J|A|C|S
ARTICLES
Published on Web 11/09/2006

# An Efficient Computational Method for Predicting Rotational Diffusion Tensors of Globular Proteins Using an Ellipsoid Representation

Yaroslav E. Ryabov,[†] Charles Geraghty,[†] Amitabh Varshney,[‡] and David Fushman*[,†]

*Contribution from the Department of Chemistry and Biochemistry, Center for Biomolecular Structure and Organization, University of Maryland, 1115 Biomolecular Sciences Building, College Park, Maryland 20742, and Department of Computer Science, UMIACS, University of Maryland, 4407 A.V. Williams Building, College Park, Maryland 20742*

Received May 10, 2006; E-mail: fushman@umd.edu

***Abstract:*** We propose a new computational method for predicting rotational diffusion properties of proteins in solution. The method is based on the idea of representing protein surface as an ellipsoid shell. In contrast to other existing approaches this method uses principal component analysis of protein surface coordinates, which results in a substantial increase in the computational efficiency of the method. Direct comparison with the experimental data as well as with the recent computational approach (Garcia de la Torre; et al. *J. Magn. Reson.* **2000**, *B147*, 138−146), based on representation of protein surface as a set of small spherical friction elements, shows that the method proposed here reproduces experimental data with at least the same level of accuracy and precision as the other approach, while being approximately 500 times faster. Using the new method we investigated the effect of hydration layer and protein surface topography on the rotational diffusion properties of a protein. We found that a hydration layer constructed of approximately one monolayer of water molecules smoothens the protein surface and effectively doubles the overall tumbling time. We also calculated the rotational diffusion tensors for a set of 841 protein structures representing the known protein folds. Our analysis suggests that an anisotropic rotational diffusion model is generally required for NMR relaxation data analysis in single-domain proteins, and that the axially symmetric model could be sufficient for these purposes in approximately half of the proteins.

## Introduction

Biological activity of proteins occurs mostly in solution, and it is hard to overestimate the importance of the diffusion phenomenon for protein function. Diffusion controls protein transport and could be the rate-limiting factor in a protein's interactions with its reaction counterparts. Knowledge of protein diffusion characteristics is also important for accurate interpretation of experimental data measured in protein solutions; fluorescence polarization, magnetic resonance and relaxation, dynamic light scattering, dielectric dispersion and relaxation, analytical ultracentrifugation are examples of experimental techniques susceptible to protein diffusion.

Here we focus on the rotational diffusion characteristics of proteins, which are of particular interest for applications of nuclear magnetic resonance (NMR) and fluorescence spectroscopy, techniques especially sensitive to the overall molecular tumbling in solution. In the limit of low-Reynolds number hydrodynamics, applicable to proteins in aqueous solution,[1,2] the overall tumbling of a protein molecule can be approximated by Brownian rotational diffusion of a rigid body characterized by its rotational diffusion tensor, $\underline{\underline{\mathbf{D}}}$. The rotational diffusion has tensorial properties because rotations about different directions in the molecule can proceed with different speed. The diffusion tensor can by represented by a 3-by-3 matrix[3−7] which is symmetric with respect to the transposition operation, $D_{i,j} = D_{j,i}$ ($i,j = 1,2,3$), such that only six elements of the matrix are independent. The principal values, $D_x$, $D_y$, and $D_z$, and the three corresponding principal vectors, $\mathbf{V}_x$, $\mathbf{V}_y$, and $\mathbf{V}_z$, of the tensor can be obtained from the set of three (eigenvalue) vector equations, $\underline{\underline{\mathbf{D}}}\mathbf{V}_l = D_l\mathbf{V}_l$, where $l = x,y,z$. In other words, the diffusion tensor is completely defined by its three principal values and three Euler angles ($0 \le \alpha \le 2\pi$, $0 \le \beta \le \pi$, $0 \le \gamma \le 2\pi$) that specify orientation of the reference frame (called the principal axes frame, PAF) attached to its principal vectors with respect to a given reference frame. The matrix representing the diffusion tensor is diagonal in the reference frame that coincides with the PAF of the tensor. In this case it has only

[†] Department of Chemistry and Biochemistry, Center for Biomolecular Structure and Organization.
[‡] Department of Computer Science, UMIACS.

(1) Purcell, E. M. *Am. J. Phys.* **1977**, *45*, 3−10.
(2) Berg, H. C. *Random Walks in Biology*; Princeton University Press: Princeton, NJ, 1983.
(3) Perrin, F. *J. Phys. Radium* **1934**, *5*, 497−511.
(4) Perrin, F. *J. Phys. Radium* **1936**, *7*, 1−11.
(5) Favro, D. L. *Phys. Rev.* **1960**, *119*, 53−62.
(6) Woessner, D. *J. Chem. Phys* **1962**, *37*, 647−654.
(7) Cantor, C. R.; Schimmel, P. R. *Biophysical Chemistry*; W. H. Freeman & Co: New York, 1980.


**10.1021/ja062715t CCC: $33.50** 

three nonzero components, $D_{1,1} = D_x$, $D_{2,2} = D_y$, and $D_{3,3} = D_z$. The physical meaning of the principal values of the diffusion tensor is that they are diffusion coefficients, i.e., they determine the spread of angular coordinates with time. Thus, if one considers a rotational diffusion process occurring about an axis that coincides with the principal vector $\mathbf{V}_l$, then the corresponding mean-square angular displacement, $\langle \Delta\varphi^2 \rangle_l$, in time $t$ is $\langle \Delta\varphi^2 \rangle_l = 2D_l\, t$.

Knowledge of the overall rotational diffusion tensor of a protein is critical for accurate characterization of protein dynamics from NMR measurements in solution[8,9] and analysis of protein oligomerization,[10] and could be indispensable for structure and dynamics determination in multidomain systems and molecular complexes,[11−13] when other structural methods fail. In many cases the individual components of a protein rotational diffusion tensor can be measured, using for example NMR relaxation experiments.[11,14−17] However, such measurements are not always possible because they require relatively high concentrations of isotope-enriched proteins and are limited by the size of a protein amenable to NMR measurements. Therefore, there is significant interest in methods for accurate prediction of the diffusion tensor of any protein (or any other molecule) from its three-dimensional structure.

A number of theoretical models had been developed for predicting the diffusion properties of proteins (reviewed in ref 18). Various bead and shell models[19−33] are based on a "bead concept" which represents a protein as a set of $N$ spherical friction elements, beads. The diffusion tensor is then calculated from the frictional forces derived by inverting a $3N \times 3N$ supermatrix composed of pairwise hydrodynamics interaction tensors between the beads. In some modifications of this approach, the protein's surface is treated as a set of some finite elements, e.g., triangles, and the diffusion tensor is then obtained in a similar way, as a result of summation over these elements.[34−37] There were also attempts[38] to estimate diffusion properties using molecular dynamics simulations. Besides the obvious advantage of these models, in that they provide a detailed and fairly accurate representation of the protein's shape/surface by a large number of some small elements, all of them require a significant amount of CPU time, proportional to $N^3$.

A different, much older concept, originating from the pioneering works of Einstein,[39] Debye,[40] and Perrin,[3,4] represents a solute molecule as a rigid object of simple geometrical shape, such as a sphere, ellipsoid, cylinder, etc.,[7] for which there is an exact analytical expression relating diffusion characteristics and the dimensions of such an object. In particular, the so-called ellipsoid models[41,42] represent the shape of a protein molecule as an (triaxial) ellipsoid. Because the rotational diffusion tensor for a molecule of any, even very complex, shape is determined by only six independent components—the same number of independent parameters as is necessary to define an ellipsoid—it is natural to anticipate that with regard to rotational diffusion every protein molecule can be represented using an equivalent ellipsoid. Conceptually, this establishes a direct correspondence between the tumbling of an arbitrarily shaped protein and that of an ellipsoid. Since the diffusion tensor of an ellipsoid can be easily calculated, the problem of predicting the diffusion properties of a given protein can then be reduced to the problem of finding its equivalent ellipsoid representation. The appeal of this approach is in its generality and conceptual simplicity. Its practical implementation, however, requires a method for constructing such an ellipsoid.

A typical approach to building the ellipsoidal representation of a protein shape uses the concept of inertia-equivalent ellipsoid, i.e., approximates the protein molecule with an ellipsoid that has the same components of the inertia tensor.[41,43] Such an approach provides a representation of a protein that is relevant to analysis of small-angle X-ray scattering[43] and other physical measurements dealing with the static rather than dynamic properties of proteins. It is worth emphasizing, however, that because of the small size of a protein molecule its inertia properties are practically irrelevant to its global motion in solution, since in this case molecular tumbling and translation both are fully controlled by the viscous/frictional forces (which is the essential condition of the diffusion regime).[1,2,7] Instead, it is the shape of the molecule and the area of its solvent-accessible surface that are relevant, because all interactions with the solvent responsible for the friction occur at the solute's surface. Thus, for example, a (hollow) shell that is an exact replica of the solute's surface would have the same tumbling properties in solution as the entire molecule. (Note that, as in most hydrodynamic models thus far, we consider the protein molecule as a rigid object.) Therefore, for diffusion tensor prediction, the equivalent ellipsoid model must be based on the


(8) Fushman, D.; Cowburn, D. In *Structure, Motion, Interaction and Expression of Biological Macromolecules*; Sarma, R., Sarma, M., Eds.; Adenine Press: Albany, NY, 1998; pp 63−77.
(9) Hall, J. B.; Fushman, D. *J. Biomol. NMR* **2003**, *27*, 261−275.
(10) Bernado, P.; Akerud, T.; Garcia de la Torre, J.; Akke, M.; Pons, M. *J. Am. Chem. Soc.* **2003**, *125*, 916−923.
(11) Fushman, D.; Xu, R.; Cowburn, D. *Biochemistry* **1999**, *38*, 10225−10230.
(12) Fushman, D.; Varadan, R.; Assfalg, M.; Walker, O. *Prog. Nucl. Magn. Reson. Spectrosc.* **2004**, *44*, 189−214.
(13) Ryabov, Y.; Fushman, D. *Proteins* **2006**, *63*, 787−796.
(14) Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A. *J. Am. Chem. Soc.* **1995**, *117*, 12562−12566.
(15) Dosset, P.; Hus, J. C.; Blackledge, M.; Marion, D. *J. Biomol. NMR* **2000**, *16*, 23−28.
(16) Ghose, R.; Fushman, D.; Cowburn, D. *J. Magn. Reson.* **2001**, *149*, 214−217.
(17) Walker, O.; Varadan, R.; Fushman, D. *J. Magn. Reson.* **2004**, *168*, 336−345.
(18) Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B. *Biophys. J.* **2000**, *78*, 719−730.
(19) Bloomfield, V. A.; Dalton, W. O.; Van, Holde, K. E. *Biopolymers* **1967**, *5*, 135−148.
(20) Bloomfield, V. A. *Science* **1968**, *161*, 1212−1219.
(21) Teller, D. C.; Swanson, E.; de Haen, C. *Methods Enzymol.* **1979**, *61*, 103−124.
(22) Garcia de la Torre, J.; Bloomfield, V. A. *Q. Rev. Biophys.* **1981**, *14*, 81−139.
(23) Muller, J. J. *J. Appl. Crystallogr.* **1983**, *16*, 74−82.
(24) Pavlov, M. Y.; Sinev, M. A.; Timchenko, A. A.; Ptitsyn, O. B. *Biopolymers* **1986**, *25*, 1385−1397.
(25) Venable, R. M.; Pastor, R. W. *Biopolymers* **1988**, *27*, 1001−1014.
(26) Antosiewicz, J.; Porschke, D. *J. Phys. Chem.* **1989**, *93*, 5301−5305.
(27) Antosiewicz, J.; Grycuk, T.; Porschke, D. *J. Chem. Phys.* **1991**, *95*, 1354−1360.
(28) Antosiewicz, J.; Porschke, D. *Biophys. J.* **1995**, *68*, 655−664.
(29) Antosiewicz, J.; Porschke, D. *J. Phys. Chem.* **1993**, *97*, 2767−2773.
(30) Byron, O. *Biophys. J.* **1997**, *72*, 408−415.
(31) Byron, O. *Methods Enzymol.* **2000**, *321*, 278−304.
(32) Zipper, P.; Durchschlag, H. *Biochem. Soc. Trans.* **1998**, *26*, 726−731.
(33) Hellweg, T.; Eimer, W.; Krahn, E.; Schneider, K.; Muller, A. *Biochim. Biophys. Acta* **1997**, *1337*, 311−318.
(34) Brune, D.; Kim, S. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 3835−3839.
(35) Allison, S. A.; Tran, V. T. *Biophys. J.* **1995**, *68*, 2261−2270.
(36) Chae, K. S.; Lenhoff, A. M. *Biophys. J.* **1995**, *68*, 1120−1127.
(37) Zhou, H. X. *Biophys. J.* **1995**, *69*, 2286−2297.
(38) Smith, P. E.; Vangunsteren, W. F. *J. Mol. Biol.* **1994**, *236*, 629−636.
(39) Einstein, A. *Ann. Phys. (Leipzig)* **1906**, *19*, 289−306.
(40) Debye, P. *Ber. Deutsch Phys. Ges.* **1913**, *15*, 777.
(41) Taylor, W. R.; Thornton, J. M.; Turnell, W. G. *J. Mol. Graphics* **1983**, *1*, 30−38.
(42) Harding, S. E.; Horton, J. C.; Jones, S.; Thornton, J. M.; Winzor, D. J. *Biophys. J.* **1999**, *76*, 2432−2438.
(43) Muller, J. J.; Schrauber, H. *J. Appl. Crystallogr.* **1992**, *25*, 181−191.

properties of the protein's surface rather than those of the bulk protein. Note, in this regard, that the inertia tensor is sensitive to both molecular shape and mass distribution (internal atomic packing) and, therefore, could provide an inaccurate representation of those characteristics, such as the diffusion tensor, which depend solely on the surface/shape of the molecule (see Supporting Information).

In this paper we propose an efficient and robust way of building such an ellipsoid representation, based on the principal component analysis of the protein surface. The obvious advantage of this approach is in its conceptual simplicity and computational efficiency. Naturally, when replacing the actual, irregular surface of a protein with a smooth-surface ellipsoid shell, some fine details of the molecular surface get lost; thus, one might think that such a model is too rough to provide an accurate representation of the protein's surface. It could be argued, however, that the frictional torque experienced by a tumbling protein is averaged over the many frictional elements on its surface and over the subnanosecond and slower dynamics of the solvent-exposed side chains; therefore, its resulting diffusion tensor reflects some averaged surface properties. Here we present a comparison with the atomic-resolution model based on the HYDRONMR program[44] and show that not only does our method represent the experimental data with comparable accuracy to that of the bead model, but it also provides a significant, more than 2 orders of magnitude speedup in the calculations. We also use a large representative set of known protein folds in order to compare the two methods as well as to gain insights into the distribution of rotational diffusion tensors in single-domain proteins.

## Modeling Rotational Diffusion of Anisotropic Molecule

The first comprehensive theoretical investigation of rotational diffusion properties of an anisotropic molecule was made by Perrin.[3,4] In his classical papers Perrin calculated the frictional coefficients, $f_x$, $f_y$, $f_z$, for a rigid ellipsoidal body revolving in a continuous viscous medium and related them to the corresponding principal components of the rotational diffusion tensor as:

$$D_l = \frac{k_B T}{f_l} \tag{1}$$

where $k_B$ is the Boltzmann constant, $l = x,y,z$, $T$ is the absolute temperature in Kelvins, and

$$f_x = \frac{16\pi\eta(a_y^2 + a_z^2)}{3(a_y^2 Q + a_z^2 R)} \tag{2a}$$

$$f_y = \frac{16\pi\eta(a_x^2 + a_z^2)}{3(a_z^2 R + a_x^2 P)} \tag{2b}$$

$$f_z = \frac{16\pi\eta(a_x^2 + a_y^2)}{3(a_x^2 P + a_y^2 Q)} \tag{2c}$$

where $\eta$ is the solvent viscosity, $a_x$, $a_y$, and $a_z$ are the semi-axes of the ellipsoid (in general, $a_x \neq a_y \neq a_z$), and the parameters $P$, $Q$, and $R$ are defined by the following equations:

$$P = \int_0^\infty \frac{ds}{\sqrt{(a_x^2 + s)^3(a_y^2 + s)(a_z^2 + s)}} \tag{3a}$$

$$Q = \int_0^\infty \frac{ds}{\sqrt{(a_y^2 + s)^3(a_z^2 + s)(a_x^2 + s)}} \tag{3b}$$

$$R = \int_0^\infty \frac{ds}{\sqrt{(a_z^2 + s)^3(a_x^2 + s)(a_y^2 + s)}} \tag{3c}$$

which can be expressed via elliptic integrals.[45] Importantly, Perrin's analysis also showed that the directions of the principal vectors for the rotational diffusion tensor coincide with the directions of the semi-axes of the ellipsoid.

**(a) HYDRONMR.** Since the pioneering work of Bloomfield and coauthors[19,20] the diffusion properties of biological macromolecules were modeled using bead approximation, representing the molecule or its surface by a large number of spherical frictional elements, beads. The hydrodynamic properties of the bead model of a protein can then be calculated using the theoretical formalism proposed by Garcia de la Torre and Bloomfield[22] and further developed in a number of publications.[21−33,46] A recent advancement of this approach, the so-called "shell" model,[47] focuses only on the surface of a protein, modeled using a shell of frictional "mini-beads" and extrapolated to a zero-size-bead limit. This method, developed by Garcia de la Torre and co-workers, is implemented in a number of computer programs for calculating hydrodynamic properties of proteins. HYDRONMR,[44] a program from this group available in public domain, was shown to reproduce with reasonable accuracy the overall tumbling time for a set of 14 globular proteins. This approach is now considered a well-established method for calculating hydrodynamic properties of proteins, and we will use HYDRONMR here as control for evaluating the performance of our method. While this manuscript was in preparation, a new program called FAST-HYDRONMR became available, which speeds up the calculations by replacing a rigorous treatment of the bead model with an empirical approximation.[48] In this paper, however, we compare our method primarily with HYDRONMR, because this method provides an exact solution to hydrodynamic equations.

**(b) An Approach Based on Principal Component Analysis of Protein Surface.** As already mentioned in the introduction, any diffusion tensor is completely defined by six parameters, for example, the principal values of the tensor and the Euler angles that specify the orientation of the tensor with respect to a given coordinate frame. At the same time, Perrin's equations (eqs 1−3) show that for a given ellipsoid, one can always find components of the corresponding rotational diffusion tensor. An ellipsoid is also completely defined by only six parameters: the lengths of its semi-axes and the Euler angles that specify the orientation of these three mutually orthogonal semi-axes with respect to a given coordinate frame. In other words, eqs 1−3 establish a direct mapping between an ellipsoid and a diffusion tensor. Therefore, at least in principle, for any rigid body one can always find an ellipsoid that has the same rotational diffusion tensor as a given body. This means that an ellipsoidal representation should be sufficient as a model of the rotational diffusion properties of any rigid body.

**How To Build an Equivalent Ellipsoid.** It is clear from eqs 1−3 that the only parameters that matter for determining the diffusion tensor (except the temperature and the solvent viscosity) are the ellipsoid's semi-axes. When the diffusion tensor is known, the corresponding equivalent ellipsoid can be constructed by finding the values of $a_x$, $a_y$, and $a_z$ that, when plugged into eqs 1−3, result in the desired values of

(44) Garcia de la Torre, J.; Huertas, M. L.; Carrasco, B. *J. Magn. Reson.* **2000**, *B147*, 138−146.
(45) Gradshteyn, I. S.; Ryzhik, I. M. *Table of Integrals, Series, and Products*; Academic Press: New York, 1980.
(46) Garcia de la Torre, J.; Navarro, S.; Martinez, M. C. L.; Diaz, F. G.; Cascales, J. J. L. *Biophys. J.* **1994**, *67*, 530−531.
(47) Carrasco, B.; Garcia de la Torre, J. *Biophys. J.* **1999**, *75*, 3044−3057.
(48) Ortega, A.; Garcia de la Torre, J. *J. Am. Chem. Soc.* **2005**, *127*, 12764−12765.

the tensor. This problem can be solved numerically. However, if the diffusion tensor is unknown, the problem of finding such an ellipsoid becomes nontrivial, and the critical issue here is how to build the equivalent ellipsoid directly from the atom coordinates of a protein.

In this paper we propose a solution to this problem. Our approach is based on the hypothesis that the equivalent ellipsoid can be derived directly from the shape/topography of protein's surface. A simple heuristic background for this idea is the following: since the hydrodynamic friction happens at the protein's surface, the size and shape/topology of this surface should be the determining factors for protein's diffusion properties. In addition to these common-sense considerations, this hypothesis is also supported by the observed correlation between the rotational correlation time of a protein and its solvent-accessible surface area.[49]

Our approach to building the equivalent ellipsoid consists of two steps. First, we create a three-dimensional representation of the protein surface, based on the atom coordinates. Second, we create an ellipsoid that provides the best fit to this surface. The details of this procedure are outlined below.

**Step 1.** Calculation of the solvent-accessible surface of a protein is a solved problem in molecular graphics.[50,51] We use for this purpose the algorithm created by Varshney and co-workers[52] and implemented in the SURF program. This program efficiently creates a three-dimensional model of the protein surface "seen" by the solvent molecule of a given radius, by tessellating this surface with triangles, as illustrated in Figure 1. The vertices of these triangles are then treated as the points on such a surface.

**Step 2.** Once the surface coordinates are available, the essential question is, how to create an equivalent ellipsoid given the coordinates of protein's surface? This problem of representing a cloud of points by an equivalent ellipsoid has been well studied in several fields including pattern classification.[53] It has been shown that the principal component analysis (PCA) of a set of three-dimensional points gives the three principal vectors, along which the variability of that point set is the largest (in the least-sum-of-squared-error sense), and the corresponding variances.[54] These principal directions and principal variances can be used to represent the best-fitting ellipsoid to the input set of three-dimensional points. Applied to our problem, this approach can be formulated as follows. Let $\{X_j^l\}$ ($j = 1,...N_p$, where $N_p$ is the total number of points and $l = x, y, z$ as above) be a set of coordinates of $N_p$ points that define the surface. Then the covariance matrix, **C**, is defined as

$$C_{m,n} = \frac{1}{N_p}\sum_{j=1}^{N_p}(X_j^m - \langle X^m\rangle)(X_j^n - \langle X^n\rangle) \tag{4}$$

where $m,n = x,y,z$ and

$$\langle X^m\rangle = \frac{1}{N_p}\sum_{j=1}^{N_p}X_j^m \tag{5}$$

are the mean values of the corresponding coordinates. The principal components, $E^n$, and the principal vectors, $\mathbf{S}^n$, of the covariance matrix are solutions to the eigenvalue vector equations:

$$\mathbf{C}\mathbf{S}^n = E^n\mathbf{S}^n \tag{6}$$

The principal component analysis establishes that an ellipsoid with the origin at $\{\langle X^x\rangle,\langle X^y\rangle,\langle X^z\rangle\}$ and with the semi-axes $a_n = \sqrt{3E^n}$ oriented along the principal vectors $\mathbf{S}^n$ is the best-fit ellipsoid representation for this surface. In rigorous mathematical terms, the coordinates $\{F_j^l\}$ of the corresponding points of the equivalent ellipsoidal surface minimize the following sum:

$$\sum_{j=1}^{N_p}\sum_{n=x,y,z}(G_j^n - F_j^n)^2/a_n^2$$

where coordinates $\{G_j^n\}$ are obtained from the surface coordinates $\{X_j^l\}$ by the rotation that transforms the initial reference frame into the reference frame associated with the principal vectors of the covariance matrix.[53,54]

The ellipsoid created as a result of PCA for the set of vertex coordinates of the triangulation mesh generated by SURF can then be considered as the equivalent ellipsoid representation for the protein surface. Figure 1 illustrates the construction of an equivalent ellipsoid. Once such an ellipsoid is obtained, the principal components of the diffusion tensor can be calculated directly from eqs 1−3, given the solvent viscosity and temperature.

**(c) Hydration Layer.** Proteins are always hydrated in aqueous solution, i.e., are covered with water molecules that associate with the protein in various ways, ranging from hydrogen-bonded water to water trapped in holes and cavities on the protein surface. These water molecules form the so-called hydration layer, responsible for many physical properties of proteins in solution.[7] Overall tumbling is particularly sensitive to the size of the protein; therefore, a proper account for the hydration effect is critical for accurate prediction of the rotational diffusion tensor. The amount of hydration-layer water is often estimated[7] at about 0.3−0.4 (g $H_2O$)/(g protein), and various models had been proposed to reproduce this effect. The estimates of the thickness of the hydration layer reported in the literature vary, depending on the physical nature of the measurement, the protein studied, and the model used for the interpretation of the data. For example, dielectric measurements[55] gave 1840 water molecules per albumin and 450 per lysozyme (interpreted in ref 55 as 1.7 and 1.25 water layers, respectively), while NMR measurements resulted in a hydration layer of 3.75 Å (translational diffusion of the RSV Gag M domain[56]) and 3.5 Å (rotational diffusion of ubiquitin;[14] the latter result was interpreted in ref 14 as a half-layer of water). Perhaps the simplest way to account for the hydration effect in a computer program is as follows: instead of a "dry" protein, where every atom is represented by a sphere of the corresponding van der Waals radius, one considers a "wet" protein structure where the sizes of all atoms are inflated according to a certain rule. In particular, HYDRONMR uses the so-called atomic element radius (AER), which is an adjustable parameter in that model, and all atoms in a wet protein are assumed to have the same AER. SURF program allows, in principle, any size for every atom in a protein and, in addition, allows the user to specify the size of the solvent molecule. In the current study, however, when using SURF to model the hydration layer effect, we increased the radii of all protein atoms by the same amount. The corresponding parameter, called hydration layer thickness (HLT), is an adjustable parameter in our PCA-based model. Parts a and b of Figure 1 show examples of such dry and wet protein surfaces.

The obvious consequence of a hydration layer is that not only the surface area of a wet protein but also its surface topography are different from those of a dry protein. Thus, to avoid any misunderstanding, we distinguish here between the solvent-accessible surface (SAS) and the

(49) Krishnan, V. V.; Cosman, M. *J. Biomol. NMR* **1998**, *12*, 177−182.
(50) Lee, B.; Richards, F. M. *J. Mol. Biol.* **1971**, *55*, 379−400.
(51) Connolly, M. *Science* **1983**, *221*, 709−713.
(52) Varshney, A.; Brooks, F. P., Jr.; Wright, W. V. *IEEE Comput. Graphics Appl.* **1994**, *14*, 19−25.
(53) Duda, R. O.; Hart, P. E. *Pattern Classification and Scene Analysis*; John Wiley & Sons, Inc.: New York, 1973.
(54) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986.
(55) Rejou-Michel, A.; Henry, F.; de Villardi, M.; Delmotte, M. *Phys. Med. Biol.* **1985**, *30*, 831−837.
(56) McDonnell, J. M.; Fushman, D.; Cahill, S. M.; Zhou, W.; Wolven, A.; Wilson, C. B.; Nelle, T. D.; Resh, M. D.; Wills, J.; Cowburn, D. *J. Mol. Biol.* **1998**, *279*, 921−928.

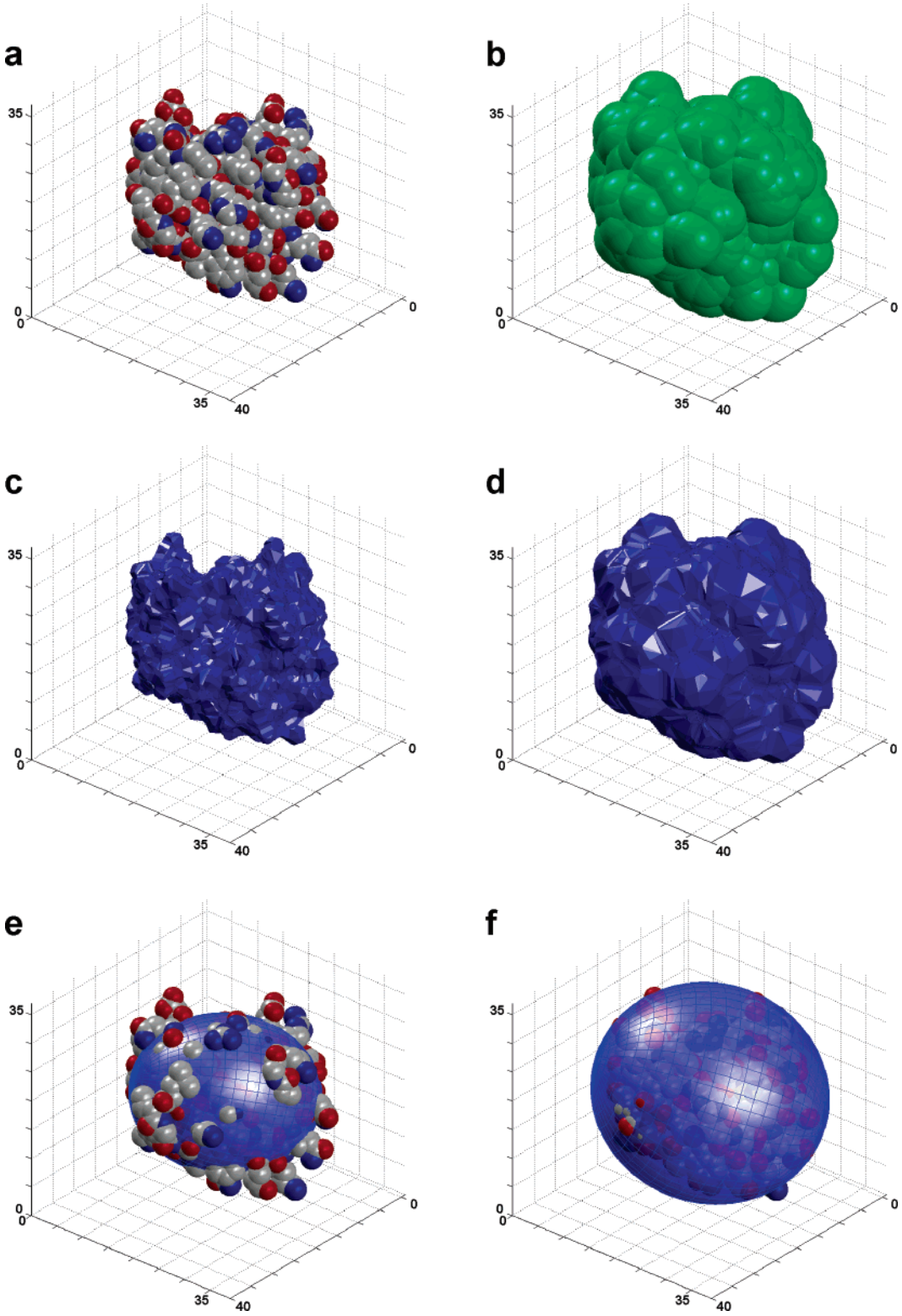


***Figure 1.*** Illustration of the construction of an equivalent ellipsoid for ubiquitin (PDB entry 1ubq). The unstructured and flexible C-terminus (residues 71−76) was not included. (a) Structure of the protein without hydration layer: the blue balls represent nitrogen atoms, the gray balls are carbons, and the red balls are oxygen atoms (Van der Waals radii are 1.75, 1.85, and 1.60 Å, respectively). (b) Structure of the protein covered with a hydration layer. The radii of all atoms were increased by HLT = 2.8 Å, which resembles a monolayer of water molecules. (c, d) Triangulation for the protein surface generated by the SURF program:[52] (c) the triangulation mesh for the protein without hydration layer consists of 17880 triangles, HLT = 0.0 Å (SAS); (d) the triangulation mesh for the protein with hydration layer consists of 9936 triangles, HLT = 2.8 Å (HSAS). When mapping these surfaces, the radius of the solvent molecule was set to 1.4 Å, which corresponds approximately to the radius of water molecule. (e, f) The positioning of equivalent ellipsoids with respect to the protein structure. The equivalent ellipsoids for proteins without (e) and with (f) hydration layer were obtained using the principal component analysis (PCA) from triangulated surfaces presented in (c) and (d).

hydrated solvent-accessible surface (HSAS). SAS is the surface of a dry protein that can be reached by a solvent molecule of a certain size (an example of such a surface is shown in Figure 1c). In contrast, HSAS is the surface of a wet protein that can be reached by a solvent molecule

***Table 1.*** Comparison between Experimentally Measured (NMR, Fluorescence) and Calculated Values of the Rotational Correlation Time $\tau_c$ [a]

| | | | PCA | | HYDRONMR | | PCA | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | variable HLT | | AER = 3.2 Å | | HLT = 2.8 Å | | HLT = 0.0 Å | |
| protein | PDB code | experiment[b] $\tau_c$ exp. [ns] | HLT Å | $\tau_c$ calc. [ns] | $\tau_c$ calc. [ns] | diff. % | $\tau_c$ calc. [ns] | diff. % | $2\tau_c$ calc. [ns] | diff. % |
| malate synthase G[67] | 1y8b | 55.3 (N) | 4.6 | 55.38 | 54.6 | −1 | 45.7 | −17 | 47.7 | −14 |
| human serum albumin[68] | 1ao6 | 41.0 (F) | 2.8 | 40.87 | 53.0 | 29 | 40.9 | 0 | 45.7 | 11 |
| maltose binding protein[69] | 1ezp | 28.6 (N) | 2.9 | 28.65 | 35.5 | 24 | 28.2 | −1 | 27.6 | −3 |
| $\beta$-lactoglobulin a (dimer)[61] | 1bsy | 23.2 (F) | 3.1 | 23.13 | 25.9 | 12 | 22.2 | −4 | 23.0 | −1 |
| $\Delta^5$-3 ketosteroid isomerase[70] | 1buq | 18.0 (F,N) | 3.2 | 17.96 | 20.4 | 13 | 16.7 | −7 | 17.6 | −2 |
| leukemia inh. factor [44] | 1lki | 14.9 (N) | 4.4 | 14.93 | 12.6 | −15 | 11.6 | −22 | 11.8 | −21 |
| trypsin[61] | 2blv | 14.8 (F) | 4.1 | 14.76 | 13.0 | −12 | 12.6 | −15 | 14.4 | −3 |
| yellow fluorescent protein[71] | 2yfp | 14.8 (F) | 3.4 | 14.9 | 14.1 | −5 | 13.5 | −9 | 13.6 | −8 |
| green fluorescent protein[72] | 1w7s | 14.2 (F) | 2.8 | 14.24 | 15.8 | 11 | 14.2 | 0 | 14.6 | 3 |
| carbonic anhydrase[61] | 2cab | 14.0 (F) | 2.4 | 13.9 | 16.2 | 16 | 14.9 | 6 | 16.5 | 18 |
| HIV-1 protease[73] | 1bvg | 13.0 (N) | 2.8 | 12.94 | 13.5 | 4 | 12.9 | −1 | 14.2 | 9 |
| savinase[44] | 1svn | 12.4 (N) | 2.3 | 12.38 | 13.6 | 10 | 13.3 | 7 | 15.2 | 23 |
| interleukin-1$\beta$[44] | 6i1b | 12.4 (N) | 3.5 | 12.43 | 12.0 | −3 | 11.1 | −10 | 11.4 | −8 |
| ribonuclease H[58] | 2rn2 | 11.7 (N) | 3.5 | 11.68 | 11.6 | −1 | 10.4 | −11 | 11.3 | −3 |
| cytochrome $c_2$[74] | 1c2n [c] | 10.4 (N) | 4.6 | 10.40 | 7.9 | −24 | 7.7 | −26 | 7.5 | −28 |
| $\beta$-lactoglobulin A (mono)[61] | 1bsy | 9.7 (F) | 2.4 | 9.79 | 11.4 | 18 | 10.5 | 8 | 10.7 | 10 |
| apomyoglobin[61] | 1bvc | 9.5 (F) | 2.4 | 9.55 | 10.2 | 9 | 10.3 | 10 | 10.5 | 12 |
| lysozyme[44] | 1hwa | 8.3 (N) | 2.2 | 8.26 | 9.6 | 16 | 9.2 | 11 | 9.2 | 11 |
| barstar C40/83A[44] | 1bta | 7.4 (N) | 3.9 | 7.33 | 6.1 | −18 | 6.0 | −19 | 6.1 | −18 |
| eglin c[44] | 1egl [d] | 6.2 (N) | 3.4 | 6.23 | 6.0 | −3 | 5.6 | −10 | 5.4 | −13 |
| cytochrome $b_5$[44] | 1wdb | 6.1 (N) | 3.0 | 6.10 | 6.5 | 7 | 5.9 | −3 | 5.4 | −11 |
| calbindin-D9k+$Ca^{2+}$ [44] | 2bca | 5.1 (N) | 2.9 | 5.07 | 5.1 | 0 | 5.0 | −2 | 4.9 | −4 |
| ubiquitin[14] | 1ubq [e] | 5.0 (N) | 2.9 | 4.98 | 5.0 | 0 | 4.9 | −2 | 4.8 | −4 |
| calbindin-D9k[44] | 1clb [d] | 4.9 (N) | 2.3 | 4.88 | 5.2 | 6 | 5.4 | 10 | 5.3 | 8 |
| BPTI[44] | 1pit [d] | 4.4 (N) | 2.6 | 4.41 | 4.8 | 9 | 4.6 | 5 | 4.7 | 7 |
| Protein G[9] | 1igd [f] | 3.7 (N) | 2.6 | 3.74 | 3.9 | 5 | 3.9 | 5 | 3.4 | −8 |
| Xfin-zinc finger DBD[44] | 1znf [d] | 2.4 (N) | 3.2 | 2.37 | 2.0 | −17 | 2.2 | −8 | 1.9 | −21 |
| mean abs. value | | | 3.1(0.7)[g] | | | 11% | | 8% | | 10% |

[a] All experimental data were rescaled to 293 K, when necessary. All calculations were done for this temperature. AER is the adjustable parameter in HYDRONMR; HLT is the thickness of hydration layer assumed in our PCA-based model. Hydrogen atoms were included here in the PCA-based calculations for all NMR-derived structures. See also Supporting Information Table 3. [b] The rescaling to 293 K was performed, assuming that $\tau_c$ scales with temperature as $\eta(T)/T$, where $\eta(T) = 1.7753 - 0.0565\cdot(T - 273) + 1.0751 \times 10^{-3}\cdot(T - 273)^2 - 9.2222 \times 10^{-6}\cdot(T - 273)^3$.[75] In this column, "N" and "F" in the parentheses indicate NMR and fluorescence data, respectively. [c] From the set of 20 NMR structures for this protein we took structure number 2. [d] The original PDB files for these proteins contain multiple structures. In this study we used the first structure. [e] Ubiquitin has a flexible C-terminus,[14] therefore for the calculations presented here we clipped the last five residues from the PDB file. [f] The protein G construct used in this study has a five-residue deletion at the N-terminus.[9] Therefore, the first five residues were deleted from the original PDB file. [g] The value in the parentheses represents the standard deviation.

of the same size. Figure 1d shows an example of such a friction-relevant accessible surface. The radius of the solvent molecule in this study was set to 1.4 Å, approximately equal to the radius of a water molecule.[57]

## Representative Sets of Protein Structures for the Analysis

Two representative sets of protein structures were used in this study: set A for testing the accuracy of the proposed method by comparison with experimental data and the other, set B, for a comprehensive comparison between HYDRONMR and our model and for statistical analysis of rotational diffusion tensors in single-domain proteins. Figure 1 depicts one particular example, a 76-amino acid (a.a.) protein ubiquitin, which is present in both data sets.

**Set A.** The accuracy of the prediction methods discussed above is validated here by comparison with experimental (NMR and fluorescence) data for a set of 27 globular proteins (Table 1). This experimental data set comprises 13 proteins used earlier for testing the HYDRONMR model,[44] augmented with more recent solution NMR data, including very high molecular weight proteins, and with nine proteins from the fluorescence literature. This collection of proteins covers the range of molecular weights from 2.9 kDa (Xfin-zinc finger DBD) to 82 kDa (malate synthase G). Most of these data include only the values of experimentally measured overall tumbling time, $\tau_c$, and only for five proteins from this set are the individual components of their diffusion tensor available (Tables 1 and 2).

**Set B.** A second, much larger set of 841 protein structures from the Protein Data Bank (PDB) is used here for a detailed comparison between the shell and ellipsoid models and for the purpose of statistical analysis of the predicted diffusion tensors for globular single-domain proteins. These structures were selected to represent the variety of known protein folds. The selection criteria, described earlier,[58,59] were as follows: these are structures of single proteins at least 30 a.a. long, with less than 40% sequence identity or more than 30% or 30 a.a. length difference from other set members. The set includes crystal structures at ≤3 Å resolution and solution (NMR) structures. A complete list can be found in the Supporting Information. This list of proteins was generously provided by Dr. A. Šali (UCSF) and covers the range of molecular weights from 3 to

(57) Eisenberg, D.; Kauzmann, W. *The Structure and Properties of Water*; Oxford University Press: New York, 1969.

(58) Sali, A.; Potterton, L.; Yuan, F.; van Vlijmen, H.; Karplus, M. *Proteins* **1995**, *23*, 318−326.

(59) Fushman, D.; Ghose, R.; Cowburn, D. *J. Am. Chem. Soc.* **2000**, *122*, 10640−10649.

***Table 2.*** Detailed Comparison between Experimentally Measured and Predicted Rotational Diffusion Parameters Using the Shell Model (HYDRONMR) and the Principal Component Analysis (PCA) in Conjunction with Perrin's eqs (1−3)[a]

| | parameters | $\tau_c$, [ns] | $D_x$, [$10^{-7}$ s$^{-1}$] | $D_y$, [$10^{-7}$ s$^{-1}$] | $D_z$, [$10^{-7}$ s$^{-1}$] | $(2D_z)/(D_x+D_y)$ | $D_y/D_x$ | α (deg) | $\beta$ (deg) | $\gamma$ (deg) |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Protein G, 1igd.pdb [b] | | | | | | |
| experiment [c] | | 3.7 | 3.73 | 4.15 | 5.63 | 1.43 | 1.12 | 85 | 68 | 179 |
| HYDRONMR | AER = 3.2 Å | 3.9 | 3.67 | 3.81 | 5.19 | 1.39 | 1.04 | 80 | 73 | 88 |
| PCA | HLT = 2.8 Å | 3.9 | 3.89 | 4.05 | 4.83 | 1.22 | 1.04 | 84 | 80 | 171 |
| PCA | HLT = 0.0 Å | 3.4 [g] | 4.26 [h] | 4.36 [h] | 6.10 [h] | 1.41 | 1.02 | 69 | 75 | 174 |
| | | | | Ubiquitin, 1ubq.pdb [d] | | | | | | |
| experiment [c] | | 5.0 | 3.12 | 3.21 | 3.67 | 1.16 | 1.03 | 47 | 40 | −17 |
| HYDRONMR | AER = 3.2 Å | 5.0 | 3.14 | 3.28 | 3.61 | 1.12 | 1.05 | 36 | 56 | −18 |
| PCA | HLT = 2.8 Å | 4.9 | 3.24 | 3.37 | 3.61 | 1.09 | 1.04 | 51 | 67 | −42 |
| PCA | HLT = 0.0 Å | 4.8 [g] | 3.29 [h] | 3.45 [h] | 3.79 [h] | 1.12 | 1.05 | 22 | 73 | −52 |
| | | | | Cytochrome *c*2, 1c2n.pdb [e] | | | | | | |
| experiment [c] | | 10.4 | 1.42 | 1.62 | 1.77 | 1.17 | 1.14 | −21 | 21 | −31 |
| HYDRONMR | AER = 3.2 Å | 7.9 | 1.81 | 1.98 | 2.52 | 1.33 | 1.09 | 16 | 3 | 16 |
| PCA | HLT = 2.8 Å | 7.7 | 1.95 | 2.09 | 2.44 | 1.21 | 1.07 | −25 | 11 | −66 |
| PCA | HLT = 0.0 Å | 7.5 [g] | 1.95 [h] | 2.08 [h] | 2.60 [h] | 1.29 | 1.07 | −11 | 4 | 7 |
| | | | | Ribonuclease H, 2rn2.pdb | | | | | | |
| experiment [c,f] | | 11.7 | 1.34 | 1.34 | 1.59 | 1.19 | 1.00 | 73 | 10 | − |
| HYDRONMR | AER = 3.2 Å | 11.6 | 1.28 | 1.38 | 1.64 | 1.24 | 1.08 | 60 | 23 | 175 |
| PCA | HLT = 2.8 Å | 10.4 | 1.45 | 1.56 | 1.78 | 1.18 | 1.07 | 61 | 30 | 183 |
| PCA | HLT = 0.0 Å | 11.3 [g] | 1.30 [h] | 1.40 [h] | 1.71 [h] | 1.27 | 1.08 | 47 | 27 | 165 |
| | | | | HIV-1 protease, 1bvg.pdb | | | | | | |
| experiment [c,i] | | 13.0 | 1.11 | 1.18 | 1.55 | 1.35 | 1.06 | 174 | 4 | 175 |
| HYDRONMR | AER = 3.2 Å | 13.8 | 1.02 | 1.05 | 1.56 | 1.51 | 1.03 | 175 | 5 | 91 |
| PCA | HLT = 2.8 Å | 13.0 | 1.12 | 1.17 | 1.57 | 1.37 | 1.04 | 176 | 6 | 179 |
| PCA | HLT = 0.0 Å | 14.2 [g] | 0.98 [h] | 1.02 [h] | 1.52 [h] | 1.52 | 1.04 | 154 | 3 | 155 |

[a] All experimental data were rescaled, when necessary, to 293 K, and all the calculations were performed for this temperature. All Euler angles are in degrees. AER is the adjustable parameter in HYDRONMR, HLT is the thickness of the hydration layer in the PCA-based method. [b] The Protein G construct studied in ref 9 had the first five residues from the PDB file clipped off. [c] The experimental data for protein G, ubiquitin, cytochrome $c_2$, ribonuclease H, HIV-1 protease are from refs 9, 14, 74, 76, 73, respectively. The rescaling to 293 K was performed, assuming that $\tau_c$ as well as $1/D_i$ ($i = x,y,z$) scale with temperature as $\eta(T)/T$, see footnote to Table 1. [d] Ubiquitin has a flexible C-terminus;[14] therefore, for the calculations presented here we clipped the last five residues from the PDB file. [e] From the set of 20 NMR structures for this protein, we took structure number 2. [f] The experimental data presented in ref 76 assume an axially symmetric diffusion tensor. [g] These $\tau_c$ values were multiplied by 2. [h] These $D_x$, $D_y$, $D_z$ values were divided by 2. [i] The orientation of the diffusion tensor for HIV-1 is reported here with respect to the PAF of the inertia tensor.[73]

85 kDa. This collection is considered here as a representative set of globular single-domain-protein folds. For most of these proteins, experimental data for their rotational diffusion tensors are not available.

Note also that HYDRONMR program does not include hydrogen atoms into calculations. Our analysis suggests that the presence of hydrogens has a small effect on the results: the overall tumbling time predicted for NMR structures from set A with hydrogen atoms included was up to 5% (3% on average) longer than when the hydrogens were removed from the structures (Supporting Information). Therefore, for a fair comparison of the performance of our method with HYDRONMR, hydrogen atoms were excluded from all structures in set B.

**Comparison with Experimental Data.** A general scientific validity criterion for any model is its agreement with experimental data. Thus, we first use set A to compare the predictions of the two models under consideration, HYDRONMR and the PCA-based method, with experimental data.

As mentioned above, HYDRONMR and the PCA-based method both have an adjustable parameter (AER and HLT, respectively) introduced to simulate the effect of hydration. It is worth mentioning that the value of such a parameter depends on several factors including the topographic (distribution of the hydrophilic and hydrophobic regions) and electrostatic properties of the protein surface and, therefore, could differ from one protein to another, in order to provide the best agreement between experimental and calculated values. A comparison of HYDRONMR predictions with the experimental data for a set of proteins examined in ref 44 showed that an optimal AER varies among the proteins, with the average value of 3.2 Å and the standard deviation of 0.6 Å. Following the recommendation by HYDRONMR authors this average AER value (3.2 Å) was used in our HYDRONMR calculations. All other vital parameters of the HYDRONMR program, the number of interpolation steps and the maximal and minimal mini-bead sizes, were kept at their default settings. Note that in all calculations in this paper the temperature was set to 293 K and the solvent viscosity to 0.01 poise.

Table 1 presents a comparison between the experimental and predicted values of the overall tumbling time, $\tau_c = [2(D_x + D_y + D_z)]^{-1}$. These results demonstrate that if HLT is adjusted individually for each protein, the PCA-based method is in remarkable agreement with experimental data. However, in order to provide a fair comparison between the PCA-based method and HYDRONMR, we decided to use a single value of this parameter for all proteins in the set. The optimal HLT value varied among the proteins in set A (Table 1) with the average value of 3.1 Å and the standard deviation of 0.7 Å. Note that the presence of hydrogen atoms in the protein structure has little effect on the optimal HLT: for the 12 NMR structures of set A the omission of hydrogens increased the average optimal HLT value from 3.2 Å to 3.35 Å; the standard deviation of the HLTs among these structures was 0.8 Å in both cases. Thus, because of the broad range of HLT values obtained for a rather limited set of experimental data, we decided not to use this average

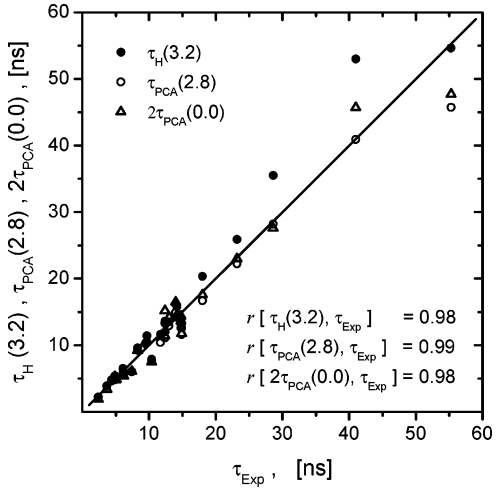


***Figure 2.*** Agreement between experimental, $\tau_{Exp}$, and calculated correlation times: $\tau_H$ calculated using HYDRONMR (solid symbols) and $\tau_{PCA}$ calculated using the PCA-based method (open symbols). Numbers in the parentheses indicate the values of AER and HLT parameters. All numerical data are presented in Table 1. The solid line is a guide for the eye, representing the case of absolute agreement. Also shown are the values of Pearson's correlation coefficient *r*.

value as a fixed size of the HLT parameter. Instead, we use a value of 2.8 Å for the thickness of the hydration layer in this paper. The rationale behind this choice is that this value approximately equals the diameter of water molecule;[57] thus, setting HLT = 2.8 Å amounts to covering the protein with a monolayer of water molecules. All other parameters of the SURF program, on which our method is based, such as the radius of the solvent molecule (1.4 Å), atomic van der Waals radii (Figure 1a), parameters of triangulation mesh, were set as specified in the SURF default settings. The results presented in Table 1 for constant HLT = 2.8 Å and AER = 3.2 Å show that the PCA-based approach and HYDRONMR both describe these experimental data with approximately the same accuracy (the average inaccuracies are 8% and 11%, respectively). This conclusion is further emphasized by the correlation plots in Figure 2 that show that both computational methods exhibit approximately the same high correlation with the experimental data. It is worth noting that both methods show similar deviations from the experimental data and exhibit an even higher level of correlation between each other (with the average difference of less than 8% and the correlation coefficient of 0.997). A similar comparison with the experimental data showed that FAST-HYDRONMR is on average 2-fold less accurate than the PCA-based approach (see Supporting Information Table 2). A more detailed comparison presented in Table 2 shows that not only the overall correlation time but also the individual components and the orientation of the diffusion tensor are reproduced with comparable accuracy by both HYDRONMR and the PCA-based method.

**Comparison between the PCA-Based Method and HYDRONMR.** In order to perform a comprehensive comparison between the proposed approach and HYDRONMR, we applied both methods to the set of 841 protein structures (Set B) described above.

**(a) Computational Efficiency.** All calculations presented in this work were conducted on a desktop computer with an Intel Xeon 1.7 GHz processor, 512 MB of ECC RDRAM, under the Red Hat Linux 7.1 operating system. We used HYDRONMR version 5a. The original version of SURF program[52] was modified to include the calculation of the covariance matrix (eq 4) and was combined with a MATLAB script that performs the numerical integration involved in eqs 1−3. The analysis of 841 protein structures from Set B has shown that with this computer setup, HYDRONMR required on average 13 min per structure, while it took only 1.6 s per structure for the PCA-based method. Thus, in its current realization, our method is 488 times faster than HYDRONMR. For example, in the case of a smooth sphere of 20 Å radius discussed below, the elapsed time was 0.7, 269, and 1.5 s, for PCA, HYDRONMR, and FAST-HYDRONMR, respectively.

In terms of the complexity of the problem, HYDRONMR inverts a $3N \times 3N$ matrix; thus, the computational time is proportional to $N^3$, where $N$ is the number of minibeads approximating the surface (a few thousands at least). FAST-HYDRONMR is based on the so-called double-sum approximation, which reduces the time complexity of the problem to $N^2$.[48] The time required for the SURF algorithm is proportional to $N_{at}k \log k$, where $N_{at}$ is the total number of atoms in a protein, and $k$ is a constant that depends on the atomic packing density and the radius of the solvent molecule.[52] For instance, for radius of the solvent molecule of 1.4 Å, $k$ is between 40 and 50 for proteins. Because the number of mini-beads necessary to accurately cover protein surface scales as $N_{at}^{2/3}$, the time required for FAST-HYDRONMR algorithm should be at least proportional to $N_{at}^{4/3}$, which is still a higher time complexity (hence slower) than that for SURF.

**(b) Overall Tumbling Time.** We first compare the predictions of the two methods for the overall rotational correlation time, $\tau_c$. This physical quantity is inversely proportional to the trace of the diffusion tensor, and is, therefore, invariant with respect to rotations of the reference frame. Thus, $\tau_c$ can be considered as a scalar characteristic of the rotational diffusion tensor.

With the hydration layer parameters set as discussed above, there is a reasonable agreement (within 10% difference) between the PCA-based method and HYDRONMR for proteins with $\tau_c < 15$ ns ($Mw < 27$ kDa) (Figure 3a), whereas for bigger proteins ($\tau_c > 15$ ns) the two methods give different $\tau_c$ values. A direct comparison with experimental data (Figure 2, Table 1) shows that in four out of five proteins with $\tau_c > 15$ ns the PCA-based predictions are in a better agreement with the measured rotational correlation time values. However, this data set is too small for a definitive conclusion concerning which of the two methods is more accurate for bigger proteins. The fact that both methods show a similar overall agreement with the experimental observations perhaps indicates that the differences between HYDRONMR and PCA-based predictions are comparable to the uncertainties in the experimental data.

What could be the reason for this discrepancy between HYDRONMR and PCA-based method? HYDRONMR calculates protein diffusion properties using the shell model which represents protein surface as closely as possible with a set of spherical friction elements. Our approach calculates the diffusion tensor for a smooth ellipsoidal representation of the protein's shape, when "the details of molecular structure are blurred in the smooth, entirely convex ellipsoidal shape".[18] Thus, it seems reasonable to assume that the observed difference between these methods can be attributed to the difference in protein surface

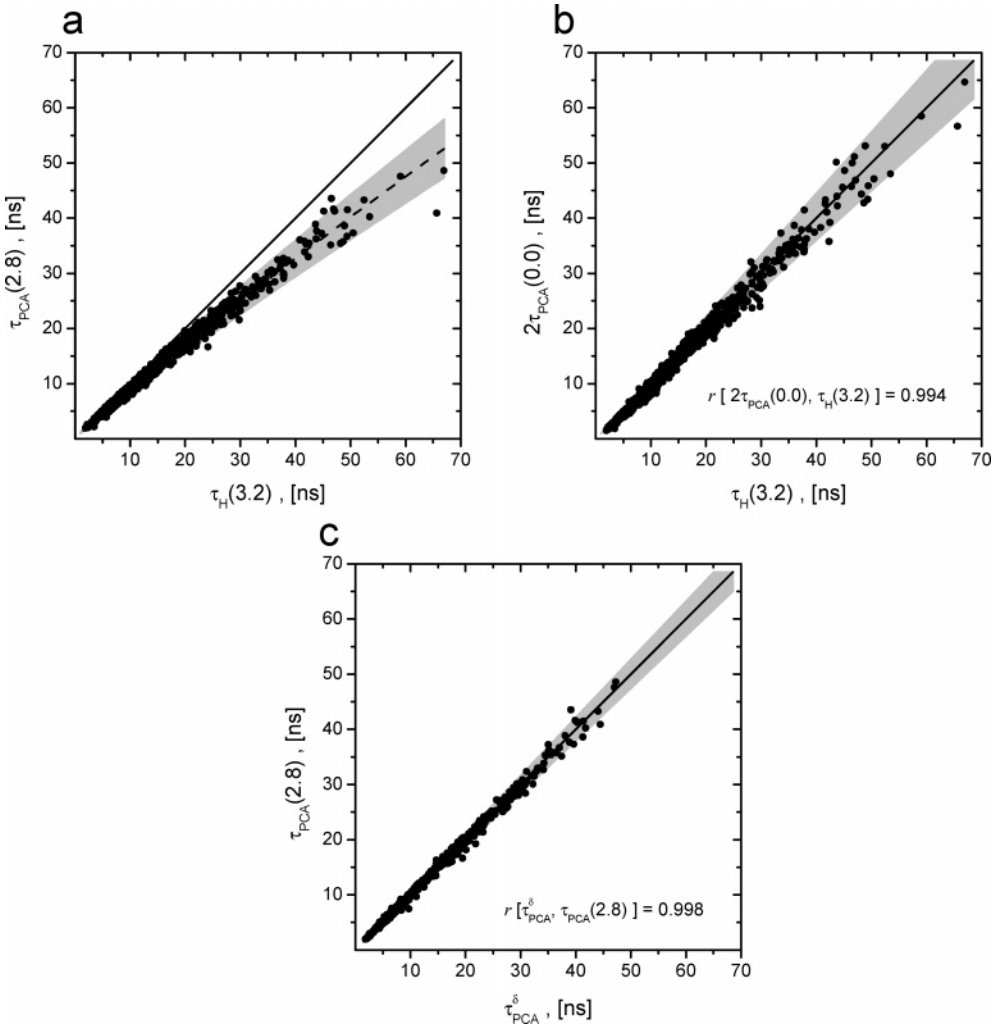


***Figure 3.*** Agreement between the values of the overall correlation time calculated using HYDRONMR ($\tau_H$) and the PCA-based method ($\tau_{PCA}$). Numbers in parentheses indicate the values of AER and HLT parameters. Lines in both panels are guides for the eye, representing the case of absolute agreement. (a) $\tau_{PCA}$ values shown in this panel were computed for "wet" proteins, i.e., including a hydration layer with HLT = 2.8 Å. Dash line in this panel corresponds to the power law, $\tau_{PCA}(\text{HLT} = 2.8\ \text{Å}) = K{\cdot}\tau_H^p$, with $K = 1.086 \pm 0.014$ and $p = 0.923 \pm 0.004$. Shaded area in this panel indicates the region of 10% deviation from the power law. (b) $\tau_{PCA}$ data were calculated for "dry" proteins (no hydration layer, HLT = 0.0 Å). Shaded area in panel b indicates the regions of 10% deviation from the absolute agreement. (c) The agreement between $\tau_{PCA}$ values computed for wet proteins (HLT = 2.8 Å) directly and estimated from $\tau_{PCA}$ for dry proteins (HLT = 0.0 Å) corrected for the volume of hydration layer ($\tau^{\delta}_{PCA}$, eq 8). Shaded area in panel c indicates the regions of 5% deviation from the absolute agreement. Also shown in b,c are the values of Pearson's correlation coefficient $r$.

representation. As illustrated in Figure 1f, for a protein covered with a hydration layer, the PCA method generates an equivalent ellipsoid that almost entirely encapsulates the protein. However, a more detailed representation of the actual rough protein surface by HYDRONMR leads to longer correlation times, as shown in Figure 3a. This effect increases with the size of the protein.

Interestingly, there is a striking linear relationship (Figure 3b) between the PCA-calculated correlation time for a dry protein and the HYDRONMR prediction for a wet protein. Another intriguing observation illustrated in Figure 3b is that the overall tumbling time, $\tau_H$(AER = 3.2 Å), from HYDRONMR calculation for a wet protein is almost exactly *twice* the PCA prediction for a dry protein, $\tau_{PCA}$(HLT = 0.0 Å). A linear regression analysis of the data in Figure 3b gives the following relationship between these correlation times:

$$\tau_H(\text{AER} = 3.2\ \text{Å}) = C\tau_{PCA}(\text{HLT} = 0.0\ \text{Å}) \qquad (7)$$

with $C = 2.055 \pm 0.005$ for all proteins analyzed. The agreement between $\tau_H$(AER = 3.2 Å) and $2\tau_{PCA}$(HLT = 0.0 Å) improves with the size of the protein: the root-mean-square (rms) percent difference is 10% for all proteins and only 6% for proteins with $\tau_c > 15$ ns. These results indicate that the PCA-based method can capture the complexity of protein surfaces equally as well as the mini-beads shell model used in HYDRONMR, and the actual reason for the "blurring" of the details of the protein structure is not the ellipsoidal representation *per se* but rather the added hydration layer that increases the effective volume of the protein and also smoothens its surface/shape substantially.

**(c) The Effect of Hydration.** In order to gain insight into the effect of hydration on the calculated diffusion tensors, it is instructive to discuss the observed differences in the predicted correlation times (Figure 3a,b) for a "dry" and a "hydrated" protein. For this purpose we compare the tumbling time, $\tau_{PCA}$(HLT = 0.0 Å), predicted by the PCA-based method for a protein without hydration layer with $\tau_{PCA}$(HLT = 2.8 Å) for the same protein covered with a hydration layer. Note that the overall correlation time of a molecule is generally proportional

to its molecular weight (hence, volume). For a smooth object, such as an ellipsoid generated by PCA, this follows from Perrin's eqs 1−3; the same is generally true for the proteins analyzed here (not shown). Therefore, the presence of a hydration layer is expected to increase the correlation time, compared to that for a dry protein, by the amount reflecting the corresponding volume increase, which we estimate as SAS·$\delta$:

$$\tau_{PCA}^{\delta} = \tau_{PCA}(\mathrm{HLT} = 0.0\ \text{Å}) + B{\cdot}\mathrm{SAS}{\cdot}\delta \qquad (8)$$

Here $\delta$ is the effective thickness of the added layer, and $B$ is the proportionality coefficient converting volume into time units: $B = \langle \tau_{PCA}(\mathrm{HLT} = 0.0\ \text{Å})/Mw \rangle \cdot N_A \cdot \bar{\nu}$, where $Mw$ is the molecular weight of a protein (in g·mol$^{-1}$), $\bar{\nu}$ is the specific volume of a protein[7] (set here to 0.76 cm$^3$·g$^{-1}$), $N_A = 6.02{\cdot}10^{23}$ mol$^{-1}$ is the Avogadro's number, and <...> denotes averaging over all proteins. A value of $\delta = 2.5$ Å was found optimal and resulted in excellent quantitative agreement between $\tau_{PCA}^{\delta}$ and the overall correlation time for a wet protein for all 841 proteins studied here (Figure 3c), with the rms difference of 3.7%. The small difference between 2.5 Å and the HLT value of 2.8 Å in the PCA calculations for wet proteins is likely due to the approximate estimation of the volume increase used in eq 8, which assumes an infinitesimal thickness of the hydration layer. This result suggests that the slower tumbling of a hydrated molecule is primarily due to the increase in its effective volume, which in turn reflects the properties of the protein's surface.

To illustrate this point, consider a particular example of ubiquitin shown in Figure 1. The dry ubiquitin molecule has a rough surface with many bulges and cavities (Figure 1a), whereas the same molecule covered with a hydration layer is rather smooth (Figure 1b). In general, hydration smoothens most of surface irregularities (see also the corresponding equivalent ellipsoids in e and f of Figure 1). For a wet ubiquitin structure, HYDRONMR predicts the rotational correlation time of 5 ns at 293 K (for AER = 3.2 Å). A very close value of 4.9 ns is obtained for wet ubiquitin using PCA-based method with HLT = 2.8 Å, whereas for the dry protein (HLT = 0.0 Å) the same program predicts a factor of 2 faster tumbling, with the correlation time of 2.4 ns (Tables 1 and 2).

As pointed out above (eq 7, Figure 3b), there is a remarkable linear relationship between the PCA prediction for a dry protein and HYDRONMR calculation for a hydrated protein, with the actual values of $\tau_c$ being different by a factor of 2. The presence of such a factor of 2 was noted a long time ago. It was the subject of discussion between Kirkwood and Kuhn at the Symposium on Macromolecules in Stockholm[60] in 1953. After the presentation by Kirkwood of his theory of irreversible processes in a solution of macromolecules Kuhn noted that the experimentally measured values of intrinsic viscosities are a factor of 2 higher than theoretical prediction. Kuhn also noted that this discrepancy vanishes for smooth objects, such as a sphere, and with increasing shear rates. In fact, for a (smooth) sphere (e.g., of 20 Å radius) tumbling in water at 293 K, the results from the PCA-based calculation ($\tau_{PCA}$(HLT = 0.0 Å) = 8.30 ns) and from HYDRONMR ($\tau_H$(AER = 20 Å) = 8.21 ns) agree with each other and with the Stokes−Einstein result ($\tau_c$ = 8.28 ns, also following from Perrin's eqs 1−3) without the necessity of introducing the factor of 2. Note also that the factor of 2 difference between the experimental rotational correlation times for proteins measured by fluorescence techniques and those calculated for an unhydrated sphere of the same molecular weight has been known in the fluorescence literature for quite some time.[61,62]

The observation that the same scaling factor holds for both small and large proteins might seem surprising because one could expect that due to the constant, finite thickness of the hydration layer it should have less effect on larger proteins. On the basis of the results of the numerical experiments presented here we propose an explanation for this fact. We argue that the abovementioned factor of 2 originates from the hydration layer effect and reflects the topography of protein surface. The surface of a dry protein (SAS presented in a and c of Figure 1) is highly irregular, and had been characterized as a fractal object.[63,64] Our analysis for protein set B shows that SAS $\propto Mw^q$, with $q = 0.81 \pm 0.01$ (see also refs 64,65), i.e. the solvent accessible surface area increases with the molecular weight (size) faster than for a smooth object, where SAS $\propto Mw^{2/3}$.[7] In other words, the larger the protein the greater is its "extra" surface area. As shown above (see eq 8, Figure 3b), protein volume increase due to hydration has a strong effect on the overall tumbling time. This increase in the volume depends on the SAS (eq 8) and, therefore, is expected to follow a similar trend. (In fact, the factor of 2 difference between $\tau_{PCA}$(HLT = 0.0 Å) and $\tau_H$(AER = 3.2 Å) can be recovered, to a reasonable approximation, for all proteins represented in Figure 3b when setting $\delta = 3.4$ Å in eq 8.) Our results suggest that protein surfaces are naturally sculpted in such a way that, when covered with a monolayer of water molecules, proteins gain an increase in their effective hydrodynamic volume (responsible for their rotational diffusion properties) by approximately a factor of 2. This model explains why the discrepancy between theoretical predictions and experiment disappears for a smooth object, such as a sphere, and in the case of large shear rates, when the hydration layer probably cannot be formed.

**(d) Detailed Comparison between Diffusion Tensors Derived by HYDRONMR and PCA-Based Methods.** Calculations with the PCA-based method for dry protein structures (HLT = 0.0 Å) show that, when including the above mentioned factor of 2 that accounts for the hydration layer effect, this method reproduces experimental data as well as the HYDRONMR predictions quite accurately (Tables 1 and 2 and Figure 2). Thus, for a detailed comparison with HYDRONMR and for further analysis the HLT value in PCA calculations will be set to zero, and the resulting diffusion tensors will be simply divided by the factor of 2. Figure 4 clearly shows that not only the overall correlation times but also the individual components of the diffusion tensors predicted by both calculation methods are in excellent agreement with each other (Pearson's $r > 0.99$).

Figure 5 demonstrates the agreement between the PCA-based method and HYDRONMR in terms of the Euler angles that specify orientations of the calculated diffusion tensors. According to the general convention, Euler angles $\alpha$ and $\beta$ specify orientation of the $z$-axis of the diffusion tensor with respect to

(60) Kirkwood, J. G. *J. Polym. Sci.* **1954**, *12*, 1−14.

(61) Yguerabide, J.; Epstein, H. F.; Stryer, L. *J. Mol. Biol.* **1970**, *51*, 573−590.
(62) Lakowicz, J. R.; Maliwal, B. P.; Cherek, H.; Balter, A. *Biochemistry* **1983**, *22*, 1741−1752.
(63) Lewis, M.; Rees, D. C. *Science* **1985**, *230*, 1163−1165.
(64) Fushman, D. *J. Biomol. Struct. Dyn.* **1990**, *7*, 1333−1344.
(65) Miller, S.; Lesk, A. M.; Janin, J.; Chothia, C. *Nature* **1987**, *328*, 834−836.

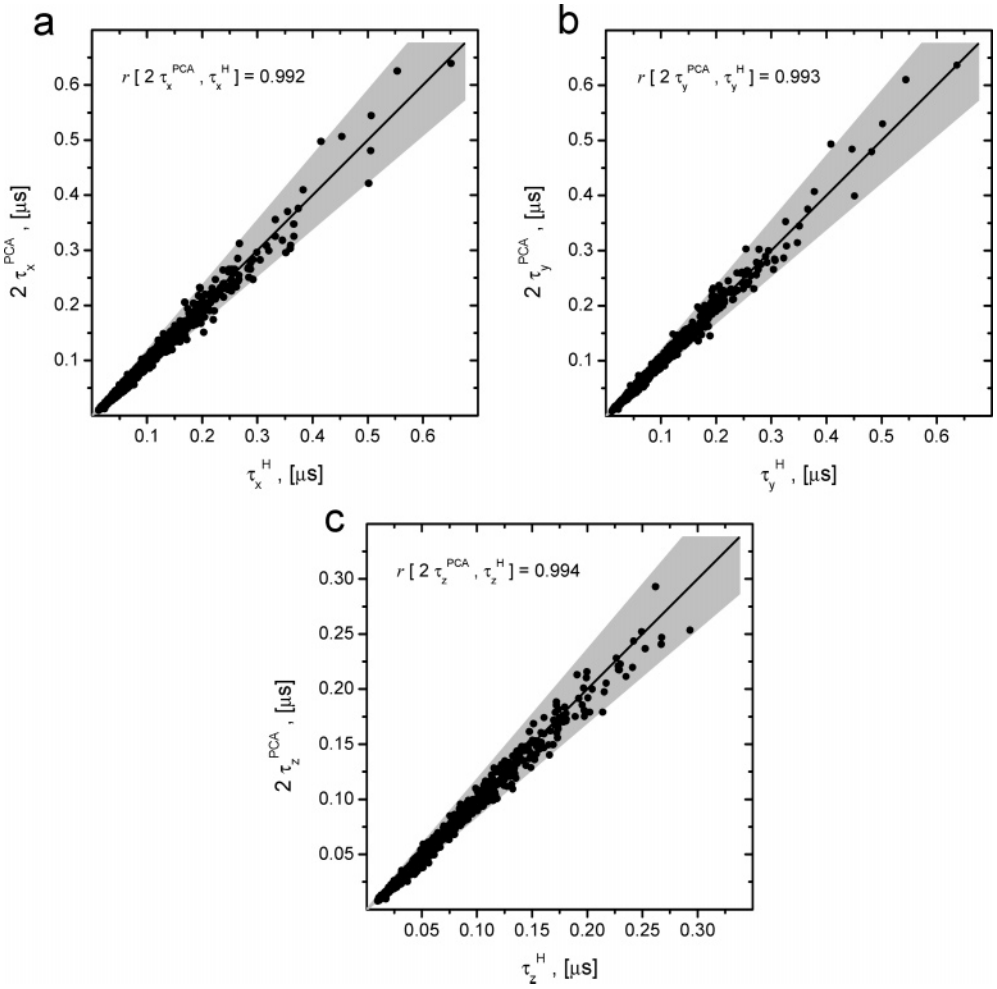


***Figure 4.*** Agreement between predictions of HYDRONMR (superscript "H") and the PCA-based method (superscript "PCA") for correlation times, $\tau_x = 1/D_x$, $\tau_y = 1/D_y$, and $\tau_z = 1/D_z$ which characterize rotations about the principal axes *x*, *y*, and *z* of diffusion tensor (panels a, b, and c, respectively). HYDRONMR calculations were performed for proteins with hydration layer (AER = 3.2 Å). Calculations using the PCA-based method were for dry proteins (HLT = 0.0 Å); the resulting diffusion tensors were scaled by a factor of 1/2. Solid lines are guides for the eye, representing the case of absolute agreement. Shaded areas indicate regions of 15% deviation from the absolute agreement. The values of Pearson's *r* are indicated.

the PDB reference frame while the angle $\gamma$ is associated with a rotation about the *z*-axis of the diffusion tensor; this latter rotation determines the orientation of the *x* and *y* axes of the tensor. It should be pointed out that the level of agreement is different for different Euler angles: the $\alpha$ and $\beta$ angles show a better agreement between the two methods than the angles $\gamma$ (the correlation coefficients $r = 0.996$, 0.986, and 0.968, respectively). In order to understand the reasons for this, consider deviations from the full symmetry (isotropy) of a tensor, which are usually characterized by the anisotropy, *A*, and rhombicity, *Rm*, that can be defined as follows:

$$A = \frac{2D_z}{D_x + D_y}; \quad Rm = \frac{3}{2}\frac{D_y - D_x}{D_z - \frac{1}{2}(D_x + D_y)}$$

for a prolate tensor (9a)

$$A = \frac{2D_x}{D_y + D_z}; \quad Rm = \frac{3}{2}\frac{D_z - D_y}{D_z - \frac{1}{2}(D_x + D_y)}$$

for an oblate tensor (9b)

where the individual components of the diffusion tensor are defined such that $D_x \le D_y \le D_z$, and the tensor is called prolate if $D_z - D_y \ge D_y - D_x$ and oblate otherwise. The anisotropy equals 1 for an isotropic tensor ($D_x = D_y = D_z$) expected for a spherically shaped molecule, while zero rhombicity corresponds to an axially symmetric tensor ($D_y = D_x$ or $D_y = D_z$), as in the case of an ellipsoid of revolution (rigid rotor), or to an isotropic tensor. Thus, $A \neq 1$ indicates a deviation from isotropy of the tensor, while $Rm \neq 0$ is an indicator of a deviation from axial symmetry (note that $0 \le Rm \le 1$).

The comparison of the anisotropies and rhombicities of the diffusion tensors calculated using HYDRONMR and the PCA-based method, presented in Figure 6, shows that the anisotropies are in a remarkably better agreement ($r = 0.986$) than the rhombicities ($r = 0.858$). This indicates a significantly better correlation in the orientation of the *z*-axes of the diffusion tensors than of their *x*- and *y*-axes, and is directly related to the above mentioned differences in the agreement for various Euler angles (Figure 5). Thus, the reason for this observation is in the generally small rhombicities (see below) reflecting rather small deviations from axial symmetry of the diffusion tensors for many

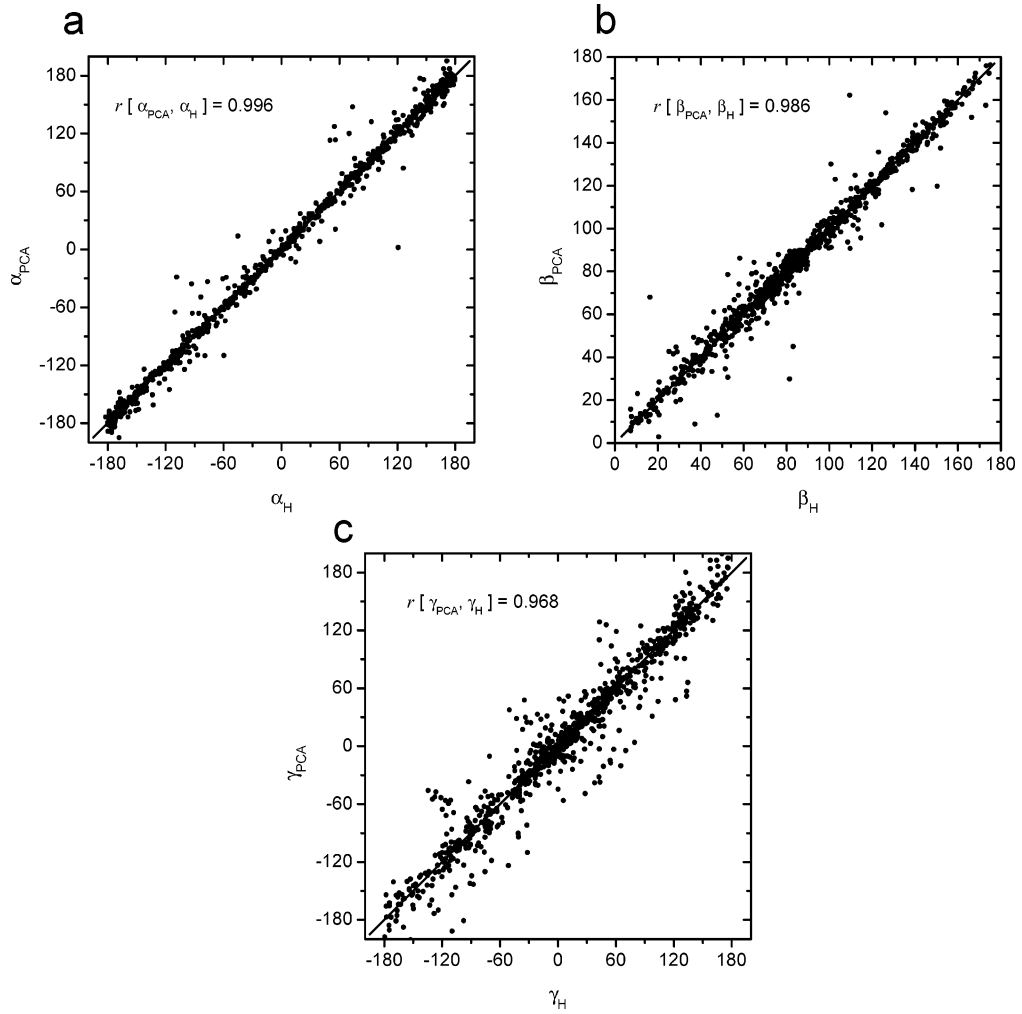


***Figure 5.*** Agreement between predictions of HYDRONMR (subscript "H") and PCA-based method (subscript "PCA") for Euler angles that specify the orientation of the PAF of the diffusion tensor with respect to the protein coordinate frame. All angles are in degrees. HYDRONMR calculations assumed that a protein is covered with a hydration layer (AER = 3.2 Å). PCA-based calculations were done for dry proteins (HLT = 0.0 Å). Solid lines are guides for the eye, representing the case of absolute agreement; 68% of all differences between the angles estimated by the HYDRONMR and the PCA-based methods are within the confidence interval of ± 5.5° for angles α, 3° for $\beta$, and 11° for angles $\gamma$.

of the 841 proteins analyzed here. Small rhombicity makes it difficult to determine accurately the *x*- and *y*-axes of the diffusion tensors for both methods.

### How Anisotropic Are Rotational Diffusion Tensors in Proteins?

The extent of deviations from the full symmetry in protein diffusion tensors is of particular interest for NMR applications to protein dynamics, where accurate analysis of experimental data requires a proper model for the overall tumbling.[8,9] The representative set of 841 protein folds considered here is sufficiently large to address this question quantitatively. Deviations of a tensor from full symmetry (isotropy) are characterized here in terms of its anisotropy *A* and rhombicity *Rm* introduced in eq 9. We have found that out of the 753 prolate tensors, only 84 (11%) have anisotropies *A* in the range between 1 and 1.17, i.e., could be safely approximated as isotropic for the purposes of NMR data analysis.[14,9] Of the prolate tensors, 509 (68%) have anisotropies in the "intermediate" range from 1.17 to 1.6, and in total 669 (89%) of the proteins have $A \geq 1.17$, which means that anisotropic diffusion model is generally required for accurate analysis of NMR relaxation data. Approximately half of prolate proteins studied here (48%) have rhombicities below 0.2. Computer simulations show[17] that for small diffusion tensor rhombicities ($Rm < 0.2$) the available level of precision in $^{15}$N relaxation data might not be sufficient to unequivocally discriminate between the fully anisotropic and axially symmetric tumbling. This suggests that a simpler, axially symmetric model could be sufficient for NMR data analysis in many monomeric proteins. This conclusion is in agreement with the known examples when the use of a more complex, fully anisotropic diffusion model for analysis of experimental $^{15}$N relaxation data was not statistically warranted.[9,14,66−76] The statistics for oblate ellipsoids are similar, although the anisotropy

(66) Hall, J. B.; Fushman, D. *J. Am. Chem. Soc.* **2006**, *128*, 7855−7870.
(67) Tugarinov, V.; Muhandiram, R.; Ayed, A.; Kay, L. E. *J. Am. Chem. Soc.* **2002**, *124*, 10025−10035.
(68) Helms, M. K.; Petersen, C. E.; Bhagavan, N. V.; Jameson, D. M. *FEBS Lett.* **1997**, *408*, 67−70.
(69) Hwang, P. M.; Skrynnikov, N. R.; Kay, L. E. *J. Biomol. NMR* **2001**, *20*, 83−88.
(70) Damberg, P.; Jarvet, J.; Allard, P.; Mets, U.; Rigler, R.; Graslund, A. *Biophys. J.* **2002**, *83*, 2812−2825.
(71) Visser, A.; Van-Hoek, A.; Visser, N.; Ghisla, S. *Photochem. Photobiol.* **1997**, *65*, 570−575.
(72) Striker, G.; Subramanian, V.; Seidel, C. A. M.; Volkmer, A. *J. Phys. Chem. B* **1999**, *103*, 8612−8617.
(73) Tjandra, N.; Wingfield, P.; Stahl, S.; Bax, A. *J. Biomol. NMR* **1996**, *8*, 273−284.

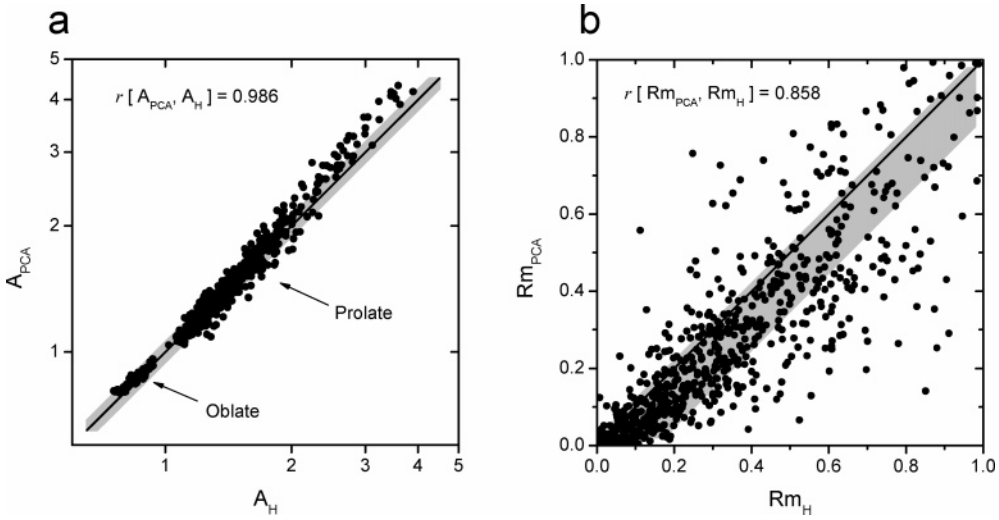


***Figure 6.*** Agreement between (a) anisotropies *A* and (b) rhombicities *Rm* of the diffusion tensors for the set of 841 proteins. These parameters were calculated from the principal values of the diffusion tensors according to eq 9. Subscripts "H" and "PCA" correspond to diffusion tensors obtained using HYDRONMR with AER = 3.2 Å and the PCA-based method with HLT = 0.0 Å, respectively. Solid line is a guide for the eye, representing the case of absolute agreement. Shaded areas indicate the region containing 68% of all data. Anisotropies of prolate diffusion tensors are greater than unity; anisotropies of oblate diffusion tensors are less than unity. The values of the Pearson's correlation coefficient *r* are indicated.

values appear somewhat smaller (68% have $0.8 < A < 0.9$) which could reflect the size of the sampling set smaller than that for prolate tensors. It is worth emphasizing that these statistics represent properties of single-domain proteins. These proteins will have different diffusion tensors in the context of a multidomain system or when complexed with other proteins or nucleic acids.

## Conclusions

The results presented in this paper can be summarized as follows:

(i) The ellipsoid representation for protein shape based on the PCA of protein surface coordinates, suggested in this work, provides a comparable level of accuracy and precision in predicting experimental data to the rigorous approach representing protein surface with a large number of small friction elements. Thus, conceptually, the ellipsoidal model can account for protein tumbling. The opinion that the ellipsoidal approximation is too rough for an accurate prediction of protein rotational diffusion tensors is incorrect. That opinion originated from earlier attempts to derive an ellipsoid approximation of a protein from the inertia tensor, which is irrelevant in the case of protein diffusion in solution.

(ii) The proposed method for predicting protein diffusion tensors from atomic-resolution structures based on the principal component analysis of the protein's surface is about 500 times faster than the rigorous method based on shell/bead representation.

(iii) Our analysis suggests that protein surfaces are naturally sculpted in such a way that a hydration layer consisting of approximately one monolayer of water molecules (HLT = 2.8 Å) increases the apparent rotational correlation time of a protein by approximately a factor of 2. Thus, a simple recipe that follows from our analysis of a large set of proteins is to calculate the diffusion tensor for a dry protein using the PCA-based method and then scale it by a factor of 1/2.

(iv) Most of the proteins analyzed in this work have anisotropic diffusion tensors, with the anisotropy parameter values greater than 1.2 in 85% of the proteins with prolate diffusion tensors. At the same time about 50% of the proteins have their tensor rhombicities small enough ($Rm < 0.2$) to be approximated by an axially symmetric diffusion tensor. These results suggest that (1) an anisotropic rotational diffusion model is generally required for NMR data analysis in single-domain proteins, but (2) the axially symmetric model could be sufficient for this purpose in approximately half of these proteins.

We anticipate that the method proposed here will find applications in various computer programs that require multiple steps of fast and accurate assessment of the diffusion tensor, for example, in protein structure determination that includes diffusion tensors or diffusion-sensitive parameters (e.g., spin-relaxation data) as additional structural restraints (Ryabov and Fushman, manuscript in preparation).

**Acknowledgment.** Supported by NIH Grant GM065334 to D.F. and NSF CCF 0429753 to A.V. We thank Dr. Andrej Šali for making available the representative set or protein folds used in this study and Ming-Yih Lai for help with SURF adaptation. A computer program package (ELM) for the PCA-based diffusion tensor prediction is available from us upon request.

**Supporting Information Available:** A list of 841 PDB files used in this study (set B); three tables presenting evaluation of the accuracy of an inertia-equivalent ellipsoid method and of the FAST-HYDRONMR program, as well as the effect of omitting hydrogen atoms from protein structures; and two figures and a table presenting statistical distributions of the predicted diffusion tensor for set B. This material is available free of charge via the Internet at http://pubs.acs.org.

JA062715T

(74) Cordier, F.; Caffrey, M.; Brutscher, B.; Cusanovich, M. A.; Marion, D.; Blackledge, M. *J. Mol. Biol.* **1998**, *281*, 341−361.
(75) Weast, R. C. *Handbook of Chemistry and Physics*, 59th ed.; CRC Press: West Palm Beach, FL, 1978.
(76) Kroenke, C. D.; Loria, J. P.; Lee, L. K.; Rance, M.; Palmer, A. G., III. *J. Am. Chem. Soc.* **1998**, *120*, 7905 -7915.