



BioMagResBank database with sets of experimental NMR constraints corresponding to the structures of over 1400 biomolecules deposited in the Protein Data Bank

Jurgen F. Doreleijers^a, Steve Mading^a, Dimitri Maziuk^a, Cassandra Sojourner^a, Lei Yin^b, Jun Zhu^c, John L. Markley^a & Eldon L. Ulrich^{a,*}

^aBioMagResBank, Department of Biochemistry, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI 53706, U.S.A.; ^bKeithley Instruments Inc., 28775 Aurora Road, Cleveland, OH 44139, U.S.A.; ^cDepartment of Animal Health and Biomedical Sciences, University of Wisconsin-Madison, 1656 Linden Drive, Madison, WI 53706, U.S.A.

Received 6 December 2002; Accepted 12 February 2003

Key words: biomolecular structure, BMRB, constraints, database, dihedral angle, nuclear magnetic resonance, nuclear Overhauser effect, residual dipolar coupling, structure

Abstract

Experimental constraints associated with NMR structures are available from the Protein Data Bank (PDB) in the form of 'Magnetic Resonance' (MR) files. These files contain multiple types of data concatenated without boundary markers and are difficult to use for further research. Reported here are the results of a project initiated to annotate, archive, and disseminate these data to the research community from a searchable resource in a uniform format. The MR files from a set of 1410 NMR structures were analyzed and their original constituent data blocks annotated as to data type using a semi-automated protocol. A new software program called Wattos was then used to parse and archive the data in a relational database. From the total number of MR file blocks annotated as constraints, it proved possible to parse 84% (3337/3975). The constraint lists that were parsed correspond to three data types (2511 distance, 788 dihedral angle, and 38 residual dipolar couplings lists) from the three most popular software packages used in NMR structure determination: XPLOR/CNS (2520 lists), DISCOVER (412 lists), and DYANA/DIANA (405 lists). These constraints were then mapped to a developmental version of the BioMagResBank (BMRB) data model. A total of 31 data types originating from 16 programs have been classified, with the NOE distance constraint being the most commonly observed. The results serve as a model for the development of standards for NMR constraint deposition in computer-readable form. The constraints are updated regularly and are available from the BMRB web site (<http://www.bmrwisc.edu>).

Introduction

The number of macromolecular structures determined from NMR spectroscopic data has increased sharply over the last five years. The Protein Data Bank (Berman et al., 2000; Bernstein et al., 1977) already contains 2986 coordinate entries of NMR origin, approximately 16% of the total number of entries. At the time of writing, it is common practice to sub-

mit NMR experimental data (chemical shifts, coupling constants, relaxation parameters, etc.) to the BioMagResBank (BMRB URL: <http://www.bmrwisc.edu>) (Ulrich et al., 1998; Seavey et al., 1991) and to deposit (as a separate submission) the atomic coordinates and constraints used for structure calculation with the PDB. Of the 2986 NMR entries, only 1410 have associated deposited constraints. The deposited constraint lists allow groups worldwide to reproduce structures and to develop and test protocols for structure calculation and structure refinement. However, two issues

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

need to be resolved before such studies can become feasible on a large scale. First, the data are not annotated by the PDB (or authors) in a standard way to provide the constraint type, the computer program associated with it, or the way the constraints were calculated. As an example, even though two sets of data might be of the same type and used in the same program, the restraining force constants and potential functions as used in the calculation may have been different. Because distinguishing features are frequently undocumented or provided as free text, a simple computer algorithm cannot assign the boundaries between separate constraint lists. Second, a multiplicity of data formats are in use nowadays, and few software packages have the ability to use an inclusive set of these as input. For example, the formats for ambiguous distance constraints used by XPLOR/CNS (Brünger et al., 1998) and ARIA (Linge et al., 2001) are recognized uniquely by these software packages. These two issues, the parsing of legacy PDB MR files and the development of a standardized interchange format for distance constraints, are addressed here through the development of the software program Wattos.

Methods

Software design

The Wattos software utilizes a multi-tiered architecture, and was programmed in Java v.1.3 Enterprise Edition, a platform independent language. The code was developed under Windows 2000 and Linux, in both cases under the Forte For Java v.3 Integrated Development Environment (IDE). The production runs were deployed on a Sun Enterprise 250 computer running Solaris 8. The user interface forms the first tier and permits Internet access to the database (which resides in the third tier). All of the business logic was implemented in the second tier, and it represents the majority of the software written. A back-end Oracle 8i database running on a separate Sun server under Solaris 8 serves as the data archive. Access to local file systems is available to the second tier components, e.g., to interact with the MR files in our local PDB mirror (kept up to date with the software package Mirror v2.9 (McLoughlin, 1998)).

Write Ahead Logging (WAL) ensures proper recovery should the database experience an unexpected shutdown. In addition, the annotated MR files in the database are checked daily for concurrency, and updates are archived by means of the regular backup

services that run at BMRB. All information, except for the annotation on the MR files, is secondary, meaning it can be fully regenerated by automated procedures as part of Wattos.

The second tier serves data back to end-users by requests from HTML (HyperText Markup Language) forms using Java Servlet v2.1 standards as implemented by the Tomcat Web Server v3.2 (for testing) and Apache Web Server + JServ v1.0 (for deployment). The servlet technique has significant advantages over more traditional techniques such as those used by CGI (Common Gateway Interface) programs. Most notably, it maintains persistency between invocations, which ensures better response times after an initial request.

Data model implementation

A simple entity-relationship data model was designed as shown in Figure 1. The initiation of the necessary database tables was programmed in an SQL (Structured Query Language) script that was run using the application program Oracle SQL*Plus. All code and data structures were designed in a database implementation independent fashion, where possible, in order to allow portability across database vendors. All data, except the molecular structure images discussed below, are kept in the generated database tables. Simple textual and numeric data types were chosen for all database attributes with the exception of the column in which the actual textual content of the blocks within each MR file is stored. For that column, the SQL99 standard CLOB (Character Large Object) data type was chosen because of its ability to hold data in sizes up to gigabytes instead of kilobytes, as usual for data types such as VARCHAR (Variable length Characters). User defined temporary database tables were used in order to allow large updates (e.g., updates of the NMR-STAR versions of all the converted files) in a single transaction. These tables have the same definitions as the tables in Figure 1 but are not shown separately.

Grammar parsing

Java Compiler Compiler (JavaCC) was chosen to parse complex grammars, such as those encountered in files generated for use in XPLOR/CNS and ARIA. JavaCC, which is fostered by WebGain, Inc. for Sun Microsystems, Inc., currently is the most popular parser generator for use with the Java programming language. A reader/parser is generated automatically on the basis

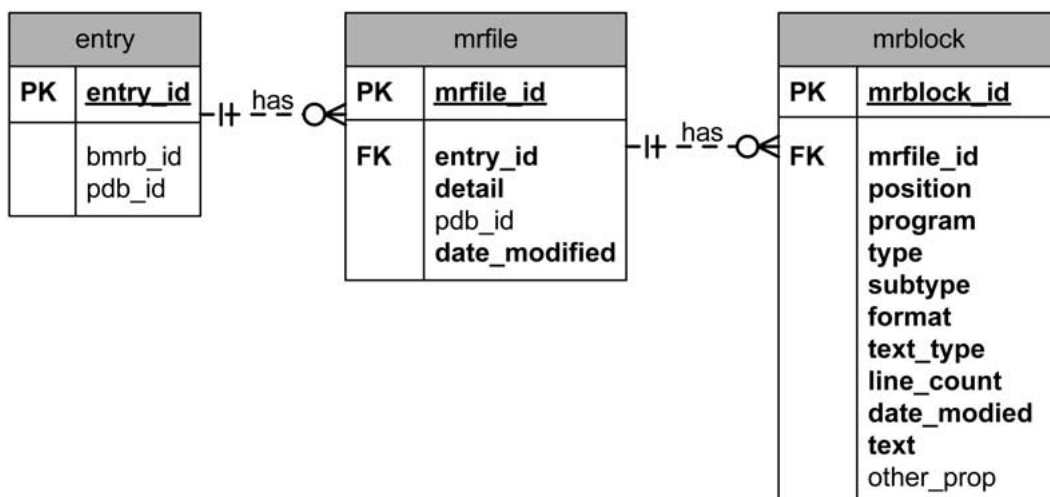


Figure 1. Diagram showing the entity-relationship data model used by the Wattos computer software package. Each Protein Data Bank (PDB) entry can have multiple Magnetic Resonance (MR) files, and each MR file can consist of one or more MR text blocks of data. The mandatory attributes are shown in bold type. For simplicity, the scheme shown omits three temporary tables (having the same definitions as the three tables shown) used for batched updates.

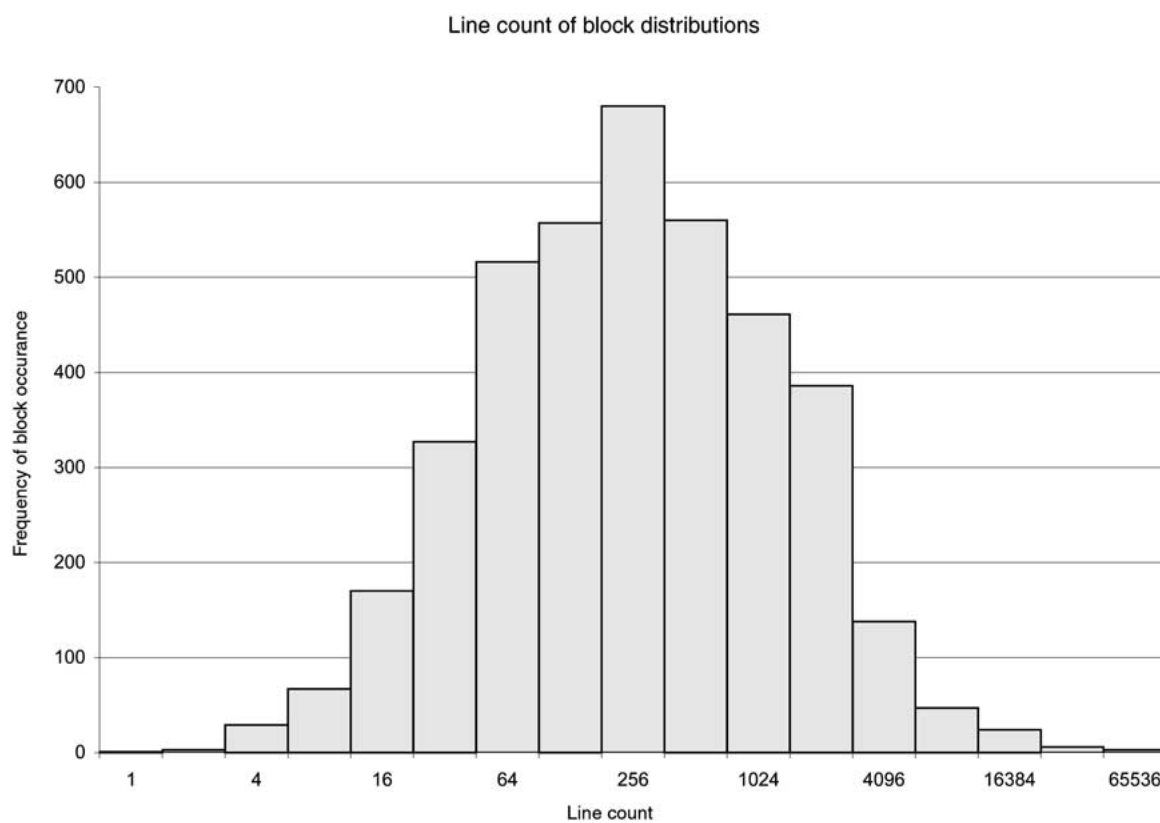



Figure 2. Histogram of the distribution of the number of lines in blocks other than comments and mapping tables in the PDB MR files parsed by the Wattos software. The x-axis is shown on a logarithmic scale.



NMR Restraints Grid

PDB code

BMRB accession ID

Block detail parsed raw

File detail classified parsed

Show Files Blocks

Program instances above

Please select the query you are interested in, leaving blank those fields to be unselected. Selections within an individual selection are combined by a logical 'OR'. These selections are then combined by a logical 'AND'. You can select to display either the number of blocks or the number of files. When you click on the icon next to the table description, you can save the table as a comma separated value formatted file that can be read by most common spreadsheets. Give the file an extension .csv so it will be recognized by your application. Please also do this when downloading sets of files representing blocks as the file name contains only the key to its content as described in the table. One PDB entry can have many associated files. Each file can have many blocks (a section of data of similar type).

The last option determines which program specific files or blocks are shown. Set it to zero to see files or blocks for all programs.

Information on file formats as classified is available [here](#).

A howto is available [here](#).

Known bugs and limitations are listed [here](#). Please report any unknown bugs you may encounter.

Figure 3. Representation of the query interface showing the selections available to the user. For brevity, the listing of available data types was omitted here, but it can be found in full in Table 1.

of a set of token definitions and grammar rules in a grammar file. In this grammar file, the full Java API can be used making it versatile and powerful. Multiple grammars can be combined into one grammar file, and JavaCC features for automatic error recovery and comment (text not part of the regular grammar) parsing were used for this project.

Annotation

The process of splitting the MR file into separate blocks and identifying the type and program for each block was carried out in a semi-automatic way. Wattos was written to identify MR files that need to be processed by comparing entries in the PDB mirror (the source of the MR files) against those that already have been annotated and put in the database and those currently still in processing (with annotation started but not finished). MR files in need of processing were manually scanned by annotators at BMRB for the beginning and end of a block of the same data type, and comments identifying these types were inserted at the beginning of each block. Wattos analyzed the annotated MR file and split and archived it according to the

inserted comments if the annotation conformed to set standards (e.g., followed allowed enumerations for the data types). Wattos also compared the annotated MR file against the original data and rejected any changes that might accidentally have been introduced.

Molecular structure images

High quality images of a structural model for all PDB entries were created from the atomic coordinates and are available from the BMRB as part of this project. The secondary structural elements of the proteins were determined by DSSP (Kabsch and Sander, 1983) as implemented in MOLMOL (Koradi et al., 1996). The molecular topology was photo-realistically rendered by the Persistence of Vision™ Ray Tracer v3.1 (POV-Ray) (Anger et al., 2002) to an image file, which was scaled, compressed, and archived. The whole process was fully automated and runs weekly after the coordinate files are updated in our local PDB mirror.



Figure 4. An individual block containing parsed RDCs. The example is from the structure of calmodulin deposited in the PDB as entry 1J7P (Chou et al., 2001). The image of the molecular structure was generated as described in the Methods section. The list of RDCs has been truncated for clarity.

Table 1. Overview of the number of PDB MR files as sorted by data type (columns 2 through 4) and separated by the associated computer program (top row in columns 6 through 23). The list was generated through the query interface by selecting for all programs, all data types, and all 'classified' files (see Figure 3)

Type	Subtype	Format	Total	AMBER	AQUA	DISCOVER	DSPACE	DYANA/DIANA	EMBOSS	GROMOS	MARDIGRAS	MR format	PDB	PIPP	STAR	SYBYL	TINKER	TRIPOS	XPLOR/CNS	n/a	Unknown		
Total			1410	26	3	140	6	183	17	2	9	1409	1	16	5	5	3	1	877	624	87		
Angle			2	1																		1	
Chemical shift		Format 1	4																			4	
Chemical shift		Format 3	23																			23	
Chemical shift			38											16	4							29	8
Comment			1409									1407										2	615
Coordinate	Initial		1										1										
Coupling constant			59					4				16										37	2
Dihedral angle			758	23		104		60	7			29					3	1			498		36
Dipolar coupling			34					3				1										30	1
Distance	NOE	Ambi	186	2		2		2														178	3
Distance	NOE	Simple	1163	23	3	134	6	165	17	2	5	37			1	5	3	1				721	47
Distance	NOE	Build-up simple	1																			1	
Distance	NOE	Not seen ambi	2																			2	
Distance	NOE	Not seen simple	2																			2	
Distance	Disulfide bond	Simple	56			5		13				5										32	1
Distance	General distance	Ambi	5			1																4	
Distance	General distance	Simple	145	1		18	3	31	1	1		6										79	5
Distance	Hydrogen bond	Ambi	16																			16	
Distance	Hydrogen bond	Simple	607	10		71	6	39	7	2		4										456	12
Distance	Symmetry	Simple	5																			5	
Exchange			3																				3
Line-broadening			1																				1
Molecular system			4																			2	2
Nomenclature mapping			1374									1374											
Peak			25					5		5		1										12	2
Planarity			9																			9	
Pseudocontact shift			8					7															1
Sequence			5	1												3							1
Stereochemistry	Chirality		50	6		44																	
Stereochemistry	Prochirality		45			39	4					1											1
Stereospecific assignment			6					1				1											4
Unknown			2																			1	1

Results

Overall makeup of the database

At the time of writing (November 21, 2002) the PDB contained 1410 entries that had associated MR files (available at <ftp://ftp.rcsb.org/pub/pdb/data/structures/>

divided/nmr_restraints/). Through manual inspection, 8383 unique data blocks were identified and annotated within the MR files. All but three MR files start with a header such as that seen in PDB formatted files, and nearly all MR files (1374) end with a list of nomenclature mappings. The latter block is often much lengthier than the actual constraint data, since the mappings are

given for each atom in all models of the NMR ensemble deposited. These blocks were set aside for now but may be used in the future for improved mapping of the original constraints. The ‘comment’ blocks (1516) and other less common non-experimental data types, such as ‘coordinate’, ‘molecular system’, ‘sequence’, ‘stereochemistry’, and ‘unknown’ were also ignored. This left 3975 blocks containing experimental constraints to be parsed.

In most file formats, the number of text lines in a block corresponded roughly to the number of constraints. On average, each MR file consisted of three blocks (2.9 ± 2.2) each of 626 text lines, although 23 MR files contained 20 or more blocks. Figure 2 shows a histogram of the number of lines in blocks. The most frequent number of lines in a block was 129–256, but many blocks contained substantially more lines. In order for more informative measures of quantity and quality of the experimental data to be calculated; the data need to be analyzed further than was attempted so far. For example, the number of constraints per residue must be corrected for the presence of redundant intra-residual NOE constraints, which can make up more than half of the total number of constraints, to become informative (Doreleijers et al., 1999).

Programs and data types observed

In total, 31 specific types of data from 16 different structure calculation or NMR data analysis programs were observed, in addition to many formats for which a particular program was not identified. A breakdown of the number of entries by programs, in alphabetical order for those with more than ten entries, reads: AMBER, 26 (Weiner et al., 1984); DISCOVER, 140 (Accelrys, San Diego, CA); DYANA/DIANA, 183 (Güntert et al., 1991); EMBOSS, 17 (Nakai et al., 1993); PIPP, 16 (Garrett, 2002); and XPLOR/CNS, 877 (Brünger, 1992; Brünger et al., 1998). Distance constraints (simple and ambiguous) are present in 94% of all MR files: 2785 blocks in 1329 MR files. Older MR files sometimes contain data in a variety of formats with PDB record identifiers, such as ‘REMARK’, ‘NOEUPP’, and ‘ANGLE’, as the first word on the line; these data blocks were classified as ‘MR format’.

NMR-STAR formatted constraints

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 (3.0) of the NMR-STAR data dictionary (Ulrich, 2002). Many programs use the NMR-STAR format for exchanging experimental NMR data.

The program Wattos was used to parse data (using JavaCC as described in Methods) to a developmental NMR-STAR predecessor of NMR-STAR v3.0. An example of an ambiguous NOE distance constraint in X-PLOR/CNS format as observed in the MR file for PDB entry 1CMZ (de Alba et al., 1999) reads: Assign (resid 104 and (name HB1 or name HB2)) (resid 105 and name HN) 3.9 1.0 1.0. This constrains a sequential distance in the target structure between either atoms HB1 or HB2 in residue 104 on the one side and atom HA in residue 105 on the other side to a distance of 0.39 nm with a 0.1 nm deviation allowed on both the upper and lower bounds. The example shows two distinct logical operations (‘AND’ and ‘OR’) and nested brackets which define ambiguity between methylene hydrogen atoms. In this grammar: The ordering of the terms is not relevant, terms are conditionally optional and may be repeated (changing: Resid 104 to resid 104 and rename ‘PHE’ would have specified the same atoms in residue Phe104), the brackets may be nested to arbitrary depths, and the constraint may be interspersed with comments. All of these features make parsing a non-trivial task.

From parsed blocks; Wattos translated 3337 constraint lists to the developmental NMR-STAR version; these corresponded to three data types (2511 distance, 788 dihedral angle, and 38 residual dipolar couplings lists) from the three most popular software packages used in NMR structure determination: XPLOR/CNS (2520 constraint lists), DISCOVER (412 constraint lists), and DYANA/DIANA (405 constraint lists). The set of lists thereby covered 84% of the total number of lists that could be used as constraints (3975). Since it takes a fair amount of time to write a robust parser, no time has been invested yet in parsing constraints in less common data formats; such as AMBER and EMBOSS, or the MR format, which contains many formats. In most cases, the chemical shifts and coupling constants are already available from the BMRB and were also not pursued in the first pass.

Query interface

A query interface was developed that for the first time provides users longitudinal access to all publicly available NMR constraint data from a relational database. This interface is available from the BMRB web site under ‘Features/Data Access’. The standard view of

the database, presented to a user when first entering the interface, shows all of the files available for the different data types and programs (Figure 3 and Table 1). The user selects a subset of the data sections by clicking on a hyperlinked number of items (files or sections) for a specific data type/program combination or by entering a PDB code. The subset of data blocks can then be downloaded all at once in a zipped file streamed from the database, or an individual block can be selected for viewing from the user's browser. Figure 4 shows an example of the latter, in which the parsed RDCs are those deposited for calmodulin (Chou et al., 2001). Upon request, direct access can be made available to a mirror of the database through SQL commands or Java Database Connectivity (JDBC) allowing programs to have direct access.

Future perspectives

The BMRB is working with the PDB to develop a new, unified deposition interface for NMR and X-ray data. The new tool is based on the ADIT (AutoDep Input Tool) software currently in use at the PDB. The goal is to provide NMR spectroscopists with a single interface called ADIT-NMR for depositing both the experimental data destined for BMRB and the coordinate data destined for PDB. Data deposited through ADIT-NMR can be annotated for data type and program and subsequently parsed, thus making the current MR files obsolete. The beta version of ADIT-NMR can be found from the BMRB web site (<http://www.bmrwisc.edu>) by selecting the links labeled 'Deposit' and subsequently 'ADIT-NMR'.

Future work for this project will include matching the atom descriptors in constraint lists to the atom descriptors in the coordinate section of mmCIF (macromolecular Crystallographic Information File; (Westbrook and Bourne, 2000)) formatted PDB files. With the constraints in a unified format, one can envision the possibility of recalculating structures under uniform conditions for longitudinal investigations of the 1400+ structures, and several research groups have expressed an interest in doing so. In addition to the data in NMR-STAR format, new conversion routines under development will allow dissemination in other formats such as XPLOR/CNS or XML.

Acknowledgements

We thank Hamid Eghbalnia and Zsolt Zolnai for careful reading of this manuscript and useful discussions. This work was supported by NIH grant LM05799 from the National Library of Medicine.

References

- Anger, S., Bayer, D., Cason, C., Dailey, C., Dilger, A., Demlow, S., Enzmann, A., Farmer, D., Wenger, T. and Young, C. (2002) Persistence of Vision Raytracer, <http://www.povray.org>
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. and Bourne, P.E. (2000) *Nucl. Acids Res.*, **28**, 235–242.
- 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.
- Brünger, A.T. (1992) In *X-PLOR, Version 3.1: A System for X-Ray Crystallography and NMR*, Yale University Press, New Haven, CT, U.S.A.
- 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.
- Chou, J.J., Li, S., Klee, C.B. and Bax, A. (2001) *Nat. Struct. Biol.*, **8**, 990–997.
- de Alba, E., De Vries, L., Farquhar, M.G. and Tjandra, N. (1999) *J. Mol. Biol.*, **291**, 927–939.
- Doreleijers, J.F., Ravest, M.L., Rullmann, T. and Kaptein, R. (1999) *J. Biomol. NMR*, **14**, 123–132.
- 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.
- Garrett, D.S. PIPP (2002) <http://spin.niddk.nih.gov/clore/Software/software.html>
- Güntert, P., Braun, W. and Wüthrich, K. (1991) *J. Mol. Biol.*, **217**, 517–530.
- Kabsch, W. and Sander, C. (1983) *Biopolymers*, **22**, 2577–2637.
- Koradi, R., Billeter, M. and Wüthrich, K. (1996) *J. Mol. Graph.*, **14**, 51–55.
- Linge, J.P., O'Donoghue, S.I. and Nilges, M. (2001) *Meth. Enzymol.*, **339**, 71–90.
- McLoughlin, L. Mirror (1998) <ftp://sunsite.org.uk/packages/mirror/>
- Nakai, T., Kidera, A. and Nakamura, H. (1993) *J. Biomol. NMR*, **3**, 19–40.
- Seavey, B.R., Farr, E.A., Westler, W.M. and Markley, J.L. (1991) *J. Biomol. NMR*, **1**, 217–236.
- Ulrich, E.L. (2002) NMR-STAR data dictionary. http://www.bmrwisc.edu/dictionary/htmldocs/nmr_star/dictionary.html
- Ulrich, E.L., Ioannidis, Y.E., Livny, M., Mading, S., Slatter, N.C. and Markley, J.L. (1998) *XVIIth International Conference on Magnetic Resonance in Biological Systems*, Tokyo, Japan.
- Weiner, S.J., Collman, P.A., Case, D.A., Singh, U.C., Ghio, C., Alagona, G., Profetar, J.S. and Weiner, P. (1984) *J. Am. Chem. Soc.*, **106**, 765–784.
- Westbrook, J.D. and Bourne, P.E. (2000) *Bioinformatics*, **16**, 159–168.