

# Reconstruction of interatomic vectors by principle component analysis of nuclear magnetic resonance data in multiple alignments

Jean-Christophe Hus and Rafael Brüschweiler<sup>a)</sup>

*Carlson School of Chemistry and Biochemistry, Clark University, Worcester, Massachusetts 01610*

(Received 25 March 2002; accepted 24 April 2002)

A general method is presented for the reconstruction of interatomic vector orientations from nuclear magnetic resonance (NMR) spectroscopic data of tensor interactions of rank 2, such as dipolar coupling and chemical shielding anisotropy interactions, in solids and partially aligned liquid-state systems. The method, called PRIMA, is based on a principal component analysis of the covariance matrix of the NMR parameters collected for multiple alignments. The five nonzero eigenvalues and their eigenvectors efficiently allow the approximate reconstruction of the vector orientations of the underlying interactions. The method is demonstrated for an isotropic distribution of sample orientations as well as for finite sets of orientations and internuclear vectors encountered in protein systems. © 2002 American Institute of Physics. [DOI: 10.1063/1.1485727]

## I. INTRODUCTION

Nuclear magnetic resonance (NMR) spectroscopy provides unique structural information on aligned and partially aligned macromolecular systems in the solid state and in the liquid state.<sup>1–12</sup> There are many ways to achieve spatial alignment such as by crystallization, by the presence of an external magnetic or electric field, by imbedding of the macromolecules in a bilayer or an anisotropic gel, or by the presence of a liquid crystalline environment.<sup>5–10</sup> Structural information is encoded in the magnitudes of anisotropic interactions of rank 2, such as the dipolar interaction, chemical shielding anisotropy, and the nuclear quadrupolar interaction, which depend on the orientation of the interaction with respect to the external magnetic field.<sup>1</sup>

Unfortunately, the generation of a three-dimensional (3D) structure from such data is not straightforward. A standard procedure starts with the measurement of couplings or shieldings that belong to the same rigid molecular fragment from which a discrete set of orientations can be derived that are compatible with the experimental data.<sup>12</sup> The degeneracy of allowed fragment orientations can be reduced by collecting additional data with the sample being differentially aligned with respect to the external magnetic field. In the case of solid-state NMR the overall orientation can be changed by tilting the sample whereas in liquid-state NMR the alignment tensor can be modified by altering or by replacing the alignment medium. Oriented molecular fragments can be assembled into larger structures using bond-angle restrictions at the interface as additional conformational constraints.<sup>12,13</sup> These structures serve as starting points for structural refinement procedures using molecular force field-based simulated annealing schemes often in combination with other NMR-derived constraints such as nuclear overhauser enhancements (NOEs) and scalar  $J$  couplings.<sup>7,9,10,14–19</sup>

We present here a general and efficient method for the conversion of orientational tensor information of rank 2 collected for different alignments in the solid state or liquid state into three-dimensional intramolecular vector orientations. The method is based on a principle component analysis of the covariance matrix of the interactions at different sites evaluated over multiple alignments. Diagonalization of the matrix yields five eigenvectors with nonzero eigenvalues that can be converted into three-dimensional vector orientations. The method is demonstrated for sets of randomly oriented vectors as well as sets of internuclear vectors extracted from 3D protein structures.

## II. METHOD

### A. General

In the following we focus on magnetic dipole–dipole couplings. Other anisotropic interactions such as chemical shielding anisotropy and nuclear quadrupolar interaction can be treated in a fully analogous way. An ensemble of identical molecules in the presence of a strong magnetic field  $B_0$  is considered where each molecule has  $N$  heteronuclear spin  $1/2$  pairs ( $IS$ ) with a fixed internuclear distance  $r_{IS}$ . For a given overall alignment  $k$  the orientations of the internuclear vectors  $\mathbf{e}_i^{(k)}$  ( $i=1,\dots,N$ ) with respect to the laboratory frame with the  $z$  axis parallel to  $B_0$  are defined by the polar angles  $\Omega_i^{(k)} = (\theta_i^{(k)}, \varphi_i^{(k)})$ . For a static system with a fixed overall orientation (alignment), such as a single crystal, the secular part of the magnetic dipole–dipole Hamiltonian is given by<sup>1</sup>

$$\mathcal{H}^{(k)} = -\frac{\mu_0}{4\pi} \frac{h}{2\pi} \gamma_I \gamma_S r_{IS}^{-3} \sum_{i=1}^N P_2(\cos \theta_i^{(k)}) 2I_{iz} S_{iz}, \quad (1)$$

where  $\mu_0$  is the magnetic field constant,  $h$  is Planck's constant,  $\gamma_I$  and  $\gamma_S$  are the gyromagnetic ratios of spin species  $I$  and  $S$ , and  $P_2(x) = (3x^2 - 1)/2$  is the Legendre polynomial of rank 2. This Hamiltonian causes a resonance splitting of size

<sup>a)</sup> Author to whom correspondence should be addressed; electronic mail: bruscheiler@nmr.clarku.edu

$$D_i^{(k)} = dP_2(\cos \theta_i^{(k)}) = d \sqrt{\frac{4\pi}{5}} Y_{20}(\theta_i^{(k)}), \quad (2)$$

where

$$d = -\frac{1}{\pi} \frac{\mu_0}{4\pi} \frac{h}{2\pi} \gamma_I \gamma_S r_{IS}^{-3}$$

has units of Hz.

For a system that is partially oriented, for example due to the presence of a liquid-crystalline environment, the dipolar couplings are partially averaged by fast overall reorientational Brownian motions leading to residual dipolar couplings (RDCs) that can be described by a symmetric and traceless alignment tensor  $\mathbf{D}$  (in units of Hz) with eigenvalues  $D_{xx}$ ,  $D_{yy}$ ,  $D_{zz}$ , where  $|D_{zz}| \geq |D_{yy}| \geq |D_{xx}|$ . In the eigenframe of tensor  $\mathbf{D}^{(k)}$  the residual dipolar coupling between spins  $I$  and  $S$  can be expressed as

$$D_i^{(k)} = D_a^{(k)} \{3 \cos^2 \theta_i^{(k)} - 1 + \frac{3}{2} R^{(k)} \sin^2 \theta_i^{(k)} \cos 2\varphi_i^{(k)}\}, \quad (3)$$

where  $D_a^{(k)} = D_{zz}^{(k)}/2$  is the axial component and  $R^{(k)} = 2/3 \cdot (D_{xx}^{(k)} - D_{yy}^{(k)})/D_{zz}^{(k)}$  is the rhombicity of  $\mathbf{D}^{(k)}$  with  $0 \leq R^{(k)} \leq 2/3$ .  $\Omega_i^{(k)} = (\theta_i^{(k)}, \varphi_i^{(k)})$  is the direction of internuclear vector  $\mathbf{e}_i^{(k)}$ . For what follows it is convenient to express the residual dipolar couplings of Eq. (3) by normalized second order spherical harmonic functions  $Y_{2M}(\theta, \varphi)$ :

$$D_i^{(k)} = D_{zz}^{(k)} \sqrt{\frac{4\pi}{5}} \left\{ Y_{20}(\theta_i^{(k)}, \varphi_i^{(k)}) + \sqrt{\frac{3}{8}} R^{(k)} [Y_{22}(\theta_i^{(k)}, \varphi_i^{(k)}) + Y_{22}^*(\theta_i^{(k)}, \varphi_i^{(k)})] \right\}. \quad (4)$$

Except for an overall scaling factor, the solid-state case of Eq. (2) is contained in Eq. (4) by setting  $R^{(k)} = 0$ .

### B. Covariance matrix of dipolar couplings

The  $N \times N$  covariance matrix  $\mathbf{C}$  of dipolar couplings or RDCs measured for  $M$  different alignments  $k$  can be calculated according to

$$C_{ij} = \frac{1}{M-1} \sum_{k=1}^M w_k (D_i^{(k)} - \langle D_i \rangle_k) (D_j^{(k)} - \langle D_j \rangle_k), \quad (5)$$

$i, j = 1, \dots, N,$

where  $\langle D_i \rangle_k = 1/M \sum_{k=1}^M D_i^{(k)}$  is the average of the couplings  $D_i^{(k)}$  measured for  $M$  alignments and  $w_k \geq 0$  is the relative weight of the dipolar couplings of alignment  $k$ . These weights can be adjusted to take into account differential  $D_{zz}^{(k)}$  values and different noise levels. For the theoretical examples discussed below the  $w_k$  are set to 1.

### C. Isotropic distribution of alignment media

In the idealized case of an isotropic distribution of alignments with fixed axial and rhombic components,  $D_a$  and  $R$ , a simple analytical expression can be found for the elements of covariance matrix  $\mathbf{C} = \mathbf{C}^{\text{iso}}$ . The effect of a change of the

alignment on the dipolar couplings can be expressed using the well-known relationships of spherical harmonics under 3D rotations  $\mathbf{R}(\alpha, \beta, \gamma)$  defined by the three Euler angles  $\alpha$ ,  $\beta$ , and  $\gamma$ :<sup>20</sup>

$$\mathbf{R}(\alpha, \beta, \gamma) Y_{2M}(\Omega) = \sum_{M'} D_{M'M}^{(2)}(\alpha, \beta, \gamma) Y_{2M'}(\Omega). \quad (6)$$

If the alignment tensor  $l$  is related to the alignment tensor  $k$  by the rotation  $\mathbf{R}(\alpha, \beta, \gamma)$ , the dipolar couplings  $D_i^{(l)}$  are obtained by applying Eq. (6) to Eq. (4):<sup>21</sup>

$$D_i^{(l)} = D_{zz} \sqrt{\frac{4\pi}{5}} \left\{ \sum_M D_{M0}^{(2)}(\alpha, \beta, \gamma) Y_{2M}(\Omega_i^{(k)}) + \sqrt{\frac{3}{8}} R \left[ \sum_M D_{M2}^{(2)}(\alpha, \beta, \gamma) Y_{2M}(\Omega_i^{(k)}) + \sum_M D_{M2}^{(2)*}(\alpha, \beta, \gamma) Y_{2M}^*(\Omega_i^{(k)}) \right] \right\}. \quad (7)$$

Insertion of Eq. (7) into Eq. (5) followed by analytical integration over the Euler angles yields

$$C_{ij}^{\text{iso}} = D_{zz}^2 (1 + \frac{3}{8} R^2) P_2(\mathbf{e}_i \cdot \mathbf{e}_j), \quad (8)$$

where  $\langle Y_{2M}(\Omega) \rangle_{\text{iso}} \propto \int d\alpha \sin \beta d\beta d\gamma Y_{2M}(\Omega) = 0$  and the orthonormality relationship of Wigner matrix elements<sup>20</sup>

$$\frac{1}{8\pi^2} \int d\alpha \sin \beta d\beta d\gamma D_{M'N'}^{(2)}(\alpha, \beta, \gamma) D_{M''N''}^{(2)*}(\alpha, \beta, \gamma) = \frac{1}{5} \delta_{M'M''} \delta_{N'N''} \quad (9)$$

were used. In the solid-state case one obtains in full analogy:

$$C_{ij}^{\text{iso}} = d^2 P_2(\mathbf{e}_i \cdot \mathbf{e}_j). \quad (10)$$

### D. Reconstruction of dipolar vectors

Directional information about the internuclear vectors  $\mathbf{e}_i$  can be extracted from covariance matrix  $\mathbf{C}$  (or  $\mathbf{C}^{\text{iso}}$ ) after performing a principal component analysis:

$$\mathbf{C}|q\rangle = \lambda_q |q\rangle, \quad (11)$$

where  $|q\rangle$  are the  $N$  normalized eigenvectors and  $\lambda_q$  are the corresponding eigenvalues. The eigenvalues fulfill  $\lambda_q \geq 0$  and in case the shape of the molecule does not depend on the alignment at most five eigenvalues differ from zero. We assume in the following that the eigenvalues are sorted in size with the largest eigenvalue being  $\lambda_5$  and the smallest nonzero eigenvalue being  $\lambda_1$ . The distribution of orientations directly affects the magnitudes of the five largest eigenvalues. The condition number  $c$  is defined as the ratio of the largest to the smallest nonzero eigenvalue,  $c = \lambda_5 / \lambda_1$ .

From the five nonzero eigenvalues and their eigenvectors  $|1\rangle, \dots, |5\rangle$  a  $N \times 5$  matrix  $\mathbf{A}$  is constructed with elements

$$A_{iq} = \sqrt{\lambda_q} |q\rangle_i, \quad i = 1, \dots, N, \quad q = 1, \dots, 5, \quad (12)$$

where  $|q\rangle_i$  is the  $i$ th element of eigenvector  $|q\rangle$ . The information of the direction of vector  $\mathbf{e}_i$  is encoded in the five

matrix elements of the  $i$ th row. Each of the five elements is considered to correspond to the real or imaginary part of the spherical harmonics of rank 2 evaluated in a molecular frame that is the same for all vectors, i.e., all rows:

$$\begin{aligned} A_{i1} &= \sqrt{\frac{4\pi}{5}} d_{z^2} = \frac{1}{2}(3z_i^2 - 1), \\ A_{i2} &= \sqrt{\frac{4\pi}{5}} d_{zx} = \sqrt{3}z_ix_i, \\ A_{i3} &= \sqrt{\frac{4\pi}{5}} d_{zy} = \sqrt{3}z_iy_i, \\ A_{i4} &= \sqrt{\frac{4\pi}{5}} d_{x^2-y^2} = \frac{\sqrt{3}}{2}(x_i^2 - y_i^2), \\ A_{i5} &= \sqrt{\frac{4\pi}{5}} d_{xy} = \sqrt{3}x_iy_i, \end{aligned} \quad (13)$$

where  $x_i$ ,  $y_i$ , and  $z_i$  are the Cartesian coordinates of unit vector  $\mathbf{e}_i$ :

$$\mathbf{e}_i = (x_i, y_i, z_i) = (\cos \varphi_i \sin \theta_i, \sin \varphi_i \sin \theta_i, \cos \theta_i). \quad (14)$$

For each vector  $\mathbf{e}_i$  a best estimate  $\mathbf{e}'_i$  can be obtained by fitting  $(\theta_i, \varphi_i)$  to the five elements  $A_{iq}$ . This can be done by a grid search over  $\theta_i, \varphi_i$  values or by a nonlinear least-squares minimization with respect to  $\theta_i, \varphi_i$ . Because the assignment of the five largest eigenvectors to the spherical harmonics is not known, each of the  $5! = 120$  permutations has to be tested. In addition, because  $|q\rangle$  and  $-|q\rangle$  are both eigenvectors to the same eigenvalue  $\lambda_q$ , the absolute sign of the eigenvectors  $|q\rangle$  is unknown. Therefore, all  $2^5 = 32$  possible sign combinations have to be considered. The  $(\theta'_i, \varphi'_i)$  direction that yields the best agreement for any of the  $120 \times 32 = 3840$  combinations represents the best estimate for  $\mathbf{e}_i$ . For each row of matrix  $\mathbf{A}$  the best fitting vector  $\mathbf{e}'_i$  of Eq. (14) can be individually determined. Because each of the 3840 fits includes only two fitting parameters,  $\theta'_i$  and  $\varphi'_i$ , it can be carried out very efficiently. In the following the method is referred to as principal component analysis of multiple alignment data (PRIMA).

It follows directly from Eq. (13) that for each vector  $\mathbf{e}'_i$  its inversion  $-\mathbf{e}'_i$  fits equally well. Hence, the PRIMA procedure yields orientations of internuclear vectors rather than absolute directions, which is a direct consequence of the fact that rank 2 interactions are invariant under inversion. This is also manifested in the mathematical form of the isotropic covariances of Eqs. (8) and (10), which are proportional to  $3(\mathbf{e}_i \cdot \mathbf{e}_j)^2 - 1$  and, therefore, invariant under inversion of either one or both directions  $\mathbf{e}_i$  and  $\mathbf{e}_j$ .

### E. Comparison of original and reconstructed vectors

A simple way for assessing the overall agreement between the original and the reconstructed vectors is by taking the difference of the angles between two original vectors  $\mathbf{e}_i$  and  $\mathbf{e}_j$  and the corresponding reconstructed vectors  $\mathbf{e}'_i$  and  $\mathbf{e}'_j$ :

$$\Delta_{ij} = |[\cos^{-1} |(\mathbf{e}_i \cdot \mathbf{e}_j)|] - \cos^{-1} |(\mathbf{e}'_i \cdot \mathbf{e}'_j)|] \in [0^\circ, 90^\circ]. \quad (15)$$

A measure for the discrepancy between the two sets of vectors  $\mathbf{e}_i$  and  $\mathbf{e}'_i$  is the angle difference averaged over all  $N(N-1)/2$  angles:

$$\langle \Delta \rangle = \frac{2}{N(N-1)} \sum_{i < j} \Delta_{ij}, \quad (16)$$

where  $\langle \Delta \rangle$  is zero if the original and the reconstructed directions are identical, otherwise  $\langle \Delta \rangle > 0$ .

Other quantities that will be used in Sec. III are the standard deviation  $\sigma_\Delta$  of the  $\Delta_{ij}$  values:

$$\sigma_\Delta = \left( \frac{1}{N(N-1)/2 - 1} \sum_{i < j} (\Delta_{ij} - \langle \Delta \rangle)^2 \right)^{1/2} \quad (17)$$

and the maximal and minimal values of  $\Delta_{ij}$ , respectively,

$$\Delta_{\max} = \max\{\Delta_{ij}\} \quad \text{and} \quad \Delta_{\min} = \min\{\Delta_{ij}\}. \quad (18)$$

It is useful to define for a given vector  $i$  the average of  $\Delta_{ij}$  for all  $N-1$  vector pairs  $ij$  with  $j \neq i$ :

$$\langle \Delta_{ij} \rangle_j = \frac{1}{N-1} \sum_{j \neq i} \Delta_{ij}. \quad (19)$$

The quality of the reconstructed vectors can also be assessed by directly superimposing them on the original vectors. Since the two structures are expressed in different frames, the two frames need to be first aligned by applying a 3D rotation defined by the Euler angles  $\alpha'$ ,  $\beta'$ , and  $\gamma'$  that minimizes the penalty function

$$\chi^2 = \sum_{i=1}^N [\sin \delta_i(\alpha', \beta', \gamma')]^2, \quad (20)$$

where  $\delta_i$  is the angle between vectors  $\mathbf{e}_i$  and  $\mathbf{e}'_i$ . The Euler angles  $\alpha'$ ,  $\beta'$ , and  $\gamma'$  are determined by a nonlinear least-squares optimization.  $\chi^2$  is zero if the original and the reconstructed directions are identical, otherwise it is larger than zero. While the optimized  $\delta_i$  values,  $\delta_i^{\min}$ , provide a more direct assessment of the differences between original and reconstructed vectors than the measures of Eqs. (16) and (19), the minimization of Eq. (20) is significantly more time consuming. A comparison between the two measures,  $\langle \Delta_{ij} \rangle_j$  and  $\delta_i^{\min}$ , is given in the following section.

## III. APPLICATIONS

### A. Isotropic distribution of alignments

We first consider the idealized situation of dipolar couplings obtained for an isotropic distribution of alignments. The covariance matrix  $\mathbf{C}^{\text{iso}}$  is then given by Eqs. (8) and (10). The PRIMA method is first tested for a set of randomly selected unit vectors and then for vector orientations extracted from protein structures.

#### 1. Randomly distributed vectors

First, random orientations were generated for 100 vectors. For each set of vectors, the covariance matrix  $\mathbf{C}^{\text{iso}}$  was constructed using Eq. (8). A principal component analysis was applied to the covariance matrix. 95 of the 100 eigen-

TABLE I.  $\Delta$  measure for random sets of vectors with isotropically distributed alignment media.

| Number of vectors                                | 100             | 500             | 1000            | 2000            | 3000            | 700 <sup>a</sup> |
|--|-----------------|-----------------|-----------------|-----------------|-----------------|------------------|
| $\langle\Delta\rangle \pm \sigma_{\Delta}$ (deg) | 5.34 $\pm$ 4.39 | 4.02 $\pm$ 3.07 | 3.27 $\pm$ 2.54 | 3.65 $\pm$ 2.98 | 3.34 $\pm$ 2.32 | 1.01 $\pm$ 0.89  |
| $\Delta_{\min} - \Delta_{\max}$ (deg)            | 0.00–71.86      | 0.00–24.24      | 0.00–13.68      | 0.00–16.27      | 0.00–11.42      | 0.00–4.67        |
| Condition number $\langle c \rangle$             | 2.17            | 1.70            | 1.71            | 1.56            | 1.53            | 2.77             |

<sup>a</sup>Distribution of 700 nearly isotropically distributed vectors taken from Ref. 20.

values can be considered as zero, within numerical precision, since they are at least a factor of  $10^{15}$  smaller than the largest five eigenvalues. Therefore, the covariance matrix is highly singular and we will focus on the nonzero eigenvalues and their eigenvectors.

The condition number averaged over 100 runs using different sets of random vectors is  $c = \lambda_5 / \lambda_1 = 2.17 \pm 0.27$ . The vector orientations were reconstructed by the procedure described in Sec. II D. A nonlinear least-squares minimization was applied for each vector using five different randomly chosen initial values for  $(\theta_i, \varphi_i)$ . The average angular deviation and its standard deviation is  $\langle\Delta\rangle = 5.34^\circ \pm 4.39^\circ$ . A control calculation, where the original vectors are compared to other sets of randomly chosen vectors, gives  $\langle\Delta\rangle = 25.08^\circ \pm 0.33^\circ$ , which corresponds to the  $\langle\Delta\rangle$  value for two uncorrelated sets of vectors.

The dependence of  $\langle\Delta\rangle$  on the number of vectors was examined for  $N = 100, 500, 1000, 2000,$  and  $3000$  with the results presented in Table I. The results for  $N = 100$  are averaged over 100 runs and for  $N = 500$  over 10 runs. For increasing  $N$ ,  $\langle\Delta\rangle$  decreases from  $5.34^\circ$  to  $3.34^\circ$ . Also given are the minimal and maximal deviations  $\Delta_{\min}$  and  $\Delta_{\max}$ . While  $\Delta_{\min}$  is close to zero in all cases,  $\Delta_{\max}$  drops from  $71.86^\circ$  for  $N = 100$  to  $11.42^\circ$  for  $N = 3000$ . Thus, both the average and the width of the  $\Delta$  distribution decrease with an increasing number of vectors. The approachment of  $\langle\Delta\rangle$  toward zero for increasing  $N$  is slow. The average condition number  $\langle c \rangle$  is correlated with  $\langle\Delta\rangle$  ( $r = 0.93$ ), whereas the correlation essentially disappears when  $c$  is compared with  $\langle\Delta\rangle$  for individual runs. When the method is applied to 700 vectors that are nearly isotropically oriented using the SHREWD method<sup>22</sup> the  $\langle\Delta\rangle$  value drops to  $1.01^\circ$  indicating that a notably accurate reconstruction can be obtained for a highly uniform distribution of vector orientations (Table I).

## 2. Internuclear vectors extracted from protein structures

Next, the PRIMA method is applied to internuclear vectors extracted from the 3D structure of the 76-amino acid protein ubiquitin. For this system, dipolar couplings are measurable, for example, in the form of residual dipolar couplings using HNCO-type NMR experiments,<sup>9,23,24</sup> which al-

low one to collect orientational information on  $C_{i-1}^\alpha - C_{i-1}'$ ,  $C_{i-1}' - N_i$ , and  $N_i - H_i^N$  vectors that belong to the peptide plane connecting amino acids  $i - 1$  and  $i$ . Using this experimental scheme, RDC data are available for all residues except for prolines and the N-terminal residue leading to a total of  $3 \times 72 = 216$  backbone  $C^\alpha - C'$ ,  $C' - N$ , and  $N - H$  vectors. These vectors were extracted from the NMR structure of ubiquitin [PDB code 1D3Z (Ref. 25)]. Application of the method to these vectors and subsets thereof yields  $\langle\Delta\rangle$  values between  $4^\circ - 6^\circ$  (Table II). While the condition number  $c$  decreases when more vectors are included,  $\langle\Delta\rangle$  does not follow this trend.

The PRIMA method was applied to five other proteins covering a molecular weight range of 8.5 to 40.7 kDa. The selected proteins, which contain variable amounts of  $\alpha$ -helix and  $\beta$ -sheet secondary structures, are Gaip [1CMZ (Ref. 26)], CenC [1ULO (Ref. 27)], myoglobin [1MYF (Ref. 28)], cutinase [1CEX (Ref. 29)] and maltose-binding protein (MBP) [1EZP (Ref. 17)]. Gaip and myoglobin are helical proteins whereas CenC is predominantly  $\beta$ -sheet. The other proteins display both  $\alpha$ -helices and  $\beta$ -sheets. The results, which are summarized in Table III, indicate that  $\langle\Delta\rangle$  values cover the range between  $4^\circ - 6^\circ$ , i.e., they vary only little for these widely differing proteins. The correlation between protein size,  $\langle\Delta\rangle$ , and  $c$  is weak; the largest protein (MBP) has the smallest  $\langle\Delta\rangle$  ( $\langle\Delta\rangle = 4.13^\circ$ ) and the smallest condition number ( $c = 1.21$ ), while the protein with the largest condition number, CenC, is not the one with the largest difference  $\langle\Delta\rangle$ . The results of Table III indicate that the PRIMA reconstruction procedure is robust and reasonably accurate for protein systems in case the number of available alignments is sufficiently large.

## B. Discrete sets of alignments

In practice, experimental data can be gathered only for a limited number of different alignment tensors that may not be isotropically distributed. Nonuniform sampling of tensor orientations is expected to lead to systematic errors for the reconstruction of dipolar vectors. From an experimental perspective, it is conceivable to measure RDCs for ten or more different alignments.<sup>30</sup> Two cases involving 6 and 10

TABLE II.  $\Delta$  measure for backbone vectors of ubiquitin with isotropically distributed alignment media.

| Type of vectors                                  | N-H             | C'-N            | C <sup>α</sup> -C' | All             |
|--|-----------------|-----------------|--------------------|-----------------|
| Number of vectors                                | 72              | 72              | 72                 | 216             |
| $\langle\Delta\rangle \pm \sigma_{\Delta}$ (deg) | 5.94 $\pm$ 4.66 | 4.38 $\pm$ 3.71 | 5.15 $\pm$ 3.61    | 5.79 $\pm$ 4.73 |
| $\Delta_{\min} - \Delta_{\max}$ (deg)            | 0.01–22.18      | 0.00–19.88      | 0.00–18.81         | 0.00–27.92      |
| Condition number $c$                             | 3.00            | 1.80            | 2.19               | 1.26            |

TABLE III.  $\Delta$  measure for different proteins.

| Protein   | Number of vectors <sup>a</sup> | $\langle\Delta\rangle\pm\sigma_\Delta$<br>(deg) | Condition<br>number $c$ |
|-----------|--------------------------------|---|-------------------------|
| Ubiquitin | 216                            | $5.79\pm 4.73$                                  | 1.26                    |
| Gaip      | 363                            | $4.29\pm 3.36$                                  | 1.25                    |
| CenC      | 423                            | $5.57\pm 4.53$                                  | 1.33                    |
| Myoglobin | 444                            | $5.37\pm 4.57$                                  | 1.26                    |
| Cutinase  | 561                            | $4.40\pm 3.35$                                  | 1.27                    |
| MBP       | 1044                           | $4.13\pm 2.88$                                  | 1.21                    |

<sup>a</sup>All C $\alpha$ -C', C'-N, and N-H vectors were used except for proline residues and the N-terminus.

alignment tensors, respectively, are analyzed in more detail. The  $D_a$  values were set to 10 Hz and the rhombicity  $R$  was set to either 0, 1/3, or 2/3. The PRIMA analysis and reconstruction procedure was applied to a system consisting of 100 randomly oriented internuclear vectors. As for the isotropic alignment distribution case, only five eigenvalues are nonzero within numerical precision. The results were averaged over 20 sets of randomly distributed alignment tensor orientations and they are given in Table IV.

When the number of alignments is increased from 6 to 10 the  $\langle\Delta\rangle$  value improves (decreases) by almost 30%. The rhombicity  $R$  plays a minor role, since  $\langle\Delta\rangle$  changes only marginally for  $R\neq 0$ . Addition of 5% Gaussian random noise to the dipolar couplings leaves the results virtually unchanged indicating that the method is insensitive to moderately small experimental errors.

The vector orientations generated for Table IV were analyzed in terms of  $\delta_i^{\min}$  angles, which give the angular difference between the reconstructed direction and the original direction in an optimized frame as explained in Sec. II E. In Fig. 1,  $\delta_i^{\min}$  is plotted vs  $\langle\Delta_{ij}\rangle_j$ , where  $\langle\Delta_{ij}\rangle_j$  denotes the average of the internuclear angles  $\Delta_{ij}$  between vector  $\mathbf{e}'_i$  and the other 99 vectors  $\mathbf{e}'_j$  ( $j\neq i$ ). It shows that the two measures exhibit a common trend despite the fact that for a given  $\langle\Delta_{ij}\rangle_j$  the distribution of the  $\delta_i^{\min}$  values has a considerable width. For 3° intervals in  $\Delta$  the median of  $\delta_i^{\min}$ ,  $\langle\delta_i^{\min}\rangle_{\text{med}}$ , is plotted in Fig. 1 (filled circles). The dependence of  $\langle\delta_i^{\min}\rangle_{\text{med}}$  on  $\langle\Delta_{ij}\rangle_j$  follows in good approximation the power law

$$\langle\delta_i^{\min}\rangle_{\text{med}}=0.7636\cdot\langle\Delta_{ij}\rangle_j^{1.2381}, \quad (21)$$

which is indicated in Fig. 1 as a solid line. The  $\delta_i^{\min}$ -angle distributions are illustrated in Fig. 2 in the form of histograms for selected  $\langle\Delta_{ij}\rangle_j$  values [Eq. (19)] that fall into the intervals  $\langle\Delta_{ij}\rangle_j\in[4.5^\circ,5.5^\circ]$ ,  $\langle\Delta_{ij}\rangle_j\in[9^\circ,11^\circ]$ , and  $\langle\Delta_{ij}\rangle_j$

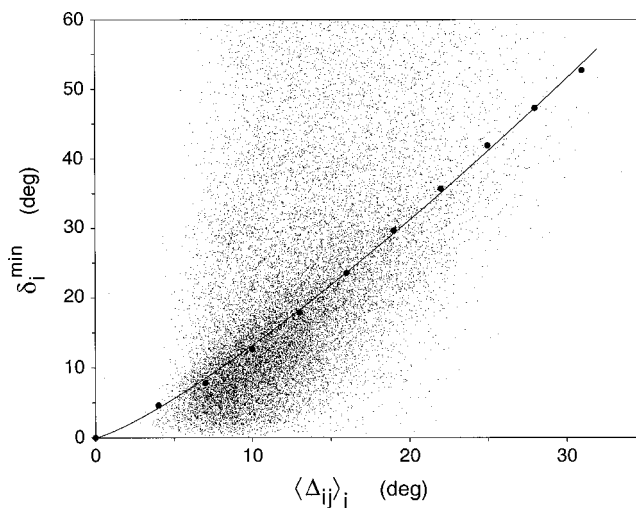


FIG. 1. Relationship between the two orientational difference measures  $\delta_i^{\min}$  and  $\langle\Delta_{ij}\rangle_j$  explained in Sec. II E for the calculations summarized in Table IV. The PRIMA analysis was applied to dipolar couplings of 100 randomly oriented internuclear vectors for 6 or 10 different alignments. The correspondence between the median  $\langle\delta_i^{\min}\rangle_{\text{med}}$  and  $\langle\Delta_{ij}\rangle_j$  (filled circles) follows the power-law of Eq. (21) (solid line).

$\in[13.5^\circ,16.5^\circ]$ . As can be seen in Fig. 2, for all three histograms the median  $\langle\delta_i^{\min}\rangle_{\text{med}}$  is within 1° of the maximum of the distribution.

In solid-state NMR it can be of advantage to reduce the number of interactions together with the corresponding line-widths by selective labeling strategies. On the other hand, in the solid state the alignment can be controlled more easily than in partially oriented systems and experiments can be performed for a larger number of different alignments. For example, if 50 different alignments distributed nearly isotropically using orientations from Ref. 22 and only 20 random vectors are used, one obtains  $\langle\Delta\rangle=10.43^\circ\pm 8.44^\circ$  where the average and the standard deviation were determined using 100 different runs.

If the 72 backbone N-H vector orientations of ubiquitin are chosen instead of a randomly oriented set of vectors, the  $\langle\Delta\rangle$  value is for 6 alignments about  $13.4^\circ$ , i.e., slightly higher than for the random vector set (Table V). The values in Table V were determined by averaging over 20 randomly chosen sets of, tensor orientations. Interestingly, PRIMA reconstruction works almost equally well if only five alignments are used, despite the fact that the covariance matrix has in this case only four nonzero eigenvalues. For 50 randomly selected alignments one finds  $\langle\Delta\rangle\cong 6^\circ$ , which is very similar to

TABLE IV.  $\Delta$  measure for random sets of 100 vectors for a finite set of alignment tensors.

| Number of<br>alignments                      | Six alignments   |                  |                  | Ten alignments |                |                |                |
|--|------------------|------------------|------------------|----------------|----------------|----------------|----------------|
|  | No noise         |                  |                  | No noise       |                | 5% noise       |                |
| Rhombicity $R$                               | 0                | 1/3              | 2/3              | 0              | 1/3            | 2/3            | 1/3            |
| $\langle\Delta\rangle\pm\sigma_\Delta$ (deg) | $12.98\pm 10.85$ | $12.48\pm 10.50$ | $12.87\pm 10.91$ | $9.45\pm 7.89$ | $9.00\pm 7.40$ | $9.15\pm 7.57$ | $9.10\pm 7.53$ |

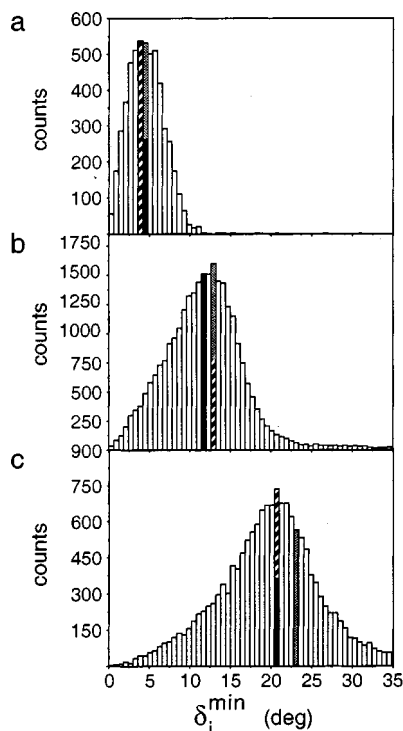


FIG. 2. Histogram representations of  $\delta_i^{\min}$  values with bin size of  $0.6^\circ$  for selected ranges of  $\langle \Delta_{ij} \rangle_j$  of Fig. 1: for panel (a)  $\langle \Delta_{ij} \rangle_j \in [4.5^\circ, 5.5^\circ]$ , for panel (b)  $\langle \Delta_{ij} \rangle_j \in [9^\circ, 11^\circ]$ , and for panel (c)  $\langle \Delta_{ij} \rangle_j \in [14.5^\circ, 16.5^\circ]$ . The arithmetic averages  $\langle \delta_i^{\min} \rangle$  are indicated as gray bars, the medians  $\langle \delta_i^{\min} \rangle_{\text{med}}$  as black bars, and the maxima of the distributions  $(\delta_i^{\min})_{\text{max}}$  as hatched bars. The numerical values for panels a, b, and c are as follows:  $\langle \delta_i^{\min} \rangle = 4.7^\circ \pm 2.4^\circ$ ,  $13.0^\circ \pm 9.0^\circ$ ,  $23.3^\circ \pm 13.4^\circ$ ;  $\langle \delta_i^{\min} \rangle_{\text{med}} = 4.5^\circ$ ,  $11.9^\circ$ ,  $20.7^\circ$ ;  $(\delta_i^{\min})_{\text{max}} = 3.9^\circ$ ,  $12.9^\circ$ ,  $20.7^\circ$ .

$\langle \Delta \rangle$  obtained for the continuous and isotropic alignment distribution case of Table II.

#### IV. DISCUSSION AND CONCLUSION

In recent years the use of dipolar couplings and other rank 2 tensor interactions for the structural characterization of macromolecules has become increasingly popular due to new methods to partially orient such systems and due to progress in NMR pulse sequence design. In liquid-state applications dipolar coupling information is mostly used in combination with other NMR parameters, such as NOEs and scalar  $J$  couplings, and an underlying molecular force field to refine a 3D structural model.

The PRIMA method presented here differs from other approaches because it requires neither knowledge of alignment tensors nor other experimental constraints. It allows the direct establishment of 3D vector orientations from experimental couplings in the absence of a molecular force field

TABLE V. Application to 72 N–H vectors of ubiquitin for variable numbers of alignments with  $R=1/3$ .

| Number of alignments                             | 5 alignments      | 6 alignments      | 50 alignments   |
|--|-------------------|-------------------|-----------------|
| $\langle \Delta \rangle \pm \sigma_\Delta$ (deg) | $14.19 \pm 12.02$ | $13.40 \pm 11.01$ | $5.97 \pm 4.94$ |
| Condition number $c$                             | infinity          | $2667 \pm 9630$   | $4.13 \pm 1.08$ |

and any knowledge of the covalent bonding properties of the molecule. In this regard, this approach can be viewed as the orientational analog of the distance geometry method.<sup>31</sup> In fact, distance geometry can be formulated in terms of a principal component analysis of the rank 1 covariance matrix of an isotropically distributed molecular ensemble.<sup>32</sup> The PRIMA method fundamentally differs from the “metric method”<sup>33</sup> used in solid-state NMR for the determination of polypeptide backbone structures, which primarily relies on the determinant property  $|\mathbf{M}|=0$  of the metric matrix of rank 1 with elements  $M_{ij} = \mathbf{e}_i \cdot \mathbf{e}_j$  for any set of four vectors.

To achieve good accuracy dipolar couplings collected for a larger number of alignments are required. The method gives the best results for macromolecular systems that possess many different dipolar interactions that are homogeneously distributed together with alignments that are as uniformly distributed as possible. If a very high resolution structure is required, further refinement of the orientations can be achieved by fitting the  $M$  alignment tensors and  $N$  vector orientations directly to the  $M \cdot N$  experimental couplings. Due to the high dimensionality and the nonlinear character of this optimization problem, the availability of a good initial model generated by the PRIMA method will be beneficial.

#### ACKNOWLEDGMENT

This work was supported by NSF Grant No. MCB-9904875.

- A. Abragam, *Principles of Nuclear Magnetism* (Clarendon, Oxford, 1961).
- K. Schmidt-Rohr and H. W. Spiess, *Multidimensional Solid-State NMR and Polymers* (Academic, London, 1994).
- R. G. Griffin, *Nat. Struct. Biol.* **5**, 508 (1998).
- S. J. Opella, C. Ma, and F. M. Marassi, *Methods Enzymol.* **339**, 285 (2001).
- J. W. Emsley, Vol. 4 of *Encyclopedia of Nuclear Magnetic Resonance*, edited by D. M. Grant and R. K. Harris (Wiley, Chichester, UK, 1996).
- A. A. Bothner-By, Vol. 4 of *Encyclopedia of Nuclear Magnetic Resonance*, edited by D. M. Grant and R. K. Harris (Wiley, Chichester, UK, 1996).
- J. H. Prestegard, J. R. Tolman, H. M. Al-Hashimi, and M. Andrec, in *Biological Magnetic Resonance*, Vol. 17: *Structure Computation and Dynamics in Protein NMR* (Plenum New York, 1999), p. 315.
- J. R. Quine and T. A. Cross, *Concepts Magn. Reson.* **12**, 71 (2000).
- A. Bax, G. Kontaxis, and N. Tjandra, *Methods Enzymol.* **339**, 127 (2001).
- E. de Alba and N. Tjandra, *Prog. Nucl. Magn. Reson. Spectrosc.* **40**, 175 (2002).
- H. M. Al-Hashimi and D. J. Patel, *J. Biomol. NMR* **22**, 1 (2002).
- J. R. Quine, M. T. Brenneman, and T. A. Cross, *Biophys. J.* **72**, 2342 (1997).
- J.-C. Hus, D. Marion, and M. Blackledge, *J. Am. Chem. Soc.* **123**, 1541 (2001).
- S. Moltke and S. Grzesiek, *J. Biomol. NMR* **15**, 77 (1999).
- J.-C. Hus, D. Marion, and M. Blackledge, *J. Mol. Biol.* **19**, 927 (2000).
- V. Tsui, L. Zhu, T.-H. Huang, P. E. Wright, and D. A. Case, *J. Biomol. NMR* **16**, 9 (2000).
- G. A. Mueller, W. Y. Choy, D. Yang, J. D. Forman-Kay, R. A. Venters, and L. E. Kay, *J. Mol. Biol.* **300**, 197 (2000).
- J. Meiler, N. Blomberg, M. Nilges, and C. Griesinger, *J. Biomol. NMR* **16**, 245 (2000).
- H. J. Sass, G. Musco, S. J. Stahl, P. T. Wingfield, and S. Grzesiek, *J. Biomol. NMR* **21**, 275 (2001).
- R. N. Zare, *Angular Momentum* (Wiley, New York, 1988).

- <sup>21</sup>J. Meiler, J. J. Prompers, W. Peti, C. Griesinger and R. Brüschweiler, *J. Am. Chem. Soc.* **123**, 6098 (2001).
- <sup>22</sup>M. Edén and M. H. Levitt, *J. Magn. Reson. Ser.* **132**, 220 (1998).
- <sup>23</sup>D. W. Yang, R. A. Venters, G. A. Mueller, W. Y. Choy, and L. E. Kay, *J. Biomol. NMR* **14**, 333 (1999).
- <sup>24</sup>P. Permi and A. Annala, *J. Biomol. NMR* **16**, 221 (2000).
- <sup>25</sup>G. Cornilescu, J. L. Marquardt, M. Ottiger, and A. Bax, *J. Am. Chem. Soc.* **120**, 6836 (1998).
- <sup>26</sup>E. de Alba, L. De Vries, M. G. Farquhar, and N. Tjandra, *J. Mol. Biol.* **291**, 927 (1999).
- <sup>27</sup>P. E. Johnson, M. D. Joshi, P. Tomme, D. G. Kilburn, and L. P. McIntosh, *Biochemistry* **35**, 14381 (1996).
- <sup>28</sup>K. Ösapay, Y. Theriault, P. E. Wright, and D. A. Case, *J. Mol. Biol.* **244**, 183 (1994).
- <sup>29</sup>S. Longhi, M. Czjzek, V. Lamzin, A. Nicolas, and C. Cambillau, *J. Mol. Biol.* **268**, 779 (1997).
- <sup>30</sup>W. Peti, J. Meiler, R. Brüschweiler, and C. Griesinger, *J. Am. Chem. Soc.* (in press).
- <sup>31</sup>G. M. Crippen and T. F. Havel, *Distance Geometry and Molecular Conformation* (Tauntun, Research Studies, United Kingdom, 1988).
- <sup>32</sup>J. J. Prompers and R. Brüschweiler, *Proteins* **46**, 177 (2002).
- <sup>33</sup>M. T. Brenneman and T. A. Cross, *J. Chem. Phys.* **92**, 1483 (1990).