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, Steve Mading, Dimitri Maziuk, Kassandra Sojourner, Lei Yin§, Jun Zhu*, John L. Markley, and Eldon L. Ulrich

BioMagResBank, UW-Madison, 433 Babcock Dr., Madison, WI, 53706, U.S.A.

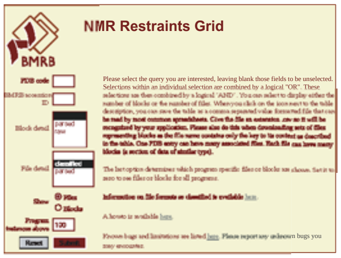
§Keithley Instruments Inc. 28775 Aurora Road, Cleveland, OH, 44139, U.S.A.

*Department of Animal Health and Biomedical Sciences, UW-Madison, 1656 Linden Dr., Madison, WI, 53706, U.S.A.

INTRODUCTION

This project[1] incorporates all the NMR data currently available from the Protein Data Bank (PDB)[2,3] into the BioMagResBank[4] (BMRB - http://bmrb.wisc.edu). The BMRB is a repository for information on biological macromolecules derived from NMR spectroscopy. The PDB has collected and archived experimental NMR data in so-called MR files (Magnetic Resonance). In these files, the different types of information are concatenated together with a PDB-like header and an often-large nomenclature mapping table at the end. There are three main problems when using this information on a large scale. First, sections (blocks) in an MR file correspond to different files that the authors provided, but the boundaries between individual files are not indicated. Second, the data are not in a standard format, which makes it hard if not impossible to use the data if the format cannot be read by the software available to a particular user. Lastly, the data have been archived "as is" and have not been validated[5].

All three problems have been addressed In this project and the BMRB is making a weekly updated archive available to the scientific community. We are currently working on matching the experimental data described here to the coordinate data available from the PDB. The experimental data themselves can then potentially be used to further refine the structures and help to reduce the time needed to solve homologous structures.



NMR Restraints Grid

PDB code

BMRB accession ID

Block detail: parsed, raw

File detail: classified, parsed

Show ◉ Files ○ Blocks

Program instances above 100

Reset | Submit

Please select the query you are interested, 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 common 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.

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

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

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).

RESULTS

At the time of writing[1] (November 21, 2002) the PDB contained 1410 entries with associated MR files. They were classified into 8383 sections. Thus on average each MR file consists of approximately 6 sections or data blocks. Almost all MR files start with a header such as that seen in PDB formatted coordinate files and end with a nomenclature mapping table (ignored in the following discussion), so on average about 4 data blocks constitute an entry.

The results are available from the BMRB web site under "Features/Data Access" and from the PDB web site where individual PDB entries have been linked to the results from this project. A new query interface at the BMRB (Fig. 1) allows a user to select files or blocks of data from individual PDB entries or to browse all of the PDB entries that have associated MR files. An overview of all classified files with raw data (Table 1) with further classification by program and data type is shown.

In total, 31 specific types of data and 16 programs were observed in addition to many formats for which a particular program has not been identified. In most file formats, the number of text lines in a block corresponds roughly to the number of constraints. The histogram of the number of lines in blocks (shown in Fig. 2) peaks for the bin containing blocks with 129 through 256 lines, but many blocks have substantially more lines. The most common data type is the distance restraint with 2511 blocks in total (combining the observed simple and ambiguous formats for NOEs, hydrogen bonds, *etc.*).

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). In parsing, invalid statements and comments are captured and put into a separate tables for future analysis. These constraints were then mapped to a developmental version of the BMRB data model.

A specific set of files (Fig. 3) or blocks (Fig. 4) can be investigated further by following specific links showing the number of text lines in the block. At the level of an individual data block: the actual data text, a molecular image (prepared using MOLMOL[6]), and links are presented (Fig. 5). Tables from any level in the hierarchy can be downloaded as a comma-separated-value file, which enables easy upload to most spreadsheets. The data themselves can be downloaded as a zipped archive of a set of data blocks.

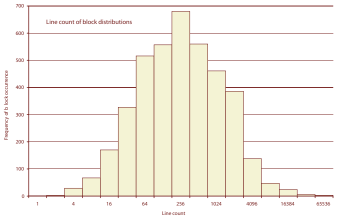


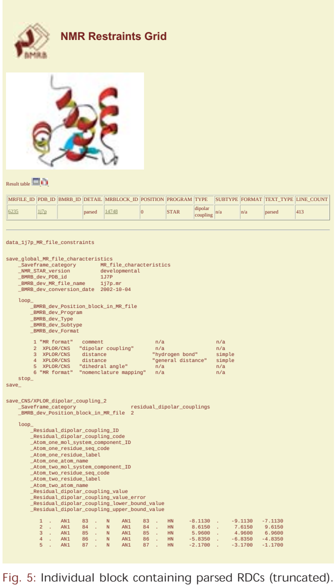
Fig. 2: Histogram of the distribution of the number of lines in blocks other than comments and mapping tables. The *x*-axis is shown on a logarithmic scale.

NMR Restraints Grid

Fig. 3: A view of a set of files sharing a particular property that was previously selected. In this case, the files all have one or more blocks with dipolar couplings (not shown).

NMR Restraints Grid

Fig. 4: The set of NOE distance restraint blocks with some additional information such as the number of lines in each block. Links to the MR file, PDB file, and block specific for each block are shown as underscored items.



NMR Restraints Grid

Result table

MRFILE_ID	PDB_ID	BMRB_ID	DETAIL	MRBLOCK_ID	POSITION	PROGRAM	TYPE	SUBTYPE	FORMAT	TEXT_TYPE	LINE_COUNT
6235	1j7p		parsed	14748	0	STAR	dipolar coupling	n/a	n/a	parsed	413

```
data_1j7p_MR_file_constraints


save_global_MR_file_characteristics
    _Saveframe_category         MR_file_characteristics
    _NMR_STAR_version           developmental
    _BMRB_dev_PDB_id            1J7P
    _BMRB_dev_MR_file_name      1j7p.mr
    _BMRB_dev_conversion_date   2002-10-04

    loop_
        _BMRB_dev_Position_block_in_MR_file
        _BMRB_dev_Program
        _BMRB_dev_Type
        _BMRB_dev_Subtype
        _BMRB_dev_Format

         1 "MR format"   comment                 n/a                 n/a
         2  XPLOR/CNS   "dipolar coupling"       n/a                 n/a
         3  XPLOR/CNS    distance               "hydrogen bond"      simple
         4  XPLOR/CNS    distance               "general distance"   simple
         5  XPLOR/CNS   "dihedral angle"         n/a                 n/a
         6 "MR format"  "nomenclature mapping"   n/a                 n/a
    stop_
save_


save_CNS/XPLOR_dipolar_coupling_2
    _Saveframe_category                   residual_dipolar_couplings
    _BMRB_dev_Position_block_in_MR_file  2

    loop_
        _Residual_dipolar_coupling_ID
        _Residual_dipolar_coupling_code
        _Atom_one_mol_system_component_ID
        _Atom_one_residue_seq_code
        _Atom_one_residue_label
        _Atom_one_atom_name
        _Atom_two_mol_system_component_ID
        _Atom_two_residue_seq_code
        _Atom_two_residue_label
        _Atom_two_atom_name
        _Residual_dipolar_coupling_value
        _Residual_dipolar_coupling_value_error
        _Residual_dipolar_coupling_lower_bound_value
        _Residual_dipolar_coupling_upper_bound_value

          1  .   AN1   83  .   N   AN1   83  .   HN    -8.1130  .   -9.1130  -7.1130
          2  .   AN1   84  .   N   AN1   84  .   HN     8.6150  .    7.6150   9.6150
          3  .   AN1   85  .   N   AN1   85  .   HN     5.9600  .    4.9600   6.9600
          4  .   AN1   86  .   N   AN1   86  .   HN    -5.8350  .   -6.8350  -4.8350
          5  .   AN1   87  .   N   AN1   87  .   HN    -2.1700  .   -3.1700  -1.1700
```

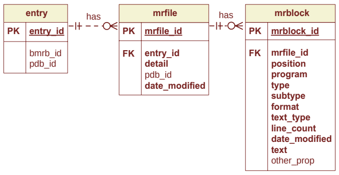
Fig. 5: Individual block containing parsed RDCs (truncated). The example is from the structure of calmodulin[7].

METHODS

All software used for this project, with the exception of the molecular image generation, is written in Java. The data files are stored in an Oracle 8i database using the entity-relationship model shown in Fig. 6. The images are actually served directly from a regular file system, because that is faster by far than by any Java servlet technique. The text files are streamed in a zipped format directly from the database.

The PDB archive including the restraints and coordinates is weekly mirrored using the mirror software (http://wuarchive.wustl.edu/packages/mirror) and the classifications and parsed data blocks will be kept up to date at BMRB.

The Java compiler compiler (javacc) program, which is freely available from http://www.webgain.com/products/java_cc, but owned by Sun Inc., allows us to express complicated grammar (for example that used by XPLOR/CNS files) in a more human readable fashion. Javacc then generates the actual parser program in Java.



entry: PK entry_id; bmrb_id, pdb_id

has

mrfile: PK mrfile_id; FK entry_id, detail, pdb_id, date_modified

has

mrblock: PK mrblock_id; FK mrfile_id, position, program, type, subtype, format, text_type, line_count, date_modified, text, other_prop

Fig. 6: Diagram showing the entity-relationship data model used by the Wattos computer software package. For simplicity, the scheme shown omits three temporary tables (having the same definitions as the three tables shown) used for batched updates.

ACKNOWLEDGEMENTS

BMRB is supported by NIH grant LM05799 from the National Library of Medicine.

REFERENCES

1. Doreleijers, J.F. *et al.*, *J.Biomol.NMR.* accepted (2003).
2. Berman, H.M. *et al.*, *Nucleic Acids Res.* **28**, 235 (2000).
3. Bernstein, F.C., *et al.*, *J.Mol.Biol.* **112**, 535 (1977).
4. Ulrich, E.L. *et al.*, XVIIth Int. Conf. Magn. Res. Biol. Systems. Tokyo, Japan (1998).
5. Doreleijers, J.F., *et al.*, *J.Mol.Biol.* **281**, 149 (1998).
6. Koradi, R. *et al.*, *J.Mol.Graph.* **14**, 51 (1996).
7. Chou, J.J. *et al.*, *Nat.Struct.Biol.* **8**, 990 (2001).