The Exact NOE as an Alternative in Ensemble Structure Determination

Beat Vögeli,1,* Simon Olsson,1,2 Peter Güntert,1,3,4 and Roland Riek1
1Laboratory of Physical Chemistry, Vladimir-Prelog-Weg 2, Swiss Federal Institute of Technology, ETH-Hönggerberg, Zürich, Switzerland; 2Institute for Research in Biomedicine, Bellinzona, Switzerland; 3Institute of Biophysical Chemistry, Center for Biomolecular Magnetic Resonance and Frankfurt Institute for Advanced Studies, J.W. Goethe-Universität, Frankfurt am Main, Germany; and 4Graduate School of Science, Tokyo Metropolitan University, Hachioji, Tokyo, Japan

ABSTRACT The structure-function paradigm is increasingly replaced by the structure-dynamics-function paradigm. All protein activity is steered by the interplay between enthalpy and entropy. Conformational dynamics serves as a proxy of conformational entropy. Therefore, it is essential to study not only the average conformation but also the spatial sampling of a protein on all timescales. To this purpose, we have established a protocol for determining multiple-state ensembles of proteins based on exact nuclear Overhauser effects (eNOEs). We have recently extended our previously reported eNOE data set for the protein GB3 by a very large set of backbone and side-chain residual dipolar couplings and three-bond J couplings. Here, we demonstrate that at least four structural states are required to represent the complete data set by dissecting the contributions to the CYANA target function, which quantifies restraint violations in structure calculation. We present a four-state ensemble of GB3, which largely preserves the characteristics obtained from eNOEs only. Due to the abundance of the input data, the ensemble and its corresponding parameters (3) obtained from relaxation measurements (4) or residual dipolar couplings (RDCs) (5). However, structural representations of the different states that are visited often offer more direct insight into protein activity (6–8). Many approaches making use of NMR data have recently emerged. Most of them rely on RDCs (9–17), paramagnetic relaxation enhancement or pseudocontact shifts (18,19), and conventional NOEs (11,12,20–23), but some rely also on three-bond J couplings (21,24), or chemical shifts (25,26).

We have established a protocol for determination of multiple-state ensembles of proteins based on the exact measurement of the nuclear Overhauser effect (eNOE) (27–29). eNOEs can be converted into exact upper and lower distance limits (30–34). For the protein GB3, we have previously shown that three states are required to back-predict an input set of 884 eNOEs supplemented with 147 backbone J couplings and 90 RDCs (35,36). Recently, we have extended the input data set to a total of 984 eNOEs by inclusion of those involving magnetically equivalent protons other than methyls, a set of 1477 backbone and side-chain RDCs obtained under eight alignment conditions, and a total of 225 J couplings by adding couplings restraining χ^1 dihedral angles (37). This data set constitutes one of the largest and most diverse to date. Importantly, we have shown that the eNOEs are able to improve the back-prediction of RDCs and J couplings, either in comparison to conventional NOEs, or upon use of more than one state in the structure determination (37).

Using this extended and self-consistent data set with an updated protocol for CYANA ensemble calculations, we present here a four-state ensemble of GB3, which largely preserves the previously obtained characteristics and those of ensembles based only on eNOE restraints. Due to the abundance of the input data, it is possible to compare the information content of the complete data set to the one of the eNOEs or conventional NOEs alone. This is particularly true for χ^1 rotamer states, which are also well suited for comparison to x-ray structures. We find that eNOE data alone offer as much information as conventional NOE data in combination with the abundant RDC and J-coupling data. Finally, we use principal component analysis (PCA) to automatically identify and validate relevant states of the ensembles.
MATERIALS AND METHODS

Input data set

A detailed account of the collection, cross-validation, alignment tensor determination, choice of Karplus coefficients, and error estimation of the data from literature and experiments is provided in our previous work (37,38). Here, we give a brief overview.

All 884 upper and 823 lower distance limits derived from eNOEs presented in our previous studies (35,36) were taken and supplemented with 100 pairs of upper and lower limits that involve either methylene groups with chemically equivalent protons or chemically equivalent methyl groups in Val and Leu. All new upper limits and lower limits were given an additional tolerance of 5% in addition to the 95% tolerance for bidirectional/unidirectional eNOEs (39).

The RDCs used for the definition of the backbone geometry were as follows. $D_{HN,N}$ and $D_{C,H_a}$ of wild-type GB3 under alignment induced by Pf1 (tensor 1) were taken from our previous work (40). For $D_{HN,N}$, the errors were uniformly set to 0.5 Hz, and for $D_{C,H_a}$, they were set to 1.5 Hz. $D_{HN,N}$ and $D_{C,C_a}$ under alignment via Pf1 of the deuterated mutants K19AD47K (tensor 2), K19ED40N (tensor 3), K19EK4A-C-His8 (tensor 4), K19EK4A-N-His8 (tensor 5), and K19AT11K (tensor 6) were taken from averages of two measurements in our previous study (41), with overall errors of (0.14, 0.17, 0.16, 0.11, 0.07) Hz for the (54, 52, 54, 54) $D_{HN,N}$ corresponding to tensors (2, 3, 4, 5, 6). If only one value was available, the error was set to twice the overall error. For the (54, 50, 54, 54) $D_{C,C_a}$, the overall errors for the sets corresponding to tensors (2, 3, 4, 5, 6) are (0.041, 0.078, 0.152, 0.073, 0.047) Hz. If only one value was available, the error was doubled. For a seventh alignment condition under Pf1 (tensor 7), mutant K19EK4A, only one data set of $D_{HN,N}$ and $D_{C,H_a}$ is available. The error was set uniformly to twice the largest error of the other five data sets (0.34 Hz and 0.30 Hz). For $H^5\cdot C^5$, one set of RDCs is available for each mutant yielding tensors 2–6 (41). The errors are uniformly set to 0.2 Hz. For each of the protonated mutants K19AD47K (tensor 2), K19ED40N (tensor 3), K19EK4A-C-His8 (tensor 4), K19EK4A-N-His8 (tensor 5), K19AT11K (tensor 6), and K19EK4A (tensor 7), one $D_{C,H_a}$ data set is available (42). The errors are uniformly set to 1.0 Hz.

The following RDCs involve at least one side-chain atom: $D_{C,H_a}$, $D_{C,G,H_2}$, $D_{C,G/H_2}$, and $D_{H_2,G,H_2}$ values obtained from alignment with Pf1 phage (tensor 1) (43). Here, the errors of these values were propagated into individual errors of the couplings of interest. $D_{C,G/H_2}$, $D_{C,G,H_2}$, $D_{C,G,H_2/1+2}$, and $D_{H_2,G/H_2/1+2}$ values were obtained from alignments with Pf1 phage (tensor 1) and polyethylene glycol (tensor 8) (10). The couplings within methyl groups were scaled by –1/3.17 such that they could be used as effective $D_{C,C_2}$ and $D_{C,G/H_2}$ in the structure calculation (44). For Pf1, the errors of $D_{C,H_a}$ and the methyl RDCs were uniformly set to 1.5 and 0.5 Hz, respectively. For polyethylene glycol, uniform errors of 2 and 0.5 Hz were chosen for $D_{C,H_a}$ and $D_{C,G,H_2}$, and methyl RDCs, respectively.

Although an initial estimate for tensor 1 was obtained from singular value decomposition of the $H^5\cdot N$ RDCs on the RDC-defined $\kappa$-ray structure (41,42,45), all measured RDCs in the backbone were used for tensors 2–7, and $D_{C,H_a}$ (obtained in the same experiment as the other RDCs and also used in structure determination) for tensor 8.

Three types of backbone $J$ couplings defining the $\varphi$ angle were used: $J_{HN,NC}$ values are averages from four data sets taken from our previous work (46). Here, we use the residue-specific averages and the standard deviation as input error (overall 0.15 Hz). If both data sets of one measurement type were missing, the error was set to 0.3 Hz. The couplings were corrected for the RDCs between $H^1$ and $H^1$ due to the natural alignment of GB3 in the magnetic field (see accompanying article). $J_{HN,NC}$ values are averages of two data sets taken from our previous work (46), with an overall error of 0.07 Hz for their averaged values. If the value of one data set was missing, 0.1 Hz was used. An additional uniform error of 0.08 Hz accounting for substituent effects was added (47). $J_{HN,NC}$ values are averages of data sets taken from our previous work (46), with an overall error of their averaged values of 0.1 Hz. If the value of one data set was missing, 0.2 Hz was used. The Karplus coefficients (A, B, C) for $J_{HC,H_2}$, $J_{HC,G}$, and $J_{HC,G}$ were set to (8.754, –1.222, 0.111), (3.693, –0.514, 0.043), and (4.516, –1.166, –0.038) Hz, respectively (37). The $J$ couplings $J_{HC,H_2}$ and $J_{HC,G}$, taken from Mietc et al. (43), define the $\chi_1$ angles. A systematic error of 1 Hz and individual random errors obtained from two data sets were propagated into an overall error. Due to inconsistencies with other data (see accompanying article), the couplings of residue 52 are not used here. The substituent-effect-corrected Karplus parameterization (A, B, C) = (7.23, –1.37, 2.40) is used as proposed for Arg, Asx, Gis, His, Leu, Lys, Met, Phe, Pro, Trp, and Tyr in Pérez et al. (48). We determined $J_{C,C_2}$ and $J_{C,G}$ couplings for aromatic residues using the pulse sequences proposed in Hu et al. (49). $J_{C,C_2}$ and $J_{C,G}$ couplings for Val, Ile, and Thr residues were taken from Chou et al. (50). The Karplus coefficients for the aromatic residues were taken from Pérez et al. (48), who proposed that (A, B, C) was (2.31, –0.87, 0.49/1.29, –0.49, 0.34); for the methyl-bearing residues, Chou et al. (50) reported Karplus coefficients of (2.76, –0.67, 0.19/0.21, 0.21, –0.12) for Thr and (3.42, –0.59, 0.17/2.64, 0.26, –0.22) for Val and Ile.

Twenty-six pairs of $\varphi$ and $\psi$ dihedral angle restraints were generated from $C^1$ chemical shifts. The allowed ranges are either $–200^\circ$ to $–80^\circ$ and $40^\circ$ to $220^\circ$ or $–120^\circ$ to $–20^\circ$ and $–100^\circ$ to $0^\circ$ for $\varphi$ and $\psi$, respectively. An exception is the C-terminus, with $–120^\circ$ to $80^\circ$ and $–100^\circ$ to $60^\circ$.

Although not used in the complete data set for the presented ensemble calculation, conventional NOEs were employed in some ensemble calculations for comparison purposes. The corresponding upper distance limits were generated similarly to the methods described in Chi et al. (51).

Ensemble determination protocol

Multistate ensembles were calculated as previously described but using all input data presented above (35,36,52). In short, a total of 984 upper and 923 lower distance limits were used for the multiple-state ensemble calculation in CYANA (53,54). This data set consists of 355 bidirectional eNOEs or averaged over multiple $H^5\cdot H^5$ data sets and 568 unidirectional eNOEs, of which 31 and 69 involve a pseudomethylene or a pseudotautomeric for both methyl groups in Val and Leu. NOEs involving pseudotautomers were processed differently than in previous studies. The CYANA protocol was executed with an individual treatment of each proton of degenerate groups such as methyls and some methylenes and aromatics by $r^2$ averaging (55). Since there is fast rotation present in methyl groups, we added an additional tolerance of ±8.5% to those based on estimations of cross-relaxation rates back-predicted from molecular dynamics simulations. In addition, we used 61 NOEs involving aromatic protons where only an upper distance limit was set (8 Å). Other input parameters included 225 scalar couplings, 1477 RDCs obtained from eight alignment conditions, and 52 angular restraints from $C^1$ chemical shifts during the initial stages of the structure calculation. All fitted alignment tensors were corrected for the rescaling due to uniform motion throughout the molecule by using redetermined tensors after the initial two-state ensemble calculation (the typical tensor increase is 4%).

One hundred conformers were calculated using 50,000 torsion-angle annealing steps, and the 20 conformers with the lowest target function values were then used to represent the calculated structure. Ensembles encompassing one to nine states of the entire protein were calculated simultaneously. Steric repulsion between atoms of different states was excluded, and the NOE distance restraints were applied to the $r^2$ averages of the corresponding distances in the individual states. The $J$ and the RDC restraints were applied to the arithmetic mean of the quantities in the individual states. Bundling restraints were applied to keep the individual structural states together in space as far as permitted by the experimental restraints (56). Weak upper distance bounds of 1.2 Å were applied to all distances between the same nitrogen and carbon atoms in different states. The weight of these bundling restraints was 100 times lower than for NOE upper distance bounds, except for the backbone atoms N, C", C", and C"", for which the
weight used was 10 times lower than that used for the NOEs. The weight of the torsion angle restraint contribution to the target function was increasingly reduced to zero during the calculation, whereas those of the $J$-couplings and RDCs were ramped up from zero (57). The effective bond lengths of $\text{H}^\alpha$-$\text{N}$ and $\text{H}^\alpha$-$\text{C}^\alpha$ were set to 1.02 and 1.09 Å, respectively. Analogous calculations with 1.041 and 1.117 Å resulted in nearly identical target-function values and ensembles.

All ensembles with reduced input data sets were calculated with the same protocol as for the complete data set.

**Target function values**

To calculate residue-specific target-function values, we used the lowest-energy ensemble obtained from the calculations with the complete eNOE, $J$-coupling, and RDC data set. Restraints involving two residues contributed half of the violation to the target-function values of each residue. If at least one of the involved atoms belongs to a side chain, the restraint is categorized as a side-chain restraint. Atoms that are fixed to the backbone are considered backbone atoms (including $\text{C}^\beta$ and $\text{H}^\text{\beta,methyl}$ in alanine, but not $\text{H}^\text{\alpha}$, $\text{H}^\text{\beta}$, and $\text{H}^\text{\gamma}$ in other amino acids).

**PCA**

We performed PCA of the four-state ensembles using all pairwise $\text{C}^\alpha$-$\text{C}^\alpha$ distances as the basis for the analysis. All 20 lowest-energy ensembles were used to perform the analysis. The singular values of the first 10 principal components are shown in Fig. S7 in the Supporting Material. Subsequently, the two first principal components where used as criteria to carry out a state decomposition, which was in turn compared to the manual decomposition.

**RESULTS AND DISCUSSION**

**Number of representative states**

The CYANA target function is a measure of the violation of the experimental restraint data by the calculated structures. Thus, it may be used to estimate the appropriate number of states for representation of the spatial sampling of a molecule. The overall target function averaged over the 20 lowest-energy conformers of $\sim117$ Å$^2$ obtained for one state decreases to approximately one-third of its value for two states (Fig. 1, left). Additional decrease is observed when going to three and four states, where it reaches a plateau value of $\sim30$ Å$^2$. Plots of the contributions to the target function from eNOEs, RDCs, and $J$-couplings (Fig. 1, right) reveal that the RDCs account for the major part of the target-function value and also for the major part of the decrease from one to two states. The decrease from two to three and again to four states is dominated by the eNOEs, whose target-function value continues to drop slightly up to nine states. The $J$-couplings have a similar profile as the eNOEs, but their target function reaches the minimum at four states.

To understand the origins of major contributions to the target function, we calculated the residue-specific target-function values of the complete eNOE, $J$-coupling, and RDC data set (Fig. 2). The largest contributions to the target functions of the single-state structure stem from residues $8$ and $35$ ($>12$ Å$^2$; Fig. 2, top). Those from residues $3$, $11$, $32$, $43$, $46$, and $47$ are also $>3$ Å$^2$. All of them undergo a dramatic decrease upon addition of a second state. As a consequence, these residues also undergo the largest decreases of the relative contributions normalized to the single-state contributions (Fig. 2, bottom). To check whether the large contributions are a simple consequence of a large number of restraints on the residue, the individual contributions were divided by the number of restraints per residue, as shown in Fig. S1. The profile is very similar to the one that is not normalized. The largest values are observed for residues $3$, $8$, $11$, $22$, $35$, and $47$ ($>0.05$ Å$^2$). If the number of restraints has an impact, it is only exerted in an indirect manner, where the requirement to satisfy all restraints simultaneously drives the violation of each restraint up.

To further dissect the residue-specific contributions to the target function, we divided them up into the contributions from the backbone and the side chains (see Figs. 3 and S2). It is striking that the profile of the target-function values from the side chains is similar to that of the total residue-specific target-function values (compare Figs. 2 and 3). Thus, the large violations are mainly caused in the side chains. Indeed, the two most violated residues, 8 and 35, are known to simultaneously populate two $\chi^1$ rotamer states (43). As most side-chain restraints of our data set strongly depend on the $\chi^1$ angle, it is expected that single-state structures cannot fulfill the experimental data well. However, one additional state enables population of both rotameric states and the target-function values undergo a dramatic decrease (see normalized target-function values). It appears to be a consequence of the large violations in the side chains that some of the elevated target function values are also observed.
in the backbone, as the same residues have the largest values (see Fig. S2).

Other segments that have larger-than-usual contributions to the target function are the highly mobile loop of residues 9–12, and also residues 43–47 in β-strand 3, which has been shown to be more mobile than the other strands (42). These residues also experience a large relative decrease of the individual target function values upon addition of further states. This indicates that the violations can be resolved with multiple states.

Although the overall target-function value is not reduced substantially when going from three to four states, the overall eNOE target function drops continuously over up to nine states, and the J-coupling target function clearly reaches the minimum at four states. Some residue-specific target-function values are reduced significantly when going to four, five, or six states. Because we wish to represent local conformational sampling appropriately rather than only globally, we choose four states for the appropriate ensemble representation in the following.

Interestingly, the optimal number of structural states to represent a set of NMR and x-ray observables obtained from GB3 was also shown to be four in a previous study (10). In that study, RDCs, Lipari-Szabo-type order
parameters as determined from relaxation measurements, and crystallographic B-factors were used for ensemble calculation. Cross-validation was done with $^3J_{HN,Ha}$ scalar couplings and backbone RDCs. Substantial improvement was observed when going from one to two states, and some more for four states. Depending on the observables, further slight improvement was achieved with 8 or 16 states. The input data set was smaller than ours, but the information content is difficult to compare to that of our set because of the very different nature of some of the data. In a prior study of GB3 by the same authors, based only on RDCs, the optimal number of states was two (17).

**Structural ensembles**

For the structural representation of GB3, we select the 20 four-state ensembles with the lowest target-function values from 100 calculations (Table 1). The ensembles are very compact, as expected for a highly rigid globular protein (Figs. 4 and 5). The squared $^1$H$_N$-N order parameters are 0.77 on average, and the root mean-square deviations (RMSDs) from the mean coordinates are 0.45 Å and 0.88 Å for the backbone and all heavy atoms, respectively. The RMSDs from an RDC-refined high-resolution x-ray structure are 0.55 Å and 1.09 Å for the backbone and all heavy atoms.
TABLE 1 Structural Statistics for GB3

<table>
<thead>
<tr>
<th>Parameters</th>
<th>eNOE-Derived Four-State Ensemble-Based Calculation</th>
<th>eNOE-Derived Three-State Ensemble-Based Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper distance limits</td>
<td>984</td>
<td>884</td>
</tr>
<tr>
<td>Bidirectional</td>
<td>355</td>
<td>324</td>
</tr>
<tr>
<td>Unidirectional</td>
<td>568</td>
<td>481</td>
</tr>
<tr>
<td>Aromatics</td>
<td>61</td>
<td>61</td>
</tr>
<tr>
<td>Lower distance limits</td>
<td>923</td>
<td>823</td>
</tr>
<tr>
<td>Bidirectional</td>
<td>355</td>
<td>324</td>
</tr>
<tr>
<td>Unidirectional</td>
<td>568</td>
<td>481</td>
</tr>
<tr>
<td>Dihedral angles</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$^3$J couplings</td>
<td>225</td>
<td>147</td>
</tr>
<tr>
<td>Backbone</td>
<td>147</td>
<td>147</td>
</tr>
<tr>
<td>Backbone-side-chain</td>
<td>78</td>
<td>–</td>
</tr>
<tr>
<td>RDCs</td>
<td>1477</td>
<td>90</td>
</tr>
<tr>
<td>Backbone</td>
<td>1335</td>
<td>90</td>
</tr>
<tr>
<td>Backbone-side-chain</td>
<td>13</td>
<td>–</td>
</tr>
<tr>
<td>Side chain</td>
<td>129</td>
<td>–</td>
</tr>
<tr>
<td>Average target-function value ($\hat{\text{A}}^2$)</td>
<td>25.03</td>
<td>9.94</td>
</tr>
<tr>
<td>Violations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Distance restraints ($&gt;0.5$ Å)</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Torsion angle restraints ($&gt;5^\circ$)</td>
<td>$^b$ 1</td>
<td></td>
</tr>
<tr>
<td>$^3$J coupling restraints ($&gt;1$ Hz)</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>RDC restraints ($&gt;4$ Hz)</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>RDC restraints ($&gt;10$% of range)</td>
<td>31 (5; 1; 19; 5; 1) $^b$ 3</td>
<td></td>
</tr>
<tr>
<td>Average RMSD from mean coordinates (Å)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backbone atoms</td>
<td>$0.45 \pm 0.04$</td>
<td>$0.47 \pm 0.05$</td>
</tr>
<tr>
<td>Heavy atoms</td>
<td>$0.88 \pm 0.04$</td>
<td>$0.86 \pm 0.04$</td>
</tr>
<tr>
<td>RMSD from RDC-refined x-ray structure $^d$ (Å)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Backbone atoms</td>
<td>0.55</td>
<td>0.72</td>
</tr>
<tr>
<td>Heavy atoms</td>
<td>1.09</td>
<td>1.31</td>
</tr>
</tbody>
</table>

$^a$Three-state ensemble-based calculations were taken from Yao et al. (41,42) and Derrick and Wigley (45).

$^b$Torsion-angle restraints were turned off during the final stage of the calculation.

$^c$H$	ext{H}^\text{H}$-$\text{N}$, 3 Hz; H$	ext{H}^\text{H}$-$\text{C}^\text{C}$, H$	ext{H}^\text{H}$(2,3)-$\text{C}^\text{C}$, and H$	ext{H}^\text{H}$(1,2)$-$\text{C}^\text{C}$, 6 Hz; C$	ext{C}^\text{C}$-$\text{C}^\text{C}$ and H$	ext{H}^\text{H}$-$\text{C}^\text{C}$, 0.7 Hz; H$	ext{H}^\text{H}$-$\text{C}$, 1.2 Hz; C$	ext{C}^\text{C}$-$\text{C}^\text{C}$methyl, 1 Hz.

$^d$RMSD values for eNOE calculations were taken from our previous work (35).

Heavy atoms, respectively (41,42,45). For comparison, the previously published three-state structure obtained from a slightly smaller number of eNOEs and small RDC and $^3$J-coupling sets had RMSDs from the mean of 0.47 Å and 0.86 Å for the backbone and all heavy atoms, respectively, and RMSDs from the RDC-refined x-ray structure of 0.72 Å and 1.31 Å (35,36). This indicates that the extended input data set does not alter the amplitude of the spatial sampling. Nevertheless, the four-state ensembles have a smaller deviation from the reference structure.

In many structural segments, it is possible to manually identify the identical four states in all 20 ensembles. Apparently, some properties of the four obtained states are well defined by the input data in combination with the geometric constraints. In Figs. 4 and 5, the states were colored accord-

ing to the segments comprising residues 9–12 and 45–46, respectively. Patterns of strong correlations in the immediate surroundings of these segments are present, but there are also long-range correlations within the β-sheet. To a lesser degree, the α-helix is also correlated with segment 45–46.

Information content of data sets

The very high density of eNOE, RDC, and $^3$J-coupling restraints is an excellent prerequisite to study the information content of these data. We wish to investigate the following. First, how much additional information is provided by RDCs and $^3$J couplings once the eNOE data set is collected? Second, how many additional data are provided by eNOEs when they replace a set of conventional NOEs? The first question is highly relevant for practical considerations, because it is a substantially smaller effort to collect an eNOE data set than the RDCs under multiple alignment conditions and $^3$J couplings. In practice, the preparation of eight alignment conditions is prohibitive for most systems, if it is possible at all (58). The second question is interesting in so much as the collection of an eNOE data set requires around four to five times more spectrometer time than conventional NOEs.

Backbone dihedral angles

First, we investigated the impact on the backbone dihedral angles, φ and ψ, of supplementing the eNOE data with RDCs and $^3$J couplings. To that end, we plotted circle diagrams of φ and ψ sampled in four-state ensembles calculated from eNOEs only or from eNOEs, RDCs, and $^3$J couplings (see Fig. 6). The mean angles and the sampling are very similar in any kind of ensemble once eNOEs are employed (averaged over all residues, there is an 8.6° and 9.9° difference for the mean φ and ψ angles, respectively, whereas the standard deviation changes by −0.9° and +0.3° upon inclusion of the RDCs and $^3$J couplings, resulting in 16.1° and 17.3°). These relatively small changes are also reflected in minor changes of the backbone heavy-atom Cartesian RMSDs (from 0.47 Å to 0.45 Å). There are, however, some exceptions. Angles with increases of the standard deviation by >5° are ψ7, ψ9, ψ10, ψ11, ψ13, ψ15, ψ16, ψ24, ψ25, ψ26, ψ38, ψ39, ψ40, ψ41, ψ46, ψ47, and ψ48, where the additional data apparently enforce a more distinct separation of states. None of the changes exceeds 20° except in the case of ψ34 and ψ39 in the highly mobile loop comprised of residues 38–41, with +56.6° and +61.1°. Many of the more flexible angles are those of the four glycines at positions 9, 14, 38, and 41, which experience less steric hindrance due to the lack of a side chain. In both ensembles, the angles have generally a smaller standard deviation in the α-helix than in other secondary elements.

The angles of both ensembles also coincide very closely with those extracted from a high-resolution x-ray structure (45). Almost all x-ray-derived angles fall within the ranges
sampled by the ensemble, the only exceptions being $\varphi_2$, $\psi_14$, $\psi_38$, $\varphi_39$, $\psi_40$, and $\varphi_41$, all of which are located in the hinge regions of the two flexible loops or the N-terminus. Note, however, that the differences are small in all cases ($<30^\circ$).

To see if a good definition of the ensemble is a general property of NOEs, we generated analogous plots for ensembles obtained from conventional NOEs with or without RDCs and $J$ couplings (see Fig. S3). Here, the changes are significantly more pronounced. Although the means change

FIGURE 4 Four-state ensembles of GB3, showing the heavy-atom representations (top) and the backbone heavy-atom representations (bottom). The 20 ensembles with the lowest target-function values are selected and the four states colored gold, red, green, and blue are obtained by grouping the loop comprising residues 9–12. To see this figure in color, go online.
only slightly more (15.8° and 17.1° for the $\phi$ and $\psi$ angles, respectively), the standard deviations decrease by $-11.2°$ and $-10.3°$ upon inclusion of the RDCs and $J$ couplings, resulting in $19.2°$ and $19.8°$. First, this indicates that definition of the ensemble backbone by conventional NOEs is not as tight as that by eNOEs. This is also reflected in the overall backbone heavy-atom RMSD from the average of 0.77 Å of the ensembles obtained only from conventional NOEs. Furthermore, it demonstrates that even using the RDCs and $J$ couplings, the backbone trace is not equally compact when conventional NOEs are used instead of eNOEs. Interestingly, this difference is not seen in the overall

FIGURE 5 Four-state ensembles of GB3, showing heavy-atom representations (top) and backbone heavy-atom representations (bottom). The 20 ensembles with the lowest target-function values are selected and the four states colored gold, red, green, and blue are obtained by grouping residues 45 and 46. To see this figure in color, go online.
heavy-atom RMSD that decreases to 0.43 Å, which is similar to what is seen for eNOE-based ensembles.

Side-chain $\chi^1$ angles

Next, we investigate the impact of supplementing the eNOE data with RDCs and $J$ couplings on the side chains. $\chi^1$ dihedral angles are the ideal probes, because our data set contains a high density of RDCs and $J$-coupling restraints on $\chi^1$. Although a single-state bundle calculation with eNOE restraints produces considerably smaller standard deviations from the mean values of $\chi^1$ than one with conventional NOEs (averaged over all residues, the standard deviations are 19.6$^\circ$ versus 46.8$^\circ$), the obtained rotamer states are identical in nearly all cases (see a circle diagram
In the following, we investigate whether differences in the $\chi^1$ means and standard deviations are obtained for the four-state ensembles calculated with different input data sets, where a higher information density is required.

Fig. 7 shows a comparison of the $\chi^1$ torsion angle sampling in the four-state ensembles calculated either from eNOEs only or from eNOEs, RDCs, and $J$ couplings. On average, the complete data set changes the mean angles by 18.8° and the standard deviations by +7.7° for individual residues. These relatively small changes are similar to those observed for the backbone dihedral angles and are again reflected by minor changes of the overall heavy-atom RMSDs (from 0.83 Å to 0.88 Å). This observation indicates the high information content of the eNOE data set. However, some additional states appear, whereas others disappear, when RDCs and $J$ couplings are added to the input data set. This is observed for about one-third of the $\chi^1$ angles (residues 4, 6, 12, 15, 17, 18, 25, 28, 42, 47, 51, and 54; Fig. 7, light blue circles). Angles extracted from high-resolution x-ray structures are also indicated (45). The few angles in the single-state structures from the conventional NOEs as well as the eNOEs that were not in agreement with the x-ray structure (see Fig. S4) have a significant population at the x-ray angles now. For residues 7 and 21, the original isotropic x-ray angle is in agreement with the ensemble value, but not the alternative state from anisotropic evaluation. For residues 11 and 24, the additional state is sampled, but not the isotropic one. For residue 35, both the original isotropic and the different states are sampled. Overall, there is a very good agreement between the NMR and x-ray data. Those non-alanine/glycine residues with a very low percentage of solvent-accessible surface area (<10% as calculated for the GB3 (PDB: 1IGD) (45) from MolMol with a solvent molecule of radius 1.4 Å (59)), residues 3, 5, 7, 30, 39, 52, and 54, have a single rotamer state with a very small standard deviation. For those with a higher percentage, no clear pattern emerges. For example, among the three residues with >50% solvent-accessible surface area, residue 40 has a narrow distribution, whereas residues 10 and 11 sample two states.

We also compared the four-state ensemble calculated from eNOEs alone to the previously determined three-state ensemble based on a slightly smaller eNOE set but with 153 $J$ couplings from the backbone and one $\mathrm{H}^\alpha$-$\mathrm{N}$ and one $\mathrm{H}^\alpha$-$\mathrm{C}^\alpha$ RDC set (PDB: 2LUM) (35,36). This ensemble yields results very similar to those of the four-state ensemble (see Fig. S5). Notable differences are only observed for residues 15 (minor population shifts from gauche+ (g+) to trans (tr)), 37 (slightly reduced minor state...
population), 42 (addition of tr state), 45 (less well defined), 50 (tr shifts to g+ rotamer), and 54 (addition of tr state). Interestingly, all these changes bring the ensemble closer to the four-state ensemble obtained from the complete data set.

Fig. 8 shows a comparison analogous to the one in Fig. 7, where the eNOE set is replaced by the conventional NOE set. On average, the complete data set changes the mean angles by 33.8° and the standard deviations by +7.7° for individual residues. Again, some additional states appear, whereas others disappear. This is observed for approximately one-half of the angles, which is more than the one-third in the case of the eNOEs. These considerably larger changes compared to those for the eNOEs are also reflected in a relatively large change of overall heavy-atom Cartesian RMSDs (from 1.11 Å to 0.88 Å). Obviously, the RDC and J couplings contain more information that is not redundant with the NOEs than with the eNOEs.

The standard deviation of the eNOE-based angles is 3.5° smaller on average than those based on conventional NOEs. However, the residue-averaged change of the mean angles by 33.5° indicates that some states are different (compare Figs. 7, right, and 8, right). These may be grouped into the following categories. When going from conventional NOEs to eNOEs, we observe a new major state (residue 22, tr, and gauche− (g−)), loss of one or two minor states (residues 8, tr; 15, tr; 17, tr; 25, tr; 44, tr and g−; 47, g−), addition of a minor state (16, 0°; 21, tr), change of a minor state (10, g+ to tr; 19, tr to g+; 36, tr to g−; 42, tr to g+; 55, g− to g+), stronger distinction of two present states (12; 13; 31; 37), and narrower distributions (32; 40; 46; 53).

**Structural correlations**

**Backbone and χ^1 dihedral angles**

Multiple-state representations of molecular structures not only map motional amplitudes but can also reveal correlations between entities as they sample the conformational space. Here, we analyze the correlations in four-state ensembles for the ϕ, ψ, and χ^1 angles by bar plots of the states (see Figs. 9 and S6). The same color code as in Figs. 4 and 5 is used. Since manual grouping was carried out in the loop connecting β-strands 1 and 2, comprising residues 9–12 in the first case (Figs. 4 and 9), there is a high correlation between the identical states in the different ensembles in close proximity to this loop. This is particularly true for the ψ angles but is also propagated into the ϕ and χ^1 angles, and to a lesser degree into the peripheral side-chain angles χ^2, χ^3, etc. For ϕ, ψ, and χ^1 (and even χ^2), the correlation is somewhat conserved in the first two β-strands. Similar effects are observed when grouping residues 45 and 46.
and 4. When grouping the segment of residues 45 and 46, the correlations are strongest in the adjacent strands 3 and 4, but are present in the entire \( \beta \) sheet. To a lesser extent, the correlations also propagate into the \( \alpha \)-helix.

**Principal component analysis**

A more quantitative way to analyze the states in multiple-state ensembles may be by principal component analysis (PCA) \((35,60)\). We applied PCA to the 20 four-state ensembles (see Materials and Methods). Fig. S7 shows the 10 largest principal components. When the 20 four-state ensembles are projected into the first two principal components, they divide into groups with the 20 lowest (most negative), second lowest, second largest, and 20 largest (most positive) values for both principal components, and the four states per ensemble end up in the four different groups in most cases (Fig. S8). This indicates that the choice of four states in the ensemble calculation is not producing redundant structural representations. This double grouping along the first two principal components captures the two main spatial sampling features that are not correlated with each other. Interestingly, the principal modes do not coincide with the highly localized manual grouping of either the loop comprising residues 9–12 or the segment with residues 45 and 46. Instead, the modes reveal more subtle features distributed over the entire protein.

**CONCLUSIONS**

We have analyzed the information content of an extended eNOE data set of GB3 with respect to large sets of RDCs obtained under eight alignment conditions and \( J \) couplings defining backbone \( \phi \) and side-chain \( \chi^1 \) angles. To this purpose, we modified our previously established protocol for multiple-state ensemble determination to determine four-state ensembles of GB3. We justified the choice of four states by dissecting the contributions to the CYANA target function and by PCA. By omitting subsets of the input data in the structure calculation, we show that 1) eNOE distance restraints contain more information than those derived from conventional NOEs, and 2) eNOEs alone, as opposed to conventional NOEs, provide state representations similar to those of the complete input data set including the RDCs and \( J \) couplings.

These insights are of immediate significance, because the experimental effort to obtain eNOEs is substantially smaller than that required to collect RDCs under multiple alignment conditions and \( J \) couplings. In addition, we conclude that it is worth replacing conventional NOEs with eNOEs in structural analysis. The additional effort is minimal. The NOESY assignment has to be carried out only once in both cases. The more sophisticated evaluation procedure of eNOEs is largely automated in our latest software versions (to be published elsewhere). We believe that the four to five times longer measurement time is easily compensated for by the

(see Figs. 5 and S6), where the highest correlation is present in the neighborhood of the corresponding residues. It is worth noting that one strength of the eNOE-based multistate ensemble is their sensitivity to translational correlations, in some cases of entire segments, rather than only angular correlations. Such correlations may be most easily visualized in the state representations of Figs. 4 and 5. In all ensembles, a strong correlation with the four states in the loop comprising residues 9–12 is present and propagates into \( \beta \)-strands 1, 2,
gain in information and the saved experimental effort (RDCs, $J$ couplings).

Our method for multistate ensemble determination can be easily applied to other systems with masses up to 20 kDa. For example, we recently determined two states of cyclophilin A, an enzyme of 18 kDa (61). In one state, the binding loop is closed, whereas it is open in the other state while the active site is preorganized for ligand binding. As the NOE is sensitive to motions on timescales as slow as milliseconds, we expect the method to reveal further biologically relevant mechanisms when applied to other proteins or nucleic acids.

Finally, the compiled data set consisting of eNOEs, RDCs, and $J$ coupling of GB3 is one of the largest, most accurate, and most diverse for any protein. Therefore, we advocate this set to be an ideal test case for methods development in molecular dynamics simulations and structure calculations.

**SUPPORTING MATERIAL**

Eight figures are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(15)01214-X.

**AUTHOR CONTRIBUTIONS**

B.V., S.O., P.G., and R.R. designed research; B.V., S.O., and P.G. performed research, analyzed data, and wrote the article.

**ACKNOWLEDGMENTS**

B.V. was supported by the Swiss National Science Foundation through grant 140214 and by an Eidgenössische Technische Hochschule Zürich (ETH) research grant (ETH-04 13-1). P.G. gratefully acknowledges financial support by the Lichtenberg program of the Volkswagen Foundation. S.O. is funded by an independent postdoctoral grant from The Danish Council for Independent Research for Natural Sciences (DFF–4002–00151).

**REFERENCES**


