Journal of Biomolecular NMR (2005) 32: 1–12
DOI 10.1007/s10858-005-2195-0

Article

BioMagResBank databases DOCR and FRED containing converted and filtered sets of experimental NMR restraints and coordinates from over 500 protein PDB structures

Jurgen F. Doreleijers[a†], Aart J. Nederveen[b†], Wim Vranken[c†], Jundong Lin[a], Alexandre M.J.J. Bonvin[b], Robert Kaptein[b], John L. Markley[a] & Eldon L. Ulrich[a]
[a]*BioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Dr., WI, Madison, 53706, USA;* [b]*Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH, Utrecht, The Netherlands;* [c]*Macromolecular Structure Database group, European Bioinformatics Institute, Hinxton, Cambridge, CB10 1SD, UK*
†*These authors contributed equally to this work.*

Received 15 November 2004; Accepted 6 January 2005

Key words: biomolecular structure, BMRB, database, nuclear magnetic resonance, PDB, restraints

Abstract

We present two new databases of NMR-derived distance and dihedral angle restraints: the Database Of Converted Restraints (DOCR) and the Filtered Restraints Database (FRED). These databases currently correspond to 545 proteins with NMR structures deposited in the Protein Databank (PDB). The criteria for inclusion were that these should be unique, monomeric proteins with author-provided experimental NMR data and coordinates available from the PDB capable of being parsed and prepared in a consistent manner. The Wattos program was used to parse the files, and the CcpNmr FormatConverter program was used to prepare them semi-automatically. New modules, including a new implementation of Aqua in the BioMagResBank (BMRB) software Wattos were used to analyze the sets of distance restraints (DRs) for inconsistencies, redundancies, NOE completeness, classification and violations with respect to the original coordinates. Restraints that could not be associated with a known nomenclature were flagged. The coordinates of hydrogen atoms were recalculated from the positions of heavy atoms to allow for a full restraint analysis. The DOCR database contains restraint and coordinate data that is made consistent with each other and with IUPAC conventions. The FRED database is based on the DOCR data but is filtered for use by test calculation protocols and longitudinal analyses and validations. These two databases are available from websites of the BMRB and the Macromolecular Structure Database (MSD) in various formats: NMR-STAR, CCPN XML, and in formats suitable for direct use in the software packages CNS and CYANA.

Abbreviations: BMRB – BioMagResBank; CCPN – Collaborative Computing Project for NMR; DOCR – Database Of Converted Restraints; DR – Distance Restraints; EBI – European Bioinformatics Institute; FRED – Filtered REstraints Database; MSD – Macromolecular Structure Database; PDB – Protein Data Bank; RDC – residual dipolar coupling; s.d. – standard deviation.

*To whom correspondence should be addressed. E-mail: elu@bmrb.wisc.edu

Introduction

A collaborative project involving members of the BMRB, CCPN, EBI, and NMRQUAL groups has been established with the aim of improving the consistency and ease of use of publicly available experimental NMR restraint data. The first products of this collaboration, the publicly available Database Of Converted Restraints (DOCR) and Filtered REstraints Database (FRED), are reported here. The DOCR database contains restraint and coordinate data made consistent with each other and IUPAC atom nomenclature conventions. The FRED database is derived from the DOCR data sets but where various restraints defined to be redundant, inconsistent, impossible, etc. have been removed. This work builds on the BMRB MR Grid database where nearly all NMR restraint files present at the PDB have been annotated and parsed into a common format (Doreleijers et al., 2003). The utility of these new databases in recalculating and refining a large set of structures has been demonstrated and the results are described in the paper by Nederveen et al. (2004).

On March 8th, 1990 when the Protein Data Bank (PDB) (Sussman et al., 1998; Berman et al., 2000) began receiving experimental NMR data, the quantity of restraint data was small enough to fit in the header of the deposition file as remarks. As the number and complexity of the NMR restraints and associated data increased, these were placed in a separate file, named 'MR' for 'Magnetic Resonance.' In 1998, a longitudinal study investigated and interpreted a large fraction of the data available from these MR files in terms of their redundancy, completeness, agreement, and quality (Doreleijers et al., 1998, 1999a). That study examined a database of 97 structures with restraints, which we name here DB97. The DB97 database was assembled by hand, contained data for nucleic acids, and did not contain any ambiguous distance restraints (DRs). In contrast, the databases developed here (DOCR and FRED) were created using software, are restricted to proteins, and contain ambiguous distance restraints. These databases are consistent with currently accepted IUPAC atom nomenclature and other conventions (Markley et al., 1998).

The Database of REfined Solution NMR Structures (DRESS), which contains 100 protein structures (Nabuurs et al., 2004a), was the first specialized database constructed from the BMRB MR Grid Database. In creating DRESS, the BMRB staff used the Wattos software (Doreleijers et al., 2003) to parse the NMR Restraint Grid data. Next, the Utrecht group used the CcpNmr FormatConverter software (Vranken et al., 2005) to convert these to the data model developed as part of the Collaborative Computing Project for the NMR Community (CCPN) and finally exported the restraints to the CNS format.

The DOCR and FRED databases presented here include cross validation between the NMR restraints and atom coordinates. They thus go beyond prior validation efforts that have not included the NMR restraints and that are available by PDB entry code from several sources including but not limited to the PDB (Berman et al., 2000; Westbrook et al., 2003), PDBSum (Laskowski et al., 1997; Laskowski, 2001), WHAT IF (Vriend, 1990; Hooft et al., 1996; Doreleijers et al., 1999b) and reviewed by (Laskowski, 2003). This paper focuses on the methods used to construct DOCR and FRED and an analysis of the contents of the databases, a topic that was recently reviewed (Nabuurs et al., 2004b).

Methods

Data preparation

The Wattos software (Doreleijers et al., 2003) was used to parse the original restraints and to convert them to NMR-STAR format. The NMR-STAR files were then read into the CCPN data model via the FormatConverter software (Vranken et al., 2005). No assumptions or analysis of the original atom names in the restraint file was made at this point, i.e. all restraint information was linked to data model 'Resonance' objects (see (Vranken et al., 2005) for a more detailed description). The sequence was read from the original PDB file, and all relevant information for the covalent structure of the molecule was then separately read into the CCPN data model. At this stage, the restraint information was linked to the 'Resonance' objects, while all atom information was present as 'Atom' objects. The precise description of which 'Atom' object(s) a 'Resonance' object corresponds to, was then made via the 'linkResonances' procedure.

This procedure analyzes the original atom names, suggests the most appropriate naming system, and normalizes the 'Resonance' information (e.g. if 'atom names' HB2, HB3 and HB* occur for a particular residue, then all the HB* information is rearranged to HB2 and HB3 Resonances). The appropriate Resonance to Atom link is then made through two other objects: the 'ResonanceSet', which describes ambiguity of the atom assignment, and the 'AtomSet', which groups NMR equivalent protons (e.g. HB1, HB2 and HB3 for an alanine). When this step is completed, the connection between the 'Resonance' and the actual 'Atom' is described unambiguously.

A special dictionary was developed to automate this procedure: if necessary, it maps the residue numbering in the restraint file to the numbering in the PDB file, and maps, in addition, particular atom names that would otherwise not be recognized. In practice, this means that, given a complete dictionary the whole procedure can be repeated automatically without any user intervention.

The stereospecificity present in the original data was retained during this conversion. Original atom names that were not recognized by the reference naming system, and that were not specifically defined in the dictionary, were ignored: restraints with Resonances corresponding to such atom names are therefore not included in the database. Their information is, however, conserved in the NMR-STAR and CCPN XML files.

These converted restraints were then exported using the FormatConverter as CNS restraint files, and a violation analysis against the PDB entry coordinates was performed. The original PDB entry heavy atom coordinates were retained for this step, but proton coordinates were recalculated using CNS (Brünger et al., 1998) following the protonation state defined in the original PDB entry. The resulting coordinate files were then read into the CCPN Data Model via the FormatConverter using a default mapping for the CNS atom names, and written out with the original sequence and restraint information as NMR-STAR files. The restraints can be exported in CYANA format as well, but for technical reasons, restraints for dihedral angle restraints defined in part by hydrogen atoms were not converted to the CYANA format. This affected 92 entries. This problem will be solved in a new release of the CCPN software. These restraint and coordinate sets constitute the DOCR database.

The FormatConverter has been used to convert the final datasets for both the DOCR and FRED databases automatically and unambiguously to formats that can be read by two of the most commonly used software packages for NMR structure determination, CYANA (Güntert et al., 1997) and X-PLOR/CNS (Brünger, 1996; Brünger et al., 1998). The converted restraints are also available in the NMR-STAR and XML (CCPN data model version 1.0.107) files, which are a superset of the data expressed in the CYANA and CNS formats.

Stereospecific data interpretation

For FRED, the stereospecificity present in the original DRs was only inverted in cases where the NOE energy, calculated over all restraints involving a given stereospecific atom or group of atoms, was lower after swapping in more than 75% of the models. Stereospecific atoms were then deassigned on an individual restraint basis when giving rise to violation over 2 Å in one of the models or to violations over 1 Å in more than half the models. The sum averaging method (Nilges, 1993) was used for the analysis of the DRs. This averaging might differ from the method used by the authors of the original structures. However, sum averaging delivers relatively shorter distances than other methods, so that upper-bound violations not seen by the authors are not expected.

DR surplus

Surplus restraints in a PDB MR file are those that are exceptional, double, impossible, fixed, or redundant, and are considered to be of either of no consequence or possibly detrimental in carrying out a structure calculation. Detailed definitions for these restraints collectively described as 'surplus' are given below, and algorithms have been developed to identify these restraints. The implementation of surplus in Wattos is partly based on the redundancy checks that the Aqua software is able to perform on unambiguous restraints (Doreleijers et al., 1998). The differences with Aqua's redundancy check are addressed below.

Definitions used:

r_{low}: lower bound distance of restraint.

r_{upp}: upper bound distance of restraint.

t_{low}: smallest distance possible in theory; i.e. if only dihedral angles are rotated.
t_{upp}: largest distance possible in theory.

A node is equivalent to what is called a restraint contribution in ARIA (Linge et al., 2001) and its definition in the NMR-STAR schema is detailed elsewhere (BMRB, 2004). An example of a multi node restraint is given at the end of this section. In the DRs handled here, only one set of distances is associated with one DR, even though multiple sets could be specified in the NMR-STAR schema.

A list of author defined DRs (including any subtype for example, hydrogen bonds and disulfide bridge defining restraints) is partitioned into the following sets:

- *U* Universe of all DRs in author deposition list.
- *Q* Unparsed restraints (e.g., syntax or grammar of restraint was unclear). Note that it might be hard to know exactly how many restraints an unparsed piece of text contains. Elements in *Q* exist as parse errors in the NMR-STAR files released earlier (Doreleijers et al., 2003).
- *A* Unmatched ('unlinked' in CCPN jargon) restraints. For example, a restraint containing an atom name 'abracadabra'. Elements in *A* exist as conversion errors in the NMR-STAR files of DOCR.
- *E* Exceptional restraints that according to CCPN linking are present but could not be found by Wattos (e.g., 'H*' for the N-terminal residue's atoms named H1, H2, H3).
- *C* Restraints for which no coordinates are present in the coordinate file.
- *D* Double restraints. As a side effect of partitioning restraints into this set, all restraints that were not partitioned before are simplified and possibly combined.
  - A restraint between the same atoms, say A-A- is considered double. Perhaps a better term would be 'corresponding to a diagonal peak' in the case of an NOE based DR. The restraint is also flagged as double if it has even one such pair as above (e.g. A, B–A).
  - A–B, B to just A–B, or (A–B or A–B) to just A–B, or (A–B or A–C) to just A–B, C and many more complicated cases to simplify the number of pairs and atoms listed. This type of check is not present in Aqua because it is specific to ambiguous restraints and is new to Wattos.
  - A–B if there was already a restraint A–B before. Combine a restraint A–B with r_{low}, r_{upp} (see definitions below for the meaning of r_{low} etc.) of 3 and 5 Å, respectively with a restraint HA–HB with r_{low}, r_{upp} of 4 and 6 Å to a restraint having the tightest bounds (representing all information in both restraints) i.e. HA–HB with r_{low}, r_{upp} of 4 and 5 Å.
- *I* Impossible restraints. Restraints that are incompatible with the range of distances allowed by the molecular topology (t_{low}-t_{upp}). If the lower bound is above, the upper bound below, or the target distance outside that range, then the restraint is classified as impossible. If a model exists for the structure, the largest diameter in the first model will determine the t_{upp} for the distances not in the dictionary of known theoretical distances derived from Aqua for all twenty common amino acids (Laskowski et al., 1996). If the distance is not found in the same dictionary, the t_{low} is set to the sum of the van der Waals radii −0.2 Å (e.g. 1.8 Å for two hydrogen atoms). If the element type is not known, then t_{low} is assumed not to exist. So neither t_{low} nor t_{upp} always have to exist but in most cases are present.

Before this and subsequent classifications (fixed and redundant as described below) the following corrections are applied to overcome technical details:

- Lower bound at or below averaged sum of involved atom radii (e.g. 1.8 Å for two hydrogen atoms) is considered not to exist.
- Upper bound above the diameter of the structure is considered not to exist. These bounds are often specified as very large values for DRs with sub type: 'NOE not seen'. For this class the upper bound has no implications.
- The above two checks are repeated for the target distance.

Restraints that have none of the following: lower bound, target, upper bound, intensity, or volume will also be marked as impossible.

This part of the check is not done per node in Aqua because only one node needs to exist in an unambiguous DR. A node is impossible if: $(r_{low} > t_{upp}) \vee (r_{upp} < t_{low}) \vee (r_{low} > r_{upp}) \vee (r_{tar} < t_{low}) \vee (r_{tar} > t_{upp})$ provided that the involved terms exist

after checking the above. A restraint is impossible if any of its nodes are impossible.

- F Fixed restraints. Restraints between atoms that have no variability in their distance if only dihedral angles are allowed to rotate. In that case the target distances are the same: $t_{low} = t_{upp}$ and both terms exist.
- R Redundant restraints. Restraints that do not add restrictions on the distance between the atoms in addition to the molecular topology if only dihedral angles are allowed to rotate. A 'threshold of redundancy' parameter (TR) is introduced to allow restraints that are on the edge of being redundant to remain non-redundant. The default setting in Wattos is 5%. Only if $r_{low} \leq (1 - TR)t_{low}$ then r_{low} is redundant and only if $r_{upp} \geq (1 + TR)t_{upp}$ then r_{upp} is redundant. A restraint is only redundant if all three distances are either not present or are redundant. This is in contrast to the set of impossible restraints (see above) for which only one distance needs to be impossible in order to qualify the whole restraint as impossible.
- N Non-redundant restraints. The remaining restraints as defined below. Properties of the sets: $U = Q \cup A \cup E \cup C \cup D \cup I \cup F \cup R \cup N$ and sets are mutually disjoint except with U (e.g. No element in Q is in A). A set S is defined to consist of the restraints in all sets that do not add information to a structure calculation and are surplus ($S = E \cup C \cup D \cup I \cup F \cup R$). Note that restraints in sets Q and A are not included in set S because they could not be parsed or could not be analyzed to atomic detail in relation to the structure which makes it impossible to denote them as surplus for certain.

As part of the redundancy check the components in the restraints are reordered. Within one member, the 'smallest' atom is the first atom. Within one node, the member with the 'smallest' atom is the first member. Within one restraint, the node with the member with the smallest atom is the first node (after the logical node). An atom is ranked 'smallest' if it occurs first in the model, (e.g. HN of residue 1 comes before HA of residue 2). As a more complex example (ignoring all but atom names for brevity):

HC – HB or

HD, HA – HB

is reordered to become:

HA, HC, HD – HB

NOE distance completeness

The calculated completeness can be used in the initial phases of NMR structure determination by focusing on NOE contacts in specific regions in a biomolecule or pinpointing problems to specific residues, atoms or classes of NOE contacts. The completeness check and its application to DB97 was previously reported (Doreleijers et al., 1999a).

Based on one or more models of a protein structure, a set of contacts expected to be observable in a NOESY type NMR experiment is generated. The intersection of the set of NOE contributions (set A) and the set of observable model contacts (set B) contains the matched contacts (set M). Completeness is defined as the ratio between the number of the matched contacts and the number of observable model contacts (Completeness $= |M|/|B|$ as specified below). The values of the lower bound, target, and upper bound DRs are not considered in this analysis.

Differences with completeness check in Aqua

In the case of an ambiguous NOE, the NOE is considered by its contributions. For example, one restraint with two contributions can match two expected contacts. So in contrast to the old procedure in Aqua that does not deal with ambiguity, this procedure can lead to 100% completeness, even if there are fewer restraints than expected contacts. Since Wattos does not have access to the chemical shifts of the individual spins, it cannot check if the ambiguity is reasonable. For example, if just one ambiguous constraint contains all atoms on either side (in CNS such a DR, ignoring the distance specification, could be expressed as: assign (name*) (name*)) which will result in a 100% completeness because it is matched to any theoretically expected contribution. This type of situation should be checked and corrected using additional information like chemical shift assignments before the completeness check is done.

In Aqua, the centered position of the pseudo atom was used for deriving the expected contacts whereas here the averaging method can be selected

to be one of center, sum, or R-6 (Brünger et al., 1998). The sum averaging is the most physically correct method. For PDB entry 1BRV the completeness up to 4 Å without intra-residue contacts and using center averaging is 64% with Aqua and Wattos. For sum averaging the completeness drops to 57%.

Sets used for calculation

The completeness calculation in Wattos runs only on the selected atoms. Nuclei known to be unobservable can be excluded from the analysis. A list of DR contributions (not the restraints themselves) is partitioned into the following sets:

- U Universe of contributions in selected restraints.
- V Universe of contributions that theoretically are expected to be shorter than a threshold.
- $W = U \cup V$
- E Exceptional restraints; those restraints that contain one or more atoms that could not be matched to an atom in the coordinate section.
- O Not observable contributions (e.g. an NOE contribution with Ser HG).
- I Intra-residue contributions if not to be analyzed (optional).
- S Surplus like double contributions. A part of the effect of transforming the experimentally observed contributions into the same domain as the theoretically expected contributions will lead to double contributions that need to be taken out. If intra-residue contributions are to be analyzed for completeness, this will also filter out the redundant ones. See previous section on DR surplus. The following sets are defined:
- $A = U - (E \cup O \cup I \cup S)$ The set of observable experimental distance contributions.
- $B = V - (I \cup S)$ The set of observable theoretical distance contributions that are shorter than a threshold. Note that no elements in E and O will occur in V.
- $M = A \cap B$ The set of matched distances, i.e. those for which both an experimental and theoretical contact exists.
- $C = A - M$ Unmatched experimental NOE contributions (increasing the threshold distance decreases this set in size) but many contributions (in case of ambiguous restraints) will end up in this set as their contribution fraction is too small.
- $D = B - M$ Theoretically expected but unseen NOE contributions. Set D is useful to check against the input data to explain why contributions could be absent. Many valid reasons for completeness below 100% occur in NOE DR lists based on real spectra.

Results

MR Grid database overall composition

Figure 1 presents the breakdown of the data available in the MR files for 1937 PDB entries available on June 7, 2004 (including entries in

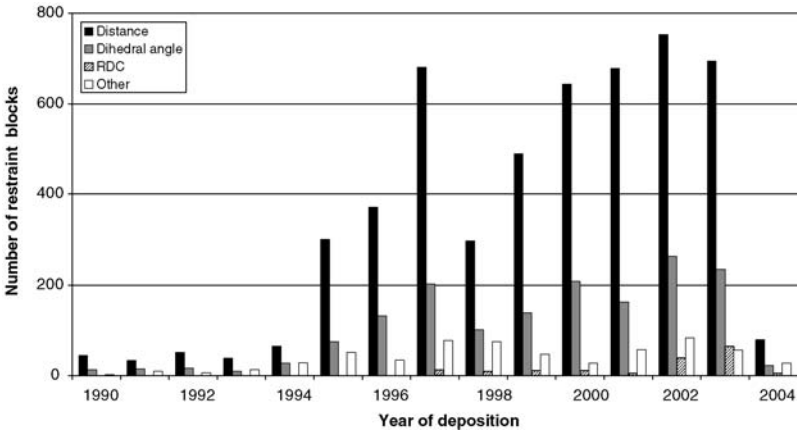

Figure 1. Breakdown of the data available in the MR files for 1937 released PDB entries available on June 7, 2004 (including entries in DOCR and FRED) by year and number of blocks for the different restraint types. Please note that this figure and Figure 2 contain many more entries than the DOCR and FRED databases do because the selection date and the criteria are different.

DOCR and FRED) by year and number of blocks representing the different restraint types. A block is a unique set of restraints of a single type recognized in the linear text of a PDB MR file. An MR file may contain a single block of restraints or a larger number of blocks representing different types of restraints, sets of similar restraints derived from different NMR experiments, etc. In general, more blocks mean more data, but the blocks might also be of smaller size. The significant decreases in years 2003 and 2004 can be attributed to the fact that the *x*-axis shows the year of deposition and only released entries are included. In 1998 the percentage of entries with MR files was anomalously low with respect to the previous year (data not shown), and hence the number of blocks was also significantly less. Interestingly, the 1998 entries have many restraints per entry as shown in Figure 2. Although distance, and to a lesser extent dihedral angle, restraint blocks dominate the distribution, it should be noted that the category of residual dipolar coupling (RDC) restraints has increased significantly over the last period (also see Figure 2) without decreasing the first types of restraints. The 'Other' category contains data types such as angles, chemical shifts, chemical shift anisotropies, planarities (mostly for nucleic acid base pairs), and pseudocontact shifts.

The average number of restraints per entry by year and restraint type is shown in Figure 2. The numbers shown have not been filtered to remove

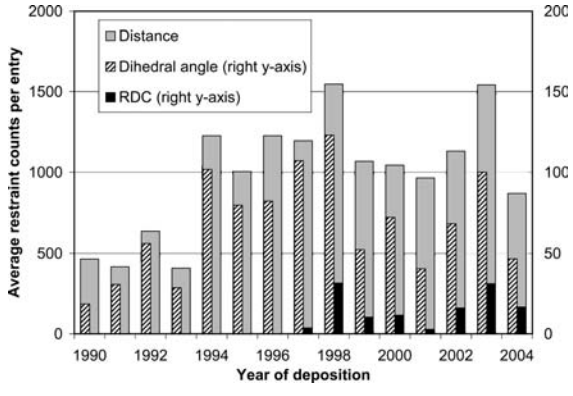

Figure 2. Average number of restraints per entry by year and restraint type. On the left y-axis the average number of DRs per entry for each year is shown whereas the same is shown on the right axis for dihedral angle and RDC restraints. Note that for the averages only entries were included for which they are present, which are relatively few for RDCs as shown in Figure 1.

surplus restraints (exceptional, double, impossible, fixed, and redundant restraints) as was done in creating the FRED database. After an approximate doubling of DRs per year for the year 1994, the average count remained more or less constant. RDCs were first deposited in the year 1997, and showed gains in 1998. They are on the rise again since 2002.

DOCR and FRED database set selection

The purpose of this project was to create a database of restraints that can be used directly for structure validation and recalculation. All entries were included that could be parsed with available software; however, the many entries not meeting these criteria were left out. Software packages used in this study were: CNS, CYANA, FormatConverter, Wattos, and in-house code developed by the Macromolecular Structure Database (MSD), CCPN, BMRB, and Utrecht groups. Out of 1083 PDB protein entries of NMR origin with parsed restraints that were available from the BMRB on October 13, 2003, the final selected set (those making up the 'DOCR' database) contained 545 entries (Table 1 in Supplementary Material). The 'FRED' database was created from the DOCR database after filtering the data as described below. In a separate parallel study, the FRED data were used to recalculate and validate the PDB entries (Nederveen et al., 2004). Here we focus on the analysis of the experimental data in DOCR (as deposited by the authors) and in FRED (following filtering).

DOCR DRs filtered for FRED

The converted distance and dihedral angle restraint lists were checked against the hydrogen positions in DOCR as recalculated from the PDB coordinates. Stereospecificity present in the original DRs was only inverted in cases where swapping lowered the NOE energy in more than 75% of the conformational models. This procedure affected the DRs in 274 entries, and resulted in a decrease of the average number of consistent violations above 0.5 Å in these entries from 14 to 10 per entry. The stereospecificity in restraints was then deassigned if violated more than 2 Å in any of the models or over 1 Å in more than half of them. This led to modifications on 117 entries and in this set resulted in a decrease in the

Table 1. Average quality indicators of proteins in the Filtered REstraints Database (FRED).

	Quantity and standard deviation
RMS DR violations (Å)	0.08 ± 0.14
RMS dihedral restraint violations (degrees)	1.6 ± 4.6
NOE completeness (% ≤ 4 Å, inter-residue)	40 ± 11

number of consistently violated restraints over 0.5 Å from 26 to 19.

The DRs were then analyzed by a new implementation of Aqua (Doreleijers et al., 1998) in Wattos that allows the analysis of ambiguous restraints (see Methods). Wattos identifies and can filter out exceptional, double, impossible, fixed, and redundant restraints, collectively denoted as surplus restraints (described above). These categories are shown in Figure 3 for all entries. They were removed because they do not contain information beyond what was already known from molecular topology or, for the category of 'impossible', was in conflict with it. The remaining non-redundant DRs, together with the unfiltered dihedral angle restraints form the basis of FRED.

FRED DR classes

In Figure 4 the entries in FRED have been sorted with respect to their percentage of DRs in various classes. The class 'mixed' denotes ambiguous restraints in which the contributions fall in different classes. The other classes are described in the legend. A large difference is observed between entries in the treatment of non-redundant intra-residue DRs. For some entries no intra-residue NOE DRs were apparently used, whereas in other entries over half of the restraints are in this class. A subset of entries shows no long range NOE DRs because they are not expected for proteins without tertiary structure. These differences were previously discussed for DB97 (Doreleijers et al., 1998).

NOE distance completeness

The NOE DRs in FRED were analyzed for NOE completeness using a new implementation of Aqua (Doreleijers et al., 1999a) in Wattos that allows ambiguous restraints to be analyzed (see Methods). The number of restraints per residue as a function of the deposition year shown in Figure 5, indicates a slight increase over the years. This trend is no longer present when considering the trend in NOE completeness up to 4 Å for inter-residue restraints (Figure 6) based on the default

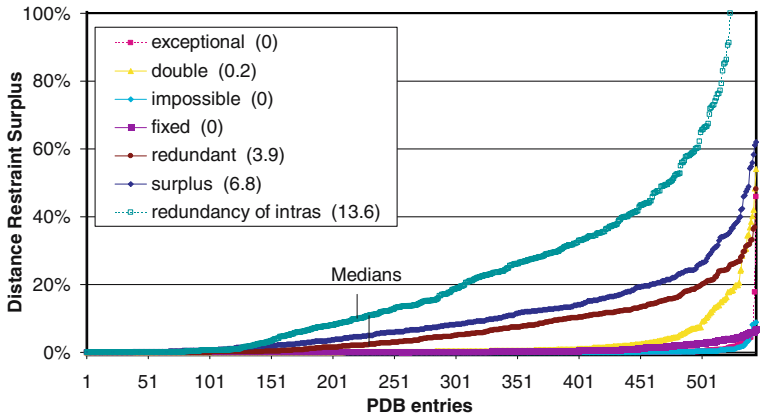

Figure 3. Surplus in the original DRs for all 545 entries sorted independently for each category. The results of the surplus check showing the top five categories (exceptional, double, impossible, fixed, and redundant) in the legend and the sum of them listed under 'surplus.' The series labeled with 'redundancy of intras' shows the redundant fraction of the intra-residue DRs for 524 of the 545 entries that have intra-residue DR. The median positions are indicated in the graph and their values are listed in the legend between brackets for each category. The PDB entry at 100% is unusual as it only has 1 intra-residue DR and it is redundant. The entry just below that has 811 intra-residue restraints at 91% redundancy because all its upper bounds were set to 5 or 6 Å. Those upper bounds are then often redundant with the molecular topology. The top entry in the 'double' series has more than half of the restraints double because the restraint list was duplicated in its entirety. The top entry in the 'exceptional' series has almost half the restraints present for large parts of the described molecule that have no coordinates in the PDB file.

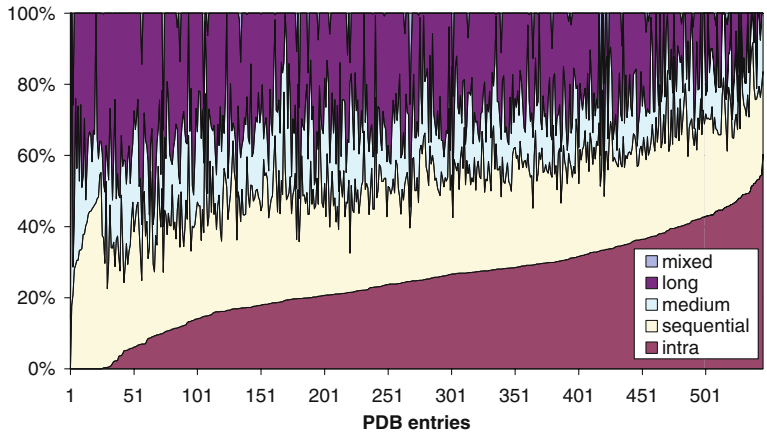

Figure 4. Classification of DRs in FRED. The entries have been sorted with respect to the percentage of DRs in the non-redundant intra-residue class. The other classes are sequential, medium ($1 < i < 5$), long ($i > 4$) and 'mixed' that denotes ambiguous restraints in which the contributions fall in different classes. A large difference between entries is observed in the treatment of non-redundant intra-residue DRs. The surplus that might have existed in the lists was removed before this analysis.

set of expected protons: overall, no clear correlation is observed between the year of deposition and NOE completeness. The NOE completeness calculated here using 'sum averaging' averages to 40 ± 11% for FRED (Table 1) which is, as expected, slightly lower than the average of 48 ± 13% from the previously calculated values using 'center averaging' for DB97. As detailed in the Methods section this is due to the larger number of expected NOEs using the 'sum averaging' method.

DR and dihedral angle restraint violations

In FRED, the average RMS violations for distance and dihedral angle restraints are 0.08 ± 0.14 Å and 1.6 ± 4.6 degrees, respectively (See Table 1). For comparison, the average RMS violation of DB97's DRs was 0.06 ± 0.04 Å. A small number of entries in FRED have significantly higher RMS violations than the entries in DB97 which causes the standard deviation (s.d.) in FRED to be higher than the average value.

Reformatted restraints

The BMRB, in collaboration with the NMR community and the Collaborative Computing Project for NMR (CCPN) (Fogh et al., 2002) is developing the next version of the NMR-STAR data dictionary (BMRB, 2004). Many programs use the NMR-STAR format for exchanging experimental NMR data. The program Wattos was used to parse data (using JavaCC) to a developmental predecessor of NMR-STAR version 3. The MR Grid database (Doreleijers et al., 2003) originally used the NMR-STAR dictionary version 2. All three databases: the parsed data sets, DOCR, and FRED, available in the NMR Restraint Grid user interface now adhere to the 'developmental predecessor of NMR-STAR version 3' and will be updated to the final version 3 data dictionary when released.

The CCPN XML files are organized by entry and by 'package.' Each 'package' contains a set of data that is logically grouped together (e.g. 'Molecule,' 'NmrConstraints,' etc.). Only the 'project' XML file in the top directory has to be read in order to find the access points to the information stored in the XML files below. The information in these files can be read automatically when needed.

Availability of data and software

The results of this study are available from the BMRB (http://www.bmrb.wisc.edu/servlets/MRGridServlet) and from the EBI (http://www.ebi.ac.uk/msd/recoord). Access to the DOCR and FRED databases at BMRB is provided through the same interface used for the parsed MR files (Doreleijers et al., 2003). The entries have been linked to related BMRB entries

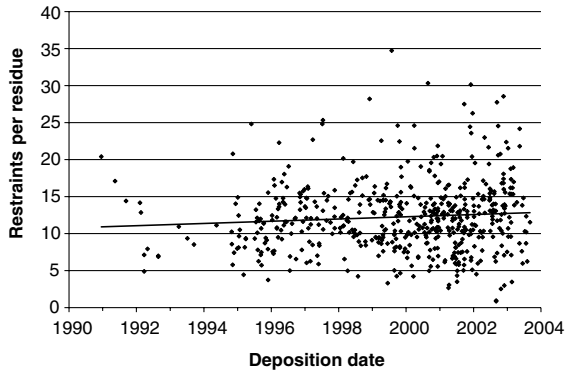

Figure 5. Trend of number of restraints per residue in FRED over deposition date. Overall there is a slightly rising number of restraints per residue (the combined sum of any distance and dihedral angle restraints) but as expected with the increase of depositions, the number of outliers on both ends has gone up too. In recent years though, we see an entry with 34.7 in 1999 and two entries in 2002 (overlapping) with less than 1 restraint per residue. For the latter entries, an incomplete set of restraints was deposited though.

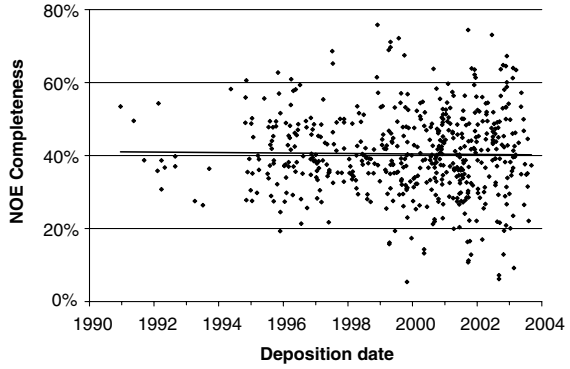

Figure 6. Trend of the NOE completeness up to 4 Å for inter-residue restraints in FRED using the default set of expected protons.

containing assigned chemical shifts and other information.

Discussion and conclusions

It is in the NMR community's best interest and crucial for the impact of many bioinformatics projects that NMR experimental data become more accessible and available in common formats. We have successfully prepared a large subset of the NMR restraint data deposited by authors. Although filtered data sets are available, users are invited to work with the unfiltered data and to critically evaluate the choices that need to be made before the data can be used for any particular study.

Comparison between the RMS distance violations in DB97 and FRED

In Figure 7 the RMS distance violations are compared for 23 entries with original and recalculated hydrogen atom positions. The original values from DB97 (Doreleijers et al., 1998) were calculated using Aqua (Laskowski et al., 1996) and the same averaging scheme used by the authors (center or sum averaging). In FRED, sum averaging has been used exclusively. The restraint data themselves are also not necessarily the same in DB97 and FRED, because they have been processed in different ways. In DB97, the restraints were deassigned based on an analysis over all

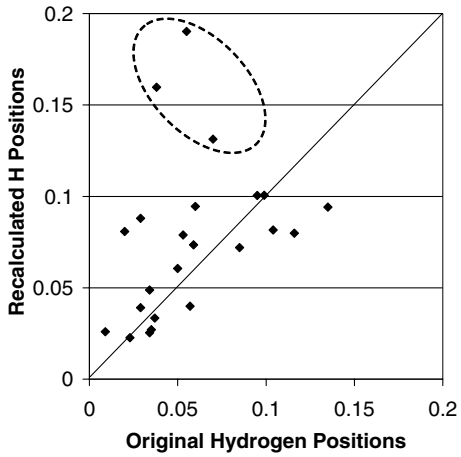

Figure 7. Comparison between the RMS distance violations calculated for 23 entries with original and recalculated hydrogen atom positions. The original values from DB97 (Doreleijers et al., 1998) were calculated using Aqua (Laskowski et al., 1996) using the same averaging scheme as the authors have used (center or sum averaging) whereas for the recalculated values the later was always used. The restraint data themselves are also not necessarily the same as they have been processed in different ways (see Methods). The three entries enclosed in the ellipse have significantly higher values when using the recalculated hydrogen positions. They were solved with three different structure calculation packages (CNS, Discover, and DYANA). It seems that the force fields used, allowed the hydrogen atoms to deviate from standard geometry under the influence of DRs so that when put back into the standard geometry the violations became significantly higher.

restraints, whereas for FRED the decision was made per restraint. Three entries (enclosed by an ellipse in Figure 7) have significantly higher values for hydrogen positions in FRED than in DB97. These three entries were solved with three different structure calculation packages (CNS, Discover, or CYANA, respectively). It seems that all three force fields used allowed the hydrogen atoms to deviate from standard geometry under the influence of DRs, so that when they were put back into the standard geometry the violations became significantly higher.

Future perspectives

For practical reasons the initial version of the DOCR and FRED databases did not use the original coordinates for hydrogen atoms when present. By using WHAT IF (Vriend, 1990; Doreleijers et al., 1999b) or the FormatConverter (Vranken et al., 2004) it would be possible to retain those coordinates and to allow for a better analysis of the restraints defined on the basis of these proton positions. The current stereospecific deassignment strategy analyzes and if needed modifies one restraint at a time. In the next iteration of the FRED database, a scheme will be used for stereospecific analyses that can make adjustments based on an overall analysis, such as the scheme utilized for the unambiguous restraints in DB97 (Doreleijers et al., 1998).

The data in the MR Grid database are organized to the level of blocks of restraints as listed by the authors. The blocks are recognized by BMRB's staff and split by the Wattos software (Doreleijers et al., 2003) in the same order as they appear in the DOCR database. In FRED however, the data are regrouped in order to have one list for NOEs and disulfide bonds, one list for hydrogen bonds, and one list for dihedral angles, if present. Although the regrouping facilitated automatic handling of the data in our setup for structure recalculations, it is clear that this ignores an important aspect of the information, and the original separation is more useful for analyses *a posteriori*, e.g. which list/spectrum has the most violations, could a different weighting scheme over the lists (disulfide bonds/NOEs) be beneficial, etc.

The BMRB plans to extend the databases presented here by increasing the number of entries included, the variety of biomolecules (inclusion of nucleic acids is of high importance), the variety of data types (e.g. RDCs which were converted but not analyzed for this project), and the variety of data sources (e.g. the data generated for use in the AMBER and EMBOSS programs). In addition, non-standard residues should be included: the 40 entries that were discarded for this study because they contained NH2/ACE as the only non-standard residues could easily be included in the next version of the database.

Supplementary material to this paper is available in electronic format at http://dx.doi.org/10.1007/s10858-005-2195-0.

Acknowledgements

Most importantly we acknowledge all authors for contributing the results of their scientific investigations to the PDB. They created the true resource that this secondary database relies on. This work was supported in part by grant LM05799 from the National Library of Medicine, National Institutes of Health (NIH), and by grant QLG2-CT-2000-0313 from the European Community program NMRQUAL.

References

Berman, H.M. Westbrook, J. Feng, Z. Gilliland, G. Bhat, T.N. Weissig, H. Shindyalov, I.N. and Bourne, P.E. (2000) *Nucleic Acids Res.* **28**, 235–242.

BMRB NMR-STAR data dictionary (2004), http://www.bmrb.wisc.edu/dictionary/htmldocs/nmr_star/dictionary.html.

Brünger, A.T. (1996) in *X-PLOR Manual* (Version 4.0). Dep of Molecular Biophysics and Biochemistry, Yale University.

Brünger, A.T. Adams, P.D. Clore, G.M. DeLano, W.L. Gros, P. Grosse-Kunstleve, R.W. Jiang, J.S. Kuszewski, J. Nilges, M. Pannu, N.S. Read, R.J. Rice, L.M. Simonson, T. and Warren, G.L. (1998) *Acta Cryst.*, **D54**, 905–921.

Doreleijers, J.F. Mading, S. Maziuk, D. Sojourner, K. Yin, L. Zhu, J. Markley, J.L. and Ulrich, E.L. (2003) *J. Biomol. NMR*, **26**, 139–146.

Doreleijers, J.F. Raves, M.L. Rullmann, T. and Kaptein, R. (1999a) *J. Biomol. NMR*, **14**, 123–132.

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

Doreleijers, J.F. Vriend, G. Raves, M.L. and Kaptein, R. (1999b) *Proteins*, **37**, 404–416.

Fogh, R.H. Ionides, J. Ulrich, E.L. Boucher, W. Vranken, W. Linge, J. Habeck, M. Rieping, W. Bhat, T.N. Westbrook, J. Henrick, K. Gilliland, G. Berman, H. Thornton, J.M. Nilges, M. Markley, J.L. and Laue, E. (2002) *Nat. Struct. Biol.*, **9**, 416–418.

Güntert, P. Mumenthaler, C. and Wüthrich, K. (1997) *J. Mol. Biol.*, **273**, 283–298.

Hooft, R.W.W. Vriend, G. Sander, C. and Abola, E.E. (1996) *Nature*, **381**, 272 .

Laskowski, R.A. (2001) *Nucleic Acids Res.*, **29**, 221–222.

Laskowski, R.A. (2003) *Methods Biochem. Anal.*, **44**, 273–303.

Laskowski, R.A. Hutchinson, E.G. Michie, A.D. Wallace, A.C. Jones, M.L. and Thornton, J.M. (1997) *Trends. Biochem. Sci.*, **22**, 488–490.

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

Linge, J.P. O'Donoghue, S.I. and Nilges, M. (2001) *Methods Enzymol.*, **339**, 71–90.

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.

Nabuurs, S.B. Nederveen, A.J. Vranken, W. Doreleijers, J.F. Bonvin, A.M. Vuister, G.W. Vriend, G. and Spronk, C.A. (2004a) *Proteins*, **55**, 483–486.

Nabuurs, S.B. Spronk, C.A. Vriend, G. and Vuister, G.W. (2004b) *Concep. Magnetic Res.*, **22**, 90–105.

Nederveen, A.J., Doreleijers, J.F., Vranken, WF., Miller, Z., Spronk, C.A.E.M., Nabuurs, S.B., Güntert, P., Livny, M., Markley, J.L., Nilges, M., Ulrich, E.L., Kaptein, R. and Bonvin, A.M.J.J. (2005) *Proteins*, **59**, 662–672.

Nilges, M. (1993) *Proteins*, **17**, 297–309.

Sussman, J.L. Lin, D. Jiang, J. Manning, N.O. Prilusky, J. Ritter, O. and Abola, E.E. (1998) *Acta. Cryst.*, **D54**, 1078–1084.

Vranken, WF., Boucher, W., Stevens, T., Fogh, R.H., Pajon, A., Llinás, M., Ulrich, E.L., Markley, J.L., Ionides, J. and Laue, E. (2005) *Proteins*, **59**, 687–696.

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

Westbrook, J. Feng, Z. Burkhardt, K. and Berman, H.M. (2003) *Methods Enzymol.*, **374**, 370–385.