PyBMRB: Data visualization tool for BioMagResBank

Kumaran Baskaran[‡§*], Jonathan R Wedell[‡§], Eldon L. Ulrich[‡§], Jeffery C. Hoch[‡§], John L. Markley[¶‖]

Abstract—The Biological Magnetic Resonance Data Bank (BioMagResBank or BMRB https://bmrb.io), founded in 1988, is the international, open archive for data generated by Nuclear Magnetic Resonance (NMR) spectroscopy of biological systems. NMR spectroscopy is unique among biophysical approaches in its ability to provide a broad range of atomic and higher-level information relevant to the structural, dynamic, and chemical properties of biological macromolecules, as well as report on metabolite and natural product concentrations in complex mixtures and their chemical structures. NMR-STAR is the official data format of BMRB and BMRB provides python parser (PyNMRSTAR https://github.com/uwbmrb/PyNMRSTAR), a data visualization tool (PyBMRB https://github.com/uwbmrb/PyBMRB) and an Application Program Interface (API)(BMRB-API https://github.com/uwbmrb/BMRB-API) to access the BMRB archive. PyBMRB displays the chemical shifts data in each entry as a simulated NMR spectrum and to generates database-wide chemical shift histograms of different atom types in proteins and nucleic acids. PyBMRB provides access to BMRB data through the API and generates portable and interactive visualizations as a single html file. It also supports data visualization workflows using Jupyter Notebooks, which can be both easily created and shared.

Index Terms—NMR Spectroscopy, chemical shifts, proteins, Biological Magnetic Resonance data Bank(BMRB),NMR-STAR, chemical shift histogram, HSQC

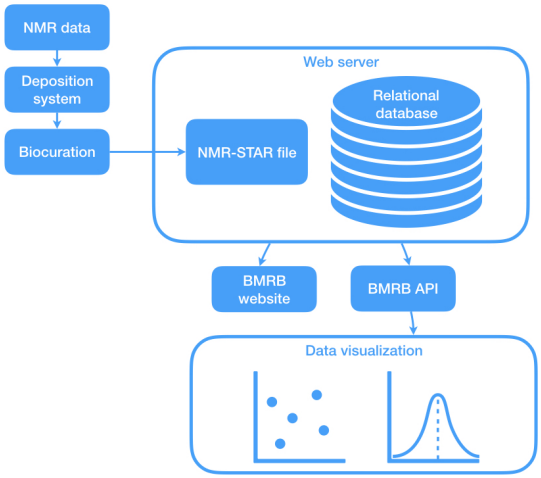

Fig. 1: *BMRB data processing workflow.*

Nuclear Magnetic Resonance (NMR) spectroscopy provides atom-level information relevant to the structural, dynamic, and chemical properties of molecules. The BioMagResBank (BMRB) [UAD+07] provides high-quality, curated NMR spectroscopic data collected from biologically important molecules such as proteins, nucleic acids, carbohydrates, and metabolites and other small compounds. BMRB, which was founded in 1988, became a core member of World Wide Protein Data Bank (wwPDB) [BBK+17] in 2007, and the BMRB Archive became a Core Archive of the wwPDB in 2018. BMRB uses the NMR-STAR [UBD+19] data format to represent experiments, spectral and derived data, and supporting metadata. NMR-STAR is constructed via an object-relational data model using a subset of the Self-defining Text Archival and Retrieval (STAR) specification [HC95]. Following validation and annotation via BMRB's biocuration pipeline (Figure 1), user-deposited data are stored as flat files in NMR-STAR format as well as in a relational database.

To achieve the full power of the BMRB database it is important to be able to retrieve and visualize the data in different scientifically relevant ways. For example, it is much more useful to compare multidimensional NMR data from the same or different BMRB entries in graphical (spectral) format rather than as lists of numerical values in text format. In addition, to understand how chemical shifts of different types of atoms are affected by structural and environmental factors, it is useful to display them as histograms. When browser vendor security policies changed to stop allowing Java Web Applets, BMRB's original visualization tool (DEVise) [LRB+97] written in Java and C++ ceased to function. BMRB originally addressed this by updating DEVise to run as a Java Web Start application. However, in mid-2015 most web browsers stopped supporting Java Web Start and some operating system made it impossible to use without changing operating system security settings.

In response to the demise of DEVise, BMRB developed graphic libraries in Python (PyBMRB) that utilize more modern interactive visualization tools, such as the Plotly visualization tool kit [Inc15] , to reproduce the most commonly used features of DEVise with interactive visualizations. PyBMRB features single-entry (peak position simulation for NMR spectrum) and database-wide visualizations (histograms).

The main motivation behind the project is to provide user friendly access to BMRB data for biologists and biochemists, who find it difficult to understand or utilize the NMR-STAR data model. NMR-STAR is a metadata rich format, which includes all necessary metadata about the NMR sample, sample condition, instrument details, author details and experimental details in addition to the measured chemical shift values. Chemical shifts

* *Corresponding author: baskaran@uchc.edu*
‡ *UCONN Health, Molecular Biology and Biophysics*
§ *263 Farmington Ave. Farmington, CT 06030-3305, USA*
¶ *Department of Biochemistry, University of Wisconsin-Madison*
‖ *433 Babcock Drive, Madison, Wisconsin 53606-1544, USA*

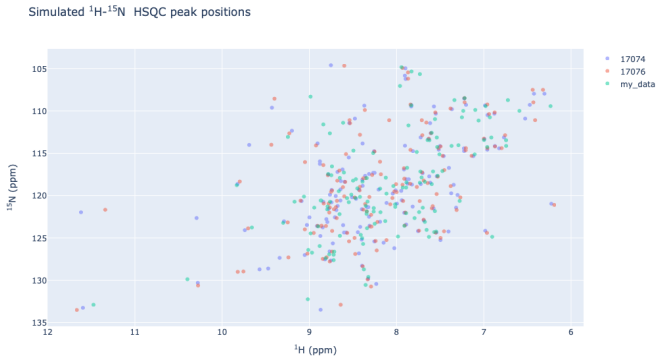

Fig. 2: *Comparison of $^{1}H-^{15}N$ HSQC spectra of arsenate reductase data from user with arsenate reductase entries in the BMRB*

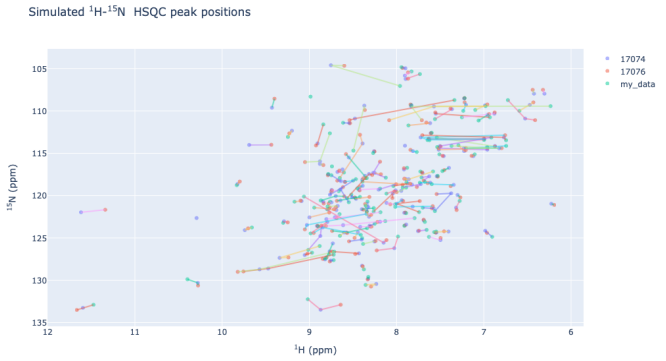

Fig. 3: *The cross peaks in the $^{1}H-^{15}N$ HSQC spectra are connected based on matching sequence order.*

are measured using several multidimensional NMR experiments and expressed one-dimensional assigned chemical shift lists in NMR-STAR data format. Biologists and biochemists prefer to view the chemical shift data graphical spectra rather than as a list of numerical values.

One of the most common and widely used NMR experiments for proteins is the $^{1}H-^{15}N$ Heteronuclear Single Quantum Coherence($^{1}H-^{15}N$ HSQC) [BR80] experiment. This 2D NMR experiment gives cross peaks between nitrogen and hydrogen for each amino acid in the sequence, whose locations strongly depend on the protein three dimensional structure. In spectroscopic perspective the $^{1}H-^{15}N$ HSQC spectrum is considered as the signature or "fingerprint" of the protein. It helps to identify whether the protein sample is in good shape or aggregated and to detect structural changes during ligand biding studies. PyBMRB library generates 2D chemical shift lists by combining the relevant chemical shift values from the given one-dimensional chemical shift list in NMR-STAR format.

The single-entry visualization method can be used, for example, to simulate $^{1}H-^{15}N$ HSQC peak positions from an NMR-STAR file (from one or more specified BMRB entries or from the user's own data) (Figures 2 and 3). It is much easier to detect the chemical shift changes by overlaying multiple $^{1}H-^{15}N$ HSQC rather than by scanning lists of chemical shifts. The most useful feature is that the user may easily compare their NMR measurements with any of the protein of interest in the BMRB database. The Figures 2 and 3 show the comparison of user data with two similar entries from BMRB database. This comparison can be done with the following code

```
from pybmrb import csviz
s=csviz.Spectra()
s.n15hsqc(bmrbid=[17074,17076],
       filename='my_data.str')
```

The chemical shift changes can be traced for each residue by using `groupbyres` option. (Figures 3)

```
s.n15hsqc(bmrbid=[17074,17076],
       filename='my_data.str',
       groupbyres=True)
```

BMRB provides rich chemical shift statistics, which are widely used by NMR spectroscopists and NMR software developers in various ways. The chemical shift histogram of a given atom type help us to understand how strongly it's position depends on the secondary structure elements like alpha helices and beta sheets.

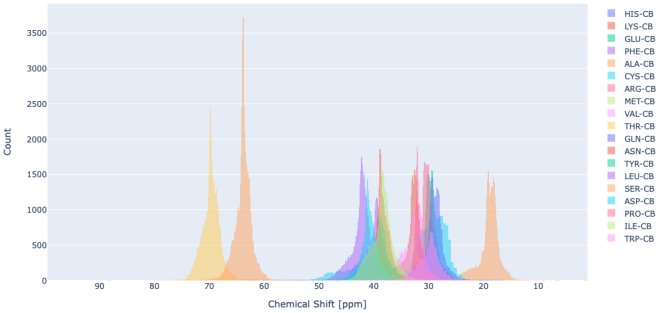

Fig. 4: *Chemical shift distribution of CB atoms in different amino acids.*

These histograms can be easily generated using a simple code using PyBMRB library

```
from pybmrb import csviz
h=csviz.Histogram()
h.hist(atom='CB')
```

Figure 4 shows the comparison of CB chemical shifts for the twenty common amino acids. The chemical shift histogram of a single atom in a given amino acid or list of atoms from different amino acids can be easily generated using PyBMRB.

PyBMRB provides options for filtering data, for example, according to chemical shift ambiguity code(used to describe different types of ambiguous chemical shift assignments https://bmrb.io/software/ambi/) or cutoff values based on standard deviation to exclude outliers. Bond correlation experiments are very common in NMR spectroscopy, and this library can be used to visualize patterns of chemical shift correlations between specified atom types in NMR spectra of proteins or nucleic acids as 2D histograms. For example the chemical shift correlation between Cysteine CB and N is shown in Figure 5.

```
h.hist2d(residue='CYS',atom1='CB',atom2='N')
```

The conditional histogram is another feature, useful during the resonance assignment process to estimate the prior probability for assigning a specific atom number to a peak. The process of labeling each cross peak in the multidimensional NMR spectra by relevant atoms is the most important step in the structure determination process. If the chemical shift values of one or more

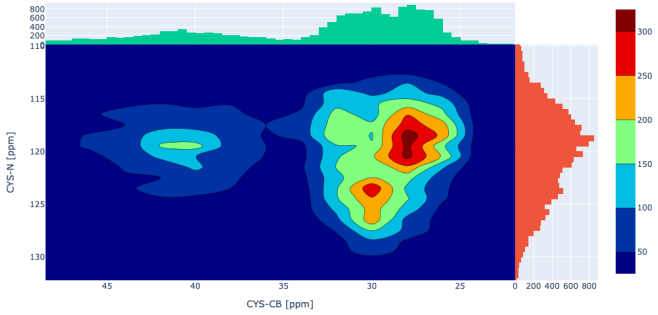

Fig. 5: *Chemical shift correlation of CYS-CB and CYS-N*

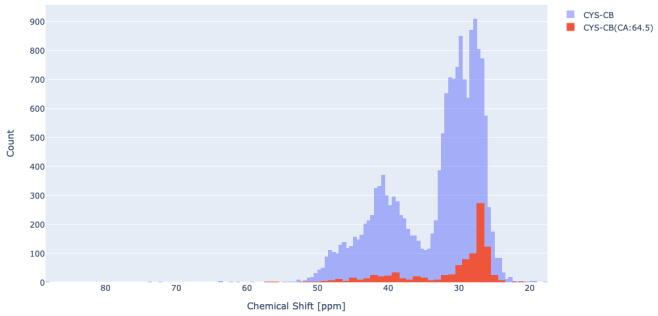

Fig. 6: *Conditional histogram of CYS-CB for CYS-CA=64.5ppm*

atoms for a given amino acid in a protein sequence are known then one can generate the distribution of the chemical shifts of the other atoms in the amino acid using the known chemical shifts as a filter. For example if the chemical shift of CA of Cysteine is known then the distribution of CB chemical shift at the BMRB database can be calculated using the following code

```
h.conditional_hist(residue='CYS',
       atom='CB',
       atomlist=['CA'],
       cslist=[64.5])
```

The overall and the filtered distribution of CYS-CB is shown in Figure 6. The overall bimodal distribution of Cysteine CB indicates that its chemical shifts are strongly depend on secondary structures and for the given value of CA (64.5 ppm) it falls into one of secondary structure element like alpha helix or beta sheet.

The visualizations generated using PyBMRB library are interactive and portable. They can be opened in any modern web browser and zoomed in and out using the mouse. The tooltip will show the peak label and some additional information when hovering over the peak. These visualizations work as a standalone web page, which can be shared via email or website. Since the visualization tools obtain data directly from the BMRB API each time they are generated, there is no need to download or parse the data, and all underlying data are fully up to date. High quality static images can be extracted from the interactive visualizations with a single click and saved or printed.

As a final note, the Jupyter Notebook [KRKP+16] [Com20] is becoming more and more popular among scientists [Per18]. Jupyter is a free, open-source, interactive web tool, known as a computational notebook, that researchers can use to combine software code, computational output, explanatory text and multimedia resources into a single document. PyBMRB can be used in a Jupyter Notebook environment, which enables one to design and document a BMRB data analysis workflow and share it with others. BMRB provides easy access to the PyBMRB library in a Jupyter Notebook environment from its homepage (https://bmrb.io/). This live BMRB Jupyter Notebook was created by using a third party software tool called Binder [PJMBJF+18], which puts PyBMRB and Jupyter Notebook together in a docker container. Examples of BMRB Jupyter Notebooks with access to PyBMRB are available for trial without the need for any installation at https://github.com/uwbmrb/PyBMRB/blob/master/jupyter.md.

BMRB is constantly working to improve the PyBMRB visualization tool. The next update aims to include simulation of more NMR experiment types and include visualization options for other data types such as distance and dihedral-angle restraints that are present in the BMRB database.

BMRB is supported by grant R01GM109046 from NIH/NIGMS.

REFERENCES

[BBK+17] Stephen K Burley, Helen M Berman, Gerard J Kleywegt, John L Markley, Haruki Nakamura, and Sameer Velankar. Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods in molecular biology (Clifton, N.J.)*, 1607:627–641, 2017. URL: https://pubmed.ncbi.nlm.nih.gov/28573592https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5823500/, doi:10.1007/978-1-4939-7000-1_26.

[BR80] Geoffrey Bodenhausen and David J. Ruben. Natural abundance nitrogen-15 nmr by enhanced heteronuclear spectroscopy. *Chemical Physics Letters*, 69(1):185–189, 1980. URL: https://www.sciencedirect.com/science/article/pii/0009261480800418, doi:https://doi.org/10.1016/0009-2614(80)80041-8.

[Com20] Executable Books Community. Jupyter book, February 2020. URL: https://doi.org/10.5281/zenodo.4539666, doi:10.5281/zenodo.4539666.

[HC95] S. Hall and A. P. Cook. Star dictionary definition language: Initial specification. *J. Chem. Inf. Comput. Sci.*, 35:819–825, 1995.

[Inc15] Plotly Technologies Inc. Collaborative data science, 2015. URL: https://plot.ly.

[KRKP+16] Thomas Kluyver, Benjamin Ragan-Kelley, Fernando Pérez, Brian Granger, Matthias Bussonnier, Jonathan Frederic, Kyle Kelley, Jessica Hamrick, Jason Grout, Sylvain Corlay, Paul Ivanov, Damián Avila, Safia Abdalla, Carol Willing, and Jupyter development team. Jupyter notebooks - a publishing format for reproducible computational workflows. In Fernando Loizides and Birgit Scmidt, editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87–90. IOS Press, 2016. URL: https://eprints.soton.ac.uk/403913/, doi:10.3233/978-1-61499-649-1-87.

[LRB+97] M. Livny, R. Ramakrishnan, K. Beyer, G. Chen, D. Donjerkovic, S. Lawande, J. Myllymaki, and K. Wenger. Devise: Integrated querying and visual exploration of large datasets. *SIGMOD Rec.*, 26(2):301–312, June 1997. URL: https://doi.org/10.1145/253262.253335, doi:10.1145/253262.253335.

[Per18] Jeffrey Perkel. Why jupyter is data scientists' computational notebook of choice. *Nature*, 563:145–146, 11 2018. doi:10.1038/d41586-018-07196-1.

[PJMBJF+18] Project Jupyter, Matthias Bussonnier, Jessica Forde, Jeremy Freeman, Brian Granger, Tim Head, Chris Holdgraf, Kyle Kelley, Gladys Nalvarte, Andrew Osheroff, M Pacer, Yuvi Panda, Fernando Perez, Benjamin Ragan Kelley, and Carol Willing. Binder 2.0 - Reproducible, interactive, sharable environments for science at scale. In Fatih Akici, David Lippa, Dillon Niederhut, and M Pacer, editors, *Proceedings of the*

17th Python in Science Conference, pages 113 – 120, 2018. doi:10.25080/Majora-4af1f417-011.

[UAD+07] Eldon L. Ulrich, Hideo Akutsu, Jurgen F. Doreleijers, Yoko Harano, Yannis E. Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, Eiichi Nakatani, Christopher F. Schulte, David E. Tolmie, R. Kent Wenger, Hongyang Yao, and John L. Markley. BioMagResBank. *Nucleic Acids Research*, 36:D402–D408, 11 2007. URL: https://doi.org/10.1093/nar/gkm957, arXiv:https://academic.oup.com/nar/article-pdf/36/suppl_1/D402/7635401/gkm957.pdf, doi:10.1093/nar/gkm957.

[UBD+19] Eldon L Ulrich, Kumaran Baskaran, Hesam Dashti, Yannis E Ioannidis, Miron Livny, Pedro R Romero, Dimitri Maziuk, Jonathan R Wedell, Hongyang Yao, Hamid R Eghbalnia, Jeffrey C Hoch, and John L Markley. NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments. *Journal of Biomolecular NMR*, 73(1):5–9, feb 2019. URL: https://doi.org/10.1007/s10858-018-0220-3, doi:10.1007/s10858-018-0220-3.