J Biomol NMR (2015) 63:141–150
DOI 10.1007/s10858-015-9964-1

ARTICLE

CONNJUR R: an annotation strategy for fostering reproducibility in bio-NMR—protein spectral assignment

Matthew Fenwick[1] · Jeffrey C. Hoch[1] · Eldon Ulrich[2] · Michael R. Gryk[1]

Received: 12 March 2015 / Accepted: 1 July 2015 / Published online: 8 August 2015
© Springer Science+Business Media Dordrecht 2015

Abstract Reproducibility is a cornerstone of the scientific method, essential for validation of results by independent laboratories and the sine qua non of scientific progress. A key step toward reproducibility of biomolecular NMR studies was the establishment of public data repositories (PDB and BMRB). Nevertheless, bio-NMR studies routinely fall short of the requirement for reproducibility that all the data needed to reproduce the results are published. A key limitation is that considerable metadata goes unpublished, notably manual interventions that are typically applied during the assignment of multidimensional NMR spectra. A general solution to this problem has been elusive, in part because of the wide range of approaches and software packages employed in the analysis of protein NMR spectra. Here we describe an approach for capturing missing metadata during the assignment of protein NMR spectra that can be generalized to arbitrary workflows, different software packages, other biomolecules, or other stages of data analysis in bio-NMR. We also present extensions to the NMR-STAR data dictionary that enable machine archival and retrieval of the "missing" metadata.

Keywords CONNJUR · Data model · Reproducibility · Analysis · NMR-STAR

Electronic supplementary material The online version of this article (doi:10.1007/s10858-015-9964-1) contains supplementary material, which is available to authorized users.

✉ Michael R. Gryk
gryk@uchc.edu

[1] Department of Molecular Biology and Biophysics, UConn Health, Farmington, CT 06030-3305, USA

[2] Department of Biochemistry, University of Wisconsin-Madison, Madison, WI 53706, USA

Introduction

Reproducibility of scientific results is an essential test of their validity. Reproducibility requires not only that empirical observations can be repeated, but also that analyses applied to those observations used to derive models or other conclusions can be repeated, independent of the researchers who performed the initial study (Ioannidis et al. 2008; Landis et al. 2012). There has recently been growing concern that much experimental science is not reproducible (Prinz et al. 2011; Ioannidis et al. 2008), and understandably, agencies responsible for public funding of science have launched initiatives to improve reproducibility (Collins and Tabek 2014). In bio-NMR, the barriers to making the computational analysis of data reproducible include incomplete reporting standards, the diversity of software employed, and missing metadata, such as information not stored by the NMR spectrometers or manual interventions not recorded. A previously suggested gold standard for computational reproducibility, making publically available the "entire computational environment required to reproduce the figures" (Buckheit and Donoho 1995; Peng 2011; Stodden and Miguez 2014), provides a well-defined target to guide efforts to improve reproducibility.

Here we consider the barriers to reproducibility posed by the assignment of protein NMR spectra as a concrete example of the difficulties in making a study reproducible to the level of the Donoho criterion. The workflows involved in protein chemical shift assignment include automated steps as well as manual interventions.

Following data collection, spectrum analysis of the time domain is used to compute frequency spectra that are subjected to peak-picking to identify and quantify features in the spectra. Analysis of the resulting peak tables to

identify correlations expected on the basis of the known protein sequence is then performed to obtain chemical shift assignments of spectral peaks to specific nuclei in the protein sequence. These assignments form the basis for subsequent analyses that are used to perform biophysical characterizations, such as structure determination. These involve additional spectra obtained using nuclear Overhauser experiments, correlated with the chemical shift assignments to quantify internuclear distances and assign these distances to specific spin pairs, or experiments performed in anisotropic media to extract residual dipolar couplings that reflect relative orientations of spin pairs. These derived NMR parameters (assigned chemical shifts, RDCs, NOEs) are then used to determine the molecular structure (Fig. 1). While the scope and applicability of automation has increased, manual interventions are necessary at various steps of the analyses to achieve high-quality results (Guerry and Herrmann 2011; Güntert 2009) because software tools are often unable to correctly analyze noisy, incomplete and ambiguous data, and their results may contain mistakes which must be rectified manually. Although incomplete metadata presents a recurring challenge to achieving reproducibility at the level of the Donoho criterion, the absence of information about the manual interventions presents a greater obstacle to reproducibility of protein NMR studies.

The BioMagResBank (BMRB) (Ulrich et al. 2008) has enabled the archival and dissemination of experimental and derived results in biomolecular NMR for more than 20 years, and as such has been a major driving force for increasing the reproducibility of NMR results. However, the full details of the computational analysis, including critical manual interventions, are not yet captured by BMRB. The procedures we describe in this work build upon and extend the utility of BMRB for fostering reproducibility in biomolecular NMR. We focus on the steps involved in chemical shift assignment of proteins based on triple-resonance spectra. The general approach proposed can be extended to other stages of computational analysis of bio-NMR data, such as structure determination.

Our strategy uses a version control system (VCS) to capture and annotate intermediate results, and a data model for metadata required to reproduce a computational analysis. A key feature of the approach is that it can be applied to any software tool or workflow, as long as intermediate data are written to files. Thus the approach supports many commonly used NMR analysis tools. VCSs have a long history in software engineering for managing changing codebases, including annotation of changes. Recently it was also proposed that such technologies could be applied to scientific databases (Dall'Olio et al. 2010). We demonstrate here that this mature technology can be applied to the annotation of analytic workflows in bio-NMR, without modification of existing software tools. However, the power of the approach for fostering reproducibility is considerably enhanced by modifications, or ancillary tools, to enable the recording of more detailed annotations. In order to facilitate exchange and dissemination of reproducible data sets, we propose extensions to the NMR-STAR data dictionary employed by BMRB. We also describe extensions to the Sparky analysis program (Goddard and Kneller 2004) to assist spectroscopists in utilizing this strategy. Finally, we provide and describe an annotated workflow illustrating our proposed method.

Materials and methods

The annotation strategy was implemented as an extension to the analysis program Sparky (Goddard and Kneller 2004), which facilitates user-defined additions to functionality by means of Python scripts. The Sparky extension was implemented using the Eclipse IDE (Eclipse Foundation 2007) and the source code was tracked using git (Loeliger and McCullough 2012). The source code for the extension is available under the MIT license (Open Source Initiative 2006) and is included with the NMRFAM distribution of Sparky (http://www.nmrfam.wisc.edu/).

The data model was created using the BMRB (Ulrich et al. 2008) and CCPN (Vranken et al. 2005) data models as starting points. The essential entities for spectral analysis and chemical shift assignment were identified and implemented inside of Sparky. Although Sparky has a built-in concept of resonances and spin systems, this does not match the CCPN semantics. A compatibility layer was implemented on top of the Sparky objects which provided CCPN-compatible semantics. All manipulation of these objects was performed through the compatibility layer.

Two different mappings of the git-based data to NMR-STAR were considered. The first was to store full snapshots of the analysis process. The second was to store a log of all the changes made. While the two approaches are both able to express the desired data, the chief concern was that the schema extensions had to be a superset of the existing NMR-STAR data dictionary, so as not to break backward compatibility. The second approach of a log of changes met this criterion, and was thus chosen. The final NMR-STAR file was constructed using a shell script that extracted all project file versions from the git repository, then checked for semantic differences between versions, and emitted the data according to the NMR-STAR schema.

The library of deductive reasoning was created in a trial-and-error approach based on analyzing the Samp3 data multiple times. The first time the analysis was performed using CCPN Analysis in conjunction with git; the snapshots were poorly focused and the annotations contained too little information to be useful or too much information

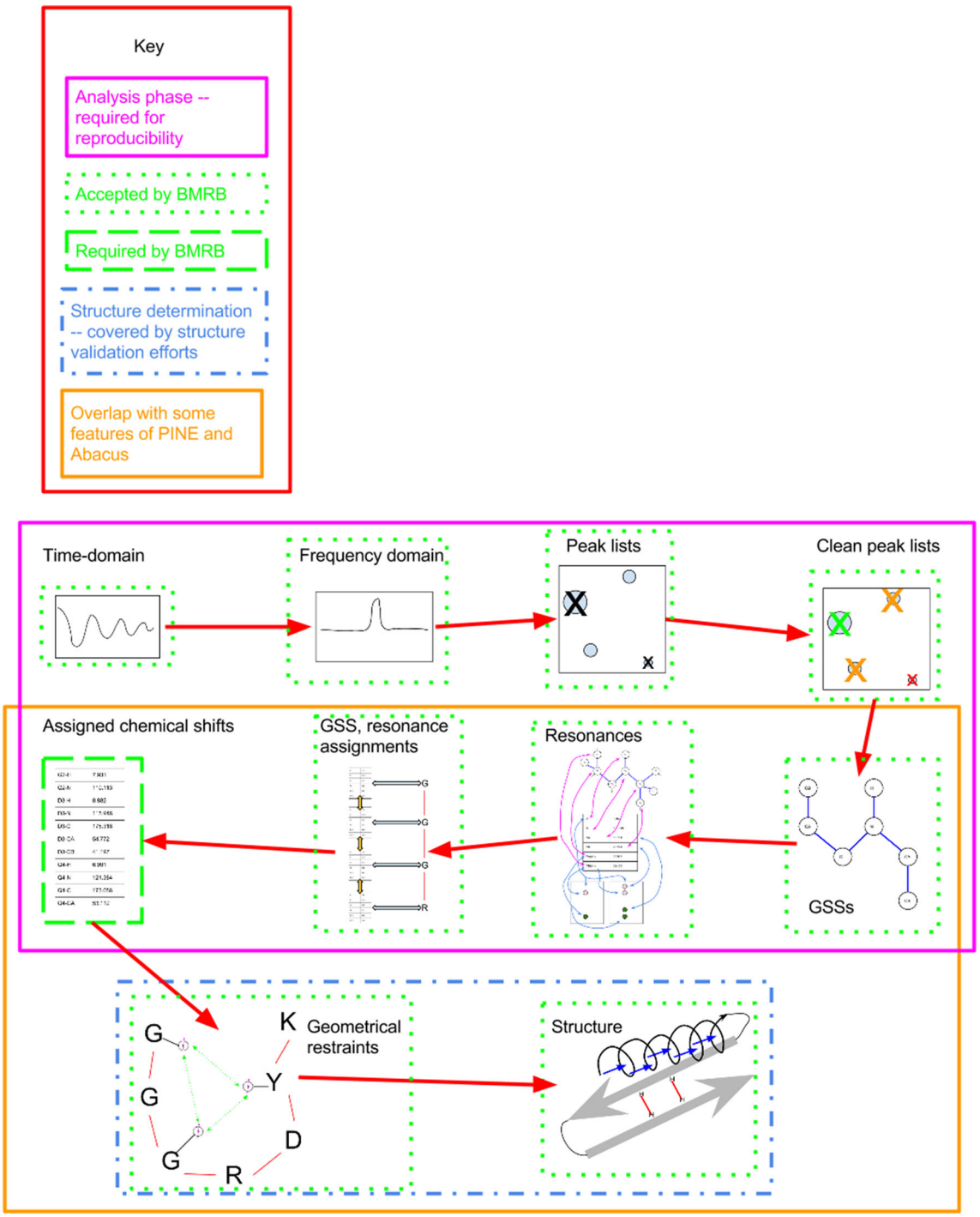

Fig. 1 Schematic of NMR data analysis. Some steps of NMR analysis are already reproducible, and some intermediate results are required to be deposited in the BMRB, while others may be deposited. Typically, intermediate peak lists, GSSs, resonances, and their assignments, and manual modifications to automated results are not deposited

to be practical and convenient. The result was that the analysis process was difficult to understand. However, as it contained many examples of annotations and problems faced during analysis, patterns and repeated elements were identified and used to re-analyze the data, being much more careful to organize the analysis and clearly annotate the progress; these were extracted into the deductive library. Principles of grouping changes and how to pace the analysis process were identified and used when the analysis was redone in the Sparky extension.

Time-domain data of the Samp3 protein was kindly provided by Dr. Mark Maciejewski. The experiments used

were the Nitrogen HSQC, HNCO (Kay et al. 1990), HNCACB (Grzesiek and Bax 1992), C(CO)NH-TOCSY (Grzesiek et al. 1993), HC(CO)NH-TOCSY (Montelione et al. 1992), HBHA(CO)NH (Grzesiek and Bax 1993), HCCH-TOCSY (Bax et al. 1990), Carbon HSQC, NOESY NHSQC (Zuiderweg and Fesik 1989), and Carbon NOESY HSQC (Marion et al. 1989).

Spectral reconstruction of the time-domain data was performed using RNMRTK processing scripts (*Rowland NMR Toolkit Script Generator*) and saved in the NMRPipe file format. The frequency spectra were then converted to Sparky format using the pipe2ucsf tool and metadata corrected using the ucsfdata tool. The spectra were then loaded into Sparky to create a new project. Continually throughout the analysis process, git was used to capture annotated snapshots of the analysis. To capture snapshots, the Sparky files were written to disk and git was invoked. Peak picking of each spectrum was initially performed using the automated peak picker built into Sparky. Manual corrections, including identifying extraneous peaks and unpicked signals, were performed immediately after automated peak picking. Additional peaks were picked during the analysis process based on Generalized Spin Systems (GSS) and resonance assignments and groupings. GSSs were initialized using signal peaks from the nitrogen HSQC experiment. GSSs were assembled by matching peaks within and across spectra based on matching chemical shifts of corresponding spectral dimensions. Ambiguities were resolved later as additional context became available. Resonance typing was initially performed using experiment definitions and BMRB chemical shift statistics, and completed during GSS typing and sequential GSS assignment. GSS typing was performed based on the resonance typing, the BMRB chemical shift statistics, and the peak patterns of intensity and sign. Sequential GSS assignment was performed based on overlap of corresponding carbon resonances with matching chemical shifts. Ambiguities were resolved with reference to the sequence specific assignments. Sequence specific assignment was performed using the matching of GSS typing and sequential GSS assignment to the primary sequence. Secondary structure predictions were made using the Talos+ (Shen et al. 2009) program included with version 2012.353.12.50 of NMRPipe, operating on the backbone chemical shift assignments. Peak picking of NOESY spectra was performed using the CCPN Analysis program. NOESY assignment and structure calculation were performed using the CYANA 3 version 3.96 program. Stereospecific assignments were made using the CYANA structure calculation and chemical shift assignment output.

The data were then extracted and exported into a single NMR-STAR file. The extraction was performed using a simple shell script based on the built-in git application programming interface (API) for accessing multiple file versions, then using a custom Python program which combined the results into a single data set, including differences between versions, and finally emitted the data in the NMR-STAR model as a single, textual file.

Results

In a fully reproducible data set, the provenance of each piece of data in the final structure—each NOE restraint used to build it, each torsion angle prediction—can be traced back to its origin in the analysis process. However, this is cumbersome to manage as the number of NOE restraints numbers in the thousands. An alternative approach, which leverages the strengths of both computers and humans, matches our use of software tools in semi-automated workflows: use software tools to perform large sets of related changes in bulk, and capture the inputs and outputs; then use manual interventions to correct any errors and omissions made by the software tool, capturing the results and the process of the manual analysis.

The core of the reproducibility approach is to capture and annotate intermediate results during the analysis process. The results are captured opportunistically, such that the full process can be recapitulated. The annotations describe the what, why, and how such that a human browsing the data can quickly understand the context of the intermediate results and how they fit into the overall process.

The end result is a data set that is a proper superset of standard analysis results, but expanded in several dimensions. First, the process of analysis is captured using snapshots of intermediate results. Second, the reasoning behind manual interventions is captured. Third, extraneous data are captured, which provide the context for data interpretation. Fourth, the rich relationships of intermediate data, which are used to obtain the final result, are captured in full throughout the process. These four tactics comprise a reproducible approach to bio-NMR analysis, and are covered in more detail.

Tactic 1: snapshots

A snapshot is a record of intermediate results from the analysis process. It contains structured data of the entire state of analysis at a specific point in time. Each snapshot has a parent snapshot. Snapshots may be compared with each other to obtain a difference. The difference between two sequential snapshots indicates the changes made in one step. The differences between sequential snapshots may be substantial or small, and applying to part or all of the data. Capturing a sequence of snapshots allows revisiting of past

analysis states and contexts, in order to understand what was done, how, and why.

Snapshots should be captured at appropriate intervals in order to create a meaningful and clean analysis history. Indiscriminately captured snapshots are hard to understand. In general, related changes should be grouped together in a single snapshot, and unrelated changes should not. When applying this approach to semi-automated analysis, snapshots are captured before and after use of an automated tool. Next, manual modifications are made as necessary and another snapshot captured. A data model for snapshots along with accompanying annotations is shown in Fig. 2.

Tactic 2: annotating rationales for human intervention in computational tasks

Each snapshot is given an annotation that provides a reason why the changes in that snapshot were made. This is a justification that the changes were appropriate and necessary in the given context. A library of annotations has been constructed which includes common and useful techniques for analyzing NMR data. This library is extensible, as missing annotations can be easily added and used, and shared among practitioners. It is presented as supplementary material. The library functions both as a tool used to describe data analysis, but also as a means of documenting and sharing analysis techniques, enhancing discoverability for newcomers.

Tactic 3: rich data model of intermediates

The richness and completeness of intermediate data determine the usefulness of snapshots. Incomplete intermediate results do not provide full context for evaluating the analysis process. Implied or missing data and relationships reduce the usefulness of intermediate results. For spectral analysis, GSSs and resonances are key components that must be captured for intermediate snapshots to be useful.

Tactic 4: preservation and identification of extraneous data

Extraneous data are not directly relatable to the final result, but rather are relevant to the quality of the process itself. Extraneous data, such as false-positive artifacts and noise peaks as well as contaminant spin systems and resonances must be appropriately identified and set aside to prevent confounding of the desired analysis results; unassigned peaks and spin systems must be captured as well. They thus introduce an element of error and bias into the analysis process. Extraneous results are naturally generated during analysis, and show how the data set was interpreted; further data sets, even collected with identical or similar techniques, may yield different extraneous results. Each annotated snapshot of an intermediate data state includes extraneous information.

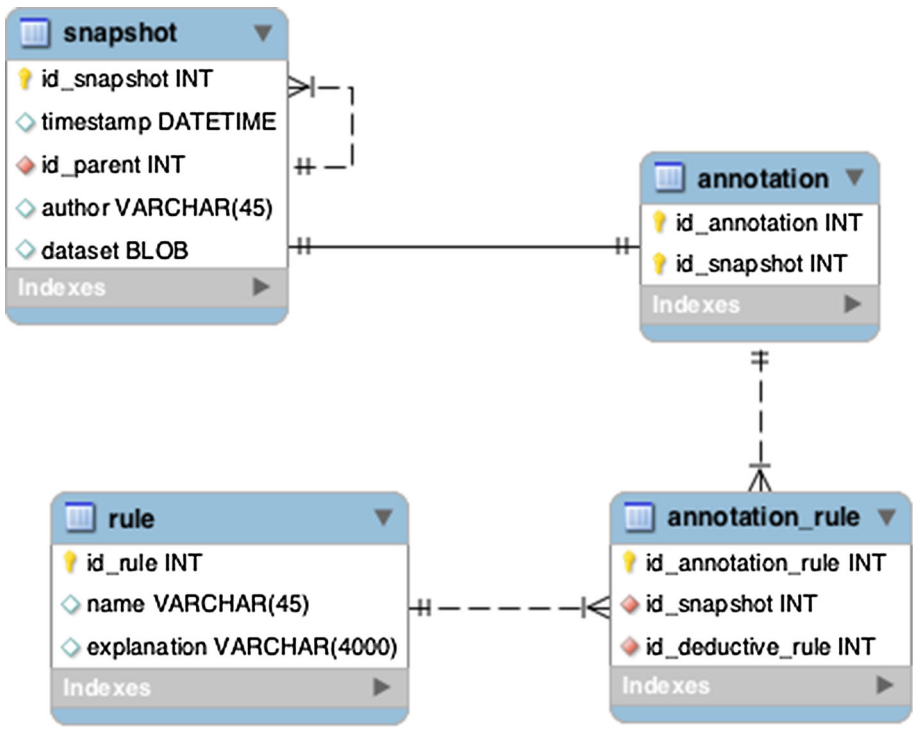

Fig. 2 Data model of snapshots and annotations. Snapshots are periodically and strategically captured during analysis in order to show the sequence of steps taken to obtain the final result. Snapshots are given annotations to provide additional context and justification of their appropriateness and applicability. The annotations may consist of one or more known rules which are commonly used to analyze NMR data. These rules can be captured and enumerated in order to provide additional explanation of their meaning and intended use, and to promote discoverability, helping newcomers to learn how to do NMR analysis more effectively and thoroughly

Capturing snapshots appropriately

The amount of changes in a snapshot should be neither too big nor too little. Snapshots that each include a single change are too little, because they create gigantic analysis histories with far too much annotation and intermediate data sets, and related changes are not grouped together into a single snapshot. A single snapshot encompassing all changes made during the entire process of analysis from start to finish is too big because unrelated changes are grouped together, many changes are lost (those in which an assignment is modified from its initial value), the contexts of specific changes are lost, and the annotations aren't applied to specific data items. The amount and similarity of changes between two snapshots should be driven by the semantics of the analysis. These should be grouped by deductive annotation; combining unrelated manual changes into a single snapshot is difficult to understand and should be avoided.

While each individual study will likely have its own idiosyncrasies, we have found that in the case of Samp3 the appropriate frequency of snapshotting is set by the nature of the common tasks. For instance, automatic peak picking followed by snapshot. Manual curating of noise and artifacts followed by snapshot. Grouping resonances into spin systems based on through-bond correlation experiment, followed by snapshot. In NMR spectral analysis, the most common situation is to perform multiple, similar inferences: identifying peaks in bulk, labeling resonances and grouping into spin systems in bulk. These large scale changes only require a single annotation as the rationale for all of the assignments is the same.

Where the NMR spectroscopist is likely to see a large benefit from the annotation strategy is in the case of ambiguous assignments, which will be unique to each study. Utilizing annotated snapshots gives the spectroscopist the option of moving forward with a low-confidence assignment while clearly documenting the rationale for later review. After further levels of analysis, these low-confidence assignments can be revisited to ascertain whether any additional data supports the original conclusion, refutes it, or whether it remains ambiguous and low-confidence.

Sparky extension which assists a spectroscopist in utilizing the strategy

In order to facilitate adoption and use of the reproducible approach, a Sparky extension was implemented which provides standard functionality for reproducibility while minimizing the additional burden of work (in terms of time and effort) placed on the user (Fig. 3). The scope of the extension is fourfold: (1) extend the core Sparky data model with the CCPN concepts of resonances and GSSs; (2) extend the core Sparky data model with the reproducibility concepts of extraneous data and notes; (3) provide facilities for the reproducibility concepts of annotated snapshots of intermediate data sets; (4) provide graphical user interface (GUI) facilities in order to simplify the correct usage of the new concepts.

The extension capitalizes on the user's familiarity with the Sparky program, pre-existing analysis strategies, and NMR concepts and data models. In order to maximize the understandability of the final, reproducible analysis, the analysis process must be carried out in an organized fashion: changes should not be made haphazardly to the data set, but rather in a principled manner. Related changes should be grouped together and snapshots taken after related groups of changes, and should be annotated appropriately.

XEasy (Bartels et al. 1995) was one of the first assignment programs to model spin systems. This model was later expanded into Generic Spin Systems (GSSs) (Zimmerman et al. 1997), and explicitly defined as part of the CCPN data model (Vranken et al. 2005) along with resonances. Our definitions for GSSs and resonances are based on the CCPN definitions; we have extended the concept into reproducibility as GSS and resonance construction and assignment are critical to spectral analysis. These two entities bridge the gap between the NMR data and the atoms and residues of the molecule. Using GSSs and resonances, the goal of assignment is to link the experimentally observed resonance signals and spin systems to the molecule. Separating these entities permits higher-fidelity results, partial interpretation, and recovery from mistakes, rendering the analysis process more tractable. For example, a peak in a through-bond spectrum indicates the presence of covalently bound resonances; matching chemical shifts across multiple spectra indicates a portion of a GSS. Peaks are first assigned to a GSS, then a GSS is assigned to a residue in the molecule. A GSS-residue assignment is easy to undo because the assignment information is not duplicated.

Annotated dataset

Samp3, a ubiquitin-like protein, was studied and chemical shift assignments performed according to the reproducibility approach. The Sparky extension described above was used, in conjunction with the VCS tool git (Loeliger and McCullough 2012), which provided facilities for annotated snapshots. The time-domain data sets were processed to frequency spectra, then peak-picked. The NHSQC peak list was manually annotated to address false positives and negatives, and the remaining peaks were used to define spin system roots, due to the number of available spectra based on H–N groups. These peaks were then used to perform restricted peak picks of the 3D experiments building on the H–N group, and the peaks merged into the

Fig. 3 Sparky reproducibility extensions. The annotation method is implemented as a Python extension to the popular spectral analysis tool Sparky. Sparky-R extends the built-in Sparky data model, allowing the user to work with GSSs and resonances. It includes functionality for identifying extraneous data such as peaks and spin systems, and capturing and annotating snapshots during analysis by integrating with the version control tool git

spin systems based on chemical shift matching. Next, the 3D peak lists were annotated, and resonances constructed based on matching of chemical shifts within spin systems. The next step was to perform sequential spin system assignment based on matching and overlap of CA and CB resonances using the HNCACB and C(CO)NH-TOCSY experiments concurrently with resonance typing where possible. This resulted in near complete backbone assignment. Next, the aliphatic sidechains were assigned using BMRB statistics and HCCH-TOCSY splitting patterns, before assigning aromatic sidechains. Finally, the NOESY spectra were peak-picked, and these peak lists imported to Cyana (Güntert 2004) along with Talos+ (Shen et al. 2009) torsion angle predictions in order to calculate a structure. The final results were in the form of a git repository, containing appropriately grouped and annotated snapshots. The extracted NMR-STAR file containing the full history of the analysis process along with appropriate metadata have been deposited to the BMRB as entry 25258.

Discussion and conclusions

Practical implications of reproducibility strategy and its implementation

The reproducibility approach requires a minimal amount of extra time and effort. We consider the additional burden imposed by reproducibility concerns to be negligible, while the ability to inspect previous data states and annotations is valuable when difficult-to-analyze data phenomena are reached. Identification of extraneous signals and snapshot annotation, both new tasks, are quick and easy due to the Sparky reproducibility extension. Furthermore, the strength of the data model—including both resonances and GSSs as well as the sequential, annotated history model—lead to data that is less confusing and easier to understand as analysis proceeds, because more context is available in which to evaluate and understand results. For example, one may query the history of the analysis process, and compare

changes, additions, and deletions between any two snapshots, or obtain a high-level overview of the complete history of the project.

Principles of the reproducible approach

Data is never deleted or omitted. Rather, data items are re-categorized as necessary. For example, whereas in a typical analysis approach, false positive peaks picked by an automated peak picker are simply deleted during the manual validation and correction intervention, in this approach, such false peaks are instead re-categorized as artifacts or noise, but not deleted. This leaves an explicit record of the results of the automated peak picker as well as the manual changes, enabling future reinterpretation in case of a mistake or ambiguity.

Future implementations in other software tools

This strategy, in principle, could be implemented in other assignment tools as well. The first two tactics of sequential snapshots and annotations can be implemented using git, orthogonally to the main program—it does not have to know or care about the git repository providing the annotated snapshot functionality. This should be possible for tools such as CCPN Analysis, XEasy (Bartels et al. 1995), NMRViewJ (Johnson 2004), and Cara (Keller 2004). The third tactic of explicit spin systems and resonances depends on the specific program in question; these were implementable in Sparky due to its flexibility and integrated Python extension language. CCPN Analysis already has this functionality. The fourth tactic of capturing extraneous results could be implemented in several ways: the program in question could be ignorant or aware of the extraneous results. The advantage of permitting the program to be ignorant is that such an approach can be easily implemented as data captured in additional files alongside the program's standard data files; the disadvantage is potentially decreased integration between the user interface and the extraneous data. On the other hand, if the program were aware of the extraneous data, better integration of such data would be possible, but the ability to and ease of such a new feature would be limited by the nature and licensing of the program in question. CCPN Analysis would be able to easily implement such an extension, due to its integrated Python interpreter, providing the capability of extensions.

Pedagogy: teaching and learning of NMR processes and results

By introducing a means to talk about and discuss what is actually carried out in an NMR analysis process, it becomes possible to point students and novices to specific examples of various analyses, types of deductions, and interpretations in data contexts of differing quality. This may also be extended to cover topics such as identification and correction of errors: an expert spectroscopist analyzes a data set, finds an error, and corrects while appropriately annotating and snapshotting the work. Later, students are able to study the example in detail, having access to the full context that enabled the expert to find the correct interpretation, as well as the reasoning used to do so. This enables a quicker and more reliable building of an intuitive understanding of how to analyze NMR data, the difficulties and problems inherent in analysis, and the ability to identify, understand, and correct mistakes.

Assessing and improving analysis quality

Capturing the full analysis process enables critical assessment of the quality of the analysis. This includes assessment of peak picker quality, with respect to false positives and negatives, spectral data quality including numbers of noise and artifact peaks, numbers and possibly sources of extraneous signals, spin systems, and resonances, ambiguity of sequential and sequence-specific assignments, and estimated error rates of resonance and GSS typing and assignments. It will also be possible to begin to understand the sources of bias during automated analysis and manual interventions, and propose strategies to eliminate or mitigate them.

This may help to improve analysis quality by offering continuous feedback to the spectroscopist of the troublesome, uncertain, and low probability areas of the analysis. This is similar to the PINE program (Bahrami et al. 2009), which provides estimates of probability along with assignment possibilities. Such probability estimates give a better indication of the actual difficulties faced in interpreting the data than that offered by "all-or-nothing" unambiguous assignments. Furthermore, a reproducible approach which makes it far easier to identify and correct mistakes thereby also reduces the cost of a negative mistake. This means it is less pressing that mistakes are avoided, thereby freeing a spectroscopist to make uncertain assignments and guesses, if necessary, secure in the knowledge that such uncertainties will be clearly marked in the data set, and will be correctable without undue effort.

Many computational tools in NMR are not able to perform as well as expert human spectroscopists in tricky, complicated cases, due to the ability of the human to bring additional context to bear as needed (Williamson and Craven 2009). On the other hand, software tools have continually improved their success rates. The annotation system provides another avenue of improvement for tools: with the creation of fully annotated, reproducible data sets, much more information will be available to use as test data

sets for algorithm development, and rules repositories for training programs. Thus, accurately annotating analysis generates data which may be used to derive patterns and rules for use in automated tools.

The general approach of capturing intermediate states of data analysis through the use of annotated snapshots maps straightforwardly to other problem domains—in fact, it is commonly used in software code development. In the scientific realm, biomolecular NMR studies are critically in need of this approach, due to the lengthy and complex data analysis pipeline coupled with the need for manual data curation at various stages (Fig. 1).

The significance of reproducibility

The key to reproducibility is to explicitly capture all relevant data and context of a scientific process. Several advantages flow naturally from such explicit data capturing; when data is explicit, it can be shared, queried, and learned from. Capturing fully reproducible analysis processes enables sharing, archival, and dissemination of results. By increasing visibility of the methods that are used, this helps avoid mistakes, errors, and bias, to improve and spread the improvements to scientists worldwide, and to increase the rate of progress by lowering the barriers to sharing. Collaboration between researchers is simplified, and differences between analyses can be identified and understood. Reproducibility facilitates maintenance of results, in which new data are collected, analyzed and merged into existing results, extending the biophysical characterization. Similarly to earlier efforts to introduce reproducibility in computational and experimental science, we aim to introduce reproducibility to NMR data analysis.

Acknowledgments This research was funded by United States National Institutes of Health Grant GM-083072. The authors would like to thank Dr. Mark Maciejewski for kindly providing time-domain data of the Samp3 protein and Dr. Woonghee Lee for adding the reproducibility extensions to the NMRFam release of Sparky.

Conflict of interest The authors declare that they have no conflict of interest.

References

Bahrami A et al (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. PLoS Comput Biol 5(3): e1000307

Bartels C et al (1995) The program XEASY for computer-supported NMR spectral analysis of biological macromolecules. J Biomol NMR 6(1):1–10

Bax A, Clore GM, Gronenborn AM (1990) 1H 1H correlation via isotropic mixing of 13C magnetization, a new three-dimensional approach for assigning 1H and 13C spectra of 13C-enriched proteins. J Magn Reson 88:425–431

Buckheit JB, Donoho DL (1995) Wavelab and reproducible research. Springer, New York

Collins FS, Tabek LA (2014) NIH plans to enhance reproducibility. Nature 505:612–613

Dall'Olio GM, Bertranpetit J, Laayouni H (2010) The annotation and the usage of scientific databases could be improved with public issue tracker software. Database 2010:baq035

Eclipse IDE (2007) The Eclipse Foundation. www.eclipse.org

Goddard TD, Kneller DG (2004) SPARKY 3. University of California, San Francisco, p 15

Grzesiek S, Bax A (1992) An efficient experiment for sequential backbone assignment of medium-sized isotopically enriched proteins. J Magn Reson 99(1):201–207

Grzesiek S, Bax A (1993) Amino acid type determination in the sequential assignment procedure of uniformly 13C/15N-enriched proteins. J Biomol NMR 3(2):185–204

Grzesiek S, Anglister J, Bax A (1993) Correlation of backbone amide and aliphatic side-chain resonances in ^{13}C ^{15}N-enriched proteins by isotropic mixing of ^{13}C magnetization. J Magn Reson Ser B 101(1):114–119

Guerry P, Herrmann T (2011) Advances in automated NMR protein structure determination. Q Rev Biophys 44(03) :257–309

Güntert P (2004) Automated NMR structure calculation with CYANA. In: Downing AK (ed) Methods in Molecular Biology, vol. 278: Protein NMR techniques. Humana Press, Totowa, pp 353–378

Güntert P (2009) Automated structure determination from NMR spectra. Eur Biophys J 38(2):129–143

Ioannidis JPA et al (2008) Repeatability of published microarray gene expression analyses. Nat Genet 41(2):149–155

Johnson BA (2004) Using NMRView to visualize and analyze the NMR spectra of macromolecules. In: Downing AK (ed) Methods in Molecular Biology, vol. 278: Protein NMR techniques. Humana Press, Totowa, pp 313–352

Kay LE et al (1990) Three-dimensional triple-resonance NMR spectroscopy of isotopically enriched proteins. J Magn Reson 89(3):496–514

Keller RLJ (2004) Optimizing the process of nuclear magnetic resonance spectrum analysis and computer aided resonance assignment. Diss ETH No. 15947. Diss. Swiss Federal Institute of Technology, Zurich

Landis SC et al (2012) A call for transparent reporting to optimize the predictive value of preclinical research. Nature 490(7419): 187–191

Loeliger J, McCullough M (2012) Version control with Git: powerful tools and techniques for collaborative software development. O'Reilly Media Inc, Sebastopol

Marion D et al (1989) Three-dimensional heteronuclear NMR of nitrogen-15 labeled proteins. J Am Chem Soc 111(4): 1515–1517

Montelione GT, Lyons BA, Emerson SD, Tashiro M (1992) An efficient triple resonance experiment using carbon-13 isotropic mixing for determining sequence-specific resonance assignments of isotopically enriched proteins. J Am Chem Soc 114(27): 10974–10975

Open Source Initiative (2006) The MIT License. http://opensource.org/licenses/MIT

Peng RD (2011) Reproducible research in computational science. Science 334(6060):1226

Prinz F, Schlange T, Asadullah K (2011) Believe it or not: how much can we rely on published data on potential drug targets? Nat Rev Drug Discov 10(9):712

Rowland NMR Toolkit Script Generator. Web. September 18, 2014. http://sbtools.uchc.edu/nmr/nmr_toolkit/

Shen Y et al (2009) TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44(4):213–223

Stodden V, Miguez S (2014) Best practices for computational science: software infrastructure and environments for reproducible and extensible research. J Open Res Softw 2(1):1–6. doi:10.5334/jors.ay

Ulrich EL et al (2008) BioMagResBank. Nucleic Acids Res 36(suppl 1):D402–D408

Vranken WF et al (2005) The CCPN data model for NMR spectroscopy: development of a software pipeline. Proteins 59(4):687–696

Williamson MP, Craven CJ (2009) Automated protein structure calculation from NMR data. J Biomol NMR 43(3):131–143

Zimmerman DE et al (1997) Automated analysis of protein NMR assignments using methods from artificial intelligence. J Mol Biol 269(4):592–610

Zuiderweg ERP, Fesik SW (1989) Heteronuclear three-dimensional NMR spectroscopy of the inflammatory protein C5a. Biochemistry 28(6):2387–2391