J|A|C|S
ARTICLES
Published on Web 09/13/2002

A Novel Approach to the Retrieval of Structural and Dynamic Information from Residual Dipolar Couplings Using Several Oriented Media in Biomolecular NMR Spectroscopy

Joel R. Tolman*

Contribution from the Institut de Chimie Moléculaire et Biologique, École Polytechnique Fédérale de Lausanne BCH, 1015 Lausanne, Switzerland

Received March 6, 2002

Abstract: The interpretation of residual dipolar couplings in terms of molecular properties of interest is complicated because of difficulties in separating structural and dynamic effects as well as the need to estimate alignment tensor parameters a priori. An approach is introduced here that allows many of these difficulties to be circumvented when data are acquired in multiple alignment media. The method allows the simultaneous extraction of both structural and dynamic information directly from the residual dipolar coupling data, in favorable cases even in the complete absence of prior structural knowledge. Application to the protein ubiquitin indicates greater amplitudes of internal motion than expected from traditional ^{15}N spin relaxation analysis.

Introduction

The exquisite sensitivity of residual dipolar couplings (RDCs) to the structure and dynamics of macromolecules in solution, coupled with the increasing variety of available alignment media, has expanded the scope of problems that can be effectively studied by NMR.[1−5] However, the interpretation of RDC data in terms of molecular parameters is complicated by several factors. Difficulties in separating contributions to RDCs arising from structural and dynamic properties as well as the need to estimate structurally dependent alignment tensor parameters continue to restrict the applicability and to place limits on the attainable accuracy of RDC-based methods. These concerns are starting to be addressed,[6−11] but RDCs remain primarily tools for structural refinement, while the dynamics are typically neglected. The development of methods for the inclusion of the effects of dynamics is nevertheless a worthwhile objective because of the expected improvements in accuracy of structural constraints as well as the prospect of exploiting the sensitivity of RDCs to motional processes occurring on time scales ranging from picoseconds to milliseconds.[7,9,11,12]

We introduce here a novel approach to these ends when data from multiple alignment media are available. This approach, referred to as Direct Interpretation of Dipolar Couplings (DIDC), provides a route whereby both the structural and dynamic content of the RDC data can be extracted without making any prior assumptions about alignment tensors. More specifically, the coupling data are converted directly into mean internuclear vector orientations and associated generalized order parameters with high accuracy. The approach also provides a description of the direction and asymmetry of motions but with much lower accuracy at present. The final objective is similar to that described recently[9,11] but removes the requirement for a priori structural information while maintaining the requirement that RDC measurements be made employing five sufficiently different (i.e., linearly independent) alignment media. Furthermore, the degree to which the requirement for five independent media has been met, referred to herein as the completeness of the measured RDC data, can be assessed in most cases by analysis of the measured data alone. This data analysis procedure also provides for signal averaging among sets of RDC measurements acquired in different alignment media, leading to gains in precision and accuracy of the determined parameters.

Although there is one recently appearing account,[11] it is not entirely clear how easily the requirement for five independent media can be met on a routine basis. For cases in which the measured data fall short of this requirement, a least-squares based approach is introduced that allows a refined model to be generated directly from an initial model and the data provided

* Current address: Department of Chemistry, Johns Hopkins University, 3400 N. Charles Street, Baltimore, MD 21218. E-mail: tolman@jhu.edu.

(1) Prestegard, J. H.; Al-Hashimi, H. M.; Tolman, J. R. *Q. Rev. Biophys.* **2000**, *33*, 371−424.
(2) Prestegard, J. H.; Kishore, A. I. *Curr. Opin. Chem. Biol.* **2001**, *5*, 584−590.
(3) Tolman, J. R. *Curr. Opin. Struct. Biol.* **2001**, *11*, 532−539.
(4) Bax, A.; Kontaxis, G.; Tjandra, N. *Methods Enzymol.* **2001**, *339*, 127−174.
(5) de Alba, E.; Tjandra, N. *Prog. Nucl. Magn. Reson. Spectrosc.* **2002**, *40*, 175−197.
(6) Moltke, S.; Grzesiek, S. *J. Biomol. NMR* **1999**, *15*, 77−82.
(7) Tolman, J. R.; Al-Hashimi, H. M.; Kay, L. E.; Prestegard, J. H. *J. Am. Chem. Soc.* **2001**, *123*, 1416−1424.
(8) Hus, J. C.; Marion, D.; Blackledge, M. *J. Am. Chem. Soc.* **2001**, *123*, 1541−1542.
(9) Meiler, J.; Prompers, J. J.; Peti, W.; Griesinger, C.; Bruschweiler, R. *J. Am. Chem. Soc.* **2001**, *123*, 6098−6107.
(10) Sass, H.-J.; Musco, G.; Stahl, S. J.; Wingfield, P. T.; Grzesiek, S. *J. Biomol. NMR* **2001**, *21*, 275−280.
(11) Peti, W.; Meiler, J.; Bruschweiler, R.; Griesinger, C. *J. Am. Chem. Soc.* **2002**, *124*, 5822−5833.
(12) Tolman, J. R.; Flanagan, J. M.; Kennedy, M. A.; Prestegard, J. H. *Nat. Struct. Biol.* **1997**, *4*, 292−297.

10.1021/ja0261123 CCC: $22.00

without the explicit appearance of terms for alignment tensors. Use of at least three independent alignment media allows estimates of generalized order parameters to be obtained in addition to refined mean internuclear vector orientations. This delivers lower accuracy and precision than can be attained using five independent media but should at present be experimentally feasible for most systems. This provides a useful tool for the study of intermediate time scale (10^{-8}–10^{-6} s) motions not easily accessible to other techniques with atomic resolution. The new formalism is illustrated with an application to the protein ubiquitin, employing amide N−H RDC data corresponding to three independent alignment tensors.

Theoretical Results

The ability to obtain complementary information by acquiring RDC data in different alignment media[13,14] provides the underlying motivation for this work. If it can be assumed that the structural and dynamic properties of the macromolecule do not change when placed in different alignment media, it becomes possible to in effect obtain a series of snapshots of the molecule from different perspectives simply by changing the alignment medium. The presence of motion complicates matters, and thus we also require that the amplitudes of internal motions are small enough that the concept of a *mean structure* remains meaningful. This amounts to requiring that internal motions do not bring about any large changes in the overall alignment tensor, so that internal motion and overall motion remain uncorrelated. Under these assumptions, the question becomes how to reconstruct an "image" of the structure and dynamics on the basis of a series of different snapshots obtained using different alignment media. The first step toward this goal is therefore to express the entire problem in terms of a single matrix equation:

$$\mathbf{D} = K\mathbf{BA};\ K = -\left(\frac{\mu_0}{4\pi}\right)\frac{\gamma_I\gamma_S h}{2\pi^2 r^3_{IS}} \qquad (1)$$

The matrix **D** is formed directly from the RDC measurements and thus has dimensions N x M, where N is the number of experimentally measured residual IS-dipolar couplings and where M is the number of different data sets. The interaction constant K is dependent on the magnetogyric ratios ($\gamma_{I,S}$) of the spins and their internuclear distance, r_{IS}. For purposes of the current work, each data set is comprised of a single set of amide N−H RDC measurements and K will be presumed constant. These RDC data sets are optimally recorded using many different alignment media but can also include couplings measured in duplicate experiments or under only slightly different aligning conditions. The matrixes **A** and **B** have dimensions 5 x M and N x 5, respectively. The columns of **A** represent the alignment tensors operative for each of the M data sets. The rows of **B** contain the structural and dynamic information that we wish to determine, that is, the mean orientation and accompanying description of dynamics for each of the N internuclear vectors for which measurements are available. Each of the columns of **A** and rows of **B** contains the five independent elements of the second rank Cartesian tensor describing the specific interaction. These five independent tensorial elements

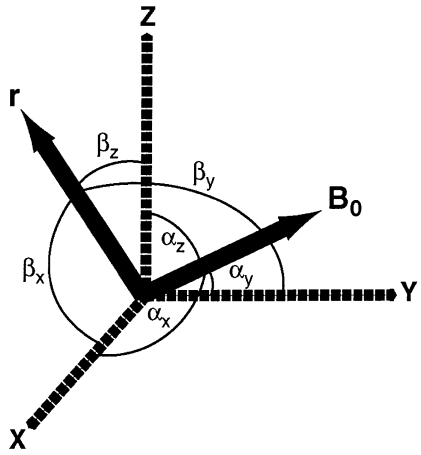

Figure 1. The angles α_k and β_k used to describe the orientation of a specific internuclear vector and the external magnetic field, respectively, relative to the axes of common reference frame.

are related to the elements of the relevant 3 × 3 Cartesian tensor as described in eq 2.

$$\mathbf{T}^{(2)} = \left[T_{zz}, \frac{1}{\sqrt{3}}(T_{xx} - T_{yy}), \frac{2}{\sqrt{3}}T_{xz}, \frac{2}{\sqrt{3}}T_{yz}, \frac{2}{\sqrt{3}}T_{xy}\right] \qquad (2)$$

These second-rank independent tensorial elements are written as row vectors to construct the matrix **B** and as column vectors in matrix **A**. The elements of the 3 × 3 Cartesian tensors, corresponding to each of the rows of **B** and columns of **A,** are averages over different angular functions and are constructed in analogy to the order tensor formalism.[15]

$$A_{mn}(j) = \left\langle \frac{1}{2}(3\cos\alpha_m \cos\alpha_n - \delta_{mn}) \right\rangle_j;\ B_{mn}(i) = \left\langle \frac{1}{2}(3\cos\beta^i{}_m \cos\beta^i{}_n - \delta_{mn}) \right\rangle \qquad (3)$$

The angle brackets denote that a time average is taken and δ_{mn} represents the Kronecker delta function. The angles α_k and β_k ($k = x, y, z$) are depicted in Figure 1 and describe the magnetic field vector (in the jth alignment medium) and the ith internuclear vector, respectively, relative to an arbitrary molecule-fixed reference frame. Similar formulations have been described previously.[1,16] The normalization scheme utilized in eq 2 produces vectors of unit length in the absence of time averaging. Thus, for a rigid molecule, each row of the matrix **B** will have a length of one.

Physical Interpretation. If the matrix **B** can be determined, then the desired structural and dynamic attributes are readily extracted. These attributes are the same as described by Meiler et al.,[9] but a different procedure for extracting the information is described here. First, recall that each of the M columns of **A** corresponds to a specific alignment tensor. For each column of **A,** the 3 × 3 order tensor can be formed on the basis of the relationship described by eq 2 and subsequently diagonalized to determine its principal values and the orientation of its principal axis system (PAS), expressed relative to some arbitrary initial molecular frame in terms of Euler angles. An identical

(13) Ramirez, B. E.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 9106−9107.
(14) Al-Hashimi, H. M.; Valafar, H.; Terrell, M.; Zartler, E. R.; Eidsness, M. K.; Prestegard, J. H. *J. Magn. Reson.* **2000**, *143*, 402−406.
(15) Saupe, A. *Z. Naturforsh.* **1964**, *19a*, 161.
(16) Prestegard, J. H.; Tolman, J. R.; Al-Hashimi, H. M.; Andrec, M. *Protein structure and dynamics from field induced residual dipolar couplings.*; Krishna, N. R., Berliner, L. J., Eds.; Plenum: New York, 1999; Vol. 17, pp 311−355.

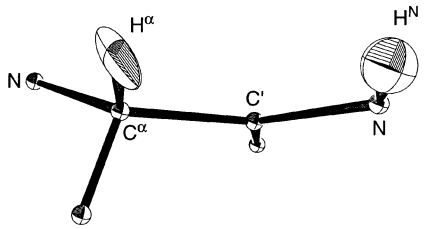

Figure 2. Conceptual illustration of the structural *and* dynamic information that can be extracted from residual dipolar couplings. The ellipsoids centered on the H^{α} and H^{N} atoms represent the nature and extent of motion of $C^{\alpha}-H^{\alpha}$ and $N-H^{N}$ internuclear bond vectors. The breadths of the ellipsoids parallel to the bond vectors do not have any significance. Generated using the program ORTEP-III.[18]

procedure can be carried out for each of the N rows of the matrix **B**, with the resulting principle values and Euler angles carrying a different physical interpretation. For a specific interaction i, these resulting Euler angles (α_i, β_i, γ_i) and principal values comprise the information schematically shown in Figure 2. The angles α_i and β_i correspond to the polar angles of the principal axis (Z) expressed in the initial arbitrary reference frame.[1] Neglecting cases in which the internal motion is highly anisotropic and of very large amplitude, these two angles describe the mean orientation of the ith interaction vector and hence correspond to the desired *structural* information. The corresponding principal values describe amplitudes of *internal motion* of the ith interaction, extending from ps to ms time scales. In analogy to the description of the alignment tensor, these two principal values may be described in terms of the principal component ($S_{zz,i}$) along with an asymmetry parameter (η_i). When present (i.e., when $\eta_i \neq 0$), the specification of motional anisotropy requires an extra angle (γ_i) for the description of its principal direction. The generalized order parameters are related to the principal values via $S_i^2 = S_{zz,i}^2\,(1 + (1/3)\eta_i^2)$ but can also be obtained independently of the described diagonalization procedure by computing, respectively, the Euclidean norm for each row of the matrix **B**. A 2-fold ambiguity remains for each vector because the dipolar interaction is invariant under permutation of the two interacting spins.

Determination of the matrix **B** is complicated because the RDC data alone are not sufficient to provide a direct solution. This can be seen by consideration of eq 1, where one would like to use the data (**D**) to determine **B** but generally will not have any information about **A** beforehand. Order tensor based approaches can bypass some of these difficulties by introducing geometric constraints which relate different interaction vectors within a putative rigid fragment.[7,17] However, for the present work we seek to obtain independent descriptions of the mean orientation and motion of individual interaction vectors and thus will consider methods for obtaining the matrix **B** directly from the data **D**.

Matrix Algebraic Background. The approach for the determination of the matrix **B** will take advantage of the powerful algebraic matrix methods which exist to deal with linear problems such as eq 1. A full development of these techniques can be found in sources dealing with the topic of generalized inverses.[19,20] Only the concepts important for this work will be discussed here, deliberately placed within the context of RDC analysis. Indeed, some of the fundamental tools are already in widespread use for determination of the alignment tensor by the method of Singular Value Decomposition (SVD).[21] We recall that the SVD of a rectangular matrix can be written[22]

$$[\mathbf{M}(k \times l)] = [\mathbf{U}_M(k \times r)][\mathbf{W}_M(r \times r)][\mathbf{V}^{\mathrm{tr}}{}_M(r \times l)] \quad (4)$$

in which $\mathbf{U}_M$ and $\mathbf{V}_M$ are column-orthogonal (and column-normalized) matrixes, $\mathbf{W}_M$ is a diagonal matrix containing the nonzero singular values, and the superscript tr is used to denote the transpose. Provided that there are no degenerate singular values, this decomposition is unique except for respective permutations of the elements of **W** and the columns of **U** and **V**. The matrixes $\mathbf{U}_M$, $\mathbf{V}_M$, and $\mathbf{W}_M$ will refer herein exclusively to the matrixes resulting from the SVD of the matrix indicated by the subscript. The number of nonzero singular values, r, determines the rank of the matrix **M** and can never exceed its smallest dimension. If the rank is equal to the smallest dimension of the matrix, then the matrix is referred to as full rank or nonsingular. For applications involving experimental data, the concept of rank can become ambiguous as a result of measurement errors. In this case, the nonsingularity of the matrix under consideration is often monitored and reported in the form of a condition number, which is obtained by computing the ratio of largest to smallest singular values. A large condition number may indicate that the matrix is effectively singular for purposes of further analysis.

It is well known that the determination of an alignment tensor **A** using SVD[21] provides the best-fit solution for **A** given some data (**D**) and a structural model (**B**). In typical applications, the matrix **B** will correspond to a rigid structure and **D** and **A** will correspond to the alignment medium under consideration and thus appear as column vectors. The solution to the problem of determining the best-fit solution for **A** proceeds by forming the Moore−Penrose generalized inverse of **B** by means of SVD. The Moore−Penrose generalized inverse of **B**, denoted $\mathbf{B}^{+}$, is constructed as follows:

$$\mathbf{B}^{+} = \mathbf{V}_B \begin{pmatrix} (1/w_1) & & & & \\ & (1/w_2) & & & \\ & & (1/w_3) & & \\ & & & (1/w_4) & \\ & & & & (1/w_5) \end{pmatrix} \mathbf{U}^{\mathrm{tr}}{}_B \quad (5)$$

with the singular values w_k obtained from the individual elements of $\mathbf{W}_B$. Provided that the internuclear vectors for which measurements have been made are at least five in number and are sufficiently independent in orientation, the singular values w_k will be of comparable magnitude and the condition number will be small. The consequent nonsingularity of the matrix **B** will manifest algebraically as $\mathbf{B}^{+}\mathbf{B} = \mathbf{1}$, where **1** refers to the identity matrix. A nonsingular matrix **B** allows a unique best-fit alignment tensor $\mathbf{A}_{\mathrm{bf}}$ (in the least-squares sense) to be obtained by left multiplication of eq 1 by $\mathbf{B}^{+}$, leading to $\mathbf{A}_{\mathrm{bf}} =$

(17) Hus, J. C.; Marion, D.; Blackledge, M. *J. Mol. Biol.* **2000**, *298*, 927−936.
(18) Burnett, M. N.; Johnson, C. K. *ORTEP−III: Oak Ridge Thermal Ellipsoid Plot Program for Crystal Structure Illustrations*; Oak Ridge National Laboratory: Oak Ridge, TN, 1996.
(19) Albert, A. *Regression and the Moore−Penrose Pseudoinverse*; Academic Press: New York, 1972.
(20) Ben-Israel, A.; Greville, T. N. E. *Generalized Inverses: Theory and Applications*; John Wiley & Sons: New York, 1974.
(21) Losonczi, J. A.; Andrec, M.; Fischer, M. W. F.; Prestegard, J. H. *J. Magn. Reson.* **1999**, *138*, 334−342.
(22) Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; Flannery, B. P. *Numerical recepies in C.*; Cambridge University Press: Cambridge, 1992.

$(1/K)\,\mathbf{B}^{+}\mathbf{D}$. If the matrix **B** were singular, then the solution for $\mathbf{A}_{bf}$ would no longer be unique. The solution must be written as

$$\mathbf{B}^{+}\mathbf{B}\mathbf{A}_{bf} = (^{1}/_{K})\mathbf{B}^{+}\mathbf{D} \tag{6}$$

where the $\mathbf{B}^{+}\mathbf{B}$ term on the left specifies that the solution obtained for **A** will correspond only to that part which lies within the range of $\mathbf{B}^{tr}$, denoted $R(\mathbf{B}^{tr})$. The range of a matrix **M**, $R(\mathbf{M})$, is defined as the vector space encompassing all solution vectors **y** for all possible vectors **x** subject to the relationship $\mathbf{Mx} = \mathbf{y}$. The range can be described explicitly by specification of a set of orthogonal basis vectors that span it. In fact, the SVD of a matrix (eq 4) provides this information. For example, the range of $\mathbf{B}^{tr}$ is spanned by the columns of $\mathbf{V}_B$ that correspond to nonzero singular values. When the range of $\mathbf{B}^{tr}$ is restricted (i.e., $R(\mathbf{B}^{tr}) \neq \mathbf{1}$), the addition of any vector drawn from the nullspace of **B** will produce an equally good solution.[19−21] The complete set of possible solutions for $\mathbf{A}_{bf}$ can be written as

$$\mathbf{A}_{bf} = (^{1}/_{K})\mathbf{B}^{+}\mathbf{D} + (\mathbf{1} - \mathbf{B}^{+}\mathbf{B})\mathbf{z} \tag{7}$$

where **z** is *any* vector of dimension 5. The purpose of the $(\mathbf{1} - \mathbf{B}^{+}\mathbf{B})$ term is to extract out the part of **z** that is orthogonal to the range of $\mathbf{B}^{tr}$ (i.e., in the nullspace of **B**). It is the significance of constructs such as $(\mathbf{1} - \mathbf{B}^{+}\mathbf{B})$ and $\mathbf{B}^{+}\mathbf{B}$ that form the basis for the approach presented here.

These constructs are referred to as orthogonal projectors and assume an important role in the mathematics underlying linear least-squares methods.[19,20] Orthogonal projectors, which are both idempotent and symmetric, are created by taking the product of a matrix with its Moore−Penrose inverse, as in $\mathbf{B}^{+}\mathbf{B}$. They function in effect as an operator which, when applied to a vector, projects out the part of the vector lying within the range of the left appearing matrix, in this case $\mathbf{B}^{+}$ ($R(\mathbf{B}^{+}) = R(\mathbf{B}^{tr})$). The relevant matrix that projects onto the orthogonal and complementary subspace is formed by difference with the identity matrix, as in $(\mathbf{1} - \mathbf{B}^{+}\mathbf{B})$. For studies of macromolecules, the matrix **B** will usually be full rank, and thus the orthogonal projector $\mathbf{B}^{+}\mathbf{B}$ holds little interest because it will then be equal to the identity matrix.

Of more interest is the orthogonal projector formed by $\mathbf{BB}^{+}$, which projects a vector onto the range of **B**, $R(\mathbf{B})$. This projector is an N x N matrix where N is the number of internuclear vectors for which measurements have been obtained. Since the matrix **B** always has five columns, $\mathbf{BB}^{+}$ will project an N-dimensional vector onto a subspace spanned by just five basis vectors (of length N). These five orthogonal basis vectors are readily obtained from the SVD of **B** in the form of $\mathbf{U}_B$ (as shown in eq 4). In fact, $\mathbf{BB}^{+}$ depends only on the matrix $\mathbf{U}_B$ according to

$$\mathbf{BB}^{+} = \mathbf{U}_B\mathbf{U}^{tr}{}_B \tag{8}$$

with the number of columns of $\mathbf{U}_B$ corresponding to the number of nonzero singular values (or the rank r) of the matrix **B**.

The orthogonal projector $\mathbf{BB}^{+}$ is employed implicitly in the calculation of RDCs on the basis of the best-fit alignment tensor. Using the notation of eq 1, a set of predicted RDCs is obtained by computing

$$\mathbf{D}_{calc} = K\mathbf{B}_{md}\mathbf{A}_{bf} \tag{9}$$

in which $\mathbf{B}_{md}$ represents the provided structural model and $\mathbf{D}_{calc}$ and $\mathbf{A}_{bf}$ represent the calculated RDCs and best-fit alignment tensor, respectively. Replacement of $\mathbf{A}_{bf}$ with the expression for its solution will produce

$$\mathbf{D}_{calc} = \mathbf{B}_{md}\mathbf{B}^{+}{}_{md}\mathbf{D} \tag{10}$$

The orthogonal projector $\mathbf{B}_{md}\mathbf{B}_{md}{}^{+}$ projects a vector of measured RDCs onto a corresponding vector of calculated RDCs on the basis of the structural model provided ($\mathbf{B}_{md}$). This formulation (eq 10) actually corresponds to an approach introduced to remove the explicit appearance of the alignment tensor during RDC-based refinement.[6,10] One can likewise write an expression for the Q value[23] from this perspective:

$$Q = \frac{||\mathbf{D} - \mathbf{B}_{md}\mathbf{B}^{+}{}_{md}\mathbf{D}||}{||\mathbf{D}||} = \sqrt{\frac{\sum_i (D_{i,meas} - D_{i,calc})^2}{\sum_i D^2{}_{i,meas}}} \tag{11}$$

with $||\,||$ indicating that the Euclidean norm is computed for the resultant vector (or matrix if data from multiple alignment media is considered).

Direct Interpretation of Dipolar Couplings. The objective of the DIDC method is to obtain the unknown matrix **B** directly from the measured RDCs, contained in **D**. We proceed by exploiting the relationship between the ranges of the matrixes **B** and **D**. It can be shown that

$$\mathbf{BB}^{+} = \mathbf{DD}^{+} (\Rightarrow R(\mathbf{B}) = R(\mathbf{D})) \tag{12}$$

provided that $\mathbf{AA}^{+} = \mathbf{1}$. In other words, the range of **B** is identical to the range of **D** if the matrix **A** is full rank (i.e., of rank 5). This is equivalent to the requirement that five *independent* alignment media be employed and is therefore a desirable experimental objective. Once this condition is met, the RDC data are complete. Further acquisition of RDCs in additional media will bring further improvements in precision and accuracy, but no additional complementary information can be obtained. For now, we consider the implications if complete RDC data are available and return momentarily to the question of errors and how one can assess whether eq 12 holds on the basis of consideration of the matrix **D** alone.

The fact that a complete set of RDC measurements (i.e., acquired using five independent alignment media) will, subject to measurement errors, provide the range of the matrix **B** in the absence of a priori information is an extremely powerful constraint. This allows **B** to be almost completely constructed on the basis of only the SVD of the data matrix **D**. The equality of the ranges of **B** and **D** does not imply that $\mathbf{U}_B = \mathbf{U}_D$. These matrixes are related by a unitary transformation

$$\mathbf{U}_B = \mathbf{U}_D\mathbf{T}, \tag{13}$$

represented by the 5×5 unitary matrix **T**. Considering the SVD of the matrix **B** (eq 4) in combination with eq 13 leads to

(23) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.

$$\mathbf{B} = \mathbf{U}_D\mathbf{T}\mathbf{W}_B\mathbf{V}^{tr}{}_B = \mathbf{U}_D\mathbf{\Lambda} \quad (14)$$

with the 5×5 matrix $\mathbf{\Lambda}$ introduced to represent collectively the remaining 25 parameters that cannot be deduced from the RDC data alone. Although we have managed to avoid needing to know **A**, the inherent underdetermination of the problem reemerges here in terms of these unknown 25 parameters. The missing information does encompass some physically concrete parameters such as the singular values of **B** ($\mathbf{W}_B$), but others (such as **T**) will contain an arbitrary component. Thus, the overall significance of $\mathbf{\Lambda}$ is probably best described simply as the 5×5 matrix which produces the solution (**B**) when provided with the range-spanning basis $\mathbf{U}_D$. Additional data from other sources could be introduced to restrict possibilities for the matrix $\mathbf{\Lambda}$. However, the possibilities for $\mathbf{\Lambda}$ are already severely restricted by the requirement that the solution **B** describes a set of internuclear vectors belonging to a presumably structured macromolecule. This suggests that a useful solution might be obtained by selecting the matrix $\mathbf{\Lambda}$ that minimizes the variation in generalized order parameters. This can be carried out using the least-squares procedure

$$||\mathrm{Diag}\{\mathbf{U}_D\mathbf{\Lambda}\mathbf{\Lambda}^{tr}\mathbf{U}_D{}^{tr}\} - \mathbf{1}^{(N)}||_{\min} \quad (15)$$

which finds the matrix $\mathbf{\Lambda}\mathbf{\Lambda}^{tr}$ that produces the best-fit solution to a molecule with all generalized order parameters equal to 1. This procedure provides only 15 of the needed parameters, namely, the matrixes **T** and $\mathbf{W}_B$ from eq 14. The remaining 10 parameters correspond to a 5×5 unitary matrix comprising any overall rotation of the molecule, which is irrelevant, and a 5×5 permutation matrix. This permutation matrix is important because the proper physical interpretation relies on the assignment of the independent elements of the tensors (eq 2) to each of the five columns of **B**. This permutation matrix can be determined to within a 4-fold degeneracy by consideration of the distributions of coefficients occurring in each of the five columns of **B**. These remaining four possibilities arise from difficulties in assigning specific pairs of columns in the matrix **B** to their correct tensorial elements (eq 2). The ambiguous pairs of tensor elements are (T_{xz}, T_{yz}) and ($T_{xx} - T_{yy}$, T_{xy}). Two of the possible choices will correspond to the correct and mirror image solutions while the other two will exhibit global distortions in the arrangement of vector orientations.

Analysis of the Matrix D. The above-described procedure can in principle be carried out as long as five sets of RDC measurements have been made. However, unless the equality of ranges stated in eq 12 holds, the results will be in part constructed entirely from random noise. It is therefore desirable to have a method whereby the validity of eq 12 can be assessed without knowledge of **B** (and hence **A**). The key to this is the fact that as a simple consequence of eq 1 the matrix **D** will be rank 5 only if *both* **A** and **B** are full rank. Although the validity of eq 12 does not require that **B** is full rank, the completeness of the data can be assessed on the basis of the matrix **D** alone only for cases in which **B** is full rank. This condition will hold if the internuclear vectors for which corresponding measurements have been obtained are sufficiently independent such that the alignment tensors *could* be uniquely determined. This will normally be satisfied easily for macromolecular systems where large numbers of RDC measurements can be made.

An additional restriction on the matrix **D** is that it cannot exceed a rank of 5, which corresponds to the maximum rank for both matrixes **A** and **B**. Therefore, the first step of the analysis should be to consider carefully the singular values of the matrix **D**. If the number of separate data sets acquired (M) does not exceed the number of RDCs measured in each dataset (N), then there will be M singular values that result from the SVD of **D**, indexed in order of decreasing magnitude. The M-5 least significant singular values correspond to errors in the measurements. Setting them equal to zero and then reconstructing the matrix **D** provides the ability to signal average between data sets acquired in completely different media. The determination of whether a complete set of RDCs has been collected requires consideration of the magnitude of the first five singular values. The condition number among these five singular values provides one indication of the level of independence of measurements. However, the best indication comes from comparison of the first five singular values with the remaining singular values arising from the measurement errors. All of the most significant five singular values (the signal) must be of distinguishably larger magnitude than the other singular values (the noise). As the difference in magnitude becomes smaller, the precision and accuracy of the results will correspondingly suffer.

Incomplete RDC Data. Although one application has now been reported which utilizes five independent alignment media,[11] this objective remains a challenge. It is therefore useful to consider how one might proceed in data-deficient cases. Since the RDC data alone provide an incomplete picture, our approach will be to supply the missing information on the basis of a rigid structural model ($\mathbf{B}_{md}$). The first step is to use the model ($\mathbf{B}_{md}$) and the data (**D**) to determine the best-fit alignment tensors ($\mathbf{A}_{bf}$) in typical fashion. Provided that $\mathbf{B}_{md}$ is full rank this results in a unique solution for $\mathbf{A}_{bf}$. These best-fit alignment tensors along with the data are then used to generate the corresponding best-fit refined model $\mathbf{B}_{ref}$ which satisfies the equation $\mathbf{D} = K\,\mathbf{B}_{ref}\,\mathbf{A}_{bf}$. The missing part of the solution for $\mathbf{B}_{ref}$, which spans the nullspace (of $\mathbf{A}_{bf}{}^{tr}$), is supplied from the initial model. This procedure is similar to that carried out for eq 7[19,20]

$$\mathbf{B}_{ref} = (1/K)\mathbf{D}\mathbf{A}^{+}{}_{bf} + \mathbf{B}_{md}(1 - \mathbf{A}_{bf}\mathbf{A}^{+}{}_{bf}) \quad (16)$$

Consider the right-hand side of the expression. The first term corresponds to the part refined using the data while the second part is simply carried over from the initial model. The explicit appearance of the alignment tensors can be removed and the equation rearranged to read

$$\mathbf{B}_{ref} = \mathbf{B}_{md} + \mathbf{D}(\mathbf{B}^{+}{}_{md}\mathbf{D})^{+} - \mathbf{B}_{md}\mathbf{B}^{+}{}_{md}\mathbf{D}(\mathbf{B}^{+}{}_{md}\mathbf{D})^{+} \quad (17)$$

This result has the desirable property of producing the best-fit refined model ($\mathbf{B}_{ref}$) that lies closest to the initial model ($\mathbf{B}_{md}$), as a function of just initial model and data. Moreover, unless the initial model is of extremely poor quality, eq 17 will produce an exact solution (i.e., with a Q value = 0) for $\mathbf{B}_{ref}$.

The most notable feature of this refinement procedure is that the dynamic properties are refined simultaneously along with the structural properties. For many cases this will be a desirable situation, but there are some associated complications with the resulting description of dynamics. Because of the incomplete

nature of the data, we will not consider the description of motional asymmetry and focus only on the estimation of generalized order parameters. The simplest method for estimating generalized order parameters is to compute them directly from the rows of the refined model $\mathbf{B}_{ref}$ according to

$$S_i = ||\mathbf{B}_{ref}||_i \tag{18}$$

in which $||\ ||_i$ denotes the Euclidean norm of the row corresponding to the *i*th interaction. The problem with computing the generalized order parameters directly from $\mathbf{B}_{ref}$ according to eq 18 is that the extent of motion will be underestimated because it is constructed in part from the initial model $\mathbf{B}_{md}$, which is likely to be rigid. This can produce particularly large distortions for very mobile regions, for which the description of motional amplitudes may be of significant interest. It would therefore be useful to exclude the contribution from the initial model when estimating generalized order parameters. By referring to the result shown in eq 17, we propose the following expression for the estimation of generalized order parameters:

$$S_i = \frac{||\mathbf{D}(\mathbf{B}^{+}{}_{md}\mathbf{D})^{+}||_i}{||\mathbf{B}_{md}\mathbf{B}^{+}{}_{md}\mathbf{D}(\mathbf{B}^{+}{}_{md}\mathbf{D})^{+}||_i} \tag{19}$$

As before, $||\ ||_i$ refers to the Euclidean norm of the *i*th row of the resultant matrix. In effect, each order parameter S_i is estimated on a row-by-row basis from the ratio of the norms of the new, refined part $\mathbf{D}(\mathbf{B}^{+}{}_{md}\mathbf{D})^{+}$ to the old part being discarded, $\mathbf{B}_{md}\mathbf{B}^{+}{}_{md}\mathbf{D}(\mathbf{B}^{+}{}_{md}\mathbf{D})^{+}$, respectively. This simple approach requires that the provided model ($\mathbf{B}_{md}$ in eq 19) represents a rigid model. Although this scheme provides a less biased estimate of order parameters, it also carries the disadvantage of being based on a ratio, which leads to occasionally unstable behavior for specific interactions.

The order parameters that result from an RDC-based analysis are inherently relative in nature. As in the spin relaxation analysis, part of the problem arises because of uncertainties in the magnitude of the dipolar interaction constant.[24] However, an additional level of uncertainty enters in because of the inability to distinguish between the magnitude of overall alignment and internal motions in an absolute manner.[9,11,12] This is underscored by eq 17, which does not even contain a term for the dipolar interaction constant. The refined matrix $\mathbf{B}_{ref}$ will exhibit raw order parameters that produce an average value of 1.0. Typically, the order parameters must be scaled in some way, for example, by requiring that no order parameter can exceed a value of 1.

Materials and Methods

Acquisition of RDC Data. ^{15}N labeled ubiquitin (VLI Research) and a suspension of purple membrane (PM) particles (Symbiotech) were purchased and used without further purification. An initial isotropic sample of ubiquitin (0.5 mM) was prepared to contain 30 mM phosphate (pH 7.2), 0.05% NaN_3, and 6% D_2O. Following acquisition of isotropic reference data, the sample was modified to contain 2.0 mg/mL PM.[25,26] Four sets of $^1J_{NH}$ couplings were measured after successive additions of NaCl (20, 40, 60, and 80 mM). The general procedure was repeated starting with a more concentrated isotropic ubiquitin sample (1 mM) under identical buffer and pH conditions. A series of three additional measurements of $^1J_{NH}$ couplings were carried out under the following PM:NaCl conditions: (1.5 mg/mL: 20 mM), (8.0 mg/mL: 90 mM), and (8.0 mg/mL: 250 mM).

NMR experiments were carried out at a temperature of 37 °C on a Bruker DRX spectrometer operating a ^{1}H resonance frequency of 600 MHz and equipped with a broad-band triple resonance probe (TBI) with three orthogonal gradient coils. Amide N−H RDCs were obtained by difference between $^1J_{NH}$ couplings measured under isotropic and aligned conditions. All $^1J_{NH}$ coupling measurements were performed using the HSQC−PEC (HSQC with phase-encoded couplings)[27,28] experiment. This experiment utilizes a constant time period for N−H coupling evolution, with coupling information encoded in the resonance intensities of a pair of resulting ^{15}N−^{1}H correlation spectra, which are created by postacquisition data shuffling. These experiments are in turn acquired in pairs, differing only in the constant time period employed, to reduce systematic errors.[28] The measured coupling is taken to be the average of values obtained from the complementary pair of experiments. Experiments on samples with low PM concentrations (<= 2.0 mg/mL PM) were acquired using constant time periods set to 64.516 and 69.892 ms. Because of significantly enhanced relaxation, the samples with high PM concentration required the use of shorter constant time periods (32.258 and 37.634 ms). Total experimental acquisition times (for each complementary pair of experiments) were ∼8 h for the 2.0 mg/mL PM samples and ∼25 h for the 1.5 and 8.0 mg/mL PM samples. Data processing was carried out using NMRPipe software.[29]

Analysis of Synthetic Data. Starting from a set of synthetic or measured RDCs, all further analysis was carried out using home-written software. According to eq 1, synthetic RDC data was generated on the basis of a provided set of alignment tensors **A** and a matrix **B** describing the structure and dynamics of a set of internuclear vectors. The alignment tensors were generated randomly with the magnitude restricted such that the maximum magnitude of the RDCs produced would range between 5 and 25 Hz. The mean orientations for **B** were taken to correspond to the 73 amide N−H internuclear vectors of a solid-state structure of ubiquitin (1UBQ). These vectors were extracted in terms of the polar angles, β_i and α_i, describing their orientation relative to the native coordinate axes. Dynamics were introduced separately for each internuclear vector by random generation of values for the generalized order parameter (S_i), as well as parameters describing the direction (γ_i) and asymmetry (η_i) of motional averaging. These five parameters were used to construct a diagonal order tensor, based on S_i and η_i, as well as a 3 × 3 unitary transformation matrix based on the other three parameters (α_i, β_i, γ_i). The transformation matrix was used to rotate each diagonal order tensor into a common molecular frame before conversion to independent tensor elements according to eq 2. Normally distributed random noise was added to the synthetic data as necessary. When required, initial imperfect model structures were generated from the "perfect" mean structure by the addition of random angular displacements for each vector. The amplitudes of the angular displacements were generated from a normal distribution and were carried out by rotation of the specific internuclear vector about a randomly selected perpendicular rotation axis.

Results

The proposed methods are illustrated using synthetic data in addition to an experimental application to the protein ubiquitin employing data that collectively represent three independent alignment tensors. The simulations are intended to provide some

(24) Case, D. A. *J. Biomol. NMR* **1999**, *15*, 95−102.
(25) Sass, J.; Cordier, F.; Hoffmann, A.; Cousin, A.; Omichinski, J. G.; Lowen, H.; Grzesiek, S. *J. Am. Chem. Soc.* **1999**, *121*, 2047−2055.
(26) Koenig, B. W.; Hu, J. S.; Ottiger, M.; Bose, S.; Hendler, R. W.; Bax, A. *J. Am. Chem. Soc.* **1999**, *121*, 1385−1386.
(27) Tolman, J. R.; Prestegard, J. H. *J. Magn. Reson., Ser. B* **1996**, *112*, 245−252.
(28) Cutting, B.; Tolman, J. R.; Nanchen, S.; Bodenhausen, G. *J. Biomol. NMR* **2002**, *23*, 195−200.
(29) Delaglio, F.; Grzesiek, S.; Vuister, G. W.; Zhu, G.; Pfeifer, J.; Bax, A. *J. Biomol. NMR* **1995**, *6*, 277−293.

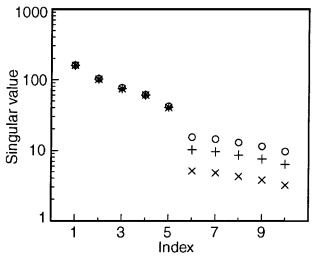

Figure 3. Singular values of the matrix **D** (of dimension 73 × 10) formed from 10 synthetic RDC datasets. It is apparent that the requirement for five independent alignment tensors is met. Singular values are plotted for three different levels of added random errors: σ (error) = 0.5 Hz (x), σ (error) = 1.0 Hz (+), σ (error) = 1.5 Hz (open circles).

general insights into the effects of experimental noise and the accuracy of initial model structures (when needed) on the precision and accuracy of extracted structural and dynamic parameters. These factors are explored using synthetic amide N−H RDC data generated from two hypothetically perfect models based on a set of ubiquitin coordinates (1UBQ). Both models have the same mean structure, but one is completely rigid while the other executes internal motions. The dynamic model exhibits values of S ranging between 0.7 and 1.0 for 90% of amide N−H vectors while the other 10% are highly mobile with $S < 0.5$. These two models, in combination with randomly generated alignment tensors, were used to create synthetic data sets that mimic the acquisition of RDC data in multiple alignment media.

Simulations Using Complete Data. The ability to extract structural and dynamic parameters in the absence of a priori information was tested using synthetic RDC data corresponding to the measurement of 73 couplings in 10 randomly generated alignment tensors. Several data matrixes **D** of dimension 73 × 10 were thus constructed containing variable levels of added random error. The independence of the data was verified on the basis of the SVD of the synthetic matrixes **D**. Shown in Figure 3 are the singular values resulting from such an analysis. As expected, singular values 6−10 are nearly constant and correlate in magnitude with the level of added errors. Even with large random errors, for example, with σ(error) = 1.5 Hz, the 5th (and smallest) singular value remains distinct from the noise by an approximate factor of 3. A minimization procedure to find $\mathbf{\Lambda\Lambda}^{tr}$ was carried out according to eq 15 for each of the matrixes $\mathbf{U}_D$ (73 × 5) resulting from the SVD analysis. This allowed the matrixes $\mathbf{U}_D\mathbf{T}\mathbf{W}_B$ to be constructed on the basis of the SVD of $\mathbf{\Lambda\Lambda}^{tr}$. The final 10 parameters were obtained by finding the 5 × 5 unitary transformation matrix producing the best fit to the matrix **B** generated from the solid-state structural coordinates of ubiquitin. This procedure produces the best-fit superposition of vector orientations to facilitate comparison.

Results are summarized in Table 1 for both dynamic and rigid protein cases at different levels of experimental error. For a truly rigid protein and perfect data, this method will produce perfect results. Otherwise, resulting accuracy and precision correlate with the quality of the data. The vector orientations of the protein core are obtained to extremely high accuracy (<5° RMSD) for all cases considered here. The results for the dynamic protein with large experimental errors present illustrate a complication that can occur for highly mobile regions. This is reflected in the large difference between RMSDs reported with and without the inclusion of highly mobile vectors. A closer examination

Table 1. Results of Simulations Using Complete RDC Data

	σ (error)[a]	Δr[b]	ΔS[c,d]	ΔS (scaled)[c,e]
rigid	0	0	0	0 (1.0)
	0.5	1.1	0.061	0.018 (1.062)
	1.0	2.2	0.117	0.035 (1.126)
	1.5	3.4	0.165	0.052 (1.187)
dynamic	0	1.3 (1.4)	0.032	0.032 (0.994)
	0.5	1.8 (2.6)	0.037	0.037 (1.009)
	1.0	3.2 (9.6)	0.051	0.048 (1.020)
	1.5	4.7 (14.2)	0.069	0.061 (1.040)

[a] Standard deviation of synthetic errors in Hz. [b] RMSD from true mean vector orientations (°). Calculation includes only residues with $S_i > 0.5$ (calculation including all residues). [c] RMSD from true generalized order parameters. [d] Apparent generalized order parameters are only scaled such that the largest value, S_i (max) = 1. [e] The apparent generalized order parameters are scaled by best fit to the true values. The best-fit scaling parameter a (in parentheses) is obtained by minimization of S_i (true) = a*S_i(apparent) over all residues.

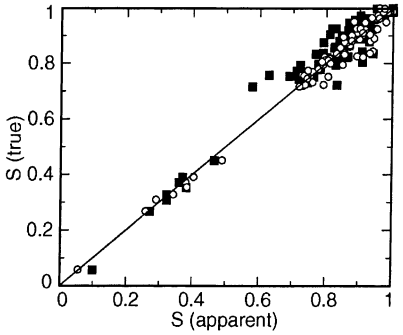

Figure 4. Comparison of true versus apparent generalized order parameters obtained using a complete set of synthetic RDC data for 73 residues and in the absence of a priori information. The comparison is shown for the case of perfect data (open circles) and with random errors added with σ (error) = 1.0 Hz (filled squares). The correlation coefficients are 0.986 and 0.968, respectively.

indicates that the difference occurs because of highly anisotropic motions of two vectors, leading to confusion in the selection of the principal axis Z, and hence angular errors of nearly 90 degrees are recorded for each of these two cases.

The generalized order parameters are also estimated to very good precision and accuracy. The resulting correlation between true and apparent generalized order parameters is shown in Figure 4 for perfect data and with added random errors of σ(error) = 1.0 Hz. Since absolute values of the apparent generalized order parameters cannot be established with certainty, the results shown are scaled to exhibit a maximum value of 1.0. In Table 1, the RMSDs for the order parameters are reported in scaled (best-fit) and unscaled (S_i (max) = 1) forms to facilitate the evaluation of overall precision and accuracy. Notably, the estimated order parameters for the rigid protein are very closely clustered with the exception of a single outlier, which dominates the reported statistics. This situation presumably arises because of the relatively flat energy landscape encountered at the end of the minimization procedure and has not been observed for any dynamic case simulated. Despite these minor anomalies, it is apparent that results of high quality can be obtained.

In comparison with the results for generalized order parameters, the ability to estimate the parameters describing the anisotropy of motion was rather poor. Using perfect data, the results obtained for the motional asymmetry parameter η are illustrated in Figure 5 for the dynamic model. Even in this idealized case, the accuracy is rather low and effectively restricts

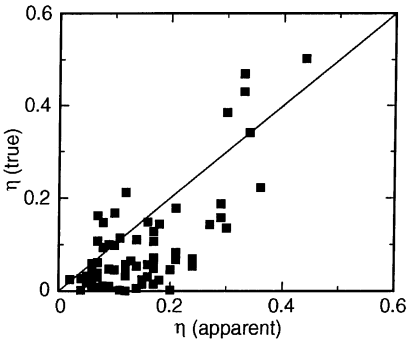

Figure 5. Comparison of true versus apparent values for the motional asymmetry parameter η obtained using a complete set of RDC data and in the absence of a priori information. The comparison is shown for the case of perfect data.

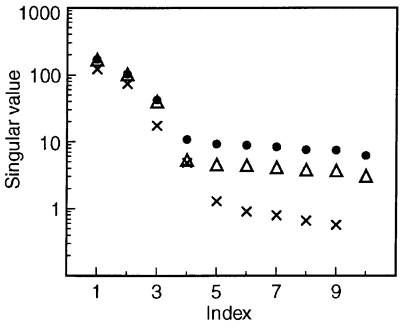

Figure 6. Singular values of the matrix **D** obtained in some data deficient cases considered in the text (synthetic and experimental). The singular values resulting from synthetic data (data matrix **D** of dimension 73 × 10) are shown for added random errors of σ (error) = 0.5 Hz (open triangles) and σ (error) = 1.0 Hz (filled circles). The singular values of the data (matrix **D** of dimension 61 × 9) employed for the experimental application to ubiquitin are indicated with x's.

the interpretation to qualitative assessments of either high or low asymmetry. In light of these initial results, the extraction of anisotropic motional parameters (η and γ) was not considered further within the context of this work.

Refinement Protocol for Incomplete Data. If the acquired RDC data is not complete, that is, the matrix **D** is less than rank 5, then the analysis must proceed on the basis of an initial model structure. Some tools for refinement were introduced in the theoretical section, but the optimal protocol for their use will depend on the extent of mobility of the molecule as well as the objectives of the investigator. For the present, we focus on exploring a simple yet robust refinement procedure. A single application of eq 17 will generate a refined dynamic model that produces perfect agreement between measured and calculated RDC data. However, this simple approach suffers from reduced accuracy of extracted parameters because it starts from a rigid model. Unless one is certain that the molecule is in reality very rigid, it would be more desirable to pre-estimate order parameters and then modify the starting model to reflect these estimates. A simple two-step refinement procedure is therefore proposed to accomplish this. In the first step, generalized order parameters are calculated directly from the matrix $\mathbf{B}_{\mathrm{ref}}$ resulting from application of eq 17. These generalized order parameters are obtained simply by computing the Euclidean norm for each row of $\mathbf{B}_{\mathrm{ref}}$ according to eq 18. Although this will underestimate motion, it has the advantage of being very resistant to the production of outliers. These initial estimates of S_i are then used to scale the initial rigid model by multiplying each row of $\mathbf{B}_{\mathrm{md}}$ by its respective estimated S_i. This implicitly assumes axially symmetric motions, but the deviations are expected to be small and, moreover, well beyond the discriminating power of incomplete RDC data. The new "dynamic" initial model created by this scaling procedure is then passed through eq 17 a second time, followed by renormalization of the N rows of $\mathbf{B}_{\mathrm{ref}}$ back to 1. This is the structurally refined model, containing no information about dynamics. The final estimates for the generalized order parameters are obtained via eq 19 using this *rigid* refined model. The final refined model can then be constructed by multiplying each row by the respective final estimates for the generalized order parameters. Although this procedure will not produce perfect agreement between measured and calculated RDCs, extensive simulations indicate that, unless the molecule is in reality very rigid, this two-step protocol will produce more accurate mean vector orientations and order parameters relative to a single-step application of the prescription given in eq 17.

Simulations Using Rank 3 Data. As a test of the proposed two-step refinement protocol, synthetic RDC data were generated corresponding to 73 RDC measurements in 10 different alignment media. After random generation of the first three alignment tensors, the remaining seven were obtained on the basis of the first three by constructing linear combinations with random coefficients. This ensures that the resulting synthetic data is of rank 3, as can be seen from the singular values plotted in Figure 6. In this case, the 4th and higher singular values are set to zero before reconstructing the data matrix **D**. The described two-step refinement protocol was then carried out using starting models of variable quality and with different levels of random error in the original data. Q values[23] were computed (eq 11) at several intermediate steps to track the improvement in agreement between the data and the model. These Q values are reported along with a summary of results in Table 2. As expected, the method produces perfect results in the event that the protein is rigid, and a perfect model and perfect data are available. Unless an extraordinarily accurate mean structure is available to start, the protocol produces refined mean orientations that exhibit rather consistent gains in accuracy (of approximately 30%) with only a small dependence on the quality of the structural model or data. Consideration on a residue-by-residue basis indicates that while the level of improvement is very inconsistent between residues, only very rarely do any individual vectors decrease in accuracy. These isolated cases arise because of the remaining presence of multiple minima for orientational solutions when three alignment media are employed. This explains the huge angular errors observed for a couple of residues when refining the starting model with an RMSD of 16.3° (see Table 2).

In contrast to its performance for structural refinement, the ability of the protocol to estimate generalized order parameters depends rather strongly on the quality of the initial model and data provided. To facilitate comparison of results in Table 2, all reported RMSDs correspond to the agreement between the true and apparent values for S after determination of a best-fit scaling factor. Figure 7 illustrates the level of precision and accuracy obtained under two different starting conditions, with the values for S (apparent) scaled such that the maximum was equal to 1.0. Provided with perfect data and a rigid model with the correct mean vector orientations can allow values of S to be estimated with rather good precision and accuracy. From

Table 2. Results of Simulations Using Incomplete RDC Data of Rank 3

	Δr (init)[a]	σ (error)[b]	Δr (refined)[c,d]	ΔS (apparent)[d,e]	*S* scaling[f]	*Q* (init)[g]	*Q* (init + *S*)[h]	*Q* (ref)[g]	*Q* (ref + *S*)[h]
rigid	0	0	0	0	1.000	0	0	0	0
		0.5	0.8	0.031	1.104	0.031	0.034	0.029	0.025
		1.0	1.6	0.061	1.209	0.063	0.067	0.057	0.049
	5.4	0	3.3	0.043	1.126	0.169	0.157	0.065	0.049
		0.5	3.5	0.051	1.152	0.169	0.157	0.071	0.055
		1.0	3.8	0.072	1.194	0.175	0.166	0.087	0.069
	10.9	0	6.8	0.080	1.232	0.308	0.283	0.122	0.091
		0.5	6.9	0.082	1.188	0.307	0.280	0.125	0.093
		1.0	7.2	0.094	1.204	0.308	0.282	0.134	0.101
dynamic	0	0	0.9 (3.3)	0.042 (0.060)	1.001	0.213	0.074	0.160	0.057
		0.5	1.3 (3.4)	0.054 (0.074)	1.038	0.215	0.084	0.162	0.065
		1.0	2.0 (3.9)	0.077 (0.096)	1.068	0.222	0.109	0.170	0.082
	5.4	0	3.5 (4.6)	0.053 (0.066)	1.012	0.260	0.163	0.177	0.075
		0.5	3.7 (4.8)	0.059 (0.075)	1.032	0.260	0.163	0.179	0.081
		1.0	4.1 (5.2)	0.079 (0.094)	1.063	0.264	0.175	0.187	0.096
	10.9	0	6.7 (7.4)	0.076 (0.082)	1.073	0.359	0.279	0.209	0.104
		0.5	6.9 (7.6)	0.078 (0.087)	1.041	0.357	0.275	0.211	0.108
		1.0	7.2 (7.9)	0.091 (0.100)	1.070	0.359	0.279	0.217	0.118
	16.3	0	9.8[j](17.5)	0.102 (0.104)	1.137	0.454	0.395	0.329	0.263
		0.5	10.2[i](14.5)	0.102 (0.104)	1.113	0.452	0.393	0.298	0.229
		1.0	10.4[i](14.8)	0.113 (0.115)	1.119	0.452	0.394	0.302	0.231

[a] RMSD of mean vector orientations for the initial model from their true orientations. [b] Standard deviation of synthetic errors in Hz. [c] RMSD of mean vector orientations for the final refined model from their true orientations. [d]Calculation includes only residues with $S_i > 0.5$ (calculation including all residues). [e] RMSD from true generalized order parameters after best-fit scaling of apparent values of S_i. [f] Best-fit scaling parameter a is obtained by minimization of $S_i(\text{true}) = a*S_i(\text{apparent})$ over all residues. [g] *Q* value for specified rigid model (all rows of the corresponding matrix **B** have a norm of 1). [h] *Q* value for the specified model scaled by final estimates for generalized order parameters (each row of the corresponding matrix **B** is multiplied by the relevant value for S_i). [i] One additional residue was excluded on the basis of a very large departure (>45°) from the initial orientation upon refinement. [j] As for *i*, but for two residues.

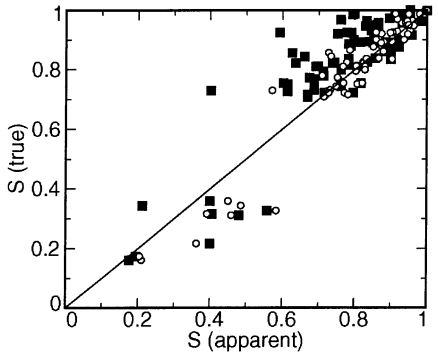

Figure 7. Comparison of true versus apparent generalized order parameters obtained using synthetic RDC data for 73 residues and corresponding to three independent alignment tensors (rank 3 data). In this case, an initial starting model is required. The open circles correspond to results using perfect data and a model with perfect mean orientations. Results using data with added errors (σ (error) = 1.0 Hz) and an imperfect starting model (10.9° RMSD from true mean vector orientations) are indicated with the filled squares. The correlation coefficients are 0.962 and 0.871 for perfect and imperfect data, respectively.

Table 2 it can be seen that, with good data and a high-quality model, results can be obtained that are comparable to this idealized case. However, with the not unrealistic case of σ (error) = 1.0 Hz and a 10.9° RMSD model, the quality of the results have already been noticeably altered. While the overall distribution of order parameters is still useful, precision and accuracy are considerably diminished and there are inevitably outliers well outside of the general precision level applicable for the majority of interaction vectors. Interestingly, the occurrence of a few outliers is in itself fairly reproducible and apparently depends strongly on the quality of the model structure provided and the specific alignment tensor configuration. This suggests that it may be possible to identify those vectors that are particularly unstable to estimation of generalized order parameter by this protocol. This is being investigated further.

Application to Ubiquitin. Amide N−H RDC data were measured for ubiquitin dissolved in purple membrane (PM) media,[25,26] under different PM and salt concentrations. The seven data sets acquired in this manner were supplemented with two data sets from the literature,[30] acquired in charged and uncharged bicelle media. From these data, a matrix **D** was constructed of dimension 61 × 9, with its singular values shown in Figure 6. Clearly, these data collectively represent at least three independent alignment tensors. Although the fourth singular value is above the noise, it was not included in the analysis because of concerns about whether it was significant given that the PM data sets are not very independent (see Supporting Information). In experimental situations, it is possible that a singular value, like the fourth one here, could correspond to the existence of systematic errors or even reflect variations in the structure or dynamics of the protein between different media. This presents some interesting implications not further explored here.

After zeroing the six least significant singular values to reduce errors, the described two-step protocol was applied using three different starting models, taken from available solid state[31,32] (1UBQ and 1UBI) and solution state[23] (1D3Z, model 1) coordinates. Refinement statistics are reported in Table 3. On the basis of the statistics for the refinement of mean vector orientations, we might infer that the NMR structure 1D3Z is in fact closer to the actual structure in solution. However, bicelle data used here was also used in refinement of that particular structure. Nevertheless, a consideration of changes in vector orientations on a residue-by-residue basis indicates that the larger RMSDs obtained for the solid-state structures upon refinement arise predominantly from internuclear vectors lying within loop

(30) Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 12334−12341.
(31) Vijaykumar, S.; Bugg, C. E.; Cook, W. J. *J. Mol. Biol.* **1987**, *194*, 531−544.
(32) Ramage, R.; Green, J.; Muir, T. W.; Ogunjobi, O. M.; Love, S.; Shaw, K. *Biochem. J.* **1994**, *299*, 151−158.

Table 3. Statistics from Refinement of Different Models of Ubiquitin

model[a]	Δr^b	Q (init)[c]	Q (init + S)[d]	Q (ref)[c]	Q (ref + S)[d]
1D3Z	3.9 (6.0)	0.240	0.139	0.112	0.059
1UBQ	9.0 (9.3)	0.251	0.216	0.132	0.100
1UBI	9.0 (9.2)	0.263	0.225	0.139	0.096

[a] PDB entry of starting model employed. [b] RMSD of refined mean vector orientations relative to the initial model (°). Calculation includes only residues with $S_i > 0.5$ (calculation including all residues). [c] Q value for specified rigid model (all rows of the corresponding matrix **B** have a norm of 1). [d] Q value for the specified model scaled by final estimates for generalized order parameters (each row of the corresponding matrix **B** is multiplied by the relevant value for S_i).

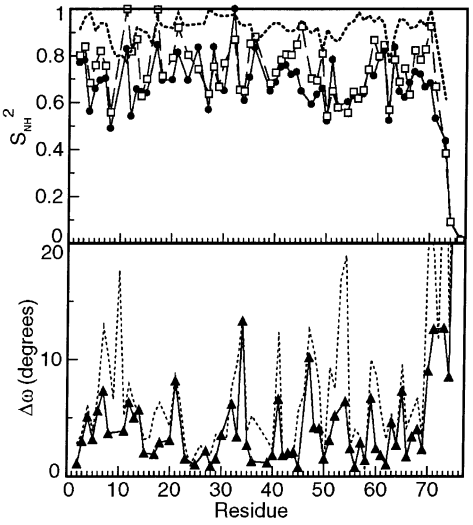

Figure 8. Top: "Dipolar" order parameters squared, S_{NH}^2, obtained using two different initial rigid structures of ubiquitin. The values shown correspond to 61 of the 73 nonproline residues in ubiquitin, for which corresponding RDC measurements were available in all nine media considered. Filled circles: starting with a solution-state NMR structure (1D3Z, model 1).[23] Open squares: starting with a solid-state X-ray structure (1UBQ).[31] For comparison, the S_{NH}^2 values derived from ^{15}N relaxation rates by Tjandra et al.[33] are indicated by a dotted line. All three sets of S_{NH}^2 have been scaled to exhibit a maximum value of 1. Bottom: Comparison of the deviation $\Delta\omega$ of amide NH internuclear vector orientations between the NMR and X-ray structures employed as starting structures. Dotted line: comparison before refinement. Solid line and filled triangles: comparison after the independent refinement of both structures using RDC data only.

regions. This remains consistent with a structure in solution that more closely corresponds to the NMR derived structure.

The results obtained for two of the initial models (1D3Z and 1UBQ) are illustrated in Figure 8. The bottom panel shows that, upon refinement, the two models have converged closer to a common structure, as one would expect. The top panel summarizes the squared generalized order parameters obtained with either of the two structures used as the initial model compared with values obtained from ^{15}N spin relaxation studies.[33] Similar results are obtained using either of the two starting models with results for residues 11, 17, and 32 standing out as probable outliers. Even with a generous accounting for limitations in precision, the results indicate the existence of additional motions occurring on time scales not probed by ^{15}N spin relaxation studies. This is further supported by the overall agreement of the present results with another, recently appearing RDC-based analysis of ubiquitin.[11] That study, which utilized more data than the one described here, also found squared generalized order parameters as low as 0.5 within the structured region of ubiquitin.

Discussion and Conclusion

The foundation for a new approach to the interpretation of multi-alignment RDC data in terms of the structural and dynamic properties of macromolecules has been presented. This approach, called DIDC, rests on a theoretical basis in which the RDC data are used in a collective manner to directly generate the desired structural and dynamic parameters, either de novo or starting from an initial structural model. The applicability of the methodology is limited primarily by the assumption that internal motions are uncorrelated with overall alignment and that different alignment media can be employed without causing changes in the structure or dynamics of the macromolecule. While it is difficult at present to fully assess the validity of these assumptions generally for macromolecules, these limitations are common to current methods for RDC-based analysis. It is anticipated that additional light will be shed on these issues as the RDC methodology evolves. Nevertheless, within these assumptions, there are some significant advantages associated with this new approach. The ability to reduce random errors in measured RDCs by means of an SVD of the data is generally applicable and can be utilized in any application where multiple sets of data have been acquired. The approach also does not require the explicit specification of alignment tensors. If desired, the best-fit alignment tensors can be computed at the end of the analysis. Most significantly, the formalism provides for the unified study of structural and dynamic properties. Not only does this allow normally unavailable dynamic information to be extracted but also should provide improvements in structural accuracy because the dynamics are at least partially accounted for. In the most favorable cases, that is, when "complete" RDC data are available corresponding to five independent alignment media, it is expected that generalized order parameters and mean internuclear vector orientations can be extracted de novo from the data with high accuracy.

It is probably reasonable to expect that future advances in experimental techniques will eventually make the acquisition of complete sets of RDC data a routine matter. However, at present this objective remains demanding. The results from the simulations and the experimental application to ubiquitin suggest that useful dynamic information in the form of generalized order parameters can still be obtained provided that a structure and RDC data acquired in at least three independent alignment media are available. This reduced requirement should now be experimentally feasible for a wide range of systems. This opens up the exciting prospect of exploring the nanosecond to microsecond motional time scales using these new techniques. The results of simulations underscore the importance of using accurate structural coordinates and accurate RDC data to obtain optimal precision of estimated order parameters. Nevertheless, these demands can be satisfied reasonably well provided that a high-resolution structure from the PDB is available and RDC data can be acquired with reasonable precision.

The results for ubiquitin are consistent with this assessment. The singular values resulting from the SVD of the ubiquitin data (Figure 6) indicate that the RDC data are of high quality,

(33) Tjandra, N.; Feller, S. E.; Pastor, R. W.; Bax, A. *J. Am. Chem. Soc.* **1995**, *117*, 12562−12566.

indeed considerably better than any of the noise-containing synthetic data sets generated. It is more difficult to assess the accuracy of the structures used as initial models. However, the agreement between the solution and solid-state structure to begin with and the improvement upon refinement (to $<5°$ RMSD for the protein core), along with the similarity in extracted order parameters (Figure 8), all point toward a level of accuracy of approximately $5-10°$ RMSD for the initial models used. Despite the high-quality structural information and RDC data employed for the application to ubiquitin, a relevant concern is the reliability of results indicating substantially greater amplitudes of internal motion than expected on the basis of ^{15}N spin relaxation studies. The dearth of outside information on this question, with the exception of a couple of studies,[7,11] in combination with the introductory nature of the formalism presented here precludes a firm conclusion. Nevertheless, the analytical tools introduced here provide least-squares best-fit solutions based on the information provided. The fact that the overall variation in generalized order parameters estimated for ubiquitin exceeds by a considerable margin even more pessimistic assessments of precision is thus very significant. While subject to the validity of some underlying assumptions, the results obtained here are clearly not consistent with a model of ubiquitin in line with ^{15}N spin relaxation derived order parameters. It is anticipated that further studies will allow further clarification.

The account presented here has been concerned with introducing the basic formalism along with some illustrations of its application. There are many aspects that remain uninvestigated. Although not explored in the current work, a comparison of Q values for the initial, intermediate, and final structures provide some indication concerning the overall extent of dynamics present. This sort of information might be used to improve the performance of the refinement protocol in data deficient cases. Furthermore, while the methodology was only illustrated in the context of RDC data acquired in a minimum of three alignment tensors, the approach is in principle applicable to situations in which less data is available. In particular, the removal of terms for alignment tensors could lead to improvements in the convergence properties during simulated annealing refinement, similar to the approach of Moltke et al.[6,10] The question is how much data is required at minimum to meaningfully support the simultaneous refinement of dynamics along with structure.

Even in the case of complete RDC data, the interpretation in terms of molecular parameters is fundamentally underdetermined (by 25 parameters). At present, it appears that the uncertainty in these parameters is sufficient to significantly inhibit the reliable description of motional asymmetry. Likely, the use of additional constraints will enable the extraction of parameters with improved accuracy and precision. This includes the addition of RDC data corresponding to different interactions. Indeed, the formalism is well suited to the unified analysis of data from multiple different dipolar interactions. By choosing interactions (such as $C^{\alpha}-C'$) with an effectively fixed geometric relationship to the N−H interaction vectors examined here, additional angular constraints may be obtained that could lift some of the ambiguities. Along these lines, this approach might be extended to allow a backbone fold accompanied by a description of dynamics to be constructed with high accuracy.

Acknowledgment. The author is grateful to Brian Cutting for help with sample preparation and data acquisition and Geoffrey Bodenhausen for his support and encouragement. This research was supported by the FNRS of Switzerland.

Supporting Information Available: Summary of the nine experimental best-fit alignment tensors. Derivation of eqs 12 and 17. This material is available free of charge via the Internet at http://pubs.acs.org.

JA0261123