BioMagResBank (BMRB) Data Deposition and Entry Annotation Requirements

Eldon L. Ulrich
Presentation topics

- New ADIT-NMR deposition system – Steve Mading and Dimitri Maziuk in collaboration with the Rutgers RCSB team (Monica Sundd, Monica Sekharan, Zukang Feng, John Westbrook, Helen Berman, and Jasmine Young)

- Restraints processing – Jurgen Doreleijers and Jundong Lin in collaboration with the EBI/CCPN (Wim Vranken), Utrecht University (Aart Nederveen, Alexandre Bonvin, and Robert Kaptein, and Radboud University (Gert Vriend and Chris Spronk)

- Assigned chemical shift validation systems – David Tolmie, Kent Wenger, and Dimitri Maziuk in collaboration with NESG (Hunter Moseley) and NMRFAM (Gabriel Cornilesca, Hamid Eghlbania, and Liya Wang)

- Future issues
Mission: Gather and distribute in the public domain as much biological NMR data as possible to further research and education

Goals:  
1) Create efficient data deposition systems  
   Requires minimal user effort  
   Complete – follows IUPAC recommendations  
   Free of obvious errors  
   Promotes uniformity  
2) Through annotation improve the usefulness of the data  
   Identify anomalous data and communicate issues with depositors  
   Migrate data into standard formats  
   Carry out value added processes  
3) Develop data query systems  
   Entry retrieval  
   Longitudinal database searches
Experimentally derived data

**NMR spectral data**
- Chemical shift assignments
- Chemical shift reference
- Theoretical chemical shifts
- Chemical shift isotope effects
- Chemical shift anisotropy
- Coupling constants
- Residual dipolar couplings
- T1 relaxation
- T1rho relaxation
- T2 relaxation
- Heteronuclear NOE
- Homonuclear NOE/ROE
- Dipole-dipole relaxation
- Cross correlation
- Spectral density values
  - Spectral peak lists
  - Time-domain data sets

**Kinetic parameters**
- H-exchange rates
- H-exchange protection factors
- Order parameters
  - (isotropic and anisotropic)

**Thermodynamic parameters**
- pKa values
- D/H-fractionation factors

**Secondary structure features**
- Helix/sheet/turn/loop
- Deduced H-bonds
- Author interpretation

**Three-dimensional structures**
- NMR constraints
- Constraint statistics
- Coordinates for structure models
- Representative model coordinates
- Structure quality parameters
One-stop RCSB BMRB/PDB ADIT-NMR deposition system
(URL: deposit.bmrb.wisc.edu/bmrb-adit/)

- BMRB and RCSB-PDB depositions can be generated from a joint interface
- BMRB interface has been streamlined
- RCSB-PDB interface has been extended with optional fields for conformer and constraint statistics
- Files in PDB format, mmCIF, and NMR-STAR can be uploaded to pre-populate a deposition

- Many fields (i.e., experiment name, software name, software author, etc.) have pull-down lists to choose from for convenience and to improve uniformity (controlled vocabulary)
- Fields common to multiple forms are linked to eliminate the need to retype information (i.e., uploaded data file names, author names, molecule names and others)
- Help and examples have been improved
- You can start with either BMRB or PDB and switch between the two as you go along
ADIT-NMR architecture

PDB

- PDB
- PDBx
- mmCIF

MAXIT

NMR-IF

- PDBx
- coordinates
- restraints
- pdbx2nmrif

ADIT-NMR

- NMR-STAR v2.1
- NMR-STAR v3

NMR-STAR v2.1

- NMR-STAR v3

s2nmr

NMR-STAR v3

- experimental data

BMRB

- deposition

BMRB deposition
Precheck and validation of coordinate files

Precheck/Validate:
Performs the same checks as are available via PDB's ADIT tool.

Uses the Coordinate file given above.
ADIT-NMR validation report

Structure Summary

- Atlas summary
- Validation summary letter

NMR Protein Validation Report

- PROCHECK Validation
  - Ramachandran plot (PS) (GIF) (PDF)
  - Ramachandran plots for all residue types (PS) (GIF) (PDF)
  - Chi1-Chi2 plots (PS) (GIF) (PDF)
  - Chi1 frequency distributions (PS) (GIF) (PDF)
  - Chi2 frequency distributions (PS) (GIF) (PDF)
  - Ensemble Ramachandran plots (PS) (GIF) (PDF)
  - Residue properties (PS) (GIF) (PDF)
  - Equivalent resolution (PS) (GIF) (PDF)
  - Model secondary structures (PS) (GIF) (PDF)

The above figures were obtained using PROCHECK V 3.4.4
ADIT-NMR constraint statistics

First enter the number and type of constraints present in each constraint file uploaded for this deposition. After clicking 'save changes', new text boxes will appear near the top of the page. Please enter the requested constraint statistical details.
BMRB depositions by year
(~260 PDB depositions received)
ADIT-NMR dual session latency

For sessions that are deposited to both PDB and BMRB, how close together do the two depositions occur in time?

BMRB first, PDB second

<table>
<thead>
<tr>
<th>Time Range</th>
<th>Number of depositions</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0..20min)</td>
<td>32</td>
</tr>
<tr>
<td>[20min..1hour)</td>
<td>2</td>
</tr>
<tr>
<td>[1..2) hours</td>
<td>1</td>
</tr>
<tr>
<td>[2..4) hours</td>
<td>1</td>
</tr>
<tr>
<td>[4..8) hours</td>
<td>0</td>
</tr>
<tr>
<td>[8..16) hours</td>
<td>0</td>
</tr>
<tr>
<td>[16..32) hours</td>
<td>0</td>
</tr>
<tr>
<td>(32..64) hours</td>
<td>5</td>
</tr>
<tr>
<td>&gt;= 64 hours</td>
<td>0</td>
</tr>
</tbody>
</table>

PDB first, BMRB second

<table>
<thead>
<tr>
<th>Time Range</th>
<th>Number of depositions</th>
</tr>
</thead>
<tbody>
<tr>
<td>[0..20min)</td>
<td>19</td>
</tr>
<tr>
<td>[20min..1hour)</td>
<td>3</td>
</tr>
<tr>
<td>[1..2) hours</td>
<td>2</td>
</tr>
<tr>
<td>[2..4) hours</td>
<td>1</td>
</tr>
<tr>
<td>[4..8) hours</td>
<td>0</td>
</tr>
<tr>
<td>[8..16) hours</td>
<td>0</td>
</tr>
<tr>
<td>[16..32) hours</td>
<td>3</td>
</tr>
<tr>
<td>[32..64) hours</td>
<td>1</td>
</tr>
<tr>
<td>&gt;= 64 hours</td>
<td>9</td>
</tr>
</tbody>
</table>
ADIT-NMR development

- Phase I – completed/retired
  BMRB only depositions

- Phase II – completed/being refined
  BMRB-PDB combined depositions

- Phase III – being designed
  Access to NMR atomic coordinate and restraint validation tools
  Access to assigned chemical shift validation tools
  Improved data import functions (PDB Extract, CCPN data harvesting tools, others)
**NMR Restraints Grid**

What is the NMR Restraints Grid?

The NMR Restraints Grid contains the original NMR data as collected for over 2500 protein and nucleic acid structures with corresponding PDB entries. In addition to the original restraints, most of the distance, dihedral angle and RDC restraint data (>85%) were parsed, and those in over 500 entries were converted and filtered. The converted and filtered data sets constitute the DOCR and FRED databases respectively as described in these references.

For tips on using this interface and how to link to it, check the howto. A block is a section of data of similar type such as hydrogen bond distance restraints or RDCs.

There are 3,247,928 parsed constraints in 3110 entries.

<table>
<thead>
<tr>
<th>Type</th>
<th>Subtype</th>
<th>Subsubtype</th>
<th>Total</th>
<th>AMBER</th>
<th>DISCOVER</th>
<th>DYANA/DIANA</th>
<th>MR format</th>
<th>MR</th>
<th>STAR</th>
<th>unknown</th>
<th>Wattos</th>
<th>XML</th>
<th>XPLOR/CNS</th>
</tr>
</thead>
<tbody>
<tr>
<td>angle</td>
<td></td>
<td></td>
<td>73515</td>
<td>282</td>
<td>748</td>
<td>7451</td>
<td>6303</td>
<td>2449</td>
<td>22058</td>
<td>311</td>
<td>10389</td>
<td>6210</td>
<td>17148</td>
</tr>
<tr>
<td>check</td>
<td>completeness</td>
<td>distance</td>
<td>2592</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>check</td>
<td>stereo assignment</td>
<td>distance</td>
<td>2604</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>check</td>
<td>surplus</td>
<td>distance</td>
<td>2598</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>check</td>
<td>violation</td>
<td>distance</td>
<td>2595</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chemical shift</td>
<td>format 1</td>
<td></td>
<td>7</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chemical shift</td>
<td>format 3</td>
<td></td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>chemical shift anisotropy</td>
<td></td>
<td></td>
<td>135</td>
<td>4</td>
<td>10</td>
<td>23</td>
<td>77</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>comment</td>
<td></td>
<td></td>
<td>11014</td>
<td></td>
<td>1</td>
<td>3099</td>
<td>2403</td>
<td>5597</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coordinate</td>
<td>alignment tensor</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coordinate</td>
<td>ensemble</td>
<td></td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coordinate</td>
<td>initial</td>
<td></td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>coupling constant</td>
<td></td>
<td></td>
<td>118</td>
<td>9</td>
<td>19</td>
<td>6</td>
<td>84</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>dihedral angle</td>
<td></td>
<td></td>
<td>9122</td>
<td>106</td>
<td>183</td>
<td>2976</td>
<td>34</td>
<td>2</td>
<td>1637</td>
<td>61</td>
<td>4104</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dipolar coupling</td>
<td></td>
<td></td>
<td>1162</td>
<td>4</td>
<td>1</td>
<td>374</td>
<td>2</td>
<td>2</td>
<td>200</td>
<td>10</td>
<td>567</td>
<td></td>
<td></td>
</tr>
<tr>
<td>distance</td>
<td>disulfide bond</td>
<td>simple</td>
<td>242</td>
<td>2</td>
<td>9</td>
<td>39</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Submitted by [User] on [Date]
NMR Restraints processing

E-MSD
- Refine molecular system (FormatConverter)
- Synchronize models
- Correct atom nomenclature
- Add missing hydrogen atoms (Wattos)

PDB
- Nomenclature Correction (WHAT IF)
- Parse restraints (Wattos)
- NMR Restraints Grid
- Link parsed restraints and coordinates
- NMR-STAR data

DOCR
- Remove surplus, calculate violations, completeness, information content (Wattos/Queen)

FRED
- Convert to structure calculation programs (FormatConverter)
- Structure recalculation (Amber/CNS/CYANA/Gromacs/Yasara etc.)

Analysis of coordinates and all NMR constraint types

Development site
- BMRB
- EBI
- Other
‘Surplus’ distance restraints categories

- Exceptional – atoms not present in the PDB entry
- Double – duplicate restraints
- Impossible – do not match topology provided
- Fixed – atoms have fixed distances
- Redundant – upper bounds greater than what is allowed by topology
<table>
<thead>
<tr>
<th>Set of entries</th>
<th>Counts</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>August, 2006</td>
</tr>
<tr>
<td>PDB entries</td>
<td>38,086</td>
</tr>
<tr>
<td>NMR with or without restraints</td>
<td>5,606</td>
</tr>
<tr>
<td>DOCR/FRED (including Set U see below)</td>
<td>3,060</td>
</tr>
<tr>
<td>With parsable restraints</td>
<td>2,706</td>
</tr>
<tr>
<td>With parsable distance restraints</td>
<td>2,684</td>
</tr>
<tr>
<td>Set 1 &lt; 80 % restraints linked</td>
<td>352</td>
</tr>
<tr>
<td>Set 2 &lt; 33 % restraints after filtering left</td>
<td>212</td>
</tr>
<tr>
<td>Set 3 Max distance violation &gt; 2 Å</td>
<td>471</td>
</tr>
<tr>
<td>Set 4 Rms distance violation &gt; 0.25 Å</td>
<td>325</td>
</tr>
<tr>
<td>Set Union of 1-4: (Union 1-2: 369/198, 3-4: 527/239)</td>
<td>786</td>
</tr>
<tr>
<td>DOCR/FRED without Set U</td>
<td>2,274</td>
</tr>
</tbody>
</table>
Conclusions

- Fraction of good converted MR files highly increased from ~500 to 2,271/3,057 and remaining issues documented
- Contacted authors since April 2006 for entry failing any of four criteria. About 10% of the entries; mostly larger violations that are ok by author
  
  Processing the entries has been further automated; only entries failing criteria get checked manually at this stage

- Results from RECOORD database analysis
  Water refinement improved structure quality
  Low correlations were found between various quality indicators
  Surprisingly, quality indicators did not correlate well with the number of distance restraints
Assigned chemical shift checks

- TALOS/NMRPipe/MolMol implementation
  Delaglio, F., unpublished

- AVS (Assignment validation software)

- LACS (Linear analysis of carbon-13 chemical shift)

- Shifts (planned)

- SHIFTX/SHIFTCOR (planned)
AVS chemical shift validation report

Anomalous Chemical Shift Assignments:

The assigned chemical shifts in the following table have been reported as anomalous, suspicious, or duplicate (A, S or D respectively, in the Error Msg. column) by the software employed by BMRB to check for chemical shift outliers. Please verify these assignments by replacing the question marks in the 'Code' column of the table with the appropriate code. The codes to use are: V = verified, D = delete, and R = replace. Where R is indicated, please supply the revised chemical shift value in the Replace C.S. column of the table. If there are a large number of revised chemical shifts, it may be more convenient to edit the full NMR-Star file. Please inform the annotator in charge of the entry of your modifications.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17</td>
<td>GLU N</td>
<td></td>
<td>108.138</td>
<td>A</td>
<td>120.68</td>
<td>3.68</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>21</td>
<td>LEU H D2</td>
<td></td>
<td>-0.446</td>
<td>S</td>
<td>0.76</td>
<td>0.28</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>30</td>
<td>LYS H</td>
<td></td>
<td>10.556</td>
<td>S</td>
<td>8.22</td>
<td>0.64</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>34</td>
<td>ILE H D1</td>
<td></td>
<td>-0.297</td>
<td>S</td>
<td>0.7</td>
<td>0.3</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>1</td>
<td>53</td>
<td>THR CB</td>
<td></td>
<td>62.879</td>
<td>S</td>
<td>69.64</td>
<td>1.7</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
Protein delta chemical shift values (observed – calculated)
Histogram of delta chemical shifts (observed – calculated) for protein helix and sheet residues
LACS visualization
BMRB data visualization service

This service allows you to create DEVise visualizations of NMR data. Click on Create session button below to proceed to file upload screen.

Note:

- Uploaded data must be in NMR-STAR format.
- Only polypeptide(L) data (proteins with L amino acids) are supported.
- Currently only chemical shift data are supported.

Create session

If you have an ID of a previously created session you can type it in the box below and click on Continue session button. The next screen will allow you to upload another data file, delete existing data file(s), or view the visualizations.

Note: we will automatically delete sessions older than a few weeks.

Session ID: [input field]  Continue session
International structure genomics task force committees

- Structural genomics informatics task force
- Task force on numerical criteria for evaluating and assuring structure quality
- Task force on tracking and registration of targets
- Task force on deposition, archiving, and curation of the primary information
- Task force on mechanisms for publication and recording of methods
- Task force on intellectual property rights
Structural genomics NMR-STAR dictionary development collaborations

Cheryl Arrowsmith
Michael Kennedy
John Markley
Guy Montelione
James Prestegard
David Wemmer
NMR dictionary general discussion topics

- Alignment with mmCIF - Yes
  Use identical tags and definitions whenever possible

- Human readable export data format – Yes

- Reproduction of the experiment and data derivation - No
  Input data, derived data, description of the tools, protocol files, and parameter files used in the derivation
  Explicit links between the input data items and individual derived data items, including the possibility of capturing intermediate results in the derivation process

- Application specific data items - No
  Capture in the protocol and parameter files

- Software must be in place to meet higher deposition requirements - Yes
BMRB Time-domain data summary
(www.bmrb.wisc.edu/data_library/timedomain/)

Entries: released 66

NMR experiments:
- Total >600
- Unique 75-100
- Reduced dimensionality 5 (entries) (entries 5596, 5844, 5859, 7170, 7191)

Pulse sequences and processing parameters are often provided

Other information:
- Peak lists 15 (entries)
- Structure calculation (with all intermediate results) 3 (entries 6128, 6176, 6318)
<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1xeasy/</td>
<td>28-Oct-2004</td>
<td>10:28</td>
<td></td>
</tr>
<tr>
<td>2spscan</td>
<td>28-Oct-2004</td>
<td>10:28</td>
<td></td>
</tr>
<tr>
<td>3garant</td>
<td>28-Oct-2004</td>
<td>10:28</td>
<td></td>
</tr>
<tr>
<td>4assign</td>
<td>28-Oct-2004</td>
<td>10:28</td>
<td></td>
</tr>
<tr>
<td>5noesy/</td>
<td>28-Oct-2004</td>
<td>10:28</td>
<td></td>
</tr>
<tr>
<td>7candid</td>
<td>28-Oct-2004</td>
<td>10:30</td>
<td></td>
</tr>
<tr>
<td>8cyana</td>
<td>28-Oct-2004</td>
<td>10:34</td>
<td></td>
</tr>
<tr>
<td>9valid</td>
<td>28-Oct-2004</td>
<td>10:37</td>
<td></td>
</tr>
<tr>
<td>10pdb_dep/</td>
<td>28-Oct-2004</td>
<td>10:37</td>
<td></td>
</tr>
<tr>
<td>At3g51030_analysis.tar.gz</td>
<td>28-Oct-2004</td>
<td>10:41</td>
<td>694M</td>
</tr>
<tr>
<td>summary.txt</td>
<td>28-Oct-2004</td>
<td>10:37</td>
<td>1.1K</td>
</tr>
</tbody>
</table>

Apache/2.0.54 (Fedora) Server at www.bmrb.wisc.edu Port 80
Future issues

- Probabilistic approaches to structure determination
- Structure determination from chemical shift data
- Use of multiple kinds of NMR data in structure calculations
- Structure refinement with data from non-NMR techniques
Acknowledgements

BMRB Madison
John L. Markley
Jurgen F. Doreleijers
Jundong Lin
Steve Mading
Dimitri Maziuk
David Tolmie
Hongyang Yao
Christopher Schulte

BMRB Osaka
Hideo Akutsu
Eiichi Nakatani
Yoko Harano

BMRB Florence
Antonio Rosato

BMRB Advisory Board

CCPN/EBI
Wayne Boucher
Rasmus Fogh
John Ionides
Ernest Laue
Tim Stevens
Wim Vranken

NMRFAM
Arash Bahrami
Gabriel Cornilescu
Hamid Eghbalnia
Liya Wang
William M. Westler

Utrecht University
Alexandre Bonvin
Aart Nederveen
Robert Kaptein

Rutgers University
Mike Baran
Dehua Hang
Guy Montelione
Hunter Moseley

Members of the NMR Community
Structural genomics groups
Catherine Bougault
Bruce Johnson
David Wishart
Zsolt Zolnai

Computer sciences collaborators
Yannis Ioannidis
Miron Livny
Zachary Miller
R. Kent Wenger

RCSB
Helen Berman
Zukang Feng
Monica Sekharan
Monica Sundd
John Westbrook
Jasmine Young

Funding