{solvation}}^{\text{morph}})/F_{\text{solvation}}^{\text{3D-HNC}}$ of

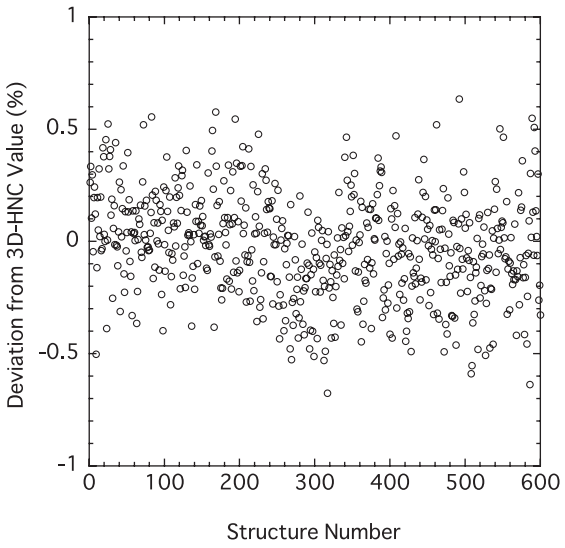

FIG. 2. The deviation D between $F_{\text{solvation}}^{\text{morph}}$ (the solvation free energy calculated by our morphometric approach) and $F_{\text{solvation}}^{\text{3D-HNC}}$ (the benchmark data from the 3D integral-equation theory) for 600 different structures shown in Fig. 1. Although the theories we compare are quite different, we find excellent agreement between their results. If we employ a even finer 3D grid, the magnitude of D reduces further.

the solvation free energy obtained by our morphometric approach from that presented in Fig. 1 obtained from the integral-equation theory for 600 structures of protein G. Although our two approaches are quite different in nature, the agreement between the 3D integral-equation theory and the morphometric approach is extremely good: the deviation $|D|$ is less than 0.7% in all cases. For most structures, the deviation is significantly smaller than this. This agreement is very encouraging because the computing time for $F_{\text{solvation}}$ is reduced by more than 4 orders of magnitude: The calculation for one structure of protein G is finished in less than 1 sec on a small personal computer.

To test if this deviation has a systematic origin or is numerical due to the grid size, we perform additional 3D integral-equation calculations for two representative structures. These calculations are done with $N_x = N_y = N_z = 512$ and $\Delta x = \Delta y = \Delta z = 0.1 d_S$, which is the largest possible grid resolution on our workstation. We find that the agreement between the results of our morphometric approach and the integral-equation theory improves significantly. For example, the deviation decreases from 0.41% to 0.05% for one structure and from 0.29% to -0.14% for the other.

In the model system we have employed here, all the system configurations share the same energy, and the behavior is purely entropic in origin. Nevertheless, the native structure is one of the most stable structures. This indicates the crucial importance of the solvent entropy in the structural stability of a protein [6]. We have recently found, however, that the hard-sphere solvent does not lead to the lowest solvation free energy of the native structure. There are some structures with lower solvation free energies. Clearly, water behaves differently than a hard-sphere fluid. It is easy to change the solvent properties in the morphometric approach, by recalculating the thermodynamic coefficients for different solvent-solvent and solute-solvent interactions. Preliminary calculations indicate that an additional attraction between solvent particles has significant effects on the solvation free energy and favors the native structure the most.

Having demonstrated the power of the morphometric approach for calculating $F_{\text{solvation}}$ for a complexly shaped molecule like a protein in different structures, we believe that further progress in understanding protein folding is within reach. Furthermore, this approach sheds new light on the extremely difficult problem of predicting the native structure of a protein.

This work was supported by Grants-in-Aid for Scientific Research on Priority Areas (No. 15076203) from the Ministry of Education, Culture, Sports, Science and Technology of Japan and by NAREGI Nanoscience Project. We wish to thank Professor S. Dietrich for promoting our collaboration.

[1] C. M. Dobson, Nature (London) **426**, 884 (2003).
[2] G. Salvi and P. de los Rios, Phys. Rev. Lett. **91**, 258102 (2003).
[3] A. Mitsutake, M. Kinoshita, Y. Okamoto, and F. Hirata, J. Phys. Chem. B **108**, 19 002 (2004).
[4] D. Beglov and B. Roux, J. Chem. Phys. **103**, 360 (1995); M. Ikeguchi and J. Doi, J. Chem. Phys. **103**, 5011 (1995); C. M. Cortis, P. J. Rossky, and R. A. Friesner, J. Chem. Phys. **107**, 6400 (1997); A. Kovalenko, F. Hirata, and M. Kinoshita, J. Chem. Phys. **113**, 9830 (2000).
[5] M. Kinoshita, J. Chem. Phys. **116**, 3493 (2002); M. Kinoshita, Chem. Phys. Lett. **387**, 47 (2004).
[6] Y. Harano and M. Kinoshita, Biophys. J. **89**, 2701 (2005).
[7] J.-P. Hansen and I. R. McDonald, *Theory of Simple Liquids* (Academic, London, UK 1986), 2nd ed.
[8] P. M. König, R. Roth, and K. R. Mecke, Phys. Rev. Lett. **93**, 160601 (2004).
[9] M. Kinoshita, S. Iba, K. Kuwamoto, and M. Harada, J. Chem. Phys. **105**, 7177 (1996); M. Kinoshita, J. Chem. Phys. **116**, 3493 (2002).
[10] Y. Okamoto, Recent Res. Dev. Pure & Appl. Chem. **2**, 1 (1998).
[11] B. Lee and F. M. Richards, J. Mol. Biol. **55**, 379 (1971).
[12] M. L. Connolly, J. Appl. Crystallogr. **16**, 548 (1983).
[13] M. L. Connolly, J. Am. Chem. Soc. **107**, 1118 (1985).
[14] K. M. Mecke, T. Buchert, and H. Wagner, Astron. Astrophys. **288**, 697 (1994).