Journal of Biomolecular NMR, **14:** 123–132, 1999.
KLUWER/ESCOM

Completeness of NOEs in protein structures: A statistical analysis of NMR data

Jurgen F. Doreleijers, Mia L. Raves, Ton Rullmann* & Robert Kaptein**
Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, the Netherlands

Received 11 January 1999; Accepted 15 April 1999

Key words: experimental data, NOE completeness, PDB, protein structure, quality assessment, validation

Abstract

The completeness of experimentally observed NOE restraints of a set of 97 NMR protein structures deposited in the PDB has been assessed. Completeness is defined as the ratio of the number of experimentally observed NOEs and the number of ‘expected NOEs’. A practical definition of ‘expected NOEs’ based on inter-proton distances in the structures up to a given cut-off distance is proposed. The average completeness for the set of 97 structures is 68, 48, and 26% up to 3, 4, and 5 Å cut-off distances, respectively. For recent state-of-the-art structures these numbers are approximately 90, 75, and 45%. Almost 20% of the observed NOEs are between atoms that are further than 5 Å apart in the final structures. The completeness is independent of the relative surface accessibility and does not depend strongly on residue type, secondary structure or local precision, although the number of observed NOEs in these classes varies considerably. The completeness of NOE restraints is a useful quality criterion in the course of structure refinement. The completeness per residue is more informative than the number of NOEs per residue, which makes it a useful tool to assess the quality of the NMR data set in relation to the resulting structures.

Abbreviations: cv, circular variance; NOE, nuclear Overhauser enhancement; NOESY, nuclear Overhauser enhancement spectroscopy; PDB, Protein Data Bank.

Introduction

Information on experimentally determined restraints is publicly available for an increasing number of NMR structures in the Protein Data Bank (Bernstein et al., 1977). The NMR measurable proton-proton distances obtained from nuclear Overhauser enhancements (NOEs) still provide the single most important source of information for solution structure determination (Clore et al., 1993). Other experimental information from J-couplings (Garrett et al., 1994), chemical shifts (Kuszewski et al., 1995) and residual dipolar couplings (Tjandra et al., 1997) can further improve the quality of NMR structures.

Here, a new quality indicator is introduced that evaluates the level of completeness to which the NOEs have been observed, with respect to the expected NOEs calculated from the inter-proton distances as present in the deposited structures. To define what should be included in the ‘expected NOEs’ is not trivial, and a practical definition is developed, which is consistent with current practice of protein NMR spectroscopy. Correlations of the completeness with the cut-off distance, type of protons and residues, NOE class, structural variance, secondary structure, relative surface accessibility, protein size, Ramachandran map, and year of publication are analysed for a set of 97 proteins previously studied (Doreleijers et al., 1998). One of these, the HU protein (Vis et al., 1995), serves as an example to illustrate the properties and behaviour of the completeness.

In X-ray crystallography, the completeness of reflections is usually close to 100%. Completeness for

*Present address: Molecular Design & Informatics, NV Organon, P.O. Box 20, 5340 BH Oss, the Netherlands.

**To whom correspondence should be addressed. E-mail: kaptein@nmr.chem.uu.nl

Supplementary material: The completenesses at cut-off distances of 3 to 9 Å for the 97 individual entries are presented on 3 pages.

reflections of at least 80% in the highest resolution shell is one of the criteria commonly used to decide whether the reflections should be included in the data set (Kleywegt and Jones, 1997). The completeness of NOE data is quite different in nature. An important difference is that there is no unequivocal way to define which NOEs are to be expected based on the experimental data alone; the expected NOEs can only be determined from the final structure. NOEs corresponding to short distances (< 3 Å) can be observed nearly completely. At large distances (> 5 Å) many NOEs are too weak to be observed. Other reasons for missing NOEs are overlap of frequencies and exchange broadening.

This study is part of a larger project aimed at the validation of biomolecular structures involving a group of X-ray crystallography, NMR spectroscopy and modelling laboratories. In this group, a number of quality indicators have been developed that are based on X-ray structures (e.g. Morris et al., 1992; Vriend and Sander, 1993; Pontius et al., 1996; Wilson et al., 1998). In a previous paper, we have shown how these quantities apply to a set of 97 NMR structures and we have added NMR-specific indicators such as NOE violations and precision of an ensemble of NMR structures (Doreleijers et al., 1998). For this purpose we have developed procedures implemented in the program AQUA (Rullmann, 1996) for converting restraint files into a standard file format using the atom names as recommended by the IUPAC NMR Task Group (Markley et al., 1998).

It will be shown that the completeness is a more informative quantity than the commonly quoted number of NOEs (per residue), as it is normalised for the number of expected NOEs. The completeness analysis is useful for both flexible and rigid parts of the molecule. To distinguish between true disorder in the solution structure and under-determination, which might be caused by insufficient analysis, experimental data from NMR relaxation studies (Palmer III et al., 1996) should preferentially be obtained. NOE completeness checks have previously been applied by Vis et al. (1995) and Gardner et al. (1997) using somewhat different definitions. The calculations shown here were performed using a new module in the AQUA program (Laskowski et al., 1996; Rullmann, 1996). The completeness analysis has been added to the Biotech validation servers (EBI server at http://biotech.ebi.ac.uk:8400), allowing easy public access.

Definitions

Definition of completeness

NOE completeness is defined as the ratio, expressed as a percentage, between the number of matched observed NOEs and the number of expected NOEs:

$$\text{Completeness} = 100\% \frac{\text{number of matched observed NOEs}}{\text{number of expected NOEs}}$$

The expected NOEs are determined from the structure as inter-proton distances below a given cut-off. The distances are averaged over all available models, since the whole ensemble provides the best possible representation of NMR data (Sutcliffe, 1993; Pearlman, 1994). A normal average is used instead of an r^{-6} average because in practice the latter gives too much weight to the shorter distances. A more detailed definition of expected NOEs is discussed below. The observed NOEs are extracted from the deposited restraint file. Ambiguous restraints, as well as restraints that only consist of lower bounds, are discarded. The intra-residual NOEs are filtered for fixed distances (e.g. alanine $H^{\alpha}-MB$) using the AQUA redundancy module (Doreleijers et al., 1998).

The observed and expected sets of NOEs are then matched to one another, discarding the observed NOEs for which the inter-proton distance is larger than the given cut-off. Of course, since the completeness is evaluated for several cut-off distances, eventually all observed NOEs are taken into account. The completeness can be decomposed with respect to: atom type, residue type, NOE class and residue number. In this study, the completeness was calculated over all residues in the structure, including those in the disordered regions that were left out in our previous study (Doreleijers et al., 1998).

Definition of expected NOEs

The simplest approach would be to consider all protons. However, some protons are rarely observed in NMR experiments. The goal is to define ‘expected NOEs’ in such a way that all structure determinations, including those using state-of-the-art techniques, are properly scored and comparable. The completeness of some types of protons is an order of magnitude lower than that of other protons, therefore only the latter are included here (see Table 1). The rarely observed protons that we will exclude here are the following: the hydroxyl (Ser, Thr, Tyr), sulph-hydryl (Cys), carboxyl (Asp, Glu, C-terminus), and amino (Lys, N-terminus) groups and the protons attached to the nitrogen atoms in the imidazole group (His) and to the terminal nitrogen atoms in the guanidinium group (Arg). A lower completeness is also expected for this group of protons, as they are normally in exchange with water. The methylene α- and β-protons, the side-chain amide protons of asparagine and glutamine, and the methyl protons of the valine and leucine isopropyl group are defined as individually observable atoms or groups. It is common practice to stereospecifically assign these protons experimentally (Wagner et al., 1987; Neri et al., 1989) or computationally (Folmer et al., 1997). All other prochiral protons of side chains were taken into account as pseudo atoms. Methyl protons and phenyl and tyrosyl δ- and ε-protons are also included as pseudo atoms (Wüthrich et al., 1983).

Two cases of stereospecificity need to be considered. In the first case, an observed NOE involves a prochiral atom (e.g. $H^{\delta1}$) and the list of expected atoms contains the representing pseudo atom (c.q. QD). In this case, the NOE is referred to the pseudo-atom position and a pseudo-atom correction is added to the upper bound of the restraint. For example, NOEs between X-$H^{\delta1}$ and between X-$H^{\delta2}$ collapse to a single restraint X-QD. In the second case, an observed NOE contains only the pseudo atom (e.g. QB) whereas the list of expected atoms contains the prochiral atoms (c.q. $H^{\beta2}$ and $H^{\beta3}$). The pseudo atom of the restraint is then matched to the prochiral proton with the shortest distance. Only half of the maximum completeness can thus be obtained for these atoms due to the lack of stereospecific assignment. NOEs involving individually listed prochiral atoms are interpreted as if they were assigned, even if a floating-assignment calculational strategy was used.

Results and discussion

NOE completeness of the HU protein

The DNA-binding protein HU (Vis et al., 1995) was used as an example to illustrate the features of the completeness quantity. This protein consists of a symmetric homo-dimer with 90 residues per chain. Three helices and a three-stranded anti-parallel β-sheet of each monomer form the core. The second and third β-strand extend from the core; these so-called β-ribbon 'arms' are able to wrap around DNA. Figure 1 shows the dependency of the number of NOEs and the completeness on the cut-off distance for this protein. The increase in the number of expected NOEs per shell is approximately quadratic up to 6 Å, after which the increase is levelled off due to boundary effects. The lowest distance shell (2.0 to 2.5 Å) has approximately 90% NOE completeness, but for the 4.5 to 5.0 Å shell the completeness is only slightly over 20%. The cumulative completeness, i.e. up to a certain cut-off distance, is still more than 40% for a 5 Å cut-off distance.

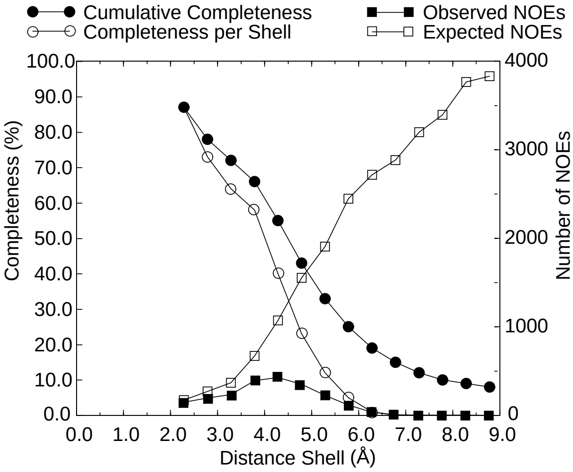

Figure 1. Completeness and number of NOEs of the HU protein. Cumulative and per-shell completeness (% on left *y*-axis) and the number of inter-residual observed and expected NOEs (right *y*-axis) are plotted versus the distance of the shell for the HU protein (Vis et al., 1995). The thickness of the shells is 0.5 Å.

Figure 2 depicts the NOE characteristics and structural variance of the HU protein sequence. As can be seen from the plot of the circular variance (Hyberts et al., 1992) versus the sequence in the bottom panel, the backbone variability is mainly concentrated in a small N-terminal region and in the 'arm' residues. A side-chain circular variance larger than 0.2 shows that more than one rotameric state is populated. A considerable number of side chains, many of which reside in the 'arms', have alternative conformations in solution.

Atom and residue types

On average a proton in the set of 97 NMR-solved proteins is expected to have 2.9 inter-residual NOEs with other atoms that are within 4 Å. At this cut-off distance, the expected number of NOEs per amide proton is 5.5, which is roughly two times higher. This high number is related to the many sequential and medium-range amide contacts present in helices, turns and, to a lesser extent, β-sheets. For these numbers each NOE between nuclei A and B is counted twice; once

Table 1. Expected atoms in a NOESY-type experiment

Amino acid	Amide	Alpha	Others[a]
Alanine	H^{N}	H$^{\alpha}$	MB
Arginine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, QG, QD, H$^{\varepsilon}$
Asparagine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, H$^{\delta21}$, H$^{\delta22}$
Aspartate	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$
Cysteine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$
Glutamine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, QG, H$^{\varepsilon21}$, H$^{\varepsilon22}$
Glutamate	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, QG
Glycine	H^{N}	H$^{\alpha2}$, H$^{\alpha3}$	–
Histidine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, H$^{\delta2}$, H$^{\varepsilon1}$
Isoleucine	H^{N}	H$^{\alpha}$	H$^{\beta}$, MG, QG, MD
Leucine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, H$^{\gamma}$, MD1, MD2
Lysine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, QG, QD, QE
Methionine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, QG, ME
Phenylalanine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, QD, QE, H$^{\zeta}$
Proline	–	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, QG, QD
Serine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$
Threonine	H^{N}	H$^{\alpha}$	H$^{\beta}$, MG
Tryptophan	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, H$^{\delta1}$, H$^{\varepsilon1}$, H$^{\varepsilon3}$, H$^{\zeta2}$, H$^{\zeta3}$, H$^{\eta2}$
Tyrosine	H^{N}	H$^{\alpha}$	H$^{\beta2}$, H$^{\beta3}$, QD, QE
Valine	H^{N}	H$^{\alpha}$	H$^{\beta}$, MG1, MG2

[a] The letter M as part of the pseudo-atom name indicates a methyl group; for all other pseudo-atom names the letter Q is used (Markley et al., 1998).

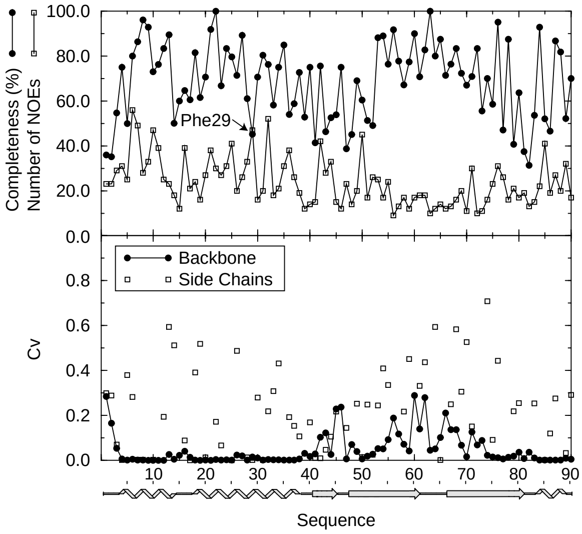

Figure 2. NOE and variability information of the HU protein. Top: completeness and the total number of observed inter-residual NOEs per residue. Bottom: the circular variances of the backbone and side-chain dihedral angles (Hyberts et al., 1992) versus the monomer sequence. All values are averaged over the two symmetric monomers. The secondary structure is adapted from Vis et al. (1995).

for A and once for B. In order of decreasing completeness, the atom types are: amide protons (59%), α-protons (55%), methyl protons (49%), aromatic ring protons (48%), stereospecifically assignable β-protons (35%), and other protons (33%). Reasons for the high completeness of the amide protons are the good dispersion and their importance for sequential assignment (Wüthrich, 1986). Most of the stereospecifically assignable β- and other protons show up in the more densely populated areas of the spectrum. Furthermore, a significant portion of the structures was solved without stereospecific assignment of the β-protons, which results in a lower completeness.

The completeness and the number of observed NOEs per residue for the 20 common amino acids are shown in Figure 3. Whereas the number of NOEs fluctuates significantly, the completeness is more or less constant for different residue types. The highest completeness is found for alanine (58%) and the lowest for proline (42%). The number of observed NOEs for tryptophan, however, is over a factor five larger than for glycine, due to the different number of protons present.

In conclusion, the completeness is well normalised for the different residue types which makes it easier

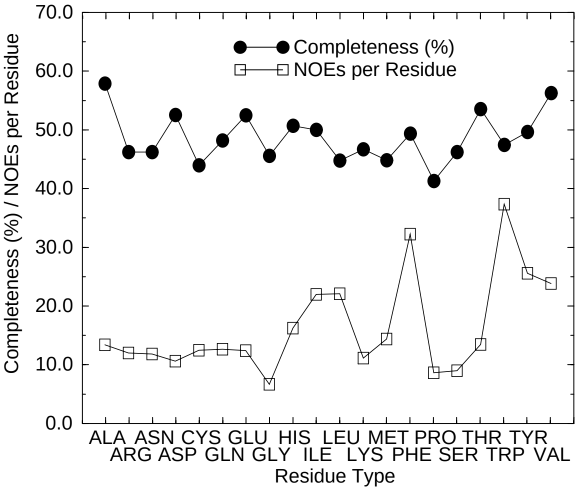

Figure 3. Average completeness and number of NOEs for the 20 common amino acids. The values are averaged over all residues in the 97 entries.

to detect problematic residues. The different nature of the completeness and the total number of observed NOEs per residue is also illustrated for the HU protein in Figure 2 (top panel). Overall, the two are correlated, but residue Phe29, for example, has more NOEs than its smaller neighbours but a significantly lower completeness.

Cut-off distance

The observed and expected NOEs of the HU protein are tabulated in Table 2. The expected NOEs are categorised according to the average distance in the models (shown vertically). The observed NOEs are matched to the expected NOEs and are further categorised on the basis of the upper-bound distance in the restraint (shown horizontally). Six NOEs (shown in bold) display a violation which is less than 1.0 Å. If the mixing times used are short enough, direct NOEs are only expected for atoms that are less than ∼5 Å apart. An upper bound in an NOE restraint involving pseudo atoms, however, is increased with one or two pseudo-atom corrections. The correction is large for phenyl-ring protons (2.2 Å), for example. There are a number of NOE upper distance restraints which are more than 3 Å larger than expected on the basis of the structure. This must be due to pseudo-atom corrections. These NOEs, situated at the top right corner of the table, have no influence in the final phase of the structure determination. There are also a significant number of observed NOEs that correspond to distances above 5 Å; these will be discussed below.

For all 97 entries, the matched percentage of observed NOEs is shown in Figure 4, for different distance shells. A significant number of the observed NOEs (17%) are between atoms that are farther than 5 Å apart. From a previous study it was clear that for the studied entries the NOEs are rarely violated by more than 1 Å and the violation rms of all restraints is 0.061 ± 0.043 Å (Doreleijers et al., 1998). Therefore, the upper-bound distance restraint for the observed NOEs must also be larger than 5 Å. These NOEs, including those of HU, must be from an indirect origin and are not severely violated because of the included pseudo-atom corrections. The structures using NOEs that correspond to distances above 9 Å have all been determined with a direct NOE refinement technique. This approach can properly account for spin diffusion, thus allowing more NOEs and longer distances to be used.

The average completeness for the 97 entries is 68 ± 14, 48 ± 13, and 26 ± 9% up to 3, 4, and 5 Å cut-off, respectively. The range of completeness for the different entries spans from 21 to 93%, 16 to 76%, and 8 to 57%, in the same order. A number of old methallothionein structures (1/2MHU, 1/2MRB, 1/2MRT) have a completeness below average. The highest completeness (up to 5 Å) was obtained for two structures; the *lac* repressor headpiece, entry 1LQC (Slijper et al., 1996) and the β chemokine hMIP-1β, entry 1HUN (Lodi et al., 1994). The completeness values up to a 3, 4, and 5 Å cut-off are 93, 75, and 57% for 1LQC, and 82, 72, and 49% for 1HUN. The fact that a direct refinement technique has been used for 1LQC may be one of the reasons of its relatively high completeness up to the 5 Å cut-off.

Taking a large cut-off distance (e.g. 5 Å) has the advantage that more NOEs are included, see Figure 4. The completeness when calculated with a 4 Å cut-off includes on average only half of the measured NOEs, and some weaker NOEs, important for the definition of the structure, are not included. In this study, however, a standard cut-off of 4 Å was used when a single value is preferred. The five structures in the current set with the highest completeness up to 4 Å have high values for all cut-off distances (∼90, 75, and 45% for 3, 4, and 5 Å, respectively) and can be regarded as state-of-the-art structures in this respect.

NOE classes

The classes of NOEs in order of increasing completeness are the intra-residual (27%), inter-subunit (37%), long-range (38%), medium-range (43%), and

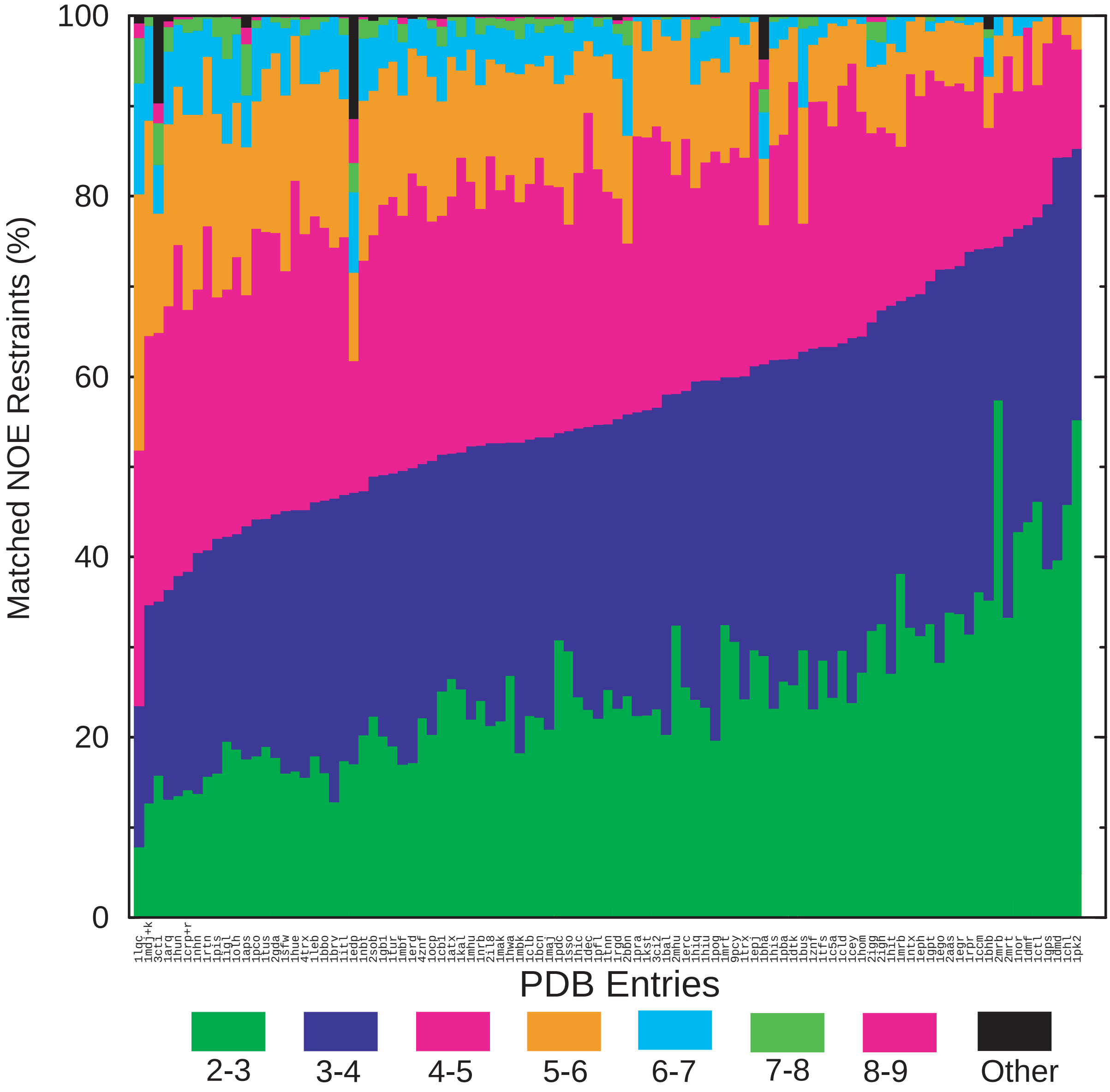

Figure 4. Overview of observed NOEs for 97 PDB entries. The observed NOEs are classified as a function of the average inter-proton distance. The PDB entry codes of the 97 structures are indicated at the bottom. The entries are sorted with respect to the total percentage of matched NOEs up to 4 Å.

sequential NOEs (59%) and are shown in Figure 5. Although intra-residual NOEs are not as informative as long-range NOEs, they are still useful as a source of information on the rotameric states of the side chains. For a number of entries the intra-residual NOE restraints were not used or were converted to dihedral-angle restraints, resulting in zero completeness. To allow a fair comparison between different entries, the class of intra-residual NOEs has been omitted from the completeness values discussed in the remainder of this paper. The inter-subunit and long-range NOEs, which are most important for structure determination by NMR, have in general a lower completeness than the medium-range and sequential NOEs.

Structural variance

In our previous study (Doreleijers et al., 1998) we have used the structural variation of the backbone di-

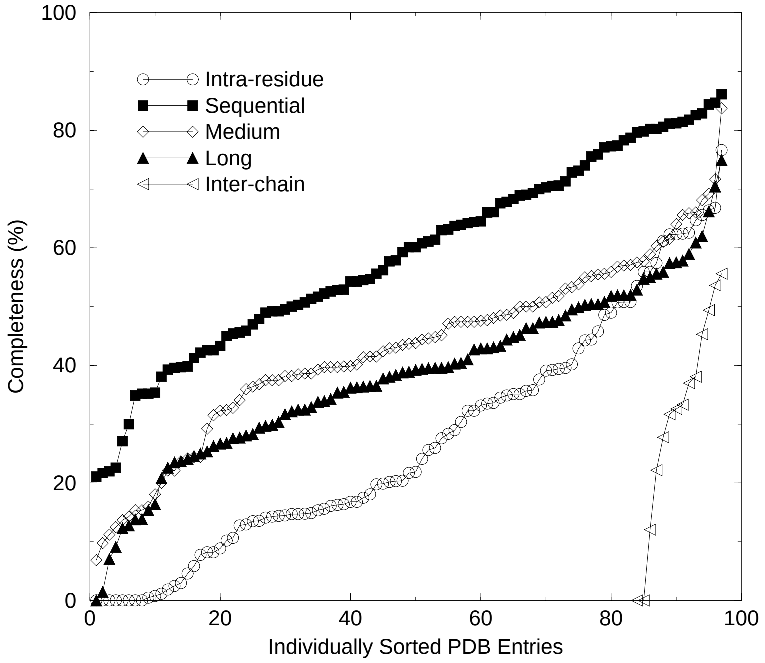

Figure 5. Overview of the completeness per class of NOEs for 97 PDB entries. The values are sorted by increasing completeness for each NOE class individually. The intra-residual NOEs have been filtered for fixed distances only. For two entries with zero completeness of the inter-chain NOEs, no NOEs in this class had been submitted to the PDB when this study was started.

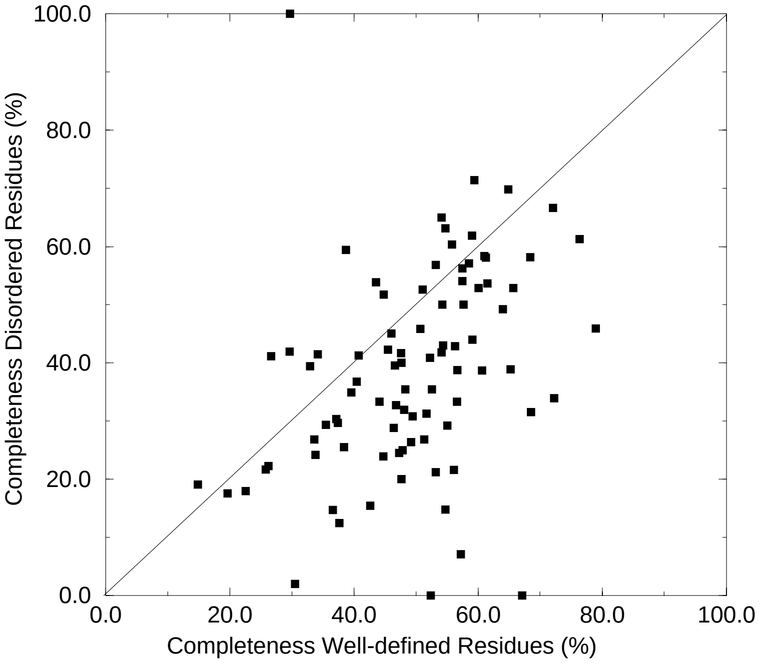

Figure 6. Correlation between completeness of well-defined and disordered residues. Well-defined residues have a backbone circular variance ($cv_{\varphi,\psi}$) averaged over three residues of less than 0.2 (Doreleijers et al., 1998). Ten structures are fully well defined based on these criteria and are thus not shown. The outliers, one single structure with the highest completeness and four structures with the lowest completeness (for the disordered residues), have only a small number of disordered residues.

Table 2. Observed and expected NOEs for the HU protein[a]

Distance shells (Å)	Expected NOEs	Observed NOEs 2.0–3.0	3.0–4.0	4.0–5.0	5.0–6.0	6.0–7.0	7.0–8.0	8.0–9.0	Total	Completeness %[b]	%[c]
2.0–3.0	442	173	66	49	53	5	2	0	348	79	79
3.0–4.0	1040	**1**	171	183	202	52	10	2	621	60	65
4.0–5.0	2624	0	**3**	140	488	109	26	17	783	30	43
5.0–6.0	4354	0	0	0	211	98	25	9	343	8	25
6.0–7.0	5603	0	0	0	**2**	22	11	4	39	1	15
7.0–8.0	6593	0	0	0	0	0	4	2	6	0	10
8.0–9.0	7598	0	0	0	0	0	0	2	2	0	8
Total	28254	174	240	372	956	286	78	36	2142		

[a] The NOEs are categorised into rows according to the average distance in the structure. The observed NOEs are divided into columns with respect to the measured upper-bound distance including pseudo-atom corrections. A similar table was shown in the paper describing the solution structure of the HU protein (Vis et al., 1995). The number of violated NOEs are shown in bold.

[b] Completeness per shell.

[c] Cumulative completeness.

hedral angles φ and Ψ as measured by the circular variance (Hyberts et al., 1992) to objectively distinguish between the well-defined and the disordered regions in the proteins. Leaving out these disordered regions in the previous analysis allowed a more correct comparison of the proteins. The correlation plot between the completeness for the well-defined and for the disordered residues is shown in Figure 6. For most structures the well-defined residues have a somewhat higher completeness than their disordered counterparts. The average number of observed NOEs per residue for the well-defined residues is almost twice as large as that for the disordered residues. However, the completeness when including the disordered residues is only a few percent lower than the completeness without the disordered residues. We have therefore included all residues in the calculation of the completeness in this study, as the completeness is relatively insensitive to local precision.

The arm residues in HU have a rather low number of NOEs per residue but the completeness for these residues is higher than that of the core residues (see Figure 2). The trend that disordered regions have a slightly lower completeness, as mentioned above, does apparently not hold for the flexible arm of this particular protein but is valid for the disordered N-terminus. In the case of the arm residues, the flexibility improved the linewidth of the NOE peaks resulting in a more complete set of NOEs (Vis et al., 1996).

The HU structure was recalculated in order to investigate the effect on the completeness of leaving out an increasing percentage of observed NOEs. Using 25% instead of all the observed NOEs led to an increase of the average pairwise rmsd from 0.53 to 2.71 Å. As a side effect of the reduction of the number of experimental restraints, the number of expected inter-residual NOEs decreased from 1677 to 1118. The completeness decreases, nevertheless, from 60 to 20% since the decrease in the number of expected NOEs is significantly smaller than the decrease in the number of observed NOEs. Therefore, a low completeness is a real indication of insufficient data even if the calculated structure is more disordered than the ‘true’ solution structure.

Secondary structure and relative surface accessibility

The completeness of all amino acids in the 97 proteins was investigated in relation to the consensus secondary structure and the relative surface accessibility, which were calculated using the WHATIF program (Vriend, 1990). A consensus secondary structure was obtained if more than half of the models had the same secondary structure. The relative surface accessibility was calculated using a tri-peptide (Gly-X-Gly) in vacuum as a reference.

Both the α-helical and β-sheet residues have an average completeness of 53%, which is only slightly higher than the global average of 47%. This increase might be due to the fact that some NOEs for these structural elements are more easily identified; e.g. the sequential and medium-range NOEs expected in an α-helix (Wüthrich, 1986). Another reason might be that these elements are more rigid than residues in a coil, variable region or turn and therefore less prone to line-broadening effects. Surface-exposed residues are expected to have a lower completeness than their

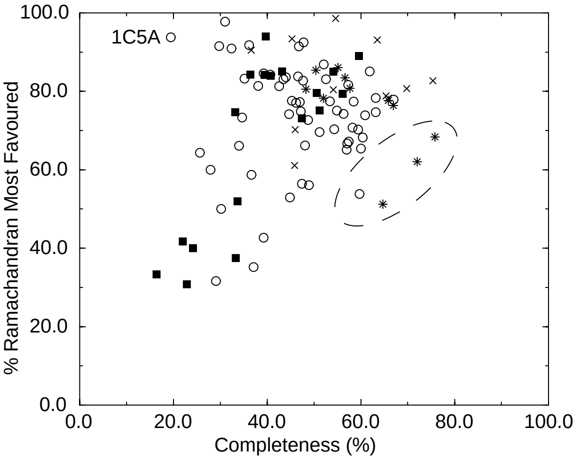

Figure 7. Completeness versus the percentage 'most favoured' in the Ramachandran plot. The percentage of residues in the most favoured region of the Ramachandran plot is calculated with PROCHECK-NMR (Laskowski et al., 1996). The three symbols, filled square, asterisk and cross, mark entries from three major laboratories. An open circle shows entries from all other laboratories.

counterparts in the core of the molecule for the same reason. However, this is not observed; the average completeness does not appear to depend on the relative surface accessibility of a residue (data not shown).

Protein mass

The smallest protein in our data set has a mass of 2.0 kDa (entry 1EDP; endothelin with 17 amino acids) and the largest protein has a mass of 19.7 kDa (entry 2BBN; calmodulin complex with 174 amino acids). Unfortunately, the completeness even for same-sized proteins displays a large variation so that no correlation of the completeness with the mass was observable. Even when differentiating for the year of the structure determination, which in turn relates to the quality of the structures, as will be shown below, no correlation was observed.

Correlation with other quality indicators

A well-established measure of the quality of protein structures (Kleywegt and Jones, 1996; Hooft et al., 1997) is the Ramachandran map, in which the (φ and Ψ) angle combinations in a structure are plotted (Ramachandran et al., 1963). The percentage of residues in the 'most favoured' area (Morris et al., 1992) is correlated with the completeness as shown in Figure 7. Although the scatter is quite pronounced, the trend is clearly positive, meaning that a structure for which the NOEs were measured more completely in general has a better Ramachandran score. The labelled entry 1C5A (porcine C5a$_{\text{desArg}}$) has severe NOE violations and low structural variance considering the number of restraints per residue (Doreleijers et al., 1998). The structure has a low NOE completeness but nevertheless a good Ramachandran score, which is an exception to the trend. There are four structures which were solved with a completeness above average but with a rather low Ramachandran score. Three of these, from the same laboratory, were solved with non-bonded interactions that do not preclude bad contacts. The fourth entry 1KST (kistrin) is somewhat unusual because the authors noted that the NOE spectrum shows little evidence for any regular secondary structure.

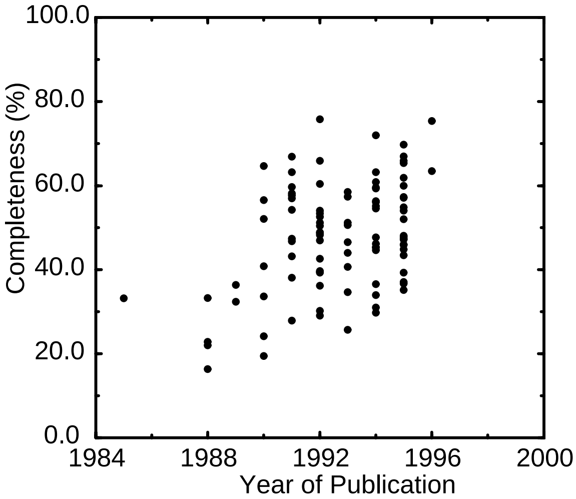

Figure 8. Correlation between completeness and year of publication of 97 structures.

NOE information is not the only source of experimental information used to obtain a good quality protein structure, as discussed above. It was found, however, that the total number of dihedral-angle and hydrogen-bond restraints per residue, averaged over the well-defined regions, does not correlate with the Ramachandran score (data not shown). This suggests that NOEs are the single most important source of information.

The completeness tends to be higher for the more recent structures, as can be seen in Figure 8. There is a significant and clear improvement up to 1992; however, the completeness does not increase much over the last four years in the current set of structures. It will be interesting to see if improved techniques and higher fields will result in higher NOE completeness.

Conclusions

The NOE completeness analysis at different cut-off distances as defined here is a simple and useful quality indicator. The average completeness values for the 97 proteins are 68, 48, and 26% up to 3, 4, and 5 Å cut-off distances, respectively. State-of-the-art structures in the current set have a high completeness of approximately 90, 75, and 45% for the same cut-off distances. These values, as well as the high completeness values for the amide protons and the class of sequential NOEs, are useful reference values for new structure determinations. Differences in residue types, secondary structure, surface accessibility and even disorder cause large differences in the number of observed NOEs but hardly influence the completeness. Hence, the completeness analysis provides another useful description of the quality of the experimental NOE data.

A considerable number of NOEs (almost 20%) correspond to inter-proton distances above 5 Å. These distances are too large to be caused by direct magnetisation transfer and are due to pseudo-atom corrections or the use of a direct refinement technique.

A positive correlation was found between NOE completeness and Ramachandran score, which favours the idea that a higher level of completeness is beneficial to the quality of the structure. The completeness of recent structures is higher than that of older structures in our data set and we expect that future structures will continue to improve as NMR techniques progress.

Acknowledgements

We acknowledge Dr. Hans Vis for providing the program SATUR and for the discussions about the HU protein. We thank the members of the EU-Validation Project for a stimulating interaction on many facets of the project. M.L.R. is funded by the BIOTECH program of DGXII of the Commission of the European Union (BIO4-CT96-0189).

References

Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.E., Jr., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) *J. Mol. Biol.*, **112**, 535–542.

Clore, G.M., Robien, M.A. and Gronenborn, A.M. (1993) *J. Mol. Biol.*, **231**, 82–102.

Doreleijers, J.F., Rullmann, J.A.C. and Kaptein, R. (1998) *J. Mol. Biol.*, **281**, 149–164.

Folmer, R.H., Hilbers, C.W., Konings, R.N. and Nilges, M. (1997) *J. Biomol. NMR*, **9**, 245–258.

Gardner, K.H., Rosen, M.K. and Kay, L.E. (1997) *Biochemistry*, **36**, 1389–1401.

Garrett, D.S., Kuszewski, J., Hancock, T.J., Lodi, P.J., Vuister, G.W., Gronenborn, A.M. and Clore, G.M. (1994) *J. Magn. Reson.*, **B104**, 99–103.

Hooft, R.W.W., Sander, C. and Vriend, G. (1997) *Comput. Appl. Biosci.*, **13**, 425–430.

Hyberts, S.G., Goldberg, M.S., Havel, T.F. and Wagner, G. (1992) *Protein Sci.*, **1**, 736–751.

Kleywegt, G.J. and Jones, T.A. (1996) *Structure*, **4**, 1395–1400.

Kleywegt, G.J. and Jones, T.A. (1997) *Methods Enzymol.*, **277**, 208–230.

Kuszewski, J., Qin, J., Gronenborn, A.M. and Clore, G.M. (1995) *J. Magn. Reson.*, **B106**, 92–96.

Laskowski, R.A., Rullmann, J.A.C., MacArthur, M.W., Kaptein, R. and Thornton, J.M. (1996) *J. Biomol. NMR*, **8**, 477–486.

Lodi, P.J., Garrett, D.S., Kuszewski, J., Tsang, M.L.S., Weatherbee, J.A., Leonard, W.J., Gronenborn, A.M. and Clore, G.M. (1994) *Science*, **263**, 1762–1767.

Markley, J.L., Bax, A., Arata, Y., Hilbers, C.W., Kaptein, R., Sykes, B.D., Wright, P.E. and Wüthrich, K. (1998) *J. Biomol. NMR*, **12**, 1–23.

Morris, A.L., MacArthur, M.W., Hutchinson, E.G. and Thornton, J.M. (1992) *Proteins*, **12**, 345–364.

Neri, D., Szyperski, T., Otting, G., Senn, H. and Wüthrich, K. (1989) *Biochemistry*, **28**, 7510–7516.

Palmer III, A.G., Williams, J. and McDermott, A. (1996) *J. Phys. Chem.*, **100**, 13293–13310.

Pearlman, D.A. (1994) *J. Biomol. NMR*, **4**, 1–16.

Pontius, J., Richelle, J. and Wodak, S.J. (1996) *J. Mol. Biol.*, **264**, 121–136.

Ramachandran, G.N., Ramakrishna, C. and Sasisekharan, V. (1963) *J. Mol. Biol.*, **7**, 95–99.

Rullmann, J.A.C. (1996) AQUA computer program, ftp://ftp.nmr.chem.uu.nl/pub/aqua.

Slijper, M., Bonvin, A.M.J.J., Boelens, R. and Kaptein, R. (1996) *J. Mol. Biol.*, **259**, 761–773.

Sutcliffe, M.J. (1993) *Protein Sci.*, **2**, 936–944.

Tjandra, N., Omichinski, J.G., Gronenborn, A.M., Clore, G.M. and Bax, A. (1997) *Nat. Struct. Biol.*, **4**, 732–738.

Vis, H., Mariani, M., Vorgias, C.E., Wilson, K.S., Kaptein, R. and Boelens, R. (1995) *J. Mol. Biol.*, **254**, 692–703.

Vis, H., Vageli, O., Nagel, J., Vorgia, C.E. and Wilson, K.S. (1996) *Magn. Reson. Chem.*, **34**, S81–S86.

Vriend, G. (1990) *J. Mol. Graph.*, **8**, 52–56.

Vriend, G. and Sander, C. (1993) *J. Appl. Crystallogr.*, **26**, 47–60.

Wagner, G., Braun, W., Havel, T.F., Schaumann, T., Go, N. and Wüthrich, K. (1987) *J. Mol. Biol.*, **196**, 611–639.

Wilson, K.S., Dauter, Z., Lamzin, V.S., Walsh, M., Wodak, S.J., Richelle, J., Pontius, J., Vaguine, A., Hooft, R.W.W., Sander, C., Vriend, G., Thornton, J.M., Laskowski, R.A., MacArthur, M.W., Dodson, E.J., Murshudov, G., Oldfield, T.J., Kaptein, R. and Rullmann, J.A.C. (1998) *J. Mol. Biol.*, **276**, 417–436.

Wüthrich, K. (1986) *NMR of Proteins and Nucleic Acids*, Wiley, New York, NY.

Wüthrich, K., Billeter, M. and Braun, W. (1983) *J. Mol. Biol.*, **169**, 949–961.