CrossMark

ARTICLE

NMRFAM-SDF: a protein structure determination framework

Hesam Dashti[1] · Woonghee Lee[1] · Marco Tonelli[1] · Claudia C. Cornilescu[1] ·
Gabriel Cornilescu[1] · Fariba M. Assadi-Porter[1] · William M. Westler[1] ·
Hamid R. Eghbalnia[1] · John L. Markley[1]

Abstract The computationally demanding nature of automated NMR structure determination necessitates a delicate balancing of factors that include the time complexity of data collection, the computational complexity of chemical shift assignments, and selection of proper optimization steps. During the past two decades the computational and algorithmic aspects of several discrete steps of the process have been addressed. Although no single comprehensive solution has emerged, the incorporation of a validation protocol has gained recognition as a necessary step for a robust automated approach. The need for validation becomes even more pronounced in cases of proteins with higher structural complexity, where potentially larger errors generated at each step can propagate and accumulate in the process of structure calculation, thereby significantly degrading the efficacy of any software framework. This paper introduces a complete framework for protein structure determination with NMR—from data acquisition to the structure determination. The aim is twofold: to simplify the structure determination process for non-NMR experts whenever feasible, while maintaining flexibility by providing a set of modules that validate each step, and to enable the assessment of error propagations. This framework, called NMRFAM-SDF (NMRFAM-Structure Determination Framework), and its various

components are available for download from the NMRFAM website (http://nmrfam.wisc.edu/software.htm).

Introduction

NMR spectroscopy has emerged as the premier approach for obtaining information about biomolecular interactions, structural dynamics, and three-dimensional structure in solution. However, the collection, processing, interpretation, and validation of NMR data remain challenging, and present barriers to more widespread applications. Efforts in the NMR community in the past two decades have focused on the automation of discrete steps involved in analyzing NMR data. More specifically, streamlining the overall sequence of steps in the procedure of protein structure calculation has received considerable attention (Lopez-Mendez and Guntert 2006; Serrano et al. 2012). The goal of the CASD-NMR competitions has been to foster the development of automated methods that lead to structures whose quality approaches those determined by tedious manual methods (Rosato et al. 2009, 2012).

The common process for NMR protein structure calculation begins with collecting NMR data for a number of through-bond and through-space experiments that will be processed into the frequency domain representation. A peak identification step, called peak-picking, is required to identify the signals of interest in the processed data. The chemical shifts of the peaks are assigned to the atoms of the backbone and side chains, and the assigned chemical shifts are used as labels for identifying NOE cross peaks in the

✉ John L. Markley
markley@nmrfam.wisc.edu

[1] National Magnetic Resonance Facility at Madison, Biochemistry Department, University of Wisconsin-Madison, 433 Babcock Drive, Madison, WI, USA

NOESY spectra. These cross peaks provide spatial restraints for the 3D structure of the protein in the study (Clore and Gronenborn 1987, 1991; Wüthrich 1986). Spatial restraints, along with an empirical force-field, are then used to arrive at an ensemble of low energy structures that satisfy most of the restraints.

Long data acquisition times are a potential limiting factor in NMR studies, particularly with unstable targets, and a number of approaches have been developed for improving data acquisition through computational or experimental means (Bahrami et al. 2012; Brutscher 2013; Frydman et al. 2004; Hoch et al. 2007; 2014; Hyberts et al. 2012; Kim and Szyperski 2003; Kupce and Freeman 2003a; Lee et al. 2013; Lescop et al. 2007, 2009; Maciejewski et al. 2006; Orekhov et al. 2003; Orekhov and Jaravine 2011; Schanda and Brutscher 2005; Szyperski et al. 2002). Toward accelerating the data acquisition and consequently improving the sensitivity of the spectra, modifications in pulse programs have been introduced (Brutscher 2013; Frydman 2006; Lescop et al. 2007). Irregular or non-uniform sampling (NUS) schemes represent an alternative approach to conventional data collection (Bahrami et al. 2012; Hoch et al. 2007, 2014; Hyberts et al. 2012; Kim and Szyperski 2003; Kupce and Freeman 2003b; Maciejewski et al. 2006; Mobli and Hoch 2008; Orekhov et al. 2003; Orekhov and Jaravine 2011). Ultimately, the gains in time or sensitivity introduced by computational processes must be validated to ensure the robustness of signal identification—or peak picking. And, despite developments in peak picking algorithms (Alipanahi et al. 2009; Cheng et al. 2013; Chylla et al. 1998; Shin and Lee 2008; Tikole et al. 2014), the ability to deconvolve peaks in split or overlapped peaks remains unsatisfactory. Some data collection methods have the potential to distinguish between noise and peaks by employing a peak identification algorithm (Bahrami et al. 2012; Hiller et al. 2005; Kim and Szyperski 2003). However, for robust automation, validating the output from individual steps, or the combined steps of spectral processing and peak picking, remains a necessity.

Eghbalnia et al. (2005) and Bahrami et al. (2009) demonstrated that the computational problem of assigning protein chemical shifts from through-bond NMR experiments is of the class mathematicians call "NP-hard" (Bovet and Plerlulgi 1994). This infers a limitation on purely deterministic algorithms for chemical shift assignment or validation. Instead, it was proposed that automated chemical shift assignment approaches rely on non-deterministic or probabilistic algorithms (Bahrami et al. 2009, 2012; Schmidt and Guntert 2012), where a probabilistic validation process becomes optimal. Alternatively, when the chemical shift assignment method uses a deterministic algorithm in its core decision-making process (Jung and

Zweckstetter 2004; MacRaild and Norton 2014; Xu et al. 2006), validation can utilize an accept-reject criterion, an approach that is suitable only when spectral signals are nearly complete and unambiguous.

The practice of structure determination by NMR spectroscopy involves a number discrete decision making steps that give rise to a non-linear relation between the inputs and outputs. The cumulative impact of nonlinear input–output relations could lead to unexpected and unpredictable errors. Stepwise and continuous validation can inform users of potential inconsistencies early in the process and flag them for optional correction; including manual corrections by users. Among existing data acquisition methods, ADAPT-NMR (Bahrami et al. 2012) provides a supporting verification GUI (graphical user interface), named ADAPT-NMR Enhancer (Lee et al. 2012). Other methods such as the ist@HMS (Hyberts et al. 2012) are designed with the goal of improving the sensitivity and resolution of multidimensional experiments by using non-uniform sampling data collection. More recently, the NESTA program (Sun et al. 2015) was developed to speed up the reconstruction of non-uniform sampled spectra thus making it more feasible for this method to be incorporated into high-throughput and automated approaches.

Accurate chemical shift assignment plays an important role in structure determination (Jee and Guntert 2003). The PINE (Probabilistic Interaction Network of Evidence) algorithm provides a probabilistically ranked set of possible assignments for every atom that users can use to investigate different possible candidates (Bahrami et al. 2009). The computational complexity of the chemical shift assignment for large proteins motivated us to introduce the PINE-SPARKY (Lee et al. 2009) to help users explore the possible assignments and validate the assignments by visualization on designated spectra. In addition to these probabilistic methods, a second category of assignment validation methods relies on chemical shift statistics (Moseley et al. 2004; Wang et al. 2005, 2010). Although useful, methods in this category do not consider the specific characteristics of the protein under study and therefore may cause false-negative and false-positive results (Dashti et al. 2015). This limitation is addressed by our recent introduction of ARECA, a probabilistic validation method that uses the NOESY spectra (or the corresponding peak lists) of the protein to validate the chemical shift assignments. The assessment of the reliability of chemical shift assignment (ARECA) package (Dashti et al. 2015) is the first probabilistic method that uses the large body of through-space statistics to validate chemical shift assignments. The CASD-NMR (Rosato et al. 2009, 2012) provided data-sets with raw and refined peaks that were used for evaluating ARECA in determining whether the assignments provided were consistent with the given NOESY peak lists.

The difficulty of the resonance assignment problem can increase when through-space (NOESY) experiments are considered—in this case, the number of peaks depends on the protein structure as well as the length of the sequence. A significant part of automation literature in NMR is focused on through-bond experiments (Bahrami et al. 2009; Hiller et al. 2005; Jung and Zweckstetter 2004; MacRaild and Norton 2014; Wu et al. 2006; Xu et al. 2006; Zimmerman et al. 1997) or mapping through-bond assignments into short-range NOESY contacts and predicting long-range NOE assignments (Güntert 2004; Herrmann et al. 2002; Lee et al. 2011, 2014a). This is, in part, a reflection of the additional computational complexity of NOE cross peak assignments (Linge et al. 2003; Schmidt and Guntert 2012), which includes the additionally complex task of extracting the distance restraints between the atoms. The ambiguities in assignment of long-range NOE cross peaks result in a set of intricate distance restraints that include a combination of ones that are correct and incorrect. Therefore finding the most suitable set of restraints to achieve an energetically favorable structure becomes a challenging optimization problem. The search for an optimal restraint set is usually performed by validation of the calculated intermediate structures and examination of the restraints used or discarded during the structure determination process (Güntert 2004; Herrmann et al. 2002; Kuszewski et al. 2004, 2008; Linge et al. 2003; Schwieters et al. 2003). The need for expertise in multiple areas (such as spectroscopic, structural, biochemical, and biophysical fields) and familiarity with several software tools makes this one of the most challenging remaining steps in NMR structure determination. PONDEROSA (Peak-picking Of NOE Data Enabled by Restriction of Shift Assignments) (Lee et al. 2011) addresses this challenge by automatically selecting peaks in the NOESY spectra and simultaneously interfacing with TALOS + (Shen et al. 2009), STRIDE (Frishman and Argos 1995) and CYANA (Güntert 2004) in an iterative process in order to identify the most reliable set of restraints. The recent introduction of PONDEROSA-C/S (Lee et al. 2014a) adds new functionality for user convenience by providing Ponderosa Client and Ponderosa Analyzer programs as interfaces to the core computational server (Ponderosa Server). In the course of developing PONDEROSA-C/S, data sets from CASD-NMR (Rosato et al. 2009, 2012) were used to evaluate and refine the algorithms in the Ponderosa Server. Ponderosa Analyzer is a reliable validation package for both identifying restraint violations and providing tools for investigating the structure and adjusting it to better fit to the experimental data. The package provides tools for visualizing the automatically generated restraints on the 3D structure and spectra by interfacing with PyMOL (DeLano and Lam 2005) and NMRFAM-SPARKY (Lee et al. 2014b). Other methods for structure validation include those that use statistics from structures in databases (Chen et al. 2010; Davis et al. 2004; Laskowski et al. 1993, 1996; Rieping et al. 2014; Shen and Bax

2007; Vranken and Rieping 2009), and those that consider the NOESY experiments for their structure validation (Huang et al. 2005).

The scheme shown in Fig. 1 summarizes various choices and validation steps involved in conventional protein structure determination in the absence of automation. Decisions at the many steps are made according to knowledge and experience and are difficult to document and thus reproduce. User-friendly validation tools are frequently lacking for intermediate steps, and the preparation of input data for structure calculation depends on the program that will be used. If the outcome of the final structure validation is satisfactory, then the process stops. Otherwise, one needs to go back to every step of the process for more precise validation and necessary adjustments.

We introduce here a framework for the process of structure calculation, that a) provides a guideline towards simplifying the process for users with limited NMR background, b) removes the necessary human intervention in data conversion and preparing inputs for discrete steps of the process, c) accelerates the structure calculation process by interconnecting different software packages, d) incorporates validation methods to avoid error accumulation and propagation, and e) incorporates user-friendly refinement modules so the users can perform adjustments whenever needed. Validation is accomplished through statistical analysis and graphical user interfaces that allow results to be compared with underlying data. Smaller and well-behaved proteins are most amenable to full automation, but the framework can be adapted to deal with larger and less well-behaved targets.

Materials and methods

Organization

Our approach is organized into three steps: (a) data acquisition and processing (including peak picking), (b) chemical shift assignment, and (c) structure determination. NMRFAM-SDF is an object-oriented framework that implements the three steps of this process (Fig. 2), and automatically performs the necessary interconnections between each step. The organization of the modules in this framework is optimized and aimed at complete fully-automated structure determination for well-behaved proteins. After the NMR sample is inserted into the NMR spectrometer, the remaining steps are executed effortlessly leading to structure calculation and refinement. However, for more challenging protein targets, the validation tools identify problems and guide the user to modify the strategy in order to overcome them. The object-oriented organization supports utilities that enable the substitution of every

Fig. 1 Conventional steps in manual protein structure determination are shown in the *green boxes*. The *blue triangles* indicate decision making steps that user is expected to perform. The *red-lines* show feedback loops from validation steps to the prior steps
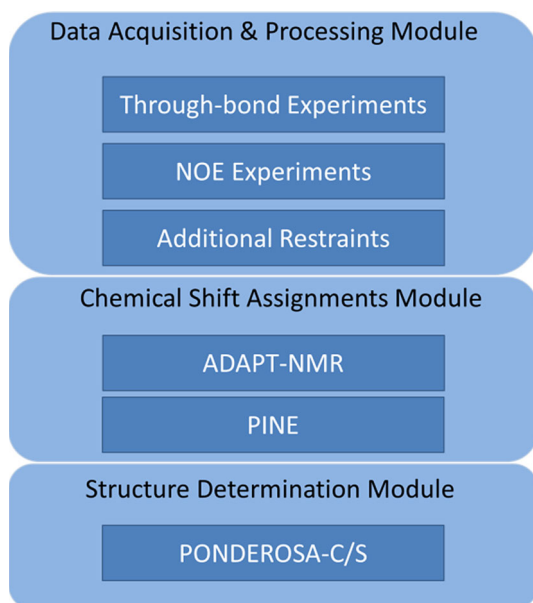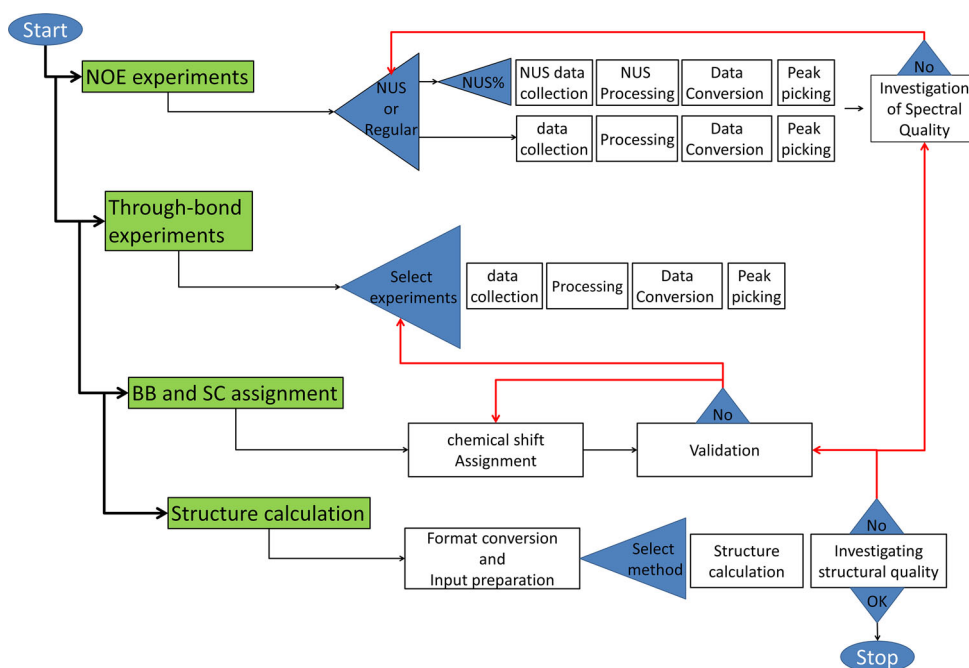
Fig. 2 Overall structure of the NMRFAM structure determination framework

module while maintaining the workflow of the framework. The modules of the framework are described in the following three sections.

Data acquisition and processing module

The 'data acquisition and processing' module consists of three units that focus, respectively, on through-bond experiments, through-space (NOE) experiments, and

additional restraints. The tools currently implemented in this module are shown in Fig. 3. Each unit of the module provides a number of options for performing the targeted task (shown as connected boxes in Fig. 3). Orange boxes identify the associated validation tools for each unit.

Through-bond experiments

NMRFAM-SDF provides three choices for through-bond experiments: (a) ADAPT-NMR, which uses a non-uniform sampling approach by collecting 3D spectra as tilted 2D planes; (b) non-uniform sampling with iterative soft thresholding (ist@HMS) (Hyberts et al. 2012) with two options for scheduling (default) (Hyberts et al. 2012) or (alternative) NUS-Score (Aoto et al. 2014), and with two options for reconstructing the spectra (default) ist@HMS or (alternative) the much faster NESTA (Sun et al. 2015); and (c) regular sampling by conventional 3D or 4D NMR experiments. Peak picking is an integrated part of ADAPT-NMR, which also achieves probabilistic chemical shift assignments. For the two other options, a peak picking step is required. For these two options, NMRFAM-SDF uses an enhanced approach to the restricted peak picking (Lee et al. 2014b). The validation component, ADAPT-NMR Enhancer, can be used for investigating and validating the results of the tilted-plane data collection and chemical shift assignment. NMRFAM-SPARKY (Lee et al. 2014b) can be used for validating the resolution and sensitivity of spectra collected by options (b) or (c).

Fig. 3 The data acquisition and processing module
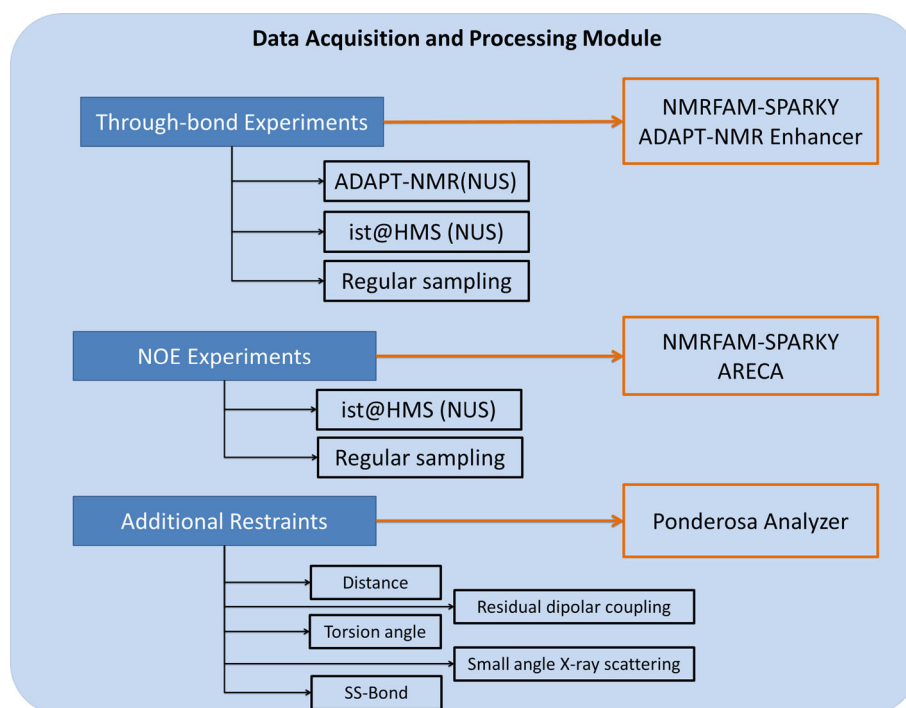
Through-space (NOE) experiments

NMRFAM-SDF provides two options for collecting NOE experiments: non-uniform sampling with ist@HMS, or regular sampling. Although these options are suitable for well-behaved proteins, the importance of NOESY experiments to achieve proper structural folds makes the validation of through-space experiments crucial. NMRFAM-SPARKY is equipped with tools that map and transfer the chemical shift assignments from the through-bond experiments onto NOESY spectra (two-letter code: ta). The resulting map can be visualized and used to evaluate the quality (resolution and sensitivity) of the spectra. Additionally, our chemical shift validation software, ARECA (Dashti et al. 2015), is used to evaluate the consistency between the assignments and the NOESY spectra (or the corresponding peak lists).

Additional restraints

Additional restraints can be incorporated on the basis of the user's knowledge of the protein under investigation, from manually analyzed experiments (disulfide bonds, residual dipolar coupling, small-angle scattering, or other sources). These additional restraints can be used as auxiliary information to help with the structure determination and/or to validate the final structure.

Chemical shift assignment module

The chemical shift assignment module consists of two packages for assigning backbone and side chain atoms. Figure 4 illustrates these packages and their validation tools. When the user selects ADAPT-NMR, assignments are generated automatically during the Bayesian NUS data acquisition. The PINE package facilitates chemical shift assignments from the alternative approaches that generate peak lists associated with particular NMR experiments.

ADAPT-NMR Enhancer and ARECA can be used to validate the chemical shift assignments generated by ADAPT-NMR. Validation of PINE's output can be performed by PINE-SPARKY (Lee et al. 2009) (incorporated into NMRFAM-SPARKY), or the ARECA package.

Structure determination module

The core of the structure determination module is the PONDEROSA-C/S package (Fig. 5), which uses the outcomes of the assignments module, the NOE experiments (either raw spectra, refined peak lists, or unrefined peak lists) and the additional restraints for initiating and completing the structure determination step (distance, angle, RDC and SAXS). Cyana (Güntert 2004) formatted files are required for restraints (the Ponderosa Server interconverts these between Cyana and Xplor-NIH formats) with the exception of

Fig. 4 Different computational options in the chemical shift assignment module
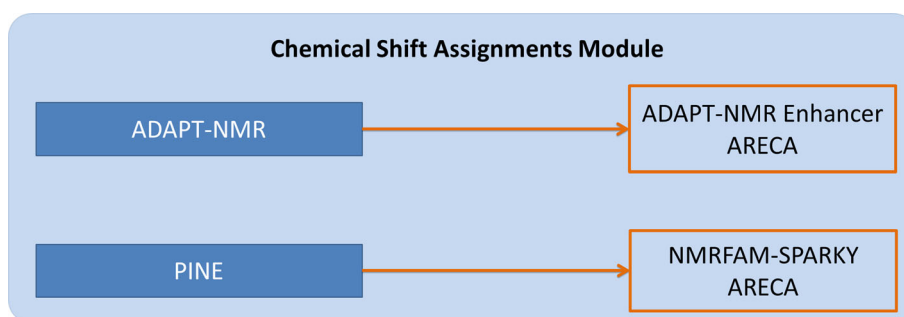
Fig. 5 Structure determination module
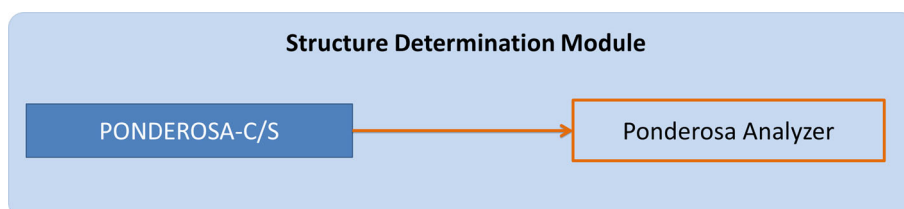
the raw output from SAXS, which is supported by Xplor-NIH (Kuszewski et al. 2004; 2008; Schwieters et al. 2003). This module is started automatically in our approach unless the user elects to deploy other methods for preparing the input data. Ponderosa Analyzer can be used to validate, evaluate, and adjust the violations in the calculated structure.

Results

In this section, we discuss applications of NMRFAM-SDF and demonstrate the use of different options within this framework. In all but one case, the proteins chosen for these illustrations are ones with manually determined structures deposited in the PDB, which could be used for comparison; they include targets used in the CASD-NMR competitions. The protein sample conditions are provided in the supplementary materials Table S1.

[U-^{13}C, U-^{15}N]-brazzein (53 amino acid residues)

The framework used in this structure determination is shown in Fig. 6.

Steps 1 and 2 (NOESY data collection): Non-uniform sampled data (at a level of 25 %) were collected on a Varian 600 MHz spectrometer; the ist@HMS package was used for scheduling, data collection, and reconstruction of both the ^{15}N- and ^{13}C-editted NOESY spectra (23 h for each experiment). The Ponderosa Client program was used for peak picking.

Step 3 (through-bond data collection and assignment): ADAPT-NMR was used for data collection and assignment of the backbone and side chain atoms. Figure 7 shows the

collected experiments and elapsed time for both data acquisition and chemical shift assignments.

Step 4 (validation with ADAPT-NMR Enhancer): ADAPT-NMR Enhancer was utilized to validate the chemical shift assignments by checking them against the spectral data.

Step 5 (validation with ARECA): The ARECA package was used to evaluate the consistency between the NOESY spectra and the assignments. ARECA flagged 133 atoms (25.3 %) of the total number of assigned atoms) with low probabilities (probabilities less than 50 % are considered low). Because more than 5 % of the atoms were flagged, inconsistency between the assignments and the NOESY spectra was considered a possibility. Figure 8a shows ARECA's report on the overall probabilities of the backbone heavy atoms.

Step 6 (NOESY data collection): Because ARECA's report on the NOESY data was unsatisfactory, the NOESY spectra were inspected manually with NMRFAM-SPARKY, and a regularly-sampled ^{13}C-edited NOESY spectrum was collected, and used to replace the ^{13}C-NOESY (NUS) data.

Step 7 (validation with ARECA): The regularly-sampled ^{13}C-edited NOESY spectrum, along with the non-uniformly sampled ^{15}N-edited NOESY spectrum, were used to recalculate ARECA's probabilities. ARECA flagged only 13 atoms (2.48 %) with low probabilities, which was a significant improvement on the consistency between the new set of NOESY spectra and the assignments. Figure 8b shows the overall probabilities of the backbone heavy atoms as reported by ARECA.

Step 8 (Structure calculation with PONDEROSA-C/S): Ponderosa Client submitted the complete validated data
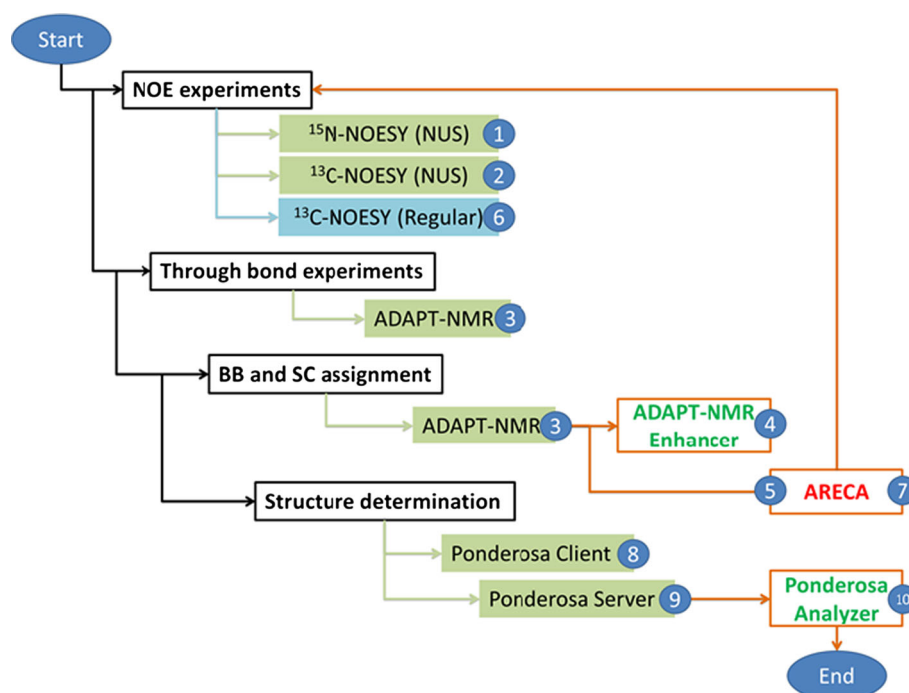
Fig. 6 Workflow from the NMRFAM-SDF used for the automated determination of the 3D structure of the protein brazzein. The *black boxes* show different modules of the workflow. The *filled boxes* show the methods used for every module; of these, the *green boxes* indicate steps performed automatically and the *blue box* indicates that ^{13}C-NOESY (*regular*) data were collected in response to a validation step. Validation methods are shown within the *orange boxes*. ADAPT-NMR Enhancer and Ponderosa Analyzer confirmed the quality of the data. ARECA reported a high number of suspicious chemical shift assignments at step (5). This prompted the collection of a ^{13}C-NOESY spectrum by regular sampling, which resulted in an acceptable ARECA score at step (7). Steps 8–10 resulted in a structure that passed validation

package to the Ponderosa Server. The refinement option was set to use Cyana for NOE assignment and structure calculation, and Xplor-NIH for water refinement (PONDEROSA refinement option).

Step 9 (Structure evaluation with Ponderosa Analyzer): Table S2(a) shows the PONDEROSA-C/S and PSVS (Bhattacharya et al. 2006) structure validation reports for this structure. These reports on the quality of the structure were satisfactory; therefore, the structure determination was considered to be successful, and the process was stopped. To further evaluate the results of this workflow, the chemical shift assignments and the calculated structure were compared with the manually derived assignments (BMRB entry 16215) and structure of the protein (PDB entry 2LY5) (Cornilescu et al. 2013). Comparison of chemical shifts assignments indicated that 84.3 % of the overall backbone and side chain assignments achieved automatically were in agreement with those deposited in BMRB. We consider the deposited assignments to be correct, because they were obtained in the course of structure determination and refinement. Despite the 15.7 % erroneous assignments, the structure calculated automatically contained the expected strands and helices and

had a backbone RMSD of 1.67 Å to the manually refined structure (Fig. 9a).

In order to test whether the early validation step was necessary for achieving a good structure, we used the non-uniformly sampled ^{15}N- and ^{13}C-edited NOESY spectra as input to the Ponderosa Server (despite the 25.3 % assignments flagged by ARECA). The resulting structure (Fig. 9b) was missing the three strands and had a backbone RMSD of 2.91Å to the manually determined structure. Table S2(b) shows the structure validation reports for this structure generated by PONDEROSA-C/S and PSVS.

To evaluate the influence of erroneous assignments on the quality of the structure, we used the regularly-sampled ^{15}N- and ^{13}C-edited NOESY spectra and correct manual assignments (BMRB entry 16215) as input to NMRFAM-SDF. The resulting structure (Fig. 9c) had a backbone RMSD of 1.22 Å from the manually refined structure (PDB entry 2LY5). From the validation report (Table S2(c)), it is clear that the overall quality of the structure is improved. However, the original structure determined with minimal human intervention (Fig. 9a) was of sufficient quality that it could have been used as a starting point for manual validation and refinement of the structure.
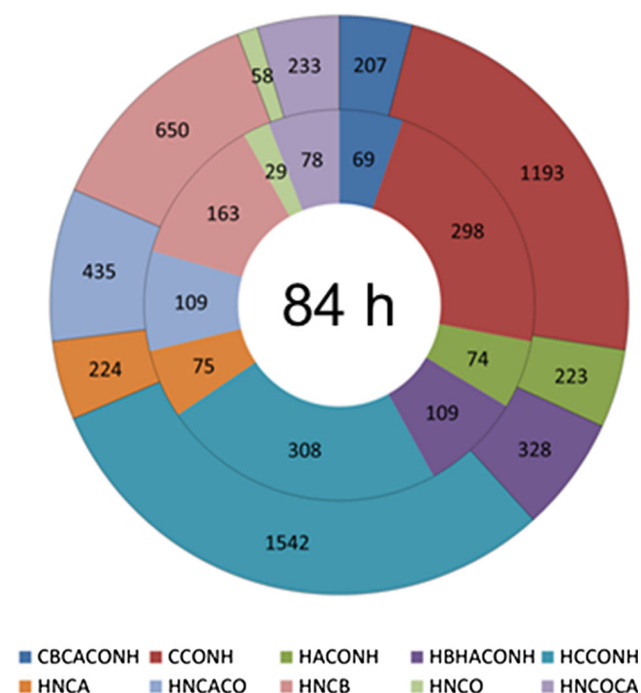
Fig. 7 Ten experiments used in ADAPT-NMR. The experiments are color-coded according to the *key at the bottom of the figure*, and the total elapsed time for data collection and chemical shift assignment was 84 h. The *inner layer of the graph* shows the elapsed time for data collection and chemical shift assignment of every tilted plane (in minutes), and the *outer layer* shows the total time for every experiment (in minutes)

[U-^{13}C, U-^{15}N]- chlorella-ubiquitin (76 amino acid residues)

A fully automated workflow (Fig. 10) was used for this protein, which was prepared by cell-free protein production.

Steps 1 and 2 (NOESY data collection): ^{15}N- and ^{13}C-edited NOESY spectra were recorded on a Varian 800 MHz spectrometer equipped with cryogenic probe and processed using the ist@HMS package. The ^{13}C-NOESY data were collected at a sampling level of 64 % (42 h), and the ^{15}N-NOESY data were collected at a sampling level of 36 % (24 h).

Step 3 (through-bond data collection and assignment): Non-uniform sampling with ADAPT-NMR was used for data collection (Fig. 11) and assignments of the backbone and side chain atoms.

Step 4 and 5 (Structure calculation with PONDEROSA-C/S): The Ponderosa Client was used for peak picking of the NOESY spectra, and for submitting the job to the Ponderosa Server with the PONDEROSA refinement option.

Step 6 (Structure evaluation with Ponderosa Analyzer): The structure validation reports generated by

PONDEROSA-C/S and PSVS are shown in Table S3. On the basis of the validation statistics, the structure was considered acceptable, and the process was stopped. Because the coordinates of the manually determined structure were not reported ((Ikeya et al. 2009) and BMRB entry 16228), we show only the structure calculated by using the NMRFAM-SDF (Fig. 12).

The two examples shown above used ADAPT-NMR for non-uniform data collection and assignments. In the following two examples, we consider a process in which through-bond experiments are collected manually, peak picking is performed with NMRFAM-SPARKY, and the PINE package is used for chemical shift assignments. The NMRFAM-SDF for this protocol (Fig. 13) was used to calculate the 3D structures of human ubiquitin and IscU (D39A).

[U-^{13}C, U-^{15}N]-human ubiquitin (76 amino acid residues)

Steps 1 and 2 (NOESY data collection): ^{15}N- and ^{13}C-edited NOESY spectra were collected with regularly-sampled time schedules.

Step 3 (through-bond data collection): Data from through-bond experiments were collected with regularly-sampled time schedules for eight experiments (2D ^{1}H-^{15}N-HSQC, 2D ^{1}H-^{13}C-HSQC, 3D CBCA(CO)NH, 3D C(CO)NH, 3D HBHA(CO)NH, 3D HCCH-TOCSY, 3D H(CCO)NH, and 3D HNCACB). NMRFAM-SPARKY was used to prepare peak lists from these experiments.

Step 4 (chemical shift assignment): These peak lists were used for chemical shift assignment with the PINE package.

Step 5 (validation with NMRFAM-SPARKY): The first step of validation was to use PINE-Sparky to evaluate the assignments. For this protein, the chemical shift assignments of 55 atoms out of 760 (7 %) were manually modified during this validation process.
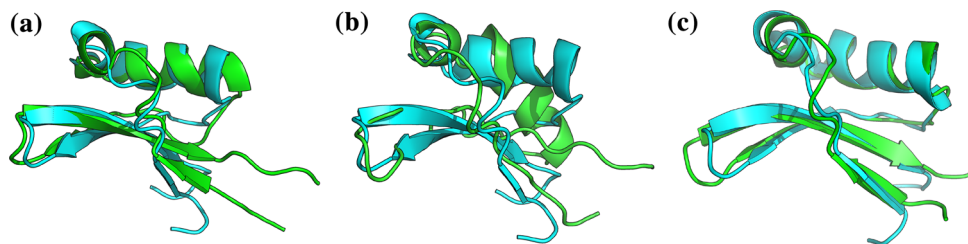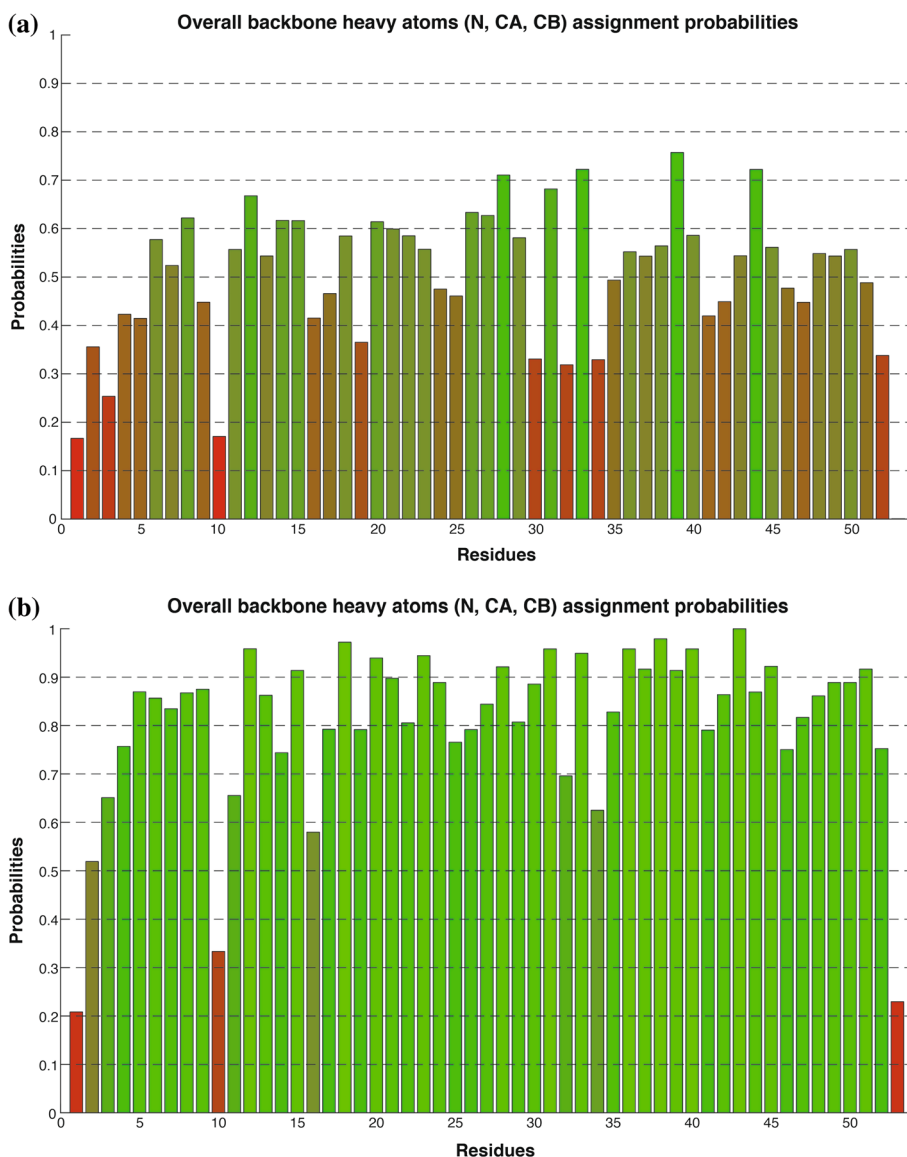
Step 6 (Validation with ARECA): The ARECA package was used to validate the assignments against NOESY spectra. ARECA reported 21 atoms (2.7 %) with low probabilities, which is considered within the acceptable range (fewer than 5 % of the total number of assigned atoms); therefore, no further data collection was needed.

Step 7 (Structure calculation with Ponderosa): Ponderosa Client was used for peak picking of the NOESY spectra and for submitting the job to PONDEROSA-C/S with the "PONDEROSA refinement option".

Step 8 (Structure evaluation with Ponderosa Analyzer): Table S4 shows validation reports for the structure generated by PODEROSA-C/S, which were considered satisfactory. For further evaluation of the structure, we compared the structure determined with this workflow

(a) Overall backbone heavy atoms (N, CA, CB) assignment probabilities

(b) Overall backbone heavy atoms (N, CA, CB) assignment probabilities

Fig. 9 Structures of brazzein protein with achieved automatically with NMRFAM-SDF (*green*) superimposed on the manually refined structure (*cyan*). **a** Structure generated with the non-uniformly sampled ^{15}N-edited NOESY spectrum and the regularly-sampled ^{13}C-edited NOESY spectrum. **b** Structure generated with the non-uniformly sampled ^{15}N- and ^{13}C-edited NOESY spectra. **c** Structure generated with the regularly-sampled ^{15}N- and ^{13}C-edited NOESY spectra and manual chemical shift assignments

against the manually-refined structure (PDB entry 1D3Z). The backbone RMSD between the two structures was 0.99 Å (Fig. 14), which indicates close match between the determined structures and shows accuracy of the framework.

[U-^{13}C, U-^{15}N]-IscU (D39A) (128 amino acid residues)

The structured variant (D38A) of the protein IscU from *Escherichia coli* (IscU (D39A)) was considered as another

Fig. 10 NMRFAM-SDF workflow used in the fully automated structure determination of ubiquitin
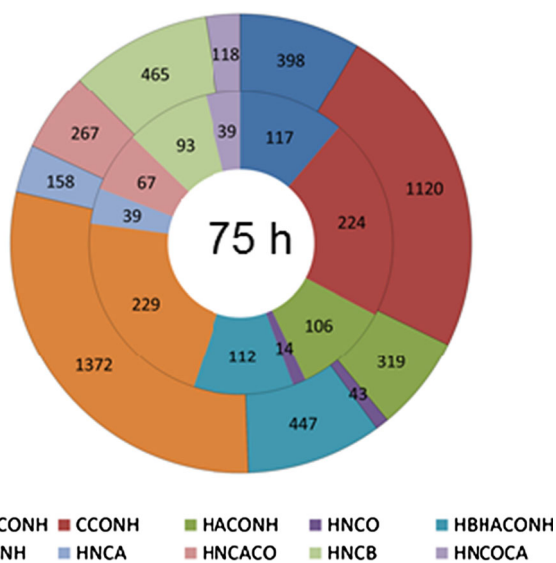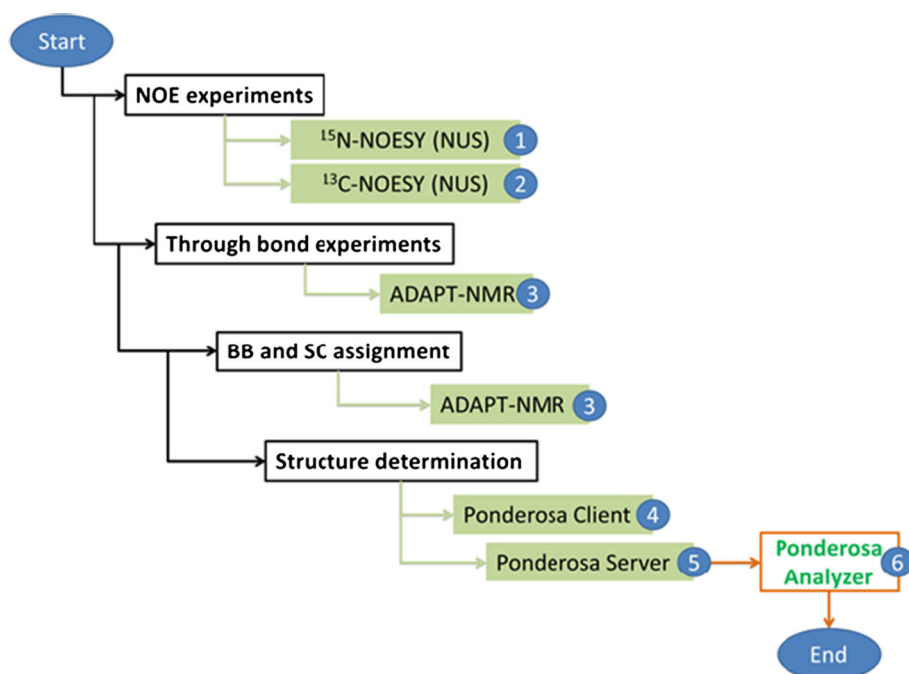
Fig. 11 The total time for data collection and chemical shift assignments by ADAPT-NMR was 75 h. For every experiment (*color-coded* according to the *key at the bottom of the figure*), the *inner layer* shows the elapsed time for data collection and chemical shift assignment of a tilted plane (in minutes) and the *outer layer* shows the total elapsed time for the experiment (in minutes)


Fig. 12 Structure of chlorella-ubiquitin obtained by using the NMRFAM-SDF

example for this alternative workflow (Fig. 13). Because of dynamics of the protein in solution (Kim et al. 2012), residual dipolar coupling (RDC) data were used as "Additional Restraints" in the framework. The Ponderosa Client was used for peak picking the NOESY spectra and

submitting a job to the Ponderosa Server. Table S5 shows the PONDEROSA-C/S and the PSVS outputs for the structure generated by the workflow. In addition to the acceptable structure validation statistics, comparison between the ordered regions (residues 19-60, 68-125) of the manually derived structure (Kim et al. 2012) (PDB entry 2KQK, BMRB entry 16603) and the structure calculated by NMRFAM-SDF resulted in a backbone RMSD of 0.99 Å (Fig. 15).
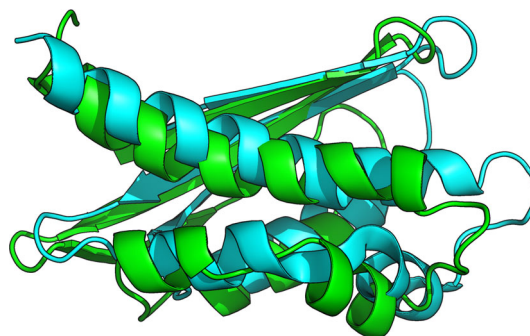
Fig. 13 NMRFAM-SDF workflow for data collected conventionally
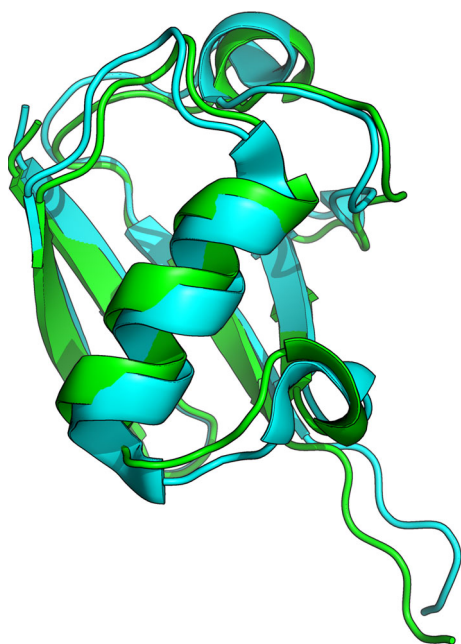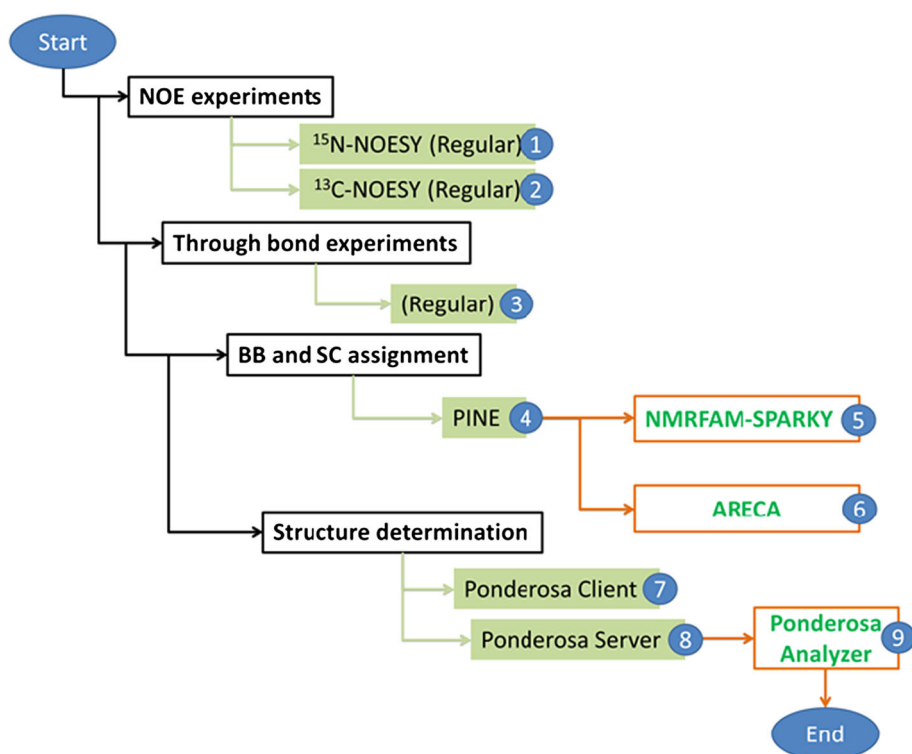
Fig. 14 Superimposition of the manual structure (*cyan*) and automated structure (*green*) of human ubiquitin

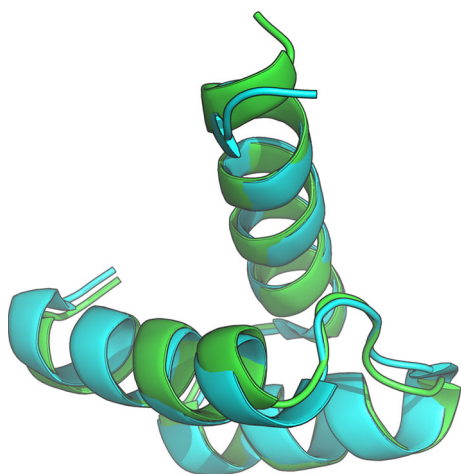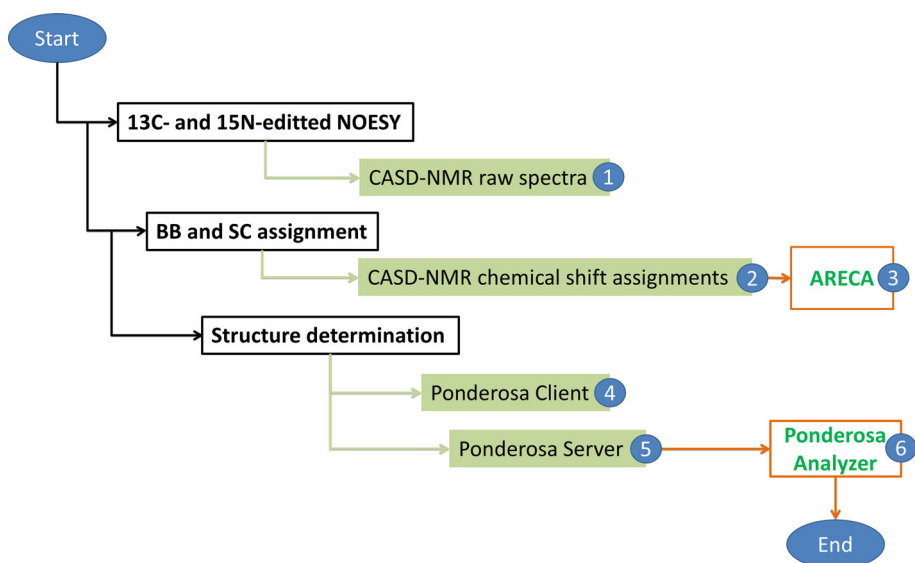**Fig. 15** Superimposition of the manual structure (*cyan*) and automated structure (*green*) of the protein IscU (D39A)

framework were the raw ^{13}C- and ^{15}N-filtered NOESY spectra and the chemical shift assignments provided for the competition. The NMRFAM-SDF workflow for this example is shown in Fig. 16.

Steps 1 and 2 (Peak lists and assignments): The raw ^{13}C- and ^{15}N-filtered NOESY spectra and the chemical shift assignments of protein HR6470A were used as the inputs to the framework. Ponderosa Client was used to peak-pick the spectra.

Step 3 (Validation with ARECA): The ARECA package was used to validate the assignments against the NOESY peak lists. ARECA reported only 6 assignments (0.70 %) with low probability, which is considered within the

[U-^{13}C, U-^{15}N]-HR6470A (69 amino acid residues)

In this final example, which involves the second round CASD-NMR target protein HR6470A, the input data to the

Fig. 16 NMRFAM-SDF
workflow for CASD-NMR
target protein HR6470A. In this
workflow, ARECA was used to
validate the chemical shift
assignments, Ponderosa Client
was used for peak picking of the
NOESY spectra and also
submitting the input files to the
Ponderosa Server for structure
calculations, and Ponderosa
Analyzer was used to validate
the structure

Fig. 17 Superposition of the structure of protein HR6470A calculated by NMRFAM-SDF (*green*) with the manually determined structure (*cyan*) deposited as PDB entry 2L9R

acceptable range (fewer than 5 % of the total number of assigned atoms); therefore, the quality of the chemical shifts assignments was considered to be satisfactory.

Step 4 (Structure calculation with Ponderosa): Ponderosa Client was used to prepare input submitted to PONDEROSA-C/S with the "PONDEROSA refinement option".

Step 5 (Structure evaluation with Ponderosa Analyzer): The statistics for structure validation generated with the Ponderosa Analyzer indicated satisfactory results (Table S6); thus the structure was deemed to be acceptable. Comparison of this structure with the manually determined structure (PDB entry 2L9R) resulted in a backbone RMSD of 0.51 Å (Fig. 17).

Conclusions

The process of protein structure determination by NMR spectroscopy consists of several computationally demanding steps. In order to develop high-throughput methods and to simplify the process into a robust approach for use by non-experts, algorithms for automation of discrete steps have been introduced. To accomplish this goal, the need for a user-friendly approach that includes several practical validation steps is inevitable. We have introduced a framework for the process of protein structure determination (NMRFAM-SDF) that is designed to achieve four goals: (a) to accelerate the structure determination process by removing human intervention, (b) to provide a workflow for fully automated structure determination for well-behaved proteins, (c) to provide unbiased validation tools for every step of the process, (d) to provide user-friendly refinement tools to prevent error propagation in the process. We have shown here that these steps can be assembled into various workflows and used to solve structures of relatively small test proteins labeled uniformly with ^{13}C and ^{15}N. The applicability of this approach to the broader landscape of structure determination remains to be tested thoroughly, although we and others have shown success in using components of the framework, such as PINE and PONDEROSA-C/S, with much larger proteins. Semi-automated inspection and validation tools will be particularly useful for more complex proteins. Additional validation tools are planned, and NMRFAM-SDF will provide a solid foundation for these extensions.

References

Alipanahi B, Gao X, Karakoc E, Donaldson L, Ming L (2009) PICKY: a novel SVD-based NMR spectra peak picking method. Bioinformatics 25:i268–i275. doi:10.1093/bioinformatics/btp225

Aoto PC, Fenwick RB, Kroon GJA, Wright PE (2014) Accurate scoring of non-uniform sampling schemes for quantitative NMR. J Magn Reson 246:31–35. doi:10.1016/j.jmr.2014.06.020

Bahrami A, Assadi AH, Markley JL, Eghbalnia HR (2009) Probabilistic interaction network of evidence algorithm and its application to complete labeling of peak lists from protein NMR spectroscopy. PLoS Comput Biol 5 doi:10.1371/journal.pcbi.1000307

Bahrami A, Tonelli M, Sahu SC, Singarapu KK, Eghbalnia HR, Markley JL (2012) Robust, integrated computational control of NMR experiments to achieve optimal assignment by ADAPT-NMR. PLoS Comput Biol 7 doi:10.1371/journal.pone.0033173

Bhattacharya A, Tejero R, Montelione GT (2006) Evaluating protein structures determined by structural genomics consortia. Proteins 66:778–795. doi:10.1002/prot.21165

Bovet DPB, Plerlulgi CD (1994) Introduction of the theory of complexity. prentice hall international series in computer science

Brutscher B (2013) SOFAST HMQC. Encycl Biophys, pp 2407–2407. doi:10.1007/978-3-642-16712-6_347

Chen VB et al (2010) MolProbity: all-atom structure validation for macromolecular crystallography. Acta Crystallogr D Biol Crystallogr 66:12–21. doi:10.1107/s0907444909042073

Cheng Y, Gao X, Liang F (2013) Bayesian peak picking for NMR spectra. Genomics Proteomics Bioinform 12:39–47. doi:10.1016/j.gpb.2013.07.003

Chylla RA, Volkman BF, Markley JL (1998) Practical model fitting approaches to the direct extraction of NMR parameters simultaneously from all dimensions of multidimensional NMR spectra. J Biomol NMR 12:277–297

Clore GM, Gronenborn AM (1987) Determination of three-dimensional structures of proteins in solution by nuclear magnetic resonance spectroscopy. Protein Eng 1:275–288

Clore GM, Gronenborn AM (1991) Structures of larger proteins in solution: three- and four-dimensional heteronuclear NMR spectroscopy. Science 252:1390–1399

Cornilescu CC et al (2013) Temperature-dependent conformational change affecting Tyr11 and sweetness loops of brazzein. Proteins 81:919–925. doi:10.1002/prot.24259

Dashti H, Tonelli M, Lee W, Westler WM, Cornilescu G, Ulrich EL, Markley JL (2015) Validation of protein NMR chemical shift assignments against NOE data manuscript in preparation

Davis IW, Murray LW, Richardson JS, Richardson DC (2004) MOLPROBITY: structure validation and all-atom contact analysis for nucleic acids and their complexes. Nucleic Acids Res 32:W615–W619. doi:10.1093/nar/gkh398

DeLano W, Lam J (2005) PyMOL: A communications tool for computational models Abstr Pap Am Chem S 230:U1371–U1372

Eghbalnia HR, Bahrami A, Wang L, Assadi A, Markley JL (2005) Probabilistic Identification of Spin Systems and their assignments including coil-helix inference as output (PISTACHIO). J Biomol NMR 32:219–233. doi:10.1007/s10858-005-7944-6

Frishman D, Argos P (1995) Knowledge-based protein secondary structure assignment. Proteins 23:566–579. doi:10.1002/prot.340230412

Frydman L (2006) Single-scan multidimensional NMR. C R Chim 9:336–345. doi:10.1016/j.crci.2005.06.014

Frydman L, Lupulescu A, Scherf T (2004) Principles and features of single-scan two-dimensional NMR spectroscopy. J Am Chem Soc 125:9204–9217. doi:10.1021/ja030055b

Güntert P (2004) Automated NMR structure calculation with CYANA protein NMR techniques. Methods Mol Biol 278:353–378. doi:10.1385/1-59259-809-9:353

Herrmann T, Güntert P, Wüthrich K (2002) Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA. J Mol Biol 319:209–227. doi:10.1016/S0022-2836(02)00241-3

Hiller S, Fiorito F, Wüthrich K, Wider G (2005) Automated projection spectroscopy (APSY). PNAS 102 doi:10.1073/pnas.0504818102

Hoch JC, Maciejewski MW, Mobli M, Schuyler AD, Stern AS (2007) Nonuniform sampling in multidimensional NMR. In: eMagRes. Wiley. doi:10.1002/9780470034590.emrstm1239

Hoch JC, Maciejewski MW, Mobli M, Schuyler AD, Stern AS (2014) Nonuniform sampling and maximum entropy reconstruction in multidimensional NMR. Acc Chem Res 47:708–717. doi:10.1021/ar400244v

Huang YJ, Powers R, Montelione GT (2005) Protein NMR Recall, Precision, and F-measure Scores (RPF Scores): structure quality assessment measures based on information retrieval statistics

Hyberts SG, Arthanari H, Wagner G (2012) Applications of non-uniform sampling and processing. Top Curr Chem 316:125–148. doi:10.1007/128_2011_187

Ikeya T, Takeda M, Yoshida H, Terauchi T, Jee JG, Kainosho M, Guntert P (2009) Automated NMR structure determination of stereo-array isotope labeled ubiquitin from minimal sets of spectra using the SAIL-FLYA system. J Biomol NMR 44:261–272. doi:10.1007/s10858-009-9339-6

Jee J, Guntert P (2003) Influence of the completeness of chemical shift assignments on NMR structures obtained with automated NOE assignment. J Struct Funct Genomics 4:179–189

Jung Y-S, Zweckstetter M (2004) Mars: robust automatic backbone assignment of proteins. J Biomol NMR 30:11–23

Kim S, Szyperski T (2003) GFT NMR, a new approach to rapidly obtain precise high-dimensional NMR spectral information. J Am Chem Soc 125:1385–1393. doi:10.1021/ja028197d

Kim JH, Tonelli M, Kim T, Markley JL (2012) Three-Dimensional Structure and Determinants of Stability of the Iron-Sulfur Cluster Scaffold Protein IscU from Escherichia coli†. Biochemistry 51:5557–5563. doi:10.1021/bi300579p

Kupce E, Freeman R (2003a) Fast multi-dimensional Hadamard spectroscopy. J Magn Reson 163:56–63

Kupce E, Freeman R (2003b) Projection-reconstruction of three-dimensional NMR spectra. J Am Chem Soc 125:13958–13959. doi:10.1021/ja038297z

Kuszewski J, Schwieters CD, Garrett DS, Byrd RA, Tjandra N, Clore GM (2004) Completely automated, highly error-tolerant macromolecular structure determination from multidimensional nuclear overhauser enhancement spectra and chemical shift assignments. J Am Chem Soc 126:6258–6273. doi:10.1021/ja049786h

Kuszewski JJ, Thottungal RA, Clore GM, Schwieters CD (2008) Automated error-tolerant macromolecular structure determination from multidimensional nuclear Overhauser enhancement spectra and chemical shift assignments: improved robustness and performance of the PASD algorithm. J Biomol NMR 41:221–239. doi:10.1007/s10858-008-9255-1

Laskowski RA, MacArthur MW, Moss DS, Thornton JM (1993) PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 26:283–291. doi:10.1107/S0021889892009944

Laskowski RA, Rullmannn JA, MacArthur MW, Kaptein R, Thornton JM (1996) AQUA and PROCHECK-NMR: programs for checking the quality of protein structures solved by NMR. J Biomol NMR 8:477–486

Lee W, Westler WM, Bahrami A, Eghbalnia HR, Markley JL (2009) PINE-SPARKY: graphical interface for evaluating automated probabilistic peak assignments in protein NMR spectroscopy. Bioinformatics 25:2085–2087. doi:10.1093/bioinformatics/btp345

Lee W, Kim JH, Westler WM, Markley JL (2011) PONDEROSA, an automated 3D-NOESY peak picking program, enables automated protein structure determination. Bioinformatics 27:1727–1728. doi:10.1093/bioinformatics/btr200

Lee W, Bahrami A, Markley JL (2012) ADAPT-NMR Enhancer: complete package for reduced dimensionality in protein NMR spectroscopy. Bioinformatics 29:515–517. doi:10.1093/bioinformatics/bts692

Lee W, Hu K, Tonelli M, Bahrami A, Neuhardt E, Glass KC, Markley JL (2013) Fast automated protein NMR data collection and assignment by ADAPT-NMR on Bruker spectrometers. J Magn Reson 236:83–88. doi:10.1016/j.jmr.2013.08.010

Lee W, Stark JL, Markley JL (2014a) PONDEROSA-C/S: client-server based software package for automated protein 3D structure determination. J Biomol NMR 60:73–75. doi:10.1007/s10858-014-9855-x

Lee W, Tonelli M, Markley JL (2014b) NMRFAM-SPARKY: enhanced software for biomolecular NMR spectroscopy. Bioinformatics. doi:10.1093/bioinformatics/btu830

Lescop E, Kern T, Brutscher B (2009) Guidelines for the use of band-selective radiofrequency pulses in hetero-nuclear NMR: example of longitudinal-relaxation-enhanced BEST-type 1H-15 N correlation experiments. J Magn Reson 203:190–198. doi:10.1016/j.jmr.2009.12.001

Lescop E, Schanda P, Brutscher B (2007) A set of BEST triple-resonance experiments for time-optimized protein resonance assignment. J Magn Reson 187:163–169. doi:10.1016/j.jmr.2007.04.002

Linge JP, Habeck M, Rieping W, Nilges M (2003) ARIA: automated NOE assignment and NMR structure calculation. Bioinformatics 19:315–316

Lopez-Mendez B, Guntert P (2006) Automated protein structure determination from NMR spectra. J Am Chem Soc 128:13112–13122. doi:10.1021/ja061136l

Maciejewski M, Stern A, King G, Hoch J (2006) Nonuniform Sampling in Biomolecular NMR. In: Webb G (ed) Modern magnetic resonance. Springer, Netherlands, pp 1305–1311. doi:10.1007/1-4020-3910-7_142

MacRaild CA, Norton RS (2014) RASP: rapid and robust backbone chemical shift assignments from protein structure. J Biomol NMR 58:155–163. doi:10.1007/s10858-014-9813-7

Mobli M, Hoch JC (2008) Maximum entropy spectral reconstruction of non-uniformly sampled data concepts. Magn Reson Part A Bridg Educ Res 32A:436–448. doi:10.1002/cmr.a.20126

Moseley HN, Sahota G, Montelione GT (2004) Assignment validation software suite for the evaluation and presentation of protein resonance assignment data. J Biomol NMR 28:341–355. doi:10.1023/b:jnmr.0000015420.44364.06

Orekhov VY, Jaravine VA (2011) Analysis of non-uniformly sampled spectra with multi-dimensional decomposition. Prog Nucl Magn Reson Spectrosc 59:271–292

Orekhov VY, Ibraghimov I, Billeter M (2003) Optimizing resolution in multidimensional NMR by three-way decomposition. J Biomol NMR 27:165–173

Rieping W, Department of Biochemistry UoCCCBGAUK, Vranken WF, Protein Data Bank in Europe EBIWTGCHCCBSDUK, Protein Data Bank in Europe EBIWTGCHCCBSDUK (2014) Validation of archived chemical shifts through atomic coordinates Proteins 78:2482–2489 doi:10.1002/prot.22756

Rosato A et al (2009) CASD-NMR: critical assessment of automated structure determination by NMR. Nat Methods 6:625–626. doi:10.1038/nmeth0909-625

Rosato A et al (2012) Blind testing of routine, fully automated determination of protein structures from NMR data. Structure 20:227–236. doi:10.1016/j.str.2012.01.002

Schanda P, Brutscher B (2005) Very fast two-dimensional NMR spectroscopy for real-time investigation of dynamic events in proteins on the time scale of seconds. J Am Chem Soc 127:8014–8015. doi:10.1021/ja051306e

Schmidt E, Guntert P (2012) A new algorithm for reliable and general NMR resonance assignment. J Am Chem Soc 134:12817–12829. doi:10.1021/ja305091n

Schwieters CD, Kuszewski JJ, Tjandra N, Clore GM (2003) The Xplor-NIH NMR molecular structure determination package. J Magn Reson 160:65–73

Serrano P, Pedrini B, Mohanty B, Geralt M, Herrmann T, Wuthrich K (2012) The J-UNIO protocol for automated protein structure determination by NMR in solution. J Biomol NMR 53:341–354. doi:10.1007/s10858-012-9645-2

Shen Y, Bax A (2007) Protein backbone chemical shifts predicted from searching a database for torsion angle and sequence homology. J Biomol NMR 38:289–302. doi:10.1007/s10858-007-9166-6

Shen Y, Delaglio F, Cornilescu G, Bax A (2009) TALOS + : a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts. J Biomol NMR 44:213–223. doi:10.1007/s10858-009-9333-z

Shin J, Lee W (2008) Structural proteomics by NMR spectroscopy. Exp Rev Proteom 5:589–601. doi:10.1586/14789450.5.4.589

Sun S, Gill M, Li Y, Huang M, Byrd RA (2015) Efficient and generalized processing of multidimensional NUS NMR data: the NESTA algorithm and comparison of regularization terms submitted

Szyperski T, Yeh DC, Sukumaran DK, Moseley HN, Montelione GT (2002) Reduced-dimensionality NMR spectroscopy for high-throughput protein resonance assignment. Proc Natl Acad Sci USA 99:8009–8014. doi:10.1073/pnas.122224599

Tikole S, Jaravine V, Rogov V, Dötsch V, Güntert P (2014) Peak picking NMR spectral data using non-negative matrix factorization. BMC Bioinformatics 15:46

Vranken WF, Rieping W (2009) Relationship between chemical shift value and accessible surface area for all amino acid atoms. BMC Struct Biol 9:20

Wang L, Eghbalnia HR, Bahrami A, Markley JL (2005) Linear analysis of carbon-13 chemical shift differences and its application to the detection and correction of errors in referencing and spin system identifications. J Biomol NMR 32:13–22. doi:10.1007/s10858-005-1717-0

Wang B, Wang Y, Wishart DS (2010) A probabilistic approach for validating protein NMR chemical shift assignments. J Biomol NMR 47:85–99. doi:10.1007/s10858-010-9407-y

Wu KP et al (2006) RIBRA–an error-tolerant algorithm for the NMR backbone assignment problem. J Comput Biol 13:229–244. doi:10.1089/cmb.2006.13.229

Wüthrich K (1986) NMR of proteins and nucleic acids. Wiley-Interscience

Xu Y, Wang X, Yang J, Vaynberg J, Qin J (2006) PASA–a program for automated protein NMR backbone signal assignment by pattern-filtering approach. J Biomol NMR 34:41–56. doi:10.1007/s10858-005-5358-0

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