The Effect of Finite Sampling on the Determination of Orientational Properties: A Theoretical Treatment with Application to Interatomic Vectors in Proteins

David Fushman,[†] Ranajeet Ghose, and David Cowburn*

Contribution from the The Rockefeller University, 1230 York Avenue, New York, New York 10021

Received March 30, 2000. Revised Manuscript Received August 15, 2000

Abstract: Second-rank tensor properties such as the overall rotational diffusion tensor and the alignment tensor can be determined by NMR methods measuring orientation of interatomic vectors. Here we examine the effect of incomplete sampling of orientation space by interatomic vectors in a molecule on determination of a second-rank tensor. We have developed a quantitative approach to determine (1) how well orientation space is sampled by a particular protein or substructure, (2) to what extent this particular distribution of bond vectors samples the various components of a second-rank tensor and, (3) the ability of this distribution of bond vectors to completely characterize the tensor. This approach is generally applicable to any second-rank tensor property whose determination relies on the sampling of the angular space by the structure or substructure. The theory permits assessment of the expected degree of accuracy of tensor determination using a selected set of interatomic vectors (e.g., NH or $C^{\alpha}H^{\alpha}$, etc), for a given molecular structure. The sampling properties of real proteins are analyzed using a database of 1736 structures, representing all experimentally determined protein folds. This theoretical approach is applied to the rotational diffusion and alignment tensors obtained from nuclear magnetic resonance data for several systems, including ubiquitin and βARK PH domain. Finally, the proposed sampling characteristics are related to the accuracy of the determination of the rotational diffusion tensor from spin-relaxation data, as an example of an unknown second-rank tensor. Knowing the accuracy of the tensor quantity derived from experimental data assists in optimizing experimental design.

Introduction

Recently, significant attention has been paid to the determination of several second-rank tensor properties from liquid-state NMR. These include measurement of the alignment tensor from residual dipolar coupling information in oriented systems[1−3] as well as the magnitude and orientation of the rotational diffusion tensor from relaxation data in isotropic solution.[4−9] These measurements provide valuable structural information in the form of "long-range", orientational constraints[10−12] not available from NOE measurements and are likely to improve significantly the accuracy of protein structures determined in solution. Determination of the overall tensor properties, like the alignment tensor or the rotational diffusion tensor, is critical for precise and accurate derivation of orientational constraints for structure determination. While these tensors can be directly determined from experimental measurements based on protein structure,[1,2,4−9] their determination in the absence of structural information is less straightforward. Approaches were suggested to estimate the largest principal value and the rhombicity of these tensors, assuming uniform orientational distribution for the measured internuclear vectors.[13] A structure refinement protocol suggested recently[14] avoids the necessity of prior knowledge of the orientation of alignment tensor; however, the derivation of intervector angles from residual dipolar couplings in this approach requires knowledge of the principal values of the tensor. An attractive possibility of deriving "low-resolution" orientational constraints without explicit determination of the alignment tensor was also suggested[15] and could be used as a starting step in structure determination. However, structure refinement to a high level of accuracy and precision will rely on accurate determination of the alignment tensor.

* Address correspondence to this author. Telephone: 212-327-8270. Fax: 212-327-7566. E-mail: cowburn@rockefeller.edu.

[†] Present Address: Center of Biomolecular Structure & Organization, Dept. of Chemistry and Biochemistry, University of Maryland, College Park, MD 20742.

[1] Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 9279−9283.

[2] Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111−1114.

[3] Clore, G. M.; Starich, M. R.; Gronenborn, A. M. *J. Am. Chem. Soc* **1998**, *120*, 10571−10572.

[4] Bruschweiler, R.; Liao, X.; Wright, P. E. *Science* **1995**, *268*, 886−889.

[5] Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A. *J. Am. Chem. Soc.* **1995**, *117*, 12562−12566.

[6] Lee, L. K.; Rance, M.; Chazin, W. J.; Palmer, A. G., III. *J. Biomol. NMR* **1997**, *9*, 287−298.

[7] Fushman, D.; Najmabadi-Haske, T.; Cahill, S.; Zheng, J.; LeVine, H., 3rd; Cowburn, D. *J. Biol. Chem.* **1998**, *273*, 2835−2843.

[8] Copie, V.; Tomita, Y.; Akiyama, S. K.; Aota, S.; Yamada, K. M.; Venable, R. M.; Pastor, R. W.; Krueger, S.; Torchia, D. A. *J. Mol. Biol.* **1998**, *277*, 663−682; Blackledge, M.; Cordier, F.; Dosset, P.; Marion, D. *J. Am. Chem. Soc.* **1998**, *120*, 4538−4539; McDonnell, J. M.; Fushman, D.; Milliman, C. L.; Korsmeyer, S. J.; Cowburn, D. *Cell* **1999**, *96*, 625−634.

[9] Fushman, D.; Xu, R.; Cowburn, D. *Biochemistry* **1999**, *38*, 10225−10230.

[10] Tjandra, N.; Garrett, D. S.; Gronenborn, A. M.; Bax, A.; Clore, G. M. *Nat. Struct. Biol.* **1997**, *4*, 443−449.

[11] Tjandra, N.; Omichinski, J. G.; Gronenborn, A. M.; Clore, G. M.; Bax, A. *Nat. Struct. Biol.* **1997**, *4*, 732−738.

[12] Clore, G. M.; Gronenborn, A. M. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 5891−9588.

[13] Clore, G. M.; Gronenborn, A. M.; Szabo, A.; Tjandra, N. *J. Am. Chem. Soc.* **1998**, *120*, 4889−90; Clore, G.; Gronenborn, A.; Bax, A. *J. Magn. Reson.* **1998**, *133*, 216−221.

[14] Meiler, J.; Blomberg, N.; Nilges, M.; Griesinger, C. *J. Biomol. NMR* **2000**, *16*, 245−252.

[15] Moltke, S.; Grzesiek, S. *J. Biomol. NMR* **1999**, *15*, 77−82.

10.1021/ja001128j CCC: $19.00 © 2000 American Chemical Society
Published on Web 10/13/2000

The degree to which sets of tensors, either solely, or in combination with small sets of internuclear scalars such as NOEs, can be used for structure determination remains ill-determined.[16,17] Nevertheless, these measurements of larger-scale hydrodynamic properties have provided insight into the relative orientation of multiple domains in weakly interacting multidomain systems,[9,18−20] where interdomain NOE information is scarce, or time-averaged.

The approaches mentioned above are based on measurements of orientation-dependent characteristics for a set of interatomic vectors in a molecule. Two issues need to be addressed for a critical assessment of the accuracy of these analyses: (1) how well do the results of such analyses fit the available experimental data and (2) how well can the tensor quantities of interest for a particular molecule or substructure be determined with the finite set of interatomic vector orientations available. Approaches to the former have been suggested.[6,17,21] The second issue, which is less explored, arises because only a limited set of vectors is available for analysis in any protein or substructure. The derived tensor values could then depend on the measured set of interatomic vectors, as illustrated in ref 6 for the rotational diffusion tensor determined from the NH- and $C^{\alpha}H^{\alpha}$-vectors, used separately or grouped. It is important to develop a measure of how well the various components of the tensor properties can be determined from experimental measurements for a specific protein structure. In other words, how far is it possible to determine accurately the magnitude and orientation of a given second-rank interaction from a finite set of interatomic vectors? This analysis could also help select a proper subset of interatomic vectors to provide optimal sampling of the desired characteristics and, therefore, could serve as a guide for experimental design in these kinds of studies.

When an infinite number of uniformly oriented vectors is available, then obviously all directions are represented equally, and the quality of a determined tensor is independent of the orientation of its principal axes. Real proteins in actual NMR experiments, however, differ from this hypothetical case in two ways: (a) the number of available interatomic vectors is limited, both by the finite number of atom pairs in a protein and by the type of atoms/nuclei observable in a particular experiment, and (b) the orientational distribution of the available vectors is not necessarily uniform. The latter condition, which reflects the nonuniform character of a protein structure, is an essential structural feature in a folded protein in contrast to a random polymer coil. Consider, for example, an α-helix, where all the backbone NH-vectors are aligned nearly parallel to the helix axis. If this set of vectors is the only set used for the determination of a second-rank tensor property, such as the alignment tensor or the rotational diffusion tensor, then the principal axis of the tensor in question, aligned parallel to the helix axis, would be well sampled by the vector set, whereas the axes orthogonal to it would be essentially undetermined.

In this work, we present a theoretical framework to assess the degree to which a protein or substructure samples conformation space. This theory allows, for a given protein structure, an assessment of the expected degree of accuracy using a selected set of vectors (e.g., NH or $C^{\alpha}H^{\alpha}$ etc). This provides guidelines for the design of experimental approaches that provide the best possible accuracy. We discuss its implications on the determination of second-rank tensor properties in solution or, conversely, the characterization of the structure of a particular protein given a set of measurements of certain second-rank tensor properties in solution. Although the discussion here is focused mostly on the overall rotational diffusion of protein molecules in solution and the characterization of the alignment tensor for proteins in liquid-crystalline media, the results are applicable to any overall second-rank tensor quantity, accessible by various orientation-dependent measurements, not only NMR.

As a particular example, let us consider the determination of the overall rotational diffusion tensor from heteronuclear relaxation data. The general procedure followed in this case involves the determination of a T_1/T_2 ratio at a given field for a set of backbone ^{15}NH-[5,6,8,9] or $^{13}C^{\alpha}H^{\alpha}$-vectors.[6] The parameters characterizing the diffusion tensor can then be obtained by minimization of a target function incorporating these ratios and those calculated from an available X-ray or NMR structure of the protein. In reality, a limited set of data is available for the analysis, since not all the atoms are available for the experimental observation and current experimental approaches are limited to pairs of bonded nuclei in the backbone: ^{15}NH and $^{13}C^{\alpha}H^{\alpha}$.[6] In addition, loops and the termini are usually excluded from this relaxation analysis because their structural features are ill-defined on the relevant time scale. In addition to these experimental considerations, the limited size of the protein or substructure, and its limited sampling by any internuclear pair, restrict the ultimate accuracy of any tensor determination.

Theory

The theory is developed here for the general case of an arbitrary, unspecified second-rank tensor, $\boldsymbol{D}$, determined by experimental measurements for a selected set of interatomic vectors. In the following sections, this theory will be applied to two particular cases: (a) the determination of the overall rotational diffusion tensor of a protein in solution from NMR relaxation measurements[5,6,8,9] and (b) the determination of the alignment tensor of the molecule in an ordered medium from residual dipolar couplings.[1−3]

Both the rotational diffusion and alignment tensors transform as second rank tensors and thus have the same properties under rotation. To understand the transformation properties of these tensors, we make use of the fact that an arbitrary rank-2 tensor can be decomposed into a scalar (rank 0) which corresponds to the trace of the tensor, a pseudovector (rank 1) which represents the anti-symmetric part of the tensor, and a traceless tensor of rank 2. The rank 1 part is zero for both symmetric tensors considered here, while the rank 0 part, which is direction independent, equals $(^1/_3)Tr[\boldsymbol{D}]$ (hence has to be determined) in the case of the rotational diffusion tensor and is zero in the case of the (traceless) alignment tensor. Therefore, the diffusion tensor is characterized by three independent eigenvalues and three orthogonal eigenvectors (or six independent elements in an arbitrary coordinate system). The alignment tensor or the other hand, has only two independent eigenvalues and three orthogonal eigenvectors (altogether five independent elements in an arbitrary coordinate system). The spectral manifestation of the alignment tensor is the residual dipolar coupling. The spectral manifestation of the rotational diffusion tensor can be represented by an effective diffusion constant for each bond vector. For small anisotropies, this can be expressed in the same form as the residual dipolar coupling[4,6] (compare, for example, eq 13 in ref 6 and eq 1 in ref 21).

Assume we have at hand a set of unit vectors denoting the various interatomic vectors of a particular protein in an arbitrary reference frame. In the case of heteronuclear NMR relaxation measurements, this

[16] Cross, T.; Arseniev, A.; Cornell, B.; Davis, J.; Killian, J.; Koeppe, R.; Nicholson, L.; Separovic, F.; Wallace, B. *Nat. Struct. Biol.* **1999**, *6*, 610−611.

[17] Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.

[18] Fischer, M. W. F.; Losonczi, J. A.; Weaver, L. J.; Prestegard, J. H. *Biochemistry* **1999**, *38*, 9013−9022.

[19] Markus, M.; Gerstner, R.; Draper, D.; Torchia, D. *J. Mol. Biol.* **1999**, *292*, 375−387.

[20] Skrynnikov, N.; Goto, N.; Yang, D.; Choy, W.; Tolman, J.; Mueller, G.; Kay, L. *J. Mol. Biol.* **2000**, *295*, 1265−1273.

[21] Clore, G. M.; Garrett, D. S. *J. Am. Chem. Soc.* **1999**, *121*, 9008−9012.

is typically a set of the backbone NH-bond vectors. In the case of residual dipolar coupling measurements, these are typically a much larger set of vectors, comprising the NH, $C^{\alpha}H^{\alpha}$, C′N, and $C^{\alpha}C^{\beta}$ bond vectors and C′H.[20,22] The order tensor formalism, introduced by Saupe[23] to represent orientational order in a uniaxial liquid-crystalline medium, can be applied to this system to represent the sampling of the orientational degrees of freedom along three Cartesian axes.

The Sampling Tensor Formalism. For a given a set of unit vectors, a sampling tensor, Ω, which is a traceless, symmetric tensor of rank 2 with five independent elements, can be defined as

$$\Omega_{ij} = \frac{3\langle \boldsymbol{r}_i \boldsymbol{r}_j \rangle - \delta_{ij}}{2} \tag{1}$$

where r_i is the projection of a given unit vector $\boldsymbol{r}$ on the axis i where i, $j = x'$, y', z' (an arbitrary reference frame) and δ_{ij} is the Kronecker delta. There is an obvious similarity of the Ω tensor, which reflects statistical, time-independent sampling, to the more conventional order parameter tensor $\boldsymbol{S}$, which reflects time-dependent fluctuations. The brackets denote averaging over all the vectors in the ensemble. The sampling tensor can be diagonalized to yield the principal axis frame that corresponds to the direction of best sampling. The best sampled frame is related to the original frame by a rotation $R(\varphi,\theta,\psi)$ where φ, θ, and ψ are the Euler angles relating the two frames. In the principal axis frame of the sampling tensor, the fraction of vectors oriented along the three principal directions are given by

$$f_i = \langle \boldsymbol{r}_i^{\,2} \rangle = \frac{2\Omega_i + 1}{3} \quad \text{or inversely,} \quad \Omega_i = \frac{3}{2}\left(f_i - \frac{1}{3}\right) \tag{2}$$

where $i = x, y, z$ are the principal axes and $\Omega_{\rm i}$ are the principal values (ordered as $\Omega_z \geq \Omega_y \geq \Omega_x$ thus, $f_z \geq f_y \geq f_x$) of the sampling tensor.[24] Note that

$$f_x + f_y + f_z = 1 \tag{3}$$

In the case of a uniform distribution of vectors, Ω is the null tensor and[25] $f_x = f_y = f_z = {}^1/_3$. When all the vectors are oriented along the principal z-axis of the sampling tensor, then $\Omega_x = \Omega_y = -{}^1/_2$ and Ω_z =1, and $f_x = f_y = 0$, and $f_z = 1$. In general, deviations of the principal values, Ω_i, of the sampling tensor from zero, and that of the corresponding f_i values from ${}^1/_3$, reflect deviation from a uniform distribution of the vectors.

Generalized Sampling Parameter. We define a generalized sampling parameter Ξ (in a manner similar to the generalized squared order parameter[26] used in spin-relaxation analysis) given by[27]

$$\Xi = \frac{2}{3}\sum_{i=x',y',z'} \Omega_{ij}^2 = \frac{2}{3}\sum_{i=x,y,z} (\Omega_i)^2 = \frac{1}{2}\Big(3\sum_{i=x,y,z} f_i^2 - 1\Big) = \frac{1}{4}[(3f_z - 1)^2 + 3(f_y - f_x)^2] \tag{4}$$

The generalized sampling parameter quantifies the distribution of vector orientations on a scale from 0 to 1. For a uniform distribution of vectors, $\Xi = 0$, and this represents an optimal sampling of angular space. $\Xi = 1$ if all the vectors are aligned along one direction, representing the worst possible sampling of angular space.

Average Constant $\boldsymbol{D}_{\rm av}$. What are the possible implications of this sampling on the determination of a second-rank tensor quantity such as the rotational diffusion tensor or the alignment tensor? It is quite clear that, in the case where the $\boldsymbol{D}$-tensor frame and the frame of best sampling are collinear, the principal element of the tensor which has the largest number of vectors aligned parallel to it is sampled the best and that with the least number of vectors parallel to it is sampled the worst (cf. the helical paradigm mentioned above). In the general case, the accuracy of determination of orientation may not be so apparent, because the data from which the tensor quantity is extracted usually have highly nonlinear dependencies on the vector orientations. Evidently, the best possible scenario is a uniform distribution of vectors.

In the general case, the principal axes of the tensor $\boldsymbol{D}$ are not necessarily aligned along the best-sampled directions, as determined by the sampling tensor introduced above. To quantify how well a given distribution samples the various elements of the tensor of interest, $\boldsymbol{D}$, we define an average constant $D_{\rm av}$, which in an arbitrary frame is represented by[28]

$$D_{\rm av} = \frac{1}{3}Tr[\boldsymbol{D}] + \frac{2}{3}\sum_{i,j=x',y',z'} \Omega_{ij} D_{ij} \tag{5}$$

This can be written in explicit form as

$$D_{\rm av} = \langle x^2 \rangle D_{xx} + \langle y^2 \rangle D_{yy} + \langle z^2 \rangle D_{zz} + 2\langle xy \rangle D_{xy} + 2\langle xz \rangle D_{xz} + 2\langle yz \rangle D_{yz} \tag{6}$$

This has the same form as the effective diffusion constant derived previously.[4,6] In the case of a uniform distribution, all parts of the tensor are sampled equally well, and $D_{\rm av} = {}^1/_3 Tr[\boldsymbol{D}] = D_{\rm iso}$ which is the isotropic value of $\boldsymbol{D}$.

The value of $D_{\rm av}$ quantifies how well the overall tensor is sampled. To quantify how well each principal component of the tensor is defined by a given set of vectors, we may rewrite eq 5 in the principal axis frame of the sampling tensor and utilize eq 2 to obtain

$$D_{\rm av} = \sum_{i=x_d,y_d,z_d} \Phi_i D_i \tag{7}$$

where $\Phi = \{\Phi_{x_d}, \Phi_{y_d}, \Phi_{z_d}\}$ is a three-component vector which provides a measure of how well each principal component, D_i $(= D_{x_d}, D_{y_d}, D_{z_d})$,[29] of the tensor is sampled

$$\Phi_i = f_x l_i^2 + f_y m_i^2 + f_z n_i^2 \tag{8}$$

and (l_i, m_i, n_i) are the direction cosines that determine orientation of the i-th principal axis of the $\boldsymbol{D}$-tensor ($i = x_d$, y_d, z_d) with respect to the principal axes (x, y, z) of the sampling tensor frame.

In the case of a uniform distribution, $\Phi_{\rm i} = {}^1/_3$ for all principal components of the tensor $\boldsymbol{D}$, independent of orientation of the principal axes. In the most extreme counter case, all vectors are aligned parallel to one axis as in NH-bonds in the α-helix, and the sampling of D_i is maximal ($\Phi_i = 1$) when the corresponding i-th principal axis of the tensor is parallel to the helix axis, and minimal ($\Phi_{\rm i} = 0$) when it is

[22] Wang, Y.; Marquardt, J.; Wingfield, P.; Stahl, S.; Lee-Huang, S.; Torchia, D.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 7385−7386.

[23] Saupe, A. *Z. Naturforsch.* **1964**, *19a*, 161−171.

[24] In the following, the principal values of the tensors will be designated by a single subscript.

[25] This is also the case if all of the vectors are equally distributed along any three orthogonal axes, which for the analysis used here is indistinguishable from an uniform distribution. Both cases are referred to subsequently in the text as "uniformly distributed". Note that this set of three mutually orthogonal groups of vectors could be reduced to only three vectors, the simplest case being a set of three unit vectors, {100}, {010}, and {001}, since from the point of view of orientational sampling all parallel vectors are equivalent (assuming noiseless measurements). From the point of view of orientational sampling, this set of vectors fully and equally represents all three orthogonal orientations and therefore is necessary (although it might not be sufficient) for full characterization of the orientation of a $\boldsymbol{D}$-tensor. However, in contrast to a uniform distribution (which implies substantial number of vectors), this set provides only three independent observables, and is then insufficient for full characterization of a rank-2 tensor, which in general contains six (five in the case of a traceless tensor) independent components. A set of at least six (diffusion tensor) or five (alignment tensor) noncollinear and nonplanar vectors is required for a full characterization of a rank-2 tensor.

[26] Lipari, G.; Szabo, A. *J. Am. Chem. Soc.* **1982**, *104*, 4559−4570.

[27] Sass, J.; Cordier, F.; Hoffmann, A.; Rogowski, M.; Cousin, A.; Omichinski, J. G.; Löwen, H.; Grzesiek, S. *J. Am. Chem. Soc.* **1999**, *121*, 2047−2055.

[28] Salvatore, B.; Ghose, R.; Prestegard, J. *J. Am. Chem. Soc.* **1996**, *118*, 4001−4008.

[29] These principal values are defined in the $\boldsymbol{D}$-tensor principal frame, $\{x_d, y_d, z_d\}$, and ordered as $|D_z| \geq |D_y| \geq |D_x|$. In the following, the subscript d is omitted, where possible, and D_i always refers to this principal frame.

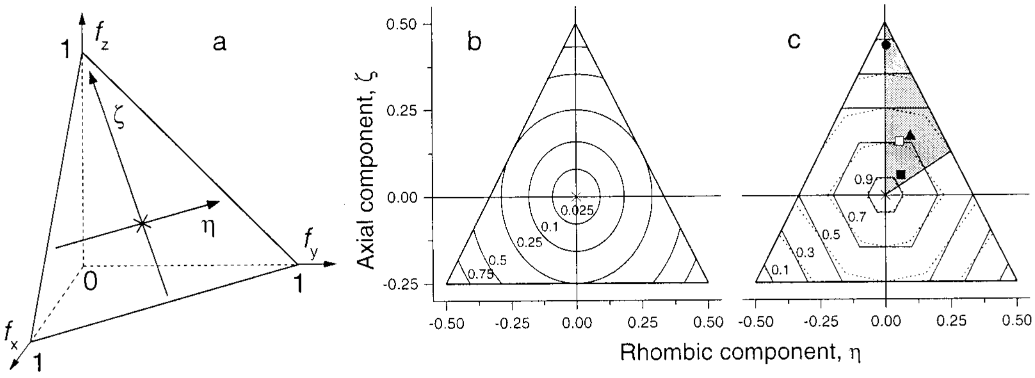

Figure 1. Geometrical representation of sampling fractions (f_i). Loci for the allowed positions in the $\{f_x, f_y, f_z\}$-space for the geometrical representation of various sets of unit vectors. (a) the orientation and 3D position of the allowed triangle, according to eq 3. (b) Two-dimensional parametrization of the triangle plane using the generalized coordinates $\{\eta,\zeta\}$, eqs 14−15, representing the rhombic and axial components of the sampling tensor. Ellipses and arcs represent contour lines corresponding to various values (indicated by the numbers) of the generalized sampling parameter, Ξ, according to eq 16. (c) The same plane representation as in (b), with the contour lines indicating various levels of the lower boundary for the quality factor, $\Lambda_{\min}$, as discussed in the text, eq 19. Solid lines represent $\Lambda_{\min}$ for an axially symmetric $\boldsymbol{D}$-tensor, and the same levels in the presence of rhombic components with $R = 0.2$ are shown with dotted lines. The shaded area is the minimal representation triangle (see text), where all allowed points are folded in, under symmetry transformations caused by ordering of the principal values of the sampling tensor. The origin, indicated by x in (a), (b), and (c), corresponds to the case of a uniform sampling, η, $\zeta = 0$. The three sides of the allowed triangle in (b, c) are described by the following equations: $\zeta + 1.5\eta = 0.5$; $\zeta - 1.5\eta = 0.5$; and $\zeta = -0.25$. Also indicated in c are four points representing the various sets of the backbone NH-vectors in the βARK PH domain: for all core amides (solid square), only for the α-helix (circle), and only for β-strands (triangle), and in ubiquitin (open square).

orthogonal to the helix. If the principal axis of the tensor makes an angle of 54.7° with the helix axis, then $\Phi_i = {}^1/_3$ in this case.

Generalized Quality Factor. We define a quality factor which reflects how efficiently a given structure samples all elements of the tensor of interest, as follows

$$\Lambda = 1 - \left|\frac{D_{\text{av}} - D_{\text{iso}}}{D_z - D_{\text{iso}}}\right| \tag{9}$$

Equation 9 is similar in form to an expression[30] representing the orientational order in lipid bilayers. Λ has a maximum value of 1 for a uniform distribution, which represents the optimal sampling, and a value of 0 when all the vectors are oriented along the z-axis of the principal frame of the tensor $\boldsymbol{D}$. Using eqs 7 and 8, the expression for the quality factor can be explicitly derived as:

$$\Lambda = 1 - \frac{1}{4}|(3f_z - 1)(3\cos^2\theta - 1) + 3(f_x - f_y)\sin^2\theta\cos 2\varphi - 3R\Delta\Phi| \tag{10}$$

Here $\{\varphi, \theta, \psi\}$ are the Euler angles describing the orientation of the principal axis frame of the $\boldsymbol{D}$-tensor with respect to the sampling tensor frame. $R = [D_y - D_x]/[D_z - (D_x + D_y)/2]$ is the degree of rhombicity of the $\boldsymbol{D}$-tensor. Note that the first two terms in eq 10 are independent of the actual values of the principal components of $\boldsymbol{D}$. The last term in eq 10 represents the effect of rhombicity of the tensor, and has the following angular dependence:

$$\Delta\Phi \equiv \Phi_{x_d} - \Phi_{y_d} = \frac{1}{2}(3f_z - 1)\sin^2\theta\cos 2\psi - (f_y - f_x)\left[\cos 2(\psi - \varphi)\sin^4\frac{\theta}{2} + \cos 2(\psi + \varphi)\cos^4\frac{\theta}{2}\right] \tag{11}$$

For an axially symmetric distribution of the vectors, that is, when $f_x = f_y$, the $\Delta\Phi$ term is not generally zero except in the case of a uniform distribution. For an axially symmetric tensor $\boldsymbol{D}$, the rhombicity $R = 0$, and the expression for the quality factor, eq 10, reduces to

$$\Lambda = 1 - \frac{1}{4}|(3f_z - 1)(3\cos^2\theta - 1) + 3(f_x - f_y)\sin^2\theta\cos 2\varphi| \tag{12}$$

Inspection of eq 12 reveals that for an axially symmetric $\boldsymbol{D}$-tensor, the value of the quality factor is independent of the principal values of $\boldsymbol{D}$. In the case of an axially symmetric distribution of vector orientations, eq 12 can be further simplified to

$$\Lambda = 1 - \frac{1}{4}|(3f_z - 1)(3\cos^2\theta - 1)| \tag{13}$$

The quality factor is maximal, $\Lambda = 1$, when all three orientations are sampled equally ($f_x = f_y = f_z = {}^1/_3$), independent of the orientation of the tensor $\boldsymbol{D}$ and of the degree of rhombicity of the latter. It also equals 1 in the case of an axially symmetric tensor $\boldsymbol{D}$, if its z-axis is oriented at the "magic" angle (54.7°) with respect to all three principal axes of the sampling tensor, independent of the f_x:f_y:f_z ratio. In the case where all of the vectors are oriented along the i-th principal axis of the diffusion tensor (i≠z), we obtain

$$\Lambda = 1 - \left|\frac{(2D_i - D_j - D_z)}{(2D_z - D_i - D_j)}\right|_{i,j\neq z} = 1 - \frac{1}{2}|1 - 3R\Delta\Phi|$$

which reduces to ${}^1/_2$ for an axially symmetric diffusion tensor (where $D_i = D_j = D_\perp$; $D_z = D_\parallel$).

Given the principal values and orientation of the $\boldsymbol{D}$-tensor, eqs 10−13 (see also eqs 17, 18 below) provide an estimate of the degree of accuracy of the derived tensor, for a particular set of vectors used for the determination. As outlined in the following sections, these equations also allow an assessment of the available accuracy of tensor determination for any given set of vectors, without prior knowledge of the tensor.

Geometric Representation of the Sampling Characteristics of a Vector Set. Equation 2 provides the basis for a useful and simple geometric representation of the distribution of vector orientations, as it allows one to represent each ensemble of unit vectors by a single point (or a single three-component vector) $\boldsymbol{f}$ with the coordinates $\{f_x, f_y, f_z\}$ in a general three-dimensional space, spanned by all values of $\boldsymbol{f}$.

Allowed Plane. According to eq 3, only two of the three components of a $\boldsymbol{f}$-vector characterizing the orientational distribution are independent. The locus of all allowed points in the $\{f_x, f_y, f_z\}$-space is then reduced to a plane triangle with the vertexes at $\{1,0,0\}$, $\{0,1,0\}$, and $\{0,0,1\}$, Figure 1a. Since the actual dimensionality of the locus is two, it is convenient to introduce a two-dimensional set of generalized coordinates $\{\eta, \zeta\}$ to parametrize the allowed plane and thus provide a direct characterization of the location of each point in this plane. We use the

[30] Sanders, C.; Hare, B.; Howard, K.; Prestegard, J. *Progr. Nucl. Magn. Reson. Spectrosc.* **1994**, *26*, 421−444.

following set of generalized coordinates:

$$\eta = \frac{1}{2}(f_y - f_x)$$

$$\zeta = \frac{1}{4}(3f_z - 1) \qquad (14)$$

where η and ζ range from -0.5 to 0.5 and from -0.25 to 0.5, respectively (Figure 1b,c). The allowed space is completely spanned by η and ζ, and can be further reduced when the fractions are ordered as $f_z \geq f_y \geq f_x$, see below.

It is worth mentioning that the generalized coordinates, ζ and η, introduced here, have certain physical meaning: they directly represent the axial and rhombic components of the sampling tensor:

$$\zeta = \frac{1}{3}\left[\Omega_z - \frac{1}{2}(\Omega_x + \Omega_y)\right]$$

$$\eta = \frac{1}{3}(\Omega_y - \Omega_x) \qquad (15)$$

These coordinates directly characterize the anisotropy and rhombicity of a given distribution of vector orientations and thus provide full characterization of the sampling tensor for a given set of vectors.

Map of the Generalized Sampling Parameter. In terms of the 3-dimensional space of the coordinates $\{f_x, f_y, f_z\}$, each value of the generalized sampling parameter Ξ can be represented by a sphere of radius $\sqrt{(2\Xi+1)/3}$ centered at the origin and described by the equation: $f_x^2 + f_y^2 + f_z^2 = (2\Xi + 1)/3$, following from eq 4. The loci of the allowed ***f***-vectors are then determined by the intersections of this sphere with the allowed triangle, which result in concentric circles (for $\Xi \leq 0.25$) or arcs ($\Xi > 0.25$) (Figure 1b). The constant-Ξ lines, in the generalized coordinate system, are described by

$$\Xi = 4\zeta^2 + 3\eta^2 \qquad (16)$$

Mapping of the Quality Factor. In the generalized coordinate system, the quality factor is given by

$$\Lambda = 1 - |\zeta(3\cos^2\theta - 1) - \frac{3}{2}\eta\sin^2\theta\cos 2\varphi - 3R\Delta\Phi| \qquad (17)$$

with the following angular dependence of the rhombic term:

$$\Delta\Phi = \frac{\zeta}{2}\sin^2\theta\cos 2\psi - \frac{\eta}{2}\left[\cos 2(\psi - \varphi)\sin^4\frac{\theta}{2} + \cos 2(\psi + \varphi)\cos^4\frac{\theta}{2}\right] \qquad (18)$$

As pointed out above, eqs 17 and 18 can be used to assess the accuracy of determination for a derived tensor value, given components of the sampling tensor. These equations also allow mapping of the quality factor establishing the relationship between sampling characteristics of vector sets and the available levels of accuracy of tensor determination. The following analysis assumes that the rhombic component in eq 17 is negligible, for simplicity.

For a particular value of the quality factor and a given relative orientation of sampling and ***D***-tensors, the corresponding values of η and ζ can be found by solving eq 17. Since the orientation of the ***D***-tensor is not known a priori, we need to find the point locations in the allowed plane corresponding to a given value of Λ for an *arbitrary* relative orientation of the two tensors, ***D*** and the sampling tensor. The obvious limiting cases are (i) $\Lambda = 0$ when the solutions to eq 17 exist only for $\theta = 0$ or $\theta = 90°$, $\varphi = 0, 90°$ and are located at the vertices of the triangle, and (ii) $\Lambda = 1$, when the solutions (in terms of η,ζ) exist for all values of $\{\theta,\varphi\}$ and are spread over the whole area of the triangle. The case of an arbitrary Λ is more complex.

Λ-Contours and the Lower Boundary Value, $\Lambda_{\min}$, of the Quality Factor. Since the values of the trigonometric functions in eq 17 are limited, for any given nonzero value of Λ there is always a region in the allowed plane centered at the origin, where there is no solution to eq 17. Contours can be plotted for any given value of Λ (Figure 1c) so that those points, within the area surrounded by the contour, have values of the quality factor greater than Λ *independent* of orientation and principal values of the ***D***-tensor. In contrast, for those points outside a given contour, the expected values of the quality factor could be greater, equal to, or less than Λ, depending on the orientation of the ***D***-tensor with respect to the sampling tensor frame. This leads to the concept of a lower boundary value, $\Lambda_{\min}$. Each point $\{\eta,\zeta\}$ in the allowed plane is characterized by a value of $\Lambda_{\min}$[31] which represents the lowest possible value of the quality factor over all possible orientations of the ***D***-tensor; $\Lambda_{\min}$ represents the "worst-case" estimate for the quality factor. The actual value of the quality factor depends on the orientation and rhombicity of the ***D***-tensor and is in the range $\Lambda_{\min} \leq \Lambda \leq 1$.[32] This geometric representation of $\Lambda_{\min}$-contours, defining areas of guaranteed high Λ values, $\Lambda \geq \Lambda_{\min}$ (Figure 1c), will help in experimental design, with the goal of selecting a vector set such that its representation on the allowed plane is as close as possible to the origin.

In the above discussion we assumed that the rhombicity effect is negligible. Significant rhombicity will perturb the clear geometrical picture outlined above. The effect is proportional to R and, therefore, is small for small degrees of rhombicity of the ***D***-tensor (Figure 1c). As can be seen from eq 18, for certain orientations of the ***D***-tensor frame, this effect could be negligible even for substantial values of R. It is also worth noticing that the rhombicity term in eq 17 is proportional to the distance from the origin in the allowed plane. Therefore, its absolute contribution is expected to be small in the target areas ($\eta,\zeta \ll 1$) close to the origin in the allowed plane, that is, those where $\Lambda \approx 1$.

"Minimal" Triangle Area. The ordering of the principal values of the sampling tensor

$$\Omega_z \geq \Omega_y \geq \Omega_x$$

introduces symmetry restrictions[33] which further reduces the allowed representation region to the "minimal triangle" (shaded in Figure 1c) described by the following conditions: $\eta \geq 0$, $\zeta \geq \eta/2$ and $\zeta + 1.5\eta \leq 0.5$. In this area, the minimal possible value of Λ (taking into account all possible orientations of the ***D***-tensor) is given by the following expression, valid for $|R| \leq 2$:[34]

$$\Lambda_{\min} = 1 - \max\left\{2\zeta + 1.5\eta\left|R\right|, \zeta + 1.5\eta + \frac{3}{2}\left|R\right|(\zeta - 0.5\eta)\right\} \qquad (19)$$

This allows a straightforward estimation of the lower bound for the quality factor for a given set of vectors, without any preexisting knowledge of the orientation of the ***D***-tensor. The lower bound for the quality factor in the presence of rhombic components of a ***D***-tensor is lower than that for an axially symmetric ***D***-tensor: $\Lambda_{\min} = 1 - \max\{2\zeta,\zeta + 1.5\eta\}$.

Derived Tensor Accuracy as a Function of Λ. To assess the relation between the accuracy of the estimated tensor quantity and the quality factor, we determined the overall rotational diffusion tensor using synthetic sets of relaxation data which include T_1, T_2, and NOEs generated from a specific distribution of vector orientations. Random errors of 2% were introduced into the T_1, T_2, and NOE data sets. To quantitate the errors in the estimated diffusion tensor, we define two quantities ϵ_d and ϵ_a, representing the relative errors in the principal values and in the orientation of the diffusion tensor, respectively. These are defined as

[31] It can be shown that, given the rhombicity factor R, each point in the allowed plane is characterized by a single value of $\Lambda_{\min}$.

[32] For any given set of vectors, there is always at least one orientation of a ***D***-tensor such that $\Lambda = 1$. This orientation is characterized by the following Euler angles: $\varphi = 45°$, $\theta = 54.7°$, and $\psi = -0.5\tan^{-1}[2\zeta/(\eta\sqrt{3})]$ (if $R \neq 0$).

[33] This ordering (hence symmetry restrictions) results from insensitivity of the sampling tensor to the directionality (sign) of the principal axes.

[34] A full expression for all values of R is

$$\Lambda_{\min} = 1 - \max\left\{2\zeta + 1.5\eta\left|R\right|, \zeta\left(1 + \frac{3}{2}\left|R\right|\right) + 1.5\eta\left|1 - \frac{1}{2}\left|R\right|\right|\right\}$$

$$\epsilon_{\mathrm{d}} = 100\sqrt{\sum_{i=x,y,z}\frac{1}{3}\left(\frac{D_i^{\mathrm{act}} - D_i^{\mathrm{calc}}}{D_i^{\mathrm{act}}}\right)^2} \tag{20}$$

$$\epsilon_a = 100\left[1 - \frac{\mathrm{Trace}|R^{\mathrm{act}}(\phi',\theta',\psi')R^{\mathrm{calc}}(\phi',\theta',\psi')^T|}{3}\right] \tag{21}$$

where D_i^{act} and D_i^{calc} are the actual and calculated values of the principal elements of the tensor, R^{act} and R^{calc} are the actual and calculated rotation matrices relating the molecular frame to the principal axis frame of the diffusion tensor and the superscript T represents the matrix transpose. Both ϵ_{d} and ϵ_{a} are zero for an exact match between the actual and estimated tensors, while ϵ_{a} has a maximal value of 100 when the actual and calculated orientations are orthogonal to each other. Absolute values are used in the evaluation of the trace in eq 21 because the signs of the *R's* are not experimentally determined. A few examples of the relationship between the quality factor Λ and the errors ϵ_{d} and ϵ_{a} are shown in Figure 4. All calculations were performed using the quadratic form of the diffusion tensor[4,6] with the tensor calculated from the relaxation data using singular value decomposition[35] (Ghose et al., in preparation).

Results and Discussion

In this section, we consider several applications of the theoretical approach developed above.

What Are the Sampling Properties of Known Protein Structures? To assess how various sets of interatomic vectors are sampled in real proteins, we performed a survey using structures from the Protein Data Bank. A set of protein structures was selected representing the whole variety of protein folds currently available. The selection criteria are:[36] the proteins are at least 30 residues long, have less than 40% sequence identity or more than 30% or 30 residues length difference from other set members, are either X-ray structures at ≤3 Å resolution or NMR structures. The set contained 1736 protein structures, of which 879 were single proteins and 857 were single chains in multisubunit proteins. Of these analyzed structures, 449 (26%) were NMR structures and the rest were X-ray structures (Supporting Information). In the latter cases, hydrogen atoms at the amide and α positions were added using standard algorithms.[37] The N- and C-terminal residues were not included. Sampling tensors were calculated for each structure for the following bond vectors in the protein backbone: NH, NC$^\alpha$, and C$^\alpha$C′ within the same residue and C′$_i$N$_{i+1}$ in the same peptide plane (i + 1), as well as for the C$^\alpha$H$^\alpha$ bonds. The distribution of the representative points in the allowed plane is shown in Figure 2. The statistics of the resulting distributions are presented in Table 1.

The NH-vectors present the least uniformly distributed sets (Figure 2a). This observation holds for both the NMR- and X-ray derived structures. To further verify this result, the same sampling tensor analysis was applied to C′O-vectors in the representative set of proteins (Figure 2e). The results were very similar to those for the NH vectors, which is expected since the directions of the C′O and NH bonds belonging to the same peptide plane are almost anti-parallel. The correlation coefficient between the distributions of the fractions of the NH- and C′O-vectors in the analyzed protein set was 0.98, 0.98, and 0.99 for f_x, f_y, and f_z values, respectively. For comparison, the corresponding values of the correlation coefficient between C′O and C$^\alpha$C′ were 0.43, 0.23, and 0.44. Because of their high correlation, the inclusion of both NH and C′O-vectors in one data set

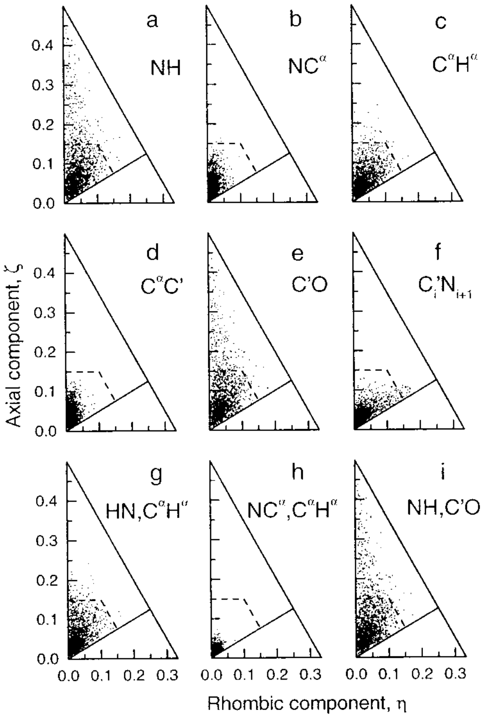

Figure 2. Sampling characteristics of the proteins represented in the PDB survey: the distributions of the representing points in the allowed plane, corresponding to sampling of the orientation space by (a) NH-, (b) NC$^\alpha$-, (c) C$^\alpha$H$^\alpha$-, (d) C$^\alpha$C′-, (e) C′O-, and (f) C′$_i$N$_{i+1}$-vectors, in the representative set of 1736 protein structures from Protein Data Base (see text). Shown in panels (g), (h), and (i) are the corresponding distributions for the grouped sets of vectors, {NH, C$^\alpha$H$^\alpha$}, {NC$^\alpha$, C$^\alpha$H$^\alpha$}, and {NH, CO}, respectively. Note that the generalized coordinates η and ζ characterize the rhombic and axial components of the orientational distribution for a given set of vectors, eq 15. Every protein structure from the representative set is represented on each panel by a single dot, with the coordinates $\{\eta,\zeta\}$ calculated for the specified set of vectors according to eqs 1, 2, 14, 15. Note that the ordering of the principal components of the sampling tensor was applied; therefore, all points are folded into the minimal triangle, indicated in Figure 1c. The dashed line is the $\Lambda_{\min}$ = 0.7 contour. The percentage of protein structures with 0.7 < $\Lambda_{\min}$ ≤ 1 is (a) 85.7, (b) 99.8, (c) 97.8, (d) 100.0, (e) 87.6, (f) 99.0, (g) 96.9, (h) 100, and (i) 86.9%.

Table 1. Statistics for the Representative Set of Protein Structures[a]

	generalized sampling parameter, Ξ		quality factor, $\Lambda_{\min}$	
bond vectors	68.3% level	90% level	68.3% level	90% level
NH	<0.05	<0.13	>0.79	>0.65
NC$^\alpha$	<0.01	<0.03	>0.90	>0.84
C$^\alpha$C′	<0.01	<0.02	>0.90	>0.85
C′O	<0.05	<0.12	>0.80	>0.67
C$^\alpha$H$^\alpha$	<0.02	<0.06	>0.85	>0.78
C′$_i$N$_{i+1}$	<0.01	<0.03	>0.90	>0.84
{NH, C$^\alpha$H$^\alpha$}	<0.03	<0.06	>0.85	>0.76
{NH, C′$_i$N$_{i+1}$}	<0.01	<0.02	>0.92	>0.87
{NH, C′O}	<0.05	<0.13	>0.79	>0.66
{C$^\alpha$H$^\alpha$, NC$^\alpha$}	<0.002	<0.004	>0.96	>0.94
all vectors	<0.005	<0.01	>0.93	>0.89

[a] The generalized sampling parameter, Ξ, and the quality factor, Λ, are assessed for each of the 1736 proteins in the representative PDB set. The levels of Ξ and Λ corresponding to 68.3 and 90% of the analyzed protein structures for each type of bond vectors are reported. These correspond to a set of 1186 or of 1563 proteins, respectively. The list of PDB structures used is available in the Supporting Information.

[35] Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical Recipes in C*; Cambridge University Press: New York, 1992.

[36] Sali, A.; Potterton, L.; Yuan, F.; van Vlijmen, H.; Karplus, M. *Proteins* **1995**, *23*, 318−326.

[37] Weiner, P. K.; Kollman, P. A. *J. Comput. Chem.* **1981,** *2,* 287−303.

Table 2. Sampling Parameters in Idealized Secondary Structural Elements[a]

vector set	Ω_z	Ω_y	Ω_x	f_z	f_y	f_x	Ξ
			β-Sheet				
NH	0.89	−0.41	−0.48	0.93	0.06	0.01	0.80
$C^{\alpha}H^{\alpha}$	0.97	−0.48	−0.49	0.98	0.01	0.00	0.95
$C'C^{\alpha}$	0.48	0.01	−0.49	0.65	0.34	0.00	0.32
{NH, $C^{\alpha}H^{\alpha}$}	0.93	−0.45	−0.47	0.95	0.03	0.02	0.86
{NH, $C^{\alpha}H^{\alpha}$, $C'C^{\alpha}$}	0.53	−0.14	−0.39	0.69	0.24	0.07	0.30
			α-helix				
NH	0.91	−0.45	−0.46	0.94	0.03	0.03	0.84
$C^{\alpha}H^{\alpha}$	0.13	0.03	−0.17	0.42	0.36	0.22	0.03
$C'C^{\alpha}$	0.24	−0.10	−0.14	0.49	0.27	0.24	0.06
{NH, $C^{\alpha}H^{\alpha}$}	0.38	−0.17	−0.21	0.59	0.22	0.19	0.15
{NH, $C^{\alpha}H^{\alpha}$, $C'C^{\alpha}$}	0.33	−0.14	−0.19	0.55	0.24	0.21	0.11
			3_{10}-helix				
NH	0.87	−0.43	−0.44	0.91	0.04	0.04	0.76
$C^{\alpha}H^{\alpha}$	0.15	0.11	−0.26	0.43	0.41	0.16	0.07
$C'C^{\alpha}$	0.34	−0.16	−0.18	0.56	0.23	0.21	0.11
{NH, $C^{\alpha}H^{\alpha}$}	0.31	−0.14	−0.16	0.54	0.24	0.22	0.10
{NH, $C^{\alpha}H^{\alpha}$, $C'C^{\alpha}$}	0.31	−0.15	−0.17	0.54	0.24	0.22	0.09

[a] Each standard secondary structural element was built for 12 residues of alanine using INSIGHT (MSI), and the structure was analyzed (see text) for the individual axial sample components.

does not improve the sampling (Figure 2i), unlike that for other pairs of sets of vectors. This highly nonuniform distribution of the NH and C′O bond orientations reflects the intrinsic feature of a folded protein, where amide hydrogens and carbonyl oxygens play essential roles in the hydrogen-bonding networks of the protein fold. Hydrogen bonding requires specific spatial and orientational arrangement in the N−H···O═C atoms,[38] resulting in orientational restrictions on the NH-bond. The hydrogen-bonding patterns characteristic for the elements of the secondary structure (α- and 3_{10}-helices, β-strands, see the examples below) then result in the distribution of the NH bond orientations being highly nonuniform. To illustrate this, we applied the same analysis to model structures of an α-helix, a 3_{10}-helix, and a β-strand, generated using INSIGHT (MSI). Table 2 shows the values of Ω_i, f_i ($i = x,y,z$) and Ξ for NH, $C^{\alpha}H^{\alpha}$, and $C'C^{\alpha}$ for these structures. In the case of an α-helix the NH-vectors are highly ordered (and aligned almost parallel to the helix axis) with $\Xi = 0.84$, whereas the $C^{\alpha}H^{\alpha}$-vectors are more evenly distributed[39] with $\Xi = 0.03$. The use of the $C^{\alpha}H^{\alpha}$-vectors would then provide a more uniform sampling of the orientational space and thus of a resulting second-rank tensor. The situation is similar in a 3_{10}-helix. In the case of β-strand, however, both NH and $C^{\alpha}H^{\alpha}$ are highly ordered, with $\Xi = 0.8$ and 0.95, respectively. Since these sets of vectors are almost anti-parallel to each other, their union does not improve sampling. However, including the $C^{\alpha}C'$-vectors in the set reduces Ξ to 0.30, indicating a more uniform sampling of the vector space.

Grouping NH- and $C^{\alpha}H^{\alpha}$-vectors in proteins from the representative database improves the sampling, compared to the NH-only data[6] (Table 1, Figure 2g). Interestingly, a much better improvement is achieved by the union of NH- and C'_iN_{i+1}-vectors. The optimal sampling results from a pairwise union were obtained for {$C^{\alpha}H^{\alpha}$, NC^{α}}-vectors (Figure 2h). In this last case, none of the structures analyzed had the $\Lambda_{\min}$ value below 0.85, and the corresponding Ξ greater than 0.024.

Rotational Diffusion Tensor from ^{15}N Relaxation Measurements. To illustrate the approach described above, we present a few examples which demonstrate its utility with respect to the rotational diffusion tensor in proteins. In the paragraphs below we discuss two specific examples, human ubiquitin and the Pleckstrin homology (PH) domain of the human β-adrenergic receptor kinase 1 (βARK1).[7]

Ubiquitin. The rotational diffusion tensor of ubiquitin has principal elements $D_{\|} = 4.43 \times 10^7$ s^{-1} and $D_{\perp}= 3.82 \times 10^7$ s^{-1}, as determined by ^{15}N relaxation (excluding those residues which exhibit large-amplitude motion as well those which are subject to conformational exchange)[5] (Ghose, Fushman, Cowburn, unpublished results). The principal axis frame of the diffusion tensor is related to the PDB-frame (1ubq.pdb) by a rotation R(48°,39°,0°). The sampling tensor (including the amide backbone ^{15}NH bond vectors only) is characterized by principal values given by $\Omega_z = 0.3148$, $\Omega_y = -0.0736$, and $\Omega_x = -0.2412$ with $f_z = 0.5432$, $f_y = 0.2843$, and $f_x = 0.1725$. The best sampled frame is related to the diffusion tensor frame by a rotation given by R(0°,82°,7°), this implies that the x-axis of the diffusion tensor plane is roughly parallel to the z-axis of the sampling frame and is the best sampled, while the z-axis of the diffusion tensor frame approximately corresponds to the x axis of the sampling frame and is the least sampled. This is confirmed by the values of Φ which are {0.5361,0.2825,0.1814}. The values of Ξ and Λ are 0.1084 and 0.7721, respectively.

***β*ARK PH Domain.** A similar analysis of ^{15}N relaxation data in the case of the βARK PH domain[7] (PDB code 1bak.pdb), which has $D_{\|} = 2.19 \times 10^7$ s^{-1} and $D_{\perp}= 1.72 \times 10^7$ s^{-1}, yields a sampling tensor characterized by $\Omega_z = 0.1090$, $\Omega_y = 0.0374$, and $\Omega_x = -0.1465$ ($f_z = 0.4060$, $f_y = 0.3583$, and $f_x = 0.2357$). The values of Ξ and Λ are given by 0.0232 and 0.9256, respectively, implying a near-optimal sampling of both the orientation space and the diffusion tensor. Note that this protein contains an extended (17 residues long) C-terminal α-helix. If only the α-helical residues are considered, the sampling tensor is characterized by $\Omega_z = 0.8722$, $\Omega_y = -0.4291$, and $\Omega_x = -0.4431$ ($f_z = 0.9148$, $f_y = 0.0473$, and $f_x = 0.0379$) and the generalized sampling parameter, Ξ, becomes 0.761 indicating a grossly inadequate sampling of orientational space, that is, the α-helix is insufficient to fully characterize the diffusion tensor of the system. Similar problems are expected in the case of helical bundles. To provide a better sampling of the orientational space, an additional set of vectors is therefore necessary. In the case of relaxation studies, this could be $C^{\alpha}H^{\alpha}$-[6] or $C^{\alpha}C'$-vectors. Although the $C'C^{\alpha}$- and $C'N$-vectors are more difficult to study by NMR relaxation than NH-vectors, some attempts have been made in this direction.[39] For the βARK PH domain considered here, a near optimal sampling of orientational space results from the NH-vectors in the β-strands. Analysis of the sampling properties of the β-strands yields the following values for the principal elements of the sampling tensor: $\Omega_z = 0.3353$, $\Omega_y = -0.0244$, and $\Omega_x = -0.3109$ ($f_z = 0.5569$, $f_y = 0.3171$, and $f_x = 0.1261$) and the generalized sampling parameter $\Xi = 0.1398$. Thus, the strands, as a substructural set, provide a better sampling of orientational space than the α-helix does. Further, by virtue of the PH domain fold, the NH-vectors in the strands are oriented approximately orthogonal to the helix axis, and therefore, a combination of the NH-vectors from the two sets of structural elements lowers the value of Ξ to the almost optimal value of 0.0232.

Alignment Tensor from Residual Dipolar Coupling Measurements. The theoretical approach presented above can be applied to molecular systems oriented in dilute liquid-crystalline media. For ubiquitin in the liquid-crystalline phase (5% w/v of DMPC:DHPC in a 3:1 ratio at 304 K), the alignment tensor has been found to be related to the PDB frame by R(42°,35°,

[38] Pauling, L.; Corey, R. B. *J. Am. Chem. Soc.* **1950**, *72*, 5349.

[39] Chiarparin, E.; Pelupessy, P.; Ghose, R.; Bodenhausen, G. *J. Am. Chem. Soc.* **1999**, *122*, 1758−1761.

Table 3. Sampling Characteristics of the Alignment Tensor for Ubiquin in Different Liquid-Crystalline Media

medium[al]	alignment tensor $\{A_x, A_y, A_z\}$[b]	PDB frame[c] $\{\alpha,\beta,\gamma\}$ [deg]	Λ[d]	Λ_{ax}[e]	sampling frame[f] $\{\varphi,\theta\}$ [deg]	$\{\Phi_x, \Phi^y, \Phi_z\}$
DMPC:DHPC = 3:1 (304 K)[2,17]	5.60, −3.60, −2.00	42.0, 35.0, 42.0	0.77	0.76	3.8, 86.2	0.45, 0.38, 0.17
DMPC:DHPC = 3:1 (310 K)[40]	2.02, 3.29, −5.31	33.1, 41.3, 50.7	0.78	0.77	2.6, 80.9	0.44, 0.38, 0.18
DMPC:DHPC = 3:1 (313 K)[27]	3.01, 6.40, −9.41	38.8, 31.5, 37.9	0.78	0.76	2.5, 89.9	0.48, 0.35, 0.17
DMPC:DHPC:A = 3:1:0.1 (310 K)[40]	1.79, 2.90, −4.69	33.2, 41.4, 49.1	0.78	0.77	2.6, 80.8	0.45, 0.37, 0.18
DMPC:DHPC:C = 3:1:0.1 (310 K)[40]	1.13, 6.02, −7.15	30.9, 29.9, 20.3	0.85	0.76	1.1, 87.8	0.54, 0.29, 0.17
PM = 1.0 (313 K)[27]	3.69, 4.70, −8.39	310.1, 128.6, 163.0	0.89	0.91	60.0, 76.8	0.22, 0.51, 0.27
PM = 1.9 (313 K)[27]	6.07, 8.48, −14.55	309.8, 127.7, 160.7	0.88	0.91	60.7, 77.5	0.21, 0.51, 0.27
PM = 1.9; N = 50 (313 K)[27]	4.91, 6.44, −11.35	310.3, 129.1, 162.3	0.89	0.91	59.6, 76.5	0.22, 0.51, 0.27
PM = 1.9; N = 50 (288 K)[27]	4.08, 5.52, −9.61	315.2, 128.3, 149.6	0.90	0.92	62.2, 73.5	0.20, 0.52, 0.28
PM = 7.0; N 50 (313 K)[27]	8.61, 12.78, −21.38	322.4, 112.2, 20.0	0.95	0.95	80.0, 75.3	0.37, 0.33, 0.30
PM = 12.0; N = 350 (313 K)[27]	1.60, 4.22, −5.82	330.8, 127.9, 47.5	0.96	0.99	67.8, 62.3	0.45, 0.22, 0.33

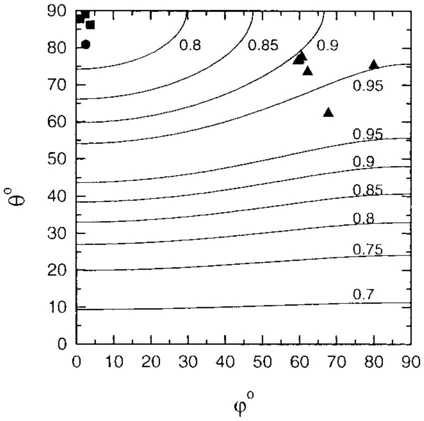

Figure 3. Dependence of the quality factor, Λ, on the orientation (φ, θ) of the unique axis of the alignment tensor with respect to the sampling tensor frame for ubiquitin. The angles φ and θ correspond to the azimuthal and polar angles, respectively. The contour lines represent various levels of Λ calculated using eq 10. An axially symmetric alignment tensor has been assumed. Only the core residues of ubiquitin have been included in the calculation of the sampling tensor. Also shown are the orientations of the alignment tensor experimentally observed in the DHPC:DMPC system (squares),[2,17,27,40] doped DHPC:DMPC system (circles)[40] and the purple membrane system (triangles).[27]

42°), with principal values $A_x = 5.6$, $A_y = -3.6$ and $A_x = -2.0$.[17] An analysis similar to that presented before yields values of 0.1084 and 0.7724 for Ξ and Λ with the *x*-axis of the alignment tensor being best sampled and the *z*-axis, the least, and $\Phi = \{0.4490,0.3763,0.1747\}$. For the backbone NH-vectors, the sampling tensor frame is related to the alignment tensor frame by $R(0°,90°,142°)$. Given the fractions of NH-vectors aligned along the three principal sampling axes in ubiquitin, eq 17 can be used to analyze how well the alignment tensor is defined, depending on the tensor orientation with respect to the sampling frame (Figure 3). Note that the lowest possible quality factor for this set of vectors is $\Lambda_{\min} = 0.69$ (eq 19). The analysis, assuming axial symmetry of the alignment tensor, indicates that the highest quality factor can be obtained in the region of $\theta = 40-60°$ for a large range in φ values. This implies that although the alignment of ubiquitin in the DMPC:DHPC medium corresponds to a rather high quality factor, $\Lambda > 0.77$, a change in the orientation of the alignment tensor by roughly 40° would result in a more optimal sampling for NH-vectors. Several methods are available to bring about a change in the orientation of the alignment tensor. These include doping the DMPC:DHPC system with ions[41] or the use of a different orienting medium, for example, phages[3] or purple membranes.[27] Table 3 shows the quality factors for ubiquitin in different liquid-crystalline environments (also depicted in Figure 3). It can be seen that in the case of ubiquitin the purple membrane system produces consistently higher quality factors, and in most cases, the *y*-axis of the alignment tensor is the best-sampled axis (Table 3). Inspection of Table 3 reveals that change in the quality factor from 0.76 (in the DHPC:DMPC system) to 0.99 (in the purple membrane system) is a result of a change of 69° in the orientation of the alignment tensor. In most cases the expected error in the quality factor due to the assumption of axial symmetry is of the order of 1% (Table 3).

Relation between the Quality Factor Λ and the Accuracy of the Diffusion Tensor Determination. How accurate are tensor values for a particular value of Λ? To answer this question, the simulation approach outlined in the Theory section was applied here to the determination of the rotational diffusion tensor, as illustrated in Figure 4. Similar results are expected for the accuracy of derivation of the alignment tensor from residual dipolar couplings, because of the same functional form as the quadratic form[4,6] used here for the rotational diffusion tensor.

Consider the simplest case of axially symmetric diffusion tensor ($D_\perp = 3.0 \times 10^7$ s^{-1} and $D_\| = 4.5 \times 10^7$ s^{-1} which corresponds to $\tau_c = 4.76$ ns and the anisotropy factor $D_\|/D_\perp$=1.5), and an axially symmetric sampling tensor. We simulated multiple sets of vectors with different values of the sampling parameter, Ξ, ranging from 0.94 to 0.03 with the values of f_z and the f_z/f_x ratio in the range from 0.94 to 0.44 and from 34.2 to 1.57, respectively. The various distributions were generated by starting with unit vectors equally partitioned between three cones of semi-angle 10°, 14° and 18° (with the cone axis aligned along the *z*-axis). For each successive step, an additional cone of semi-angle 4° greater than the largest semi-angle for the preceding distribution was added and the vectors equally partitioned between the cones. This procedure was continued until a cone semi-angle of 170° was reached. The total number of vectors varied from 195 to 216 in the various distributions. An axially symmetric distribution of vector orientations for each cone was achieved by assigning uniformly distributed values of the azimuthal angle, equi-partitioned in the range 0°−360°. For a completely anisotropic distribution of the vectors, the azimuthal angle was restricted to the 0−220° range. Synthetic relaxation data (T_1, T_2 and NOE) were generated for the above values of the diffusion tensor elements and of Ξ**.** The diffusion tensor was calculated from the relaxation data (see **Derived Tensor Accuracy as a Function of** Λ in

[40] Engelke, J.; Ruterjans, H. *J. Biomol. NMR* **1995**, *5*, 173−182; Cordier, F.; Brutcher, B.; Marion, D. *J. Biomol. NMR* **1996**, *7*, 163−168; Zheng, L.; Fischer, M.; Zuiderweg, E. *J. Biomol. NMR* **1996**, *7*, 157−162; Dayie, K. T.; Wagner, G. *J. Am. Chem. Soc.* **1997**, *119*, 7797−7806; Engelke, J.; Ruterjans, H. *J. Biomol. NMR* **1997**, *9*, 63−78; Allard, P.; Härd, T. *J. Magn. Reson.* **1997**, *126*, 48−57; Ghose, R.; Huang, K.; Prestegard, J. H. *J. Magn. Reson.* **1998**, *135*, 487−499;.Carlomagno, T.; Maurer, M.; Hennig, M.; Griesinger, C. *J. Am.Chem. Soc.* **2000**, *122*, 5105−5113.

[41] Ramirez, B. E.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 9106−9107.

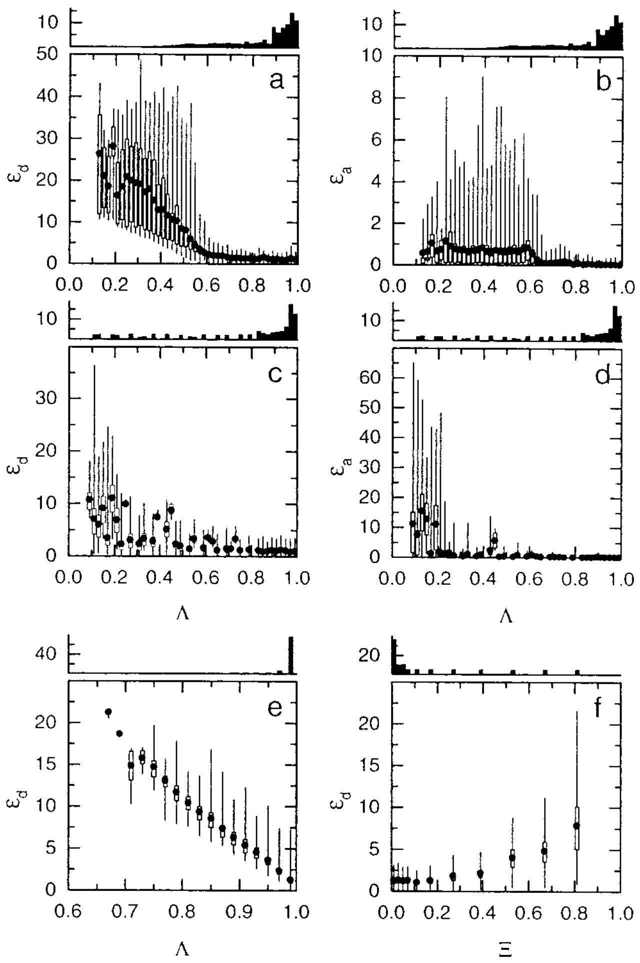

Figure 4. Illustration of the robustness and the complexity of variations in the derived values of the rotational diffusion tensor. Panels a–d illustrate that the percentile errors of magnitude (ϵ_d) and orientation (ϵ_a) of the tensor are small for any $\Lambda > 0.7$ for the simulation described in Theory. The variation in the magnitude of the errors and their distribution between the axially symmetric (a, b) and fully anisotropic (c, d) cases illustrate the complex variability of the error distribution in different models. Data shown in panels a and b were derived for an axially symmetric sampling tensor and an axially symmetric diffusion tensor ($D_\perp = 3.0 \times 10^7$ s^{-1} and $D_\| = 4.5 \times 10^7$ s^{-1}) with the unique axes of the two tensors orthogonal to each other. The corresponding cases for a fully anisotropic sampling tensor Ω and a fully anisotropic diffusion tensor ($D_x = 2.5 \times 10^7$ s^{-1}, $D_y = 3.5 \times 10^7$ s^{-1} and $D_z = 4.5 \times 10^7$ s^{-1}) are depicted in panels c and d. The variation of ϵ_d with Λ (panel e) and Ξ (panel f) was derived for axially symmetric sampling and diffusion tensors (same as Figure 4a,b) with the unique axes of the two tensors at an angle of 54.7° with respect to each other. In this particular case, Λ becomes a less accurate estimator of the quality of the tensor determination, while the observed errors in the diffusion tensor still correlate well with the generalized sampling parameter Ξ (see text). The errors in the diffusion tensor were calculated for computer-simulated relaxation data as described in the text (eqs 20–21). The total number of data points was 38 000 (a–d) and 19 000 (e–f). For each of the 38 distributions, there were 1000 and 500 Monte Carlo cases, respectively. For presentation purposes, the data were distributed between 50 bins of equal width covering the observed range of Λ values. Shown are the total ranges of ϵ_d or ϵ_a values (thin vertical lines), the range from the first to the third quartile (open bars), and the average value (solid circles) for each bin. Also indicated with bar diagrams on top of each panel is the number of data points in each bin (in percent of the total number).

Theory section). The spread in the values characterizing the diffusion tensor was determined from 1000 Monte Carlo steps using the random error in the relaxation data. Figure 4 parts a and b depict the errors in the principal elements (ϵ_d) and the orientation (ϵ_a) of the diffusion tensor in the case where the sampling tensor was taken to be axially symmetric and has its unique axis orthogonal to the unique axis of the diffusion tensor. The results indicate that although the inaccuracy in the diffusion tensor determination is high for small values of the quality factor, it becomes reasonably small (both ϵ_d and ϵ_a fall below 5%) for $\Lambda > 0.7$. Similar trends were observed for a completely anisotropic diffusion tensor ($D_x = 2.5 \times 10^7$ s^{-1}, $D_y = 3.5 \times 10^7$ s^{-1} and $D_z = 4.5 \times 10^7$ s^{-1}) and a completely anisotropic sampling tensor (Figure 4 parts c and d) (the principal axis frames of the two tensors are assumed to be co-incident), although the magnitude of the errors can be larger than in the axially symmetric case. Although the details of the relationship of Λ with ϵ_d and ϵ_a are complex in the general case, and depend on the nature of the distribution, the diffusion tensor and the relative orientation of the two tensors, extensive simulations show that the errors are expected to be within experimental error[42] for values of Λ greater than 0.7. In all of the cases we looked at, this corresponded to a $\Xi < 0.25$.

Possible Limitations of the Quality Factor Approach. The quality factor introduced here might not be an adequate estimator of the accuracy of tensor determination in a particular case of an axially symmetric ***D***-tensor, if the unique axis of the tensor is oriented at the magic angle with respect to all three principal sampling axes. The theoretical quality factor predicted from eq 12 is then 1 and is independent of the actual distribution of vectors along the principal sampling axes.[32] The "magic angle" orientation is particularly troublesome for determination of any axially symmetric rank-2 tensor, which in this case reduces to a single value. Therefore, it is important to understand the limitations of the quality factor Λ as an accurate estimator of the errors in tensor determination in this particular case.

A particular case that deserves consideration is when both the ***D***-tensor and the sampling tensor are axially symmetric with their unique axes oriented at the magic angle (54.7°) to each other. Since selection of the *x*- and *y*-axes of the sampling tensor is then arbitrary, they can always be selected (e.g., $\varphi = 45°$) so that the unique axis of the ***D***-tensor makes the magic angle to all three axes of the Ω tensor. To understand the relationship between the sampling parameters and the accuracy of tensor determination in this particular case, we also performed simulation of the rotational diffusion tensor. As indicated by the results of our simulations shown in Figure 4e, the errors in tensor determination do not correlate well with the quality factors determined from the calculated rotational diffusion tensor (Figure 4e). In particular, the errors remain large even for the estimated quality factor close to 1. Thus the quality factor becomes a less accurate estimator of the errors in tensor determination in this special case. However, even in this particular case the errors in tensor determination drop to their limiting values (within the "experimental" errors) for $\Xi < 0.25$ (Figure 4f). Thus, the generalized sampling parameter Ξ remains an accurate estimator of the expected errors of the diffusion tensor. As follows from these simulations, a second rank tensor quantity, such as the rotational diffusion tensor, can be accurately determined for vector distributions with values of Ξ <

[42] The level of "experimental" error in ϵ_d and ϵ_a was obtained as follows. Relaxation data (T_1, T_2, and NOE) were simulated for an uniformly distributed set of 1,000,000 unit vectors. The influence of the measurement errors on the diffusion tensor calculated from these relaxation data was estimated using 1000 Monte Carlo simulations utilizing the 2% random error (as in all other simulations, see Theory) added to the synthetic relaxation data. The uncertainty in the ϵ_d and ϵ_a values obtained from the resulting distributions of the principal values and orientations of the diffusion tensor, is what we term "experimental error".

0.25. Note that these values of Ξ correspond to relatively good orientational sampling, as the maximum available anisotropy of the sampling tensor at $\Xi = 0.25$ is $\zeta = 0.25$.

The example considered here is a particular case related to the magic angle orientation of the two tensors. For any other orientation and/or in the general case of the anisotropic ***D***-tensor, the quality factor treatment introduced here is valid.

Practical Guidelines for the Optimal Design of Experiments. The theory developed here makes it possible (a) to assess the quality of a second-rank tensor determined using orientational dependence of physical properties as, for example, rotational diffusion or alignment tensors, and (b) to predict the likely limitations of the vector set available for these studies prior to actual experimental measurements. This permits optimization of the experimental design, to improve accuracy. As follows from the discussion above, each set of interatomic vectors can be represented by a point on the allowed plane (Figure 1). Depending on the location of the representing point with respect to the origin, the lower-bound ("worst-case") estimate of the level of accuracy (quality factor Λ) could be performed using eqs 17−19, without prior knowledge of the orientation of the tensor to be determined. A further refinement is then possible given additional information regarding tensor magnitude and orientation. A simple rule follows from the theory developed here: the closer the representing point is to the origin (Figures 1−2) (i.e., the closer the distribution of the vectors is to a uniform distribution), the better the sampling of the tensor and the higher the accuracy of its determination. The experimenter may then either select a particular subset of the available vectors or include additional measurements (e.g., $C^{\alpha}H^{\alpha}$- or NC^{α}-vectors, in addition to NH-vectors) to ensure the desired level of tensor sampling, represented by the quality factor. Note also that the generalized sampling parameter Ξ, eqs 4, 16, provides a quantitative measure of the degree to which a particular set of vectors could be safely considered as uniformly distributed, essential for approaches based on this assumption.[13] These considerations address issues related to the intrinsic properties of a finite set of vectors available in a real experiment in real molecular systems and are unrelated to the issue of measurement precision and accuracy.

Conclusions

We have developed a quantitative approach to determine (1) how well interatomic vectors in a particular protein structure sample orientation space, (2) how well this particular distribution of bond vectors samples the various components of a second-rank tensor, and (3) the ability of this distribution of bond vectors to completely characterize the tensor. This approach is in general applicable to any second-rank tensor property whose determination relies on the sampling of the angular space by the structure. It allows optimization of the experimental design to improve accuracy. The utility of the proposed approach is demonstrated here for the overall rotational diffusion and alignment tensors. The analysis of a set of 1736 protein structures representing a variety of known protein folds, provided statistical analysis and revealed characteristic patterns in the orientational sampling by various bonds in a protein. It should be mentioned here that this method is not applicable to properties which are dependent on local structure such as, for example, the chemical shift anisotropy of a given nucleus. In this paper, the approach has been illustrated with protein structures, but it is equally applicable to other molecules, including nucleic acids, and carbohydrates.

Acknowledgment. Supported by an NIH grant (GM-47021).

Supporting Information Available: A list of PDB codes for the 1736 representative protein structures used in the PDB survey (PDF). This material is available free of charge via the Internet at http://pubs.acs.org.

JA001128J