J|A|C|S
ARTICLES
Published on Web 09/03/2003

Quantitative Evaluation of Experimental NMR Restraints

Sander B. Nabuurs,[†] Chris A. E. M. Spronk,[†] Elmar Krieger,[†] Hans Maassen,[‡] Gert Vriend,[†] and Geerten W. Vuister*[,§]

Contribution from the Center for Molecular and Biomolecular Informatics, University of Nijmegen, The Netherlands, Department of Mathematics, University of Nijmegen, The Netherlands, and Department of Biophysical Chemistry, NSRIM Center, University of Nijmegen, The Netherlands

Received April 3, 2003; E-mail: vuister@nmr.kun.nl

Abstract: Nuclear Overhauser effect (NOE) data are an indispensable source of structural information in biomolecular structure determination by NMR spectroscopy. The number and type of experimental restraints used in the structure calculation and the RMS deviation of the restraints are usually reported. We present a new method for quantifying the information contained in the experimental NMR restraints. The method is based on a description of the structure in distance space and concepts derived from information theory. It allows for an objective description of the amount of available experimental information, which we show to be related to the positional uncertainty of the NMR ensemble. The measure of information presented is not affected by redundancy in the experimental restraints. Using various examples, we show that the method successfully identifies the crucial restraints in a structure determination: those restraints that are both important and unique. Finally, we demonstrate that the method can detect a wider range of redundancy in experimental datasets when compared to currently available methods. Because our method describes the *qu*antitative *e*valuation of *e*xperimental *N*MR restraints, we propose the acronym QUEEN.

Introduction

The common method for biomolecular structure determination by NMR spectroscopy relies on the identification of a dense network of interproton distance restraints.[1] These distances can be obtained from nuclear Overhauser enhancements (NOE), which give rise to cross-peaks in NOE experiments. Since the first protein structures were solved by NMR,[2,3] other experimental information derived from *J*-couplings,[4−6] chemical shifts,[7,8] and residual dipolar couplings[9] has also been used to further improve the quality of NMR structures. Despite these new types of experimental data, distance restraints have remained the single most valuable source of information for the elucidation of high-resolution solution structures by NMR spectroscopy,[10] and it is only recently that the first backbone structure determination from residual dipolar couplings without the use of NOE data was reported.[11] Although this latest development holds great promise, most high-resolution NMR structures are still determined using distance and dihedral restraints as a predominant source of structural information.

Several techniques are available to determine the conformational space available to a molecule within the limits of the experimental data. Distance geometry[12] was the first method to be used in de novo protein structure calculations. Nowadays, simulated annealing, either in Cartesian[13] or torsion angle[14] space, is the most widely used method for calculating NMR structures. The quality of the resulting structures and the aforementioned experimental input data has been investigated,[15,16] and several computer programs are available to compare structural characteristics of the models with structural knowledge derived from biomolecular databases.[17,18] Other programs measure the agreement of the derived structures with the experimental input data.[19]

In publications reporting biomolecular structures, it is common practice to provide an overview of parameters pertaining to the experimental data and the NMR ensemble. Among these

[†] Center for Molecular and Biomolecular Informatics.
[‡] Department of Mathematics.
[§] Department of Biophysical Chemistry, NSRIM Center.

(1) Wüthrich, K. *NMR of Proteins and Nucleic Acids*; Wiley: New York, 1986.
(2) Williamson, M. P.; Havel, T. F.; Wüthrich, K. *J. Mol. Biol.* **1985**, *182*, 295−315.
(3) Kaptein, R.; Zuiderweg, E. R.; Scheek, R. M.; Boelens, R.; van Gunsteren, W. F. *J. Mol. Biol.* **1985**, *182*, 179−182.
(4) Pardi, A.; Billeter, M.; Wüthrich, K. *J. Mol. Biol.* **1984**, *180*, 741−751.
(5) Kim, Y.; Prestegard, J. H. *Proteins: Struct., Funct., Genet.* **1990**, *8*, 377−385.
(6) Torda, A. E.; Brunne, R. M.; Huber, T.; Kessler, H.; van Gunsteren, W. F. *J. Biomol. NMR* **1993**, *3*, 55−66.
(7) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson., Ser. B* **1995**, *107*, 293−297.
(8) Kuszewski, J.; Qin, J.; Gronenborn, A. M.; Clore, G. M. *J. Magn. Reson., Ser. B* **1995**, *106*, 92−96.
(9) Tjandra, N.; Bax, A. *Science* **1997**, *278*, 1111−1114.
(10) Clore, G. M.; Robien, M. A.; Gronenborn, A. M. *J. Mol. Biol.* **1993**, *231*, 82−102.
(11) Hus, J. C.; Marion, D.; Blackledge, M. *J. Am. Chem. Soc.* **2001**, *123*, 1541−1542.
(12) Havel, T. F. *Prog. Biophys. Mol. Biol.* **1991**, *56*, 43−78.
(13) Nilges, M.; Clore, G. M.; Gronenborn, A. M. *FEBS Lett.* **1988**, *239*, 129−136.
(14) Güntert, P.; Braun, W.; Wüthrich, K. *J. Mol. Biol.* **1991**, *217*, 517−530.
(15) Doreleijers, J. F.; Rullmann, J. A.; Kaptein, R. *J. Mol. Biol.* **1998**, *281*, 149−164.
(16) Doreleijers, J. F.; Vriend, G.; Raves, M. L.; Kaptein, R. *Proteins: Struct., Funct., Genet.* **1999**, *37*, 404−416.
(17) Vriend, G. *J. Mol. Graphics* **1990**, *8*, 52−56.
(18) Laskowski, R. A.; MacArthur, M. W.; Moss, D. S.; Thornton, J. M. *J. Appl. Crystallogr.* **1993**, *26*, 283−291.

10.1021/ja035440f CCC: $25.00

are the number and types of input restraints used for calculation of the structures,[20] which give an indication of the amount of experimental input data used in the structure determination. There are, however, some caveats when using the number of experimental restraints as an indicator for the quantity of experimental knowledge. For example, experimentally determined restraints between vicinal protons are often redundant given the distance limits imposed by the covalent geometry.[21] In a system in which a large number of sequential, medium-range, and long-range NOE restraints have been identified, a certain degree of redundancy between these restraints is to be expected, rendering the number of restraints only moderately useful as an indicator of the quantity of available information. These problems were partly addressed by the introduction of the NOE completeness, which is defined as the ratio of the number of experimentally observed NOEs to the number of expected NOEs.[21] The number of restraints by itself becomes more informative if redundant restraints are removed from the experimental dataset.[15] Unfortunately, this method is only able to remove intraresidual restraints and is therefore insensitive to redundancy in the structurally more informative sequential, medium-range, and long-range restraint categories.

Here, we introduce a novel method to quantify the information contained in experimental NMR distance and *J*-coupling data by evaluating their effect on the structure represented in distance space. Our method allows for the quantification of the information content of individual or groups of restraints with respect to the remainder of the dataset. We show that redundant restraints do not add to the overall information content; therefore, the method can be used to detect redundancy in any type of restraint. By evaluating the information contained in all individual restraints, we identify those that are less well supported by the remainder of the dataset and warrant further investigation.

Theory

Distance Space. An object comprised of N atoms can be described in Cartesian space by $3N$ coordinates. In distance space this object, and its mirror image, are described by a set of $N(N-1)/2$ interatomic distances. The positional uncertainty of the individual atoms translates to allowed distance ranges in distance space. Our method starts by realizing that the complete absence of knowledge about a system can be described in distance space by placing a lower bound of zero and an upper bound of infinity on all interatomic distances. If we were to have some information on an atom pair in this system, this would restrain both the position of the two atoms relative to each other in Cartesian space and the upper and lower bounds on their distance in distance space.

Because most of the experimentally derived data used in NMR structure calculations can be translated into distances between pairs of atoms, it is possible to generate a distance matrix reflecting all available distance information. A molecular structure with N atoms and $N(N-1)/2$ interatomic distances is most conveniently represented using a $N \times N$ matrix. In our convention, the elements (i,j) of this matrix with $i < j$ contain the upper bound limit, u_{ij}, for the distance D_{ij} between atoms i and j and the elements (j,i) the lower bound, l_{ij}, for this distance. All diagonal elements are zero. Initially, all elements l_{ij} are set to zero and all elements u_{ij} are set to a very large number, indicating the lack of information about any distance in the system.

In addition to the restraints originating from knowledge of the covalent geometry, experimentally derived restraints are used to describe the structure. A typical set of NMR distance restraints is neither exact nor complete. Only a small subset of the interatomic distances is restrained by experimentally determined upper and lower bounds. A *bound smoothing* algorithm[22,23] can be applied to calculate the minimum and maximum bounds on all interatomic distances given the covalent and NMR determined distances. The bound smoothing procedure applies the two triangle inequalities on all possible groups of three atoms (i,j,k). The first states that the distance between atoms i and k can be no greater than the sum of the maximum values of the distances D_{ij} and D_{jk}:

$$u_{ik} \leq u_{ij} + u_{jk} \tag{1}$$

The second states that the minimum value of the distance D_{ik} can be no less than the difference between the lower bound on D_{ij} and the upper bound on D_{jk}:

$$l_{ik} \geq l_{ij} - u_{jk} \tag{2}$$

Application of these triangle inequalities allows for the propagation of the distance limits on a limited number of atom pairs to the upper and lower bounds of all other atoms in the system.

Uncertainty and Information. The information added to a system can be defined as the difference between the uncertainty (H) of two distinct states of the system. Hence, the amount of information, I_r, that a restraint provides about a system (a biomolecular structure in our case) corresponds to the uncertainty of that system minus the uncertainty after addition of this restraint.

According to Shannon's information theory[24] the uncertainty of a probability distribution with a density function $p(x)$ is:

$$H = -\int_{-\infty}^{\infty} p(x) \log p(x)\, \mathrm{d}x \tag{3}$$

We define an analogous measure for the uncertainty of a molecular structure with N atoms. We assume that a distance D_{ij} between atoms i and j is always within the experimentally determined upper bound u_{ij} and lower bound l_{ij} for that distance. Between these bounds we make no assumptions on the magnitude of distance D_{ij}, resulting in a uniform probability distribution $p(D_{ij})$. It must hold that

$$\int_{l_{ij}}^{u_{ij}} p(D_{ij})\, \mathrm{d}D_{ij} = 1 \tag{4}$$

so the uncertainty (H_{ij}) of the distance between an atom pair (i,j) is

(19) Laskowski, R. A.; Rullmann, J. A.; MacArthur, M. W.; Kaptein, R.; Thornton, J. M. *J. Biomol. NMR* **1996**, *8*, 477−486.
(20) Markley, J. L.; Bax, A.; Arata, Y.; Hilbers, C. W.; Kaptein, R.; Sykes, B. D.; Wright, P. E.; Wüthrich, K. *J. Mol. Biol.* **1998**, *280*, 933−952.
(21) Doreleijers, J. F.; Raves, M. L.; Rullmann, T.; Kaptein, R. *J. Biomol. NMR* **1999**, *14*, 123−132.
(22) Crippen, G. M. *J. Comput. Phys.* **1977**, *26*, 449−452.
(23) Havel, T. F.; Kuntz, I. D.; Crippen, G. M. *Bull. Math. Biol.* **1983**, *45*, 665−720.
(24) Shannon, C. E.; Weaver, W. *The mathematical theory of communication*; University of Illinois Press: Champaign, IL, 1949.

$$H_{ij} = -\int_{l_{ij}}^{u_{ij}} \left(\frac{1}{u_{ij} - l_{ij}}\right) \log\left(\frac{1}{u_{ij} - l_{ij}}\right) \mathrm{d}D_{ij} = \log(u_{ij} - l_{ij}) \quad (5)$$

with all distance bounds given in Å units. We then define a measure of uncertainty (H_n) of a single atom n as the average uncertainty of the distance between atom n and all other atoms in the system:

$$H_n = \frac{1}{(N-1)} \sum_{i \neq n}^{N} H_{in} \quad (6)$$

with H_{in} equal to the uncertainty of the distance between the atom pair (i,n). The uncertainty of the complete system (H_{system}) under investigation can then be calculated by averaging the uncertainties of all interatomic distances in the system. This is the average uncertainty of all N individual atoms in the structure:

$$H_{\mathrm{structure}} = \frac{1}{N} \sum_{i=1}^{N} H_n \quad (7)$$

With this definition of uncertainty, it is possible to define a measure for the structural information, I_{total}, contained in a set of R experimental restraints:

$$I_{\mathrm{total}} = H_{\mathrm{structure}|0} - H_{\mathrm{structure}|R} \quad (8)$$

with $H_{\mathrm{structure}|0}$ equal to the uncertainty of the structure with no experimental restraints and $H_{\mathrm{structure}|R}$ equal to the uncertainty of the structure with R experimental restraints.

Similarly, the information of a single experimental restraint can be defined. The information content (I_r) of an experimental restraint r added to a structure is defined as

$$I_r = H_{\mathrm{structure}} - H_{\mathrm{structure}|r} \quad (9)$$

with $H_{\mathrm{structure}|r}$ equal to the uncertainty of the structure given restraint r and $H_{\mathrm{structure}}$ equal to the uncertainty of the structure before addition of the restraint.

When analyzing a measure of restraint information, the amount of information contained in an experimental restraint is always context dependent. For example, the addition of a restraint X that limits the distance between atoms i and j adds little or no information if this distance has already been well defined by another restraint Y. Conversely, restraint Y becomes uninformative if restraint X has already defined the distance between atoms i and j. Hence, the information measures defined here are always considered against a certain level of background information, either induced by other restraints or derived solely from our knowledge of the covalent bond lengths and angles of the system under investigation.

Distance Space and Uncertainty. The $N(N-1)/2$ interatomic distances do not represent independent degrees of freedom, and therefore the triangle inequalities alone are not sufficient to guarantee that the resulting distance matrix is embeddable in three-dimensional space. Higher-order inequalities would have to be considered to ensure this, but these are impractical because of their computational demands. As a result, the difference between the upper and lower bounds will be overestimated by the bound smoothing algorithm. However, it was previously shown that a linear relation exists between the bounds generated by bound smoothing and the actual distance,[25] even with only a few restraints per residue present in the dataset. In addition, in practice most lower bounds have near zero values. Hence, because of the logarithmic nature of H_{ij} (cf., eq 5), the overestimation of the bounds will result in a constant offset to $H_{\mathrm{structure}}$, which will cancel in the proposed information measure (cf., eq 9).

Experimental Restraints. When describing distance restraints in NMR datasets, it is important to discriminate between covalent constraints, which describe knowledge about the molecular topology, and experimental restraints, which can consist of any type of restraint that allows for a representation in distance space. A typical NMR dataset consists of interatomic distance restraints derived from NOEs, angular restraints derived from J-couplings, and possibly additional distance restraints defining hydrogen bonds. Angular restraints can be converted into distance restraints to allow for their representation in distance space. In this case, bond angles are described by their geminal distance, whereas torsion angle restraints are defined by the distance limits between the first and the fourth atom defining the dihedral angle.[23]

A bound-smoothed distance matrix containing only the covalent constraints is taken as the initial state, $H_{\mathrm{structure}|0}$ (cf., eq 8), prior to addition of any of the experimental restraints. After addition of an experimental restraint, the upper and lower bounds on the distance limits between all atoms are adjusted using the triangle inequalities to yield a consistent set of interatomic distance ranges. The uncertainty of the system can be calculated after each addition of a restraint using eq 7, indicating the information contained in that particular restraint with respect to the set of restraints already added to the distance matrix. Consequently, this information depends on the order in which the restraints are incorporated in the distance matrix as described above. Therefore, we define the unique information (I_{uni}) of a restraint, or alternatively a set of restraints, as the information it adds, given the knowledge of all other restraints ($R - 1$) in the dataset:

$$I_{\mathrm{uni},r} = H_{\mathrm{structure}|R-1} - H_{\mathrm{structure}|R} \quad (10)$$

The overall importance of a restraint is assessed by calculating its average information content sampled throughout the complete dataset (I_{ave}). The average restraint information can be calculated for a set of R restraints by averaging the information content of this restraint in every possible permutation of the restraint list:

$$I_{\mathrm{ave},r} = \langle H_{\mathrm{structure}} - H_{\mathrm{structure}|r} \rangle \quad (11)$$

However, exact calculation of $I_{\mathrm{ave},r}$ would require $R!$ determinations of the information content, which is not computationally feasible for a typical distance restraint set ($>10^3$ restraints). We have therefore chosen in these cases to approximate the average information by calculating the information content of a restraint with respect to randomly selected and sized datasets until the value for its average information converges with a standard deviation below one percent.

Inconsistent Restraints. In experimental datasets inconsistent restraints can occur, originating from several potential sources. Restraint datasets are the result of the process of collecting,

(25) Oshiro, C. M.; Thomason, J.; Kuntz, I. D. *Biopolymers* **1991**, *31*, 1049−1064.

processing, and interpreting NMR data, in which each step is susceptible to both random and systematic errors that occasionally result in the occurrence of inconsistent restraints in the dataset. Mutually inconsistent restraints can also result from conformational averaging of the NOE. These contradicting restraints can cause the distance matrix as a whole to become inconsistent. Even though the distance limits determined by bound smoothing are not as strict as those corresponding to a true three-dimensional object, the procedure is able to detect discrepancies in the experimental input data. Often these are errors which are easily overlooked in the commonly used structure calculation procedures, where the resulting structures represent a compromise between violations of both the accurate and inaccurate input data. Occurrences of mutually inconsistent restraints and other discrepancies in the data leading to divergence are reported by the *qu*antitative *e*valuation of *e*xperimental *N*MR restraints (QUEEN) procedure so that appropriate corrective action can be taken.

Materials and Methods

Datasets. We have tested the method on the following four experimental NMR datasets obtained from the Protein Data Bank:[26] the immunoglobulin G-binding domain of streptococcal protein G (IgG)[27] (56 residues, PDB entries 1GB1 and 2GB1); ubiquitin (UBI)[28] (76 residues, PDB entry 1D3Z); the second PDZ domain of PTP-BL (PDZII)[29] (93 residues, PDB entry 1GM1); the cold-shock domain of the human Y-box protein YB-1 (YBOX)[30] (79 residues, PDB entry 1H95). A second, more recent experimental dataset for the IgG-binding domain[31] (PDB entry 3GB1) was included in our analysis for comparison. Dipolar couplings restraints, if any, were excluded from the datasets because they cannot be directly expressed in distance space. Ambiguous restraints and restraints involving nonstereospecific assignments were included as their respective $<r^{-6}>^{-1/6}$ average distance.[32,33] Restraints involving identical atom groups but with different upper and lower bounds were removed from all datasets, keeping only those with the most restrictive bounds.

Bounds Matrix and Structure Calculations. The calculations of the matrices containing the upper and lower bounds were performed using the program X-PLOR.[34] In-house routines written in Python and C were used to process the available restraint files and to calculate the different restraint information values. Parameters describing the covalent interactions in all systems were taken from the PARALLHDG parameter file[35] (version 5.2) based on the CSDX parameter set.[36] All calculations were run in parallel on a Linux cluster with 1.8 GHz CPUs. Required computer time for calculation of I_{uni} for the 1GB1 dataset was 13 min using two CPUs and 3 min using 10 CPUs. Calculation of I_{ave} is computationally much more intensive and required about 7 h on 10 CPUs to reach convergence for this dataset. The QUEEN software package is freely available from the authors upon request.

(26) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. *Nucleic Acids Res.* **2000**, *28*, 235−242.
(27) Gronenborn, A. M.; Filpula, D. R.; Essig, N. Z.; Achari, A.; Whitlow, M.; Wingfield, P. T.; Clore, G. M. *Science* **1991**, *253*, 657−661.
(28) Cornilescu, G.; Marquardt, J. L.; Ottiger, M.; Bax, A. *J. Am. Chem. Soc.* **1998**, *120*, 6836−6837.
(29) Walma, T.; Spronk, C.; Tessari, M.; Aelen, J.; Schepens, J.; Hendriks, W.; Vuister, G. W. *J. Mol. Biol.* **2002**, *316*, 1101−1110.
(30) Kloks, C. P.; Spronk, C. A.; Lasonder, E.; Hoffmann, A.; Vuister, G. W.; Grzesiek, S.; Hilbers, C. W. *J. Mol. Biol.* **2002**, *316*, 317−326.
(31) Kuszewski, J.; Gronenborn, A. M.; Clore, G. M. *J. Am. Chem. Soc.* **1999**, *121*, 2337−2338.
(32) Nilges, M. *J. Mol. Biol.* **1995**, *245*, 645−660.
(33) Brünger, A. T.; Clore, G. M.; Gronenborn, A. M.; Karplus, M. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 3801−3805.
(34) Brünger, A. T. *X-PLOR, A system for X-ray crystallography and NMR*, version 3.1; Yale University Press: New Haven, CT, 1992.
(35) Linge, J. P.; Nilges, M. *J. Biomol. NMR* **1999**, *13*, 51−59.
(36) Engh, R. A.; Huber, R. *Acta Crystallogr., Sect. A* **1991**, *47*, 392−400.

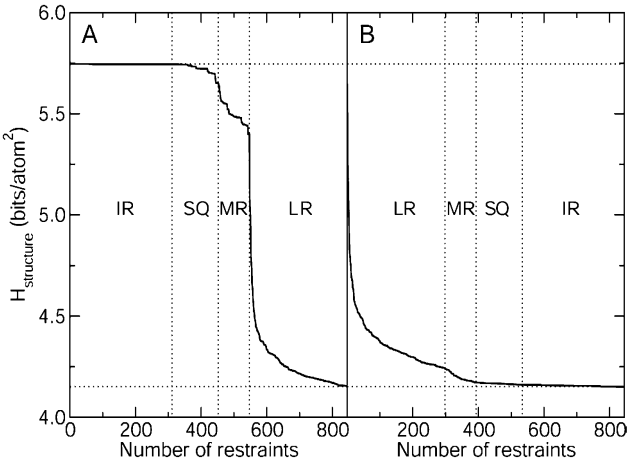

Figure 1. The structural uncertainty, $H_{structure}$, of the IgG-binding domain of protein G as a function of the number of distance restraints incorporated. The interproton distance restraints are grouped into four sets: intraresidual restraints (IR), sequential restraints (SQ), medium-range restraints (MR), and long-range restraints (LR). Two different orders of addition of the experimental data are shown: (A) IR-SQ-MR-LR and (B) LR-MR-SQ-IR.

Table 1. Structural Uncertainty and Experimental Information

	IgG (56)[a] (1GB1)	IgG (56) (3GB1)	UBI (76)	YBOX (79)	PDZII (93)
$H_{structure\|0}$ (bits/atom^2)	5.746	5.746	6.147	6.153	6.417
$H_{structure\|R}$ (bits/atom^2)	4.117	4.147	4.219	4.723	4.466
I_{total} (bits/atom^2)	1.629	1.598	1.927	1.430	1.951
RMSD (Å)[b]	2.00	2.29	2.37	6.34	3.49

[a] Numbers in parentheses indicate the number of residues in the structure. [b] Pairwise heavy-atom RMSD values of the ensembles deposited in the PDB after applying the resampling procedure of Spronk et al.[37]

Structure calculations for the IgG-binding domain were performed using the standard Cartesian dynamics simulated annealing protocol implemented in X-PLOR.[34] Forty structures were calculated in all cases, and the heavy-atom RMS deviation (all 56 residues) from the average was taken as an estimate of the ensemble precision.

Results and Discussion

Structural Uncertainty. The decrease in structural uncertainty $H_{structure}$ of the IgG-binding domain upon addition of the 845 unique interproton distance restraints is shown in Figure 1A. The set of restraints was grouped into four categories comprising the intraresidual (IR), sequential (SQ), medium-range (MR), and long-range (LR) restraints. Figure 1B demonstrates the effect on $H_{structure}$ resulting from a different order of addition of the restraints. This illustrates that the information contained in each of the sets is always dependent on the restraints that have already been added. For example, the decrease in $H_{structure}$ upon addition of the MR restraints is strongly reduced if the LR restraints have already been incorporated. As required, however, the final value for the structural uncertainty, and hence I_{total}, is independent of the order of addition of the experimental restraints.

The amount of information in the experimental datasets, I_{total}, together with the initial uncertainty, $H_{structure|0}$, the final uncertainty of the structures given the experimental restraints, $H_{structure|R}$, and the pairwise RMSD values of the corresponding structural ensembles are presented in Table 1. The $H_{structure|0}$

Table 2. Average Set Information and Dataset Sizes

	I_{ave} as fraction of I_{total} (%)					size of dataset as percentage of the total number of restraints (%)[a]				
	IgG (1GB1)	IgG (3GB1)	UBI	YBOX	PDZII	IgG (1GB1)	IgG (3GB1)	UBI	YBOX	PDZII
intraresidual restraints	0.3	0.1	0.1	0.0	0.1	30.9 (311)	21.2 (184)	24.0 (489)	34.6 (192)	30.1 (516)
sequential restraints	1.9	1.7	0.3	0.2	0.3	14.0 (141)	15.5 (135)	15.0 (305)	24.9 (138)	20.8 (356)
medium-range restraints	8.0	7.4	10.8	5.4	8.4	9.4 (95)	9.6 (83)	14.9 (304)	5.2 (29)	10.8 (185)
long-range restraints	55.7	55.2	55.2	93.7	54.5	29.6 (298)	29.6 (257)	40.0 (815)	24.7 (137)	25.3 (433)
hydrogen bond restraints	33.3	34.1	31.5	−	35.4	6.8 (68)	7.4 (64)	1.3 (27)	− (−)	4.3 (73)
dihedral angle restraints	0.7	1.6	2.0	0.7	1.3	9.2 (93)	16.8 (146)	4.8 (98)	10.6 (59)	8.8 (151)

[a] Numbers in parentheses indicate the actual number of experimental restraints.

Table 3. Average Set Information and Dataset Sizes

	I_{ave} as fraction of I_{total} (%)					size of dataset as fraction of number of restraints (%)[a]				
	IgG (1GB1)	IgG (3GB1)	UBI	YBOX	PDZII	IgG (1GB1)	IgG (3GB1)	UBI	YBOX	PDZII
intraresidual restraints	0.4	0.1	0.1	0.0	0.1	36.8 (311)	27.9 (184)	25.6 (489)	38.7 (192)	34.6 (516)
sequential restraints	2.4	2.4	0.4	0.2	0.3	16.7 (141)	20.5 (135)	15.9 (305)	27.8 (138)	23.9 (356)
medium-range restraints	11.8	11.2	14.8	5.6	11.7	11.2 (95)	12.6 (83)	15.9 (304)	5.8 (29)	12.4 (185)
long-range restraints	85.4	86.3	84.7	94.2	88.0	35.3 (298)	39.0 (257)	42.6 (815)	27.6 (137)	29.1 (433)

[a] Numbers in parentheses indicate the actual number of experimental restraints.

values become larger as the size of the system increases, reflecting the increasing number of possible random structures. The final value of the uncertainty, $H_{structure|R}$, is dependent on the available experimental data. Although of almost the same sequence length, the datasets for UBI and YBOX differ considerably in their information content, yielding 1.927 bits/atom^2 and 1.430 bits/atom^2, respectively. The much larger value for the UBI dataset reflects its relatively large number of NOE restraints (cf., Table 2). In contrast, the total number of restraints that were derived for the YBOX structure was relatively low because of a lack of restraints in its large flexible loop. The large difference in positional RMSD values of the resulting structural ensembles is consistent with the large difference in information content of the two respective restraint datasets.

The average information content, I_{ave}, of the different restraint categories in the five experimental datasets is presented together with their relative magnitudes in Table 2. Interestingly, the average amount of information per category in the different restraint sets is quite similar despite the significant differences in the actual number of restraints and their distribution over the different restraint classes. An exception to this is the YBOX dataset, which is probably caused by the rather limited amount of experimental data in this dataset.

As expected and experimentally observed,[10] Table 2 reveals that the long-range restraints contain most of the structural information, indicating the importance of these restraints. The information contributed by the intraresidual, sequential, and dihedral angle restraints to the overall information content is limited, despite their overwhelming majority in raw numbers in each of the five datasets. Hydrogen bond restraints, being tight medium-range and long-range restraints in the case of β-sheet structures, contain a significant amount of information. It is common for NMR spectroscopists to assign hydrogen bonds based on indirect experimental information, such as exchanging amide protons[38] and chemical shifts,[39] or on assumptions about regular secondary structure. Hydrogen bonds assigned without being directly detected should in principle be redundant with the remainder of the NOE restraints, whereas structural assumptions can lead to overly regular secondary structure elements. In the five datasets under investigation, the unique information, I_{uni}, of all hydrogen bond restraints varies between 1 and 2.5% of the total amount of information (data not shown).

To assess more accurately the individual contributions of the intraresidual, sequential, medium-range, and long-range NOEs, the restraint datasets have also been analyzed in the absence of other types of experimental data. The results of this analysis are shown in Table 3. With the informative but largely redundant hydrogen bond restraints removed, the importance of the long-range restraints becomes even more evident. Since we have shown that the majority of the experimental information is contained in the NOE-derived restraints, for simplicity, all subsequent analyses will be performed using only this type of experimental data.

Individual Restraints. For all individual NOE-derived restraints of the IgG dataset, we calculated both their average and the unique information content, I_{ave} and I_{uni}, respectively. The results of these calculations are shown in Figure 2. For purposes of illustration, five restraints with varying information characteristics were selected (subsequently noted as R1 through R5; see Table 4, Figures 2, 3, and 4) and are used throughout the remainder of the text. The average information per restraint, as depicted in Figure 2A, confirms that the majority of the information is indeed contained in the long-range restraints, but it also shows that there is a large variation in importance between the different restraints within one category. The subset of intraresidual restraints contains only a very limited amount of information (see above). However, a few restraints have significantly higher information content compared to the rest. A detailed analysis of these restraints, which includes the restraint between the H^{α} and $H^{\delta 1}$ of Tyr-30 in IgG (restraint R1), reveals that these all involve bulky amino acids such as tryptophan, phenylalanine, and tyrosine. For these side chains, a rearrangement would affect the local structure more strongly than would be the case for the smaller amino acids.

(37) Spronk, C. A.; Nabuurs, S. B.; Bonvin, A. M.; Krieger, E.; Vuister, G. W.; Vriend, G. *J. Biomol. NMR* **2003**, *25*, 225−234.
(38) Wagner, G.; Wüthrich, K. *J. Mol. Biol.* **1982**, *160*, 343−361.
(39) Wishart, D. S.; Sykes, B. D.; Richards, F. M. *Biochemistry* **1992**, *31*, 1647−1651.

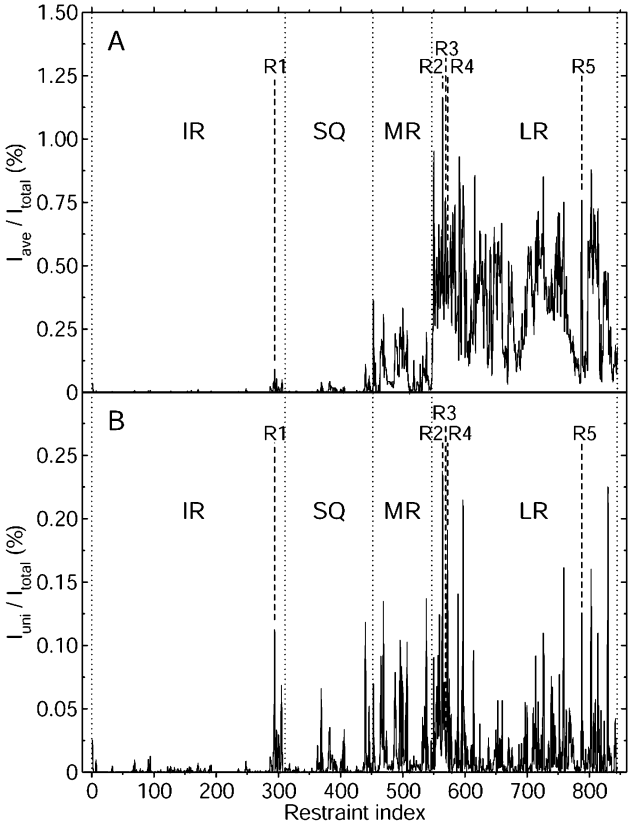

Figure 2. (A) The relative average information content, I_{ave}/I_{total}, and (B) the relative unique information content, I_{uni}/I_{total}, both plotted as a function of the NOE restraint index of the IgG-binding domain dataset. Labeled restraints are listed in Table 4 and discussed in the text; other labels are as in Figure 1.

Table 4. Five Selected Restraints from the IgG-Binding Domain Dataset

label	residue number[a]	atom name	residue number[a]	atom name	l_{ij} (Å)	u_{ij} (Å)
R1	30	H^{α}	30	$H^{\delta 1}$	1.8	2.7
R2	8	H^{N}	54	H^{N}	1.8	3.5
R3	6	H^{N}	52	H^{N}	1.8	5.0
R4	2	H^{α}	19	H^{α}	1.8	2.7
R5	26	H^{N}	45	$H^{\epsilon 2}$	1.8	5.0

[a] Residues are numbered according to PDB entry 2GB1.

The amount of unique information in the 845 restraints for 1GB1, shown in Figure 2B, shows a significantly different pattern. Despite its limited average information content (I_{ave}), restraint number R1 contains a significant amount of unique information (I_{uni}). The opposite case is demonstrated by restraint R3, which is structurally important, but contains little unique information. Restraints R2, R4, and R5 are examples of the most important restraints and contain information not found in the remainder of the dataset. These restraints are important because they are less supported by the remainder the dataset, i.e., exhibit more unique information than others, indicating that these restraints are either crucial to a structure calculation or suggest a potential error. In both instances, these restraints are interesting and definitely warrant careful investigation. This approach of placing a higher confidence in restraints that are supported by other restraints in the dataset has already been proposed and applied,[40,41] but can now be evaluated in a quantitative manner.

(40) Englander, S. W.; Wand, A. J. *Biochemistry* **1987**, *26*, 5953−5958.

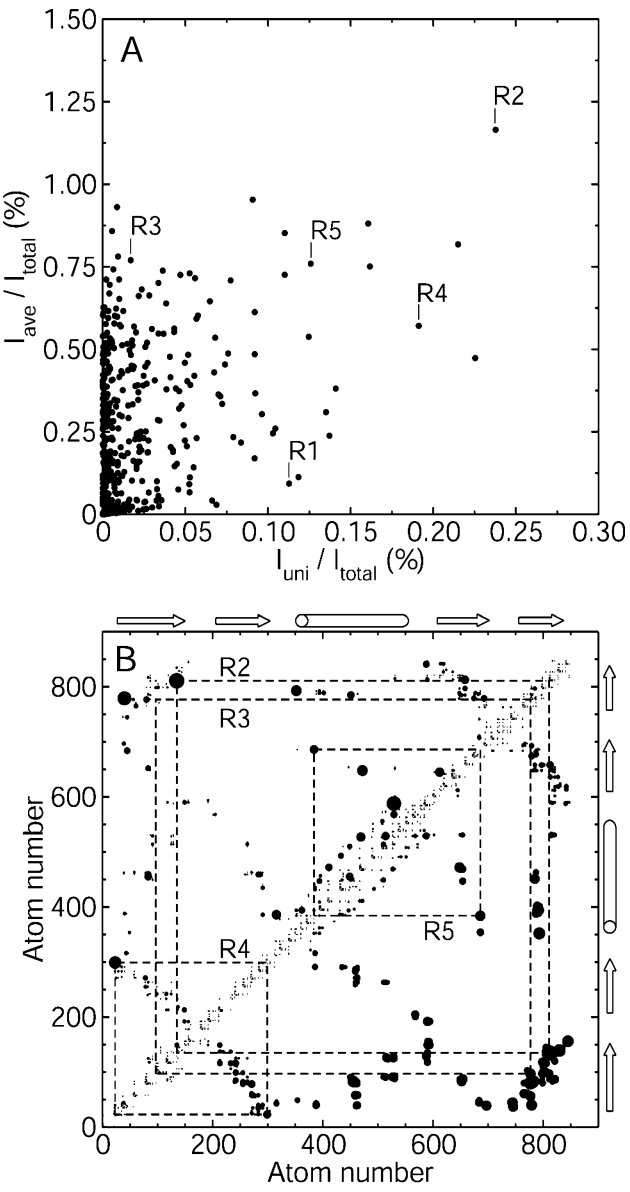

Figure 3. (A) The unique information content, I_{uni}, for the restraints of the IgG dataset versus the average information content I_{ave}. Selected restraints are indicated by R1−R5 (listed in Table 4). (B) An NOE contact plot indicating both the unique restraint information (above diagonal) and the average restraint information (below diagonal) for the IgG-binding domain. The size of the circles is scaled according to the amount of information contained in that particular restraint. Atom numbering according to PDB entry 2GB1.

To facilitate the identification of the important and less supported restraints, a plot of the average restraint information versus the unique restraint information is shown in Figure 3A. In this plot, the important and lesser supported restraints (R2, R4, and R5) are clearly separated from the less important, less supported (R1) and the important, supported restraints (R3). Figure 3B presents an alternative representation of the same phenomenon. In this NOE contact plot each circle represents a single restraint, with the size of the circles scaled according to the information content of the restraint. The section above the diagonal displays the unique restraint information, whereas the section below the diagonal displays the average restraint

(41) Herrmann, T.; Güntert, P.; Wüthrich, K. *J. Mol. Biol.* **2002**, *319*, 209−227.

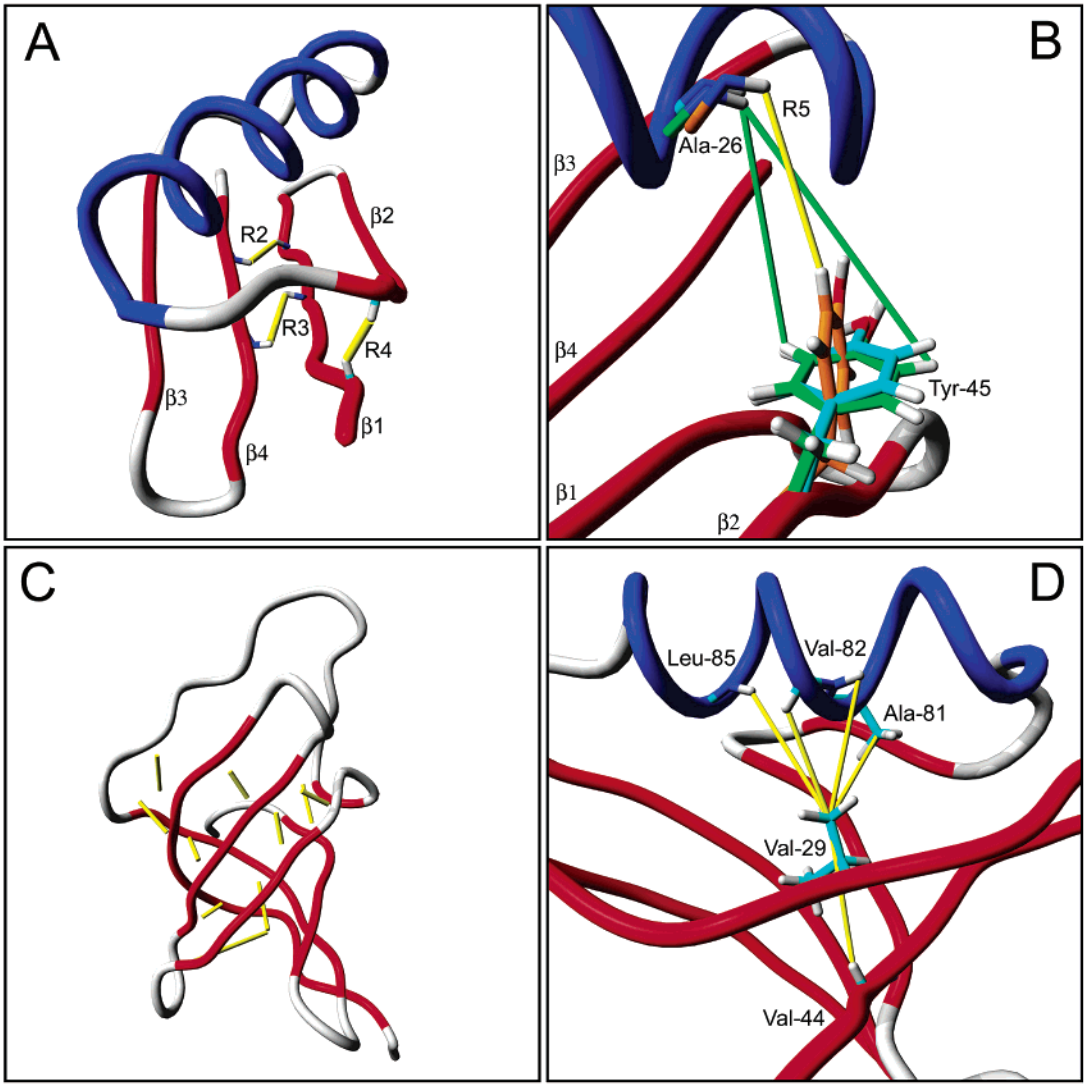

Figure 4. (A) Location of restraints R2, R3, and R4 (see Table 4) in the structure of the IgG-binding domain (PDB entry 2GB1). (B) The orientation of Tyr-45 (shown in stick representation) in the original NMR structure (PDB entry 2GB1, orange), the most recent NMR structure (PDB entry 3GB1, green), and the crystal structure (PDB entry 1PGB, cyan). Experimental NMR restraints for the 2GB1 dataset are shown in yellow, those for the 3GB1 dataset in green. (C) Location of the 10 most informative, least supported restraints in the structure of the YBOX domain (PDB entry 1H95). (D) Crucial most informative, least supported restraints (yellow) involving Val-29 (shown in stick representation) in the PDZII structure (PDB entry 1GM1). β-strands are shown in red, α-helices in blue, coil and turns in gray. Figures were made using YASARA (www.yasara.org).

information. In the lower section it can be seen that the importance of a restraint increases if it connects atoms that are further apart in the primary protein sequence. All restraints connecting the N-terminal and the C-terminal β-strands of the IgG-binding domain are clearly identified as important, together with several restraints packing the α-helix on top of the underlying β-sheet.

To translate the restraints to structural terms, the location of restraints R2, R3, and R4 in the structure of the IgG-binding domain is shown in Figure 4A. The most informative and least supported restraint in this dataset is restraint R2. This restraint is the last backbone−backbone restraint that links the two central parallel β-strands near the C-terminal end of the protein. A similar role is played by restraint R4, as it connects the first and second β-strand at the N-terminal side of the peptide chain. Restraint R3 is also among the more important restraints in the IgG-binding domain structure, but compared to restraint R2 and R4, this restraint contains much less unique information (see Figure 3). It is located in the center of the β-sheet, surrounded by many other tight interbackbone restraints and is therefore well supported.

In Figure 3 we see that restraint R5, between Ala-26 and Tyr-45, contains unique and important information. From the IgG-binding domain structure (PDB entry 2GB1), it is not obvious why this restraint would be exceptional. However, a comparison with the IgG-binding domain crystal structure (PDB entry 1PGB)[42] and a recently published IgG-binding domain NMR structure (PDB entry 3GB1)[31] yields interesting results. Panel B of Figure 4 shows that the orientation of the side chain of the tyrosine residue involved in this restraint (Tyr-45) is very different in the original structure. Tyr-45 was previously identified as the only aromatic side chain with a different rotamer in the crystal structure when compared to the NMR structure.[42] Our results indicate that restraint R5 is probably the cause of this difference. In the original dataset, the R5 NOE was assigned specifically to the $H^{\epsilon 2}$ proton of the aromatic ring,

(42) Gallagher, T.; Alexander, P.; Bryan, P.; Gilliland, G. L. *Biochemistry* **1994**, *33*, 4721−4729.

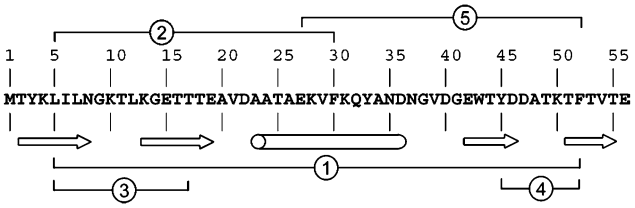

Figure 5. The first five restraints of the ordered dataset indicated in the primary sequence of the IgG-binding domain.

whereas in the latter dataset (3GB1) it is assigned, with a different distance limit, to both ϵ-protons. This latter assignment is better supported, since the amount of unique information in this restraint is drastically reduced in the updated dataset (not shown).

Analyses of the experimental restraint lists have also been performed for the YBOX domain and the PDZII domain. For the rather limited dataset of the YBOX domain, the most important and least supported restraints are spread throughout the structure (cf., Figure 4C). All restraints identified as crucial are involved in connecting the different strands in the β-barrel, therefore defining the topology of the protein.

For the PDZII domain, several residues were identified by the procedure as crucial in defining the structure of this protein. For example Val-29, which is located in PDZIIs peptide-binding groove, is involved in five out of the 25 most informative and least supported restraints (cf., Figure 4D). In this structure many of the key amino acids are located in the core of the protein and are involved in hydrophobic interactions. These results indicate that specific residues that are crucial in a protein structure determination are easily identified using the QUEEN method.

Data Redundancy. The availability of a measure of restraint information allows further investigation of the presence of redundancy in NMR datasets. For the original IgG-binding domain dataset, an ordered NOE restraint dataset was constructed by an iterative procedure in which the next most informative NOE was successively selected. Thus, in the ordered dataset the most informative restraints are present at the beginning of the restraint list, while the least informative restraints are found at the end of this dataset. Adding the restraints in this particular order will maximize the decrease in structural uncertainty as the number of incorporated restraints increases. The residues connected by the first five restraints in the ordered dataset are indicated in the primary sequence of the IgG-binding domain in Figure 5. Each of these restraints connects atoms located in separate secondary structure elements. The two β-strands most distant in primary sequence are connected first, followed by strand helix and interstrand contacts, clearly outlining the topology of the structure.

The increase in information as a function of the number of restraints from the ordered dataset is shown in Figure 6A. For reference, the trend obtained by evaluating 10 randomly ordered datasets is also indicated. The difference between the two curves illustrates the effect of adding the most informative restraints first. For the ordered dataset, approximately 50% of the total number of restraints is sufficient to describe 99.9% of the information contained in the complete dataset, whereas in a random set this level of information is normally not reached until 98% of the experimental data is used.

To validate this finding, structures where calculated using

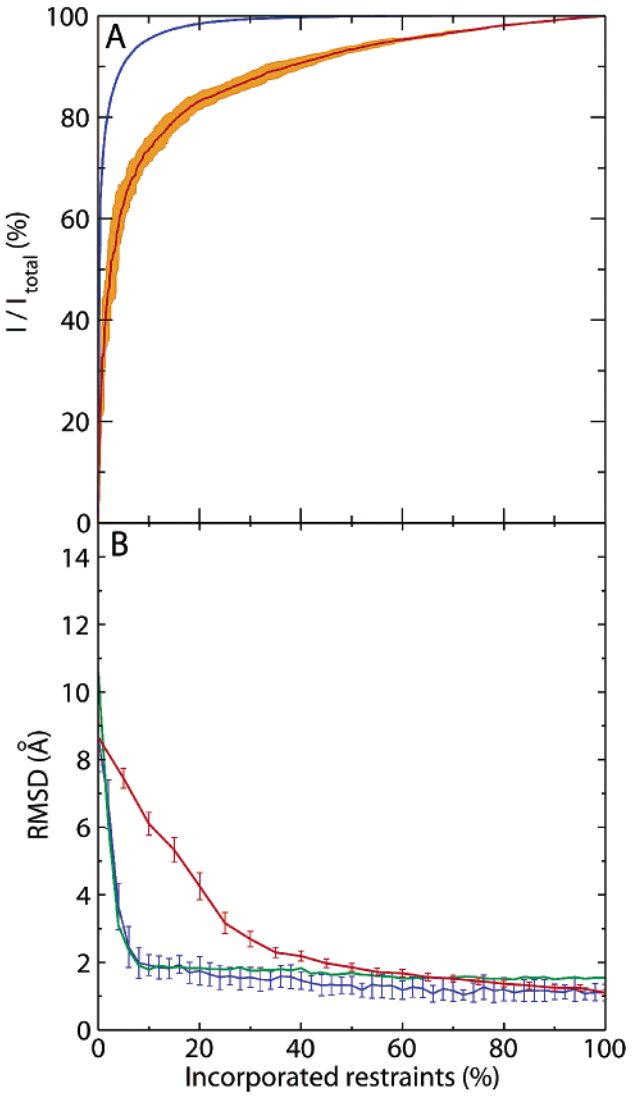

Figure 6. (A) Information content as a function of the percentage of incorporated restraints. The optimally ordered dataset is presented by a blue line. The average of 10 randomly sorted datasets is indicated by a red line; the standard deviation of these 10 sets is shown in orange. (B) Precision of the resulting structure ensemble as a function of the percentage of incorporated restraints, calculated using 50 subsets of the ordered dataset (blue line: RMSD to the mean; green line: RMSD to the crystal structure) and calculated using 10 $\times$ 20 randomly sorted datasets (red line: RMSD to the mean).

50 subsets with increasing size from the optimal dataset comprising between 0 and 100% of the total number of restraints. The heavy-atom RMSD values of the calculated structure ensembles for the 50 datasets are shown in Figure 6B. The precision of the resulting ensembles does not increase significantly on inclusion of the second half of the restraints of the ordered dataset. A similar trend is observed for the RMSD values of the structure ensemble to the crystal structure of the IgG-binding domain (PDB entry 1PGB). These data show that the similarity to the crystal structure also does not improve by incorporation of the second half of the restraint data, again indicating the redundancy of about 50% of the data in the IgG-binding domain dataset.

As expected, randomly selected subsets contain less information when compared to equally sized subsets of the ordered dataset (cf., Figure 6A), and they would be expected to yield less precise structural ensembles. This is illustrated in Figure

6B by the significantly higher structural uncertainty, as expressed by the heavy-atom RMSD values observed for the ensembles calculated from these randomly sorted subsets. The similar trend observed for the data describing the structural uncertainty as predicted by the QUEEN procedure and the positional uncertainty as expressed by the ensemble RMSD values clearly illustrates the correspondence between these two uncertainty measures.

Conclusions and Suggestions

The QUEEN method allows for a straightforward identification of the important and the unique restraints in an experimental NMR dataset. In addition to previous methods to analyze the redundancy of NMR distance restraints,[15] our method can also identify redundant inter-residual restraints. We have shown that a significant percentage of the experimental restraints is usually redundant; thus, the number of restraints can be a poor indicator of the amount of experimental information. Our proposed measure for the information content in an experimental dataset provides a quantitative way of representing the information contained in experimental input data. Our examples show that plots of I_{ave} versus I_{uni} identify critical restraints and facilitate the identification of problematic ones. We therefore hope that scientists involved in structure determination by NMR will use this tool to evaluate the experimental input data, and that the amount of information (I_{total}) and the distribution of this information over the different restraint classes will be reported.

Acknowledgment. We would like to thank Roland L. Dunbrack, Jr. (Fox Chase Cancer Center) for useful discussions, Tine Walma and David Thomas for critical reading of the manuscript, and all NMR spectroscopists who made the effort of depositing their experimental restraints. Financial support from the European community (5th Framework program NMRQUAL Contract Number QLG2-CT-2000-01313) to S.N. and E.K. and from the Netherlands Foundation for Chemical Research (NWO/CW) to C.S. is gratefully acknowledged.

JA035440F