Journal of Magnetic Resonance 254 (2015) 93–97

Contents lists available at ScienceDirect

Journal of Magnetic Resonance

journal homepage: www.elsevier.com/locate/jmr

Reliable resonance assignments of selected residues of proteins with known structure based on empirical NMR chemical shift prediction

Da-Wei Li [a,b,c], Dan Meng [c], Rafael Brüschweiler [a,b,c,*]

[a] Campus Chemical Instrumentation Center, The Ohio State University, Columbus, OH 43210, United States
[b] Department of Chemistry & Biochemistry, The Ohio State University, Columbus, OH 43210, United States
[c] Chemical Sciences Laboratory, Department of Chemistry and Biochemistry and National High Magnetic Field Laboratory, Florida State University, Tallahassee, FL 32306, United States

ARTICLE INFO

Article history:
Received 13 December 2014
Revised 12 February 2015
Available online 7 March 2015

Keywords:
Protein structure based NMR resonance assignments
Chemical shift prediction

ABSTRACT

A robust NMR resonance assignment method is introduced for proteins whose 3D structure has previously been determined by X-ray crystallography. The goal of the method is to obtain a subset of correct assignments from a parsimonious set of 3D NMR experiments of ^{15}N, ^{13}C labeled proteins. Chemical shifts of sequential residue pairs are predicted from static protein structures using PPM_One, which are then compared with the corresponding experimental shifts. Globally optimized weighted matching identifies the assignments that are robust with respect to small changes in NMR cross-peak positions. The method, termed PASSPORT, is demonstrated for 4 proteins with 100–250 amino acids using 3D NHCA and a 3D CBCA(CO)NH experiments as input producing correct assignments with high reliability for 22% of the residues. The method, which works best for Gly, Ala, Ser, and Thr residues, provides assignments that serve as anchor points for additional assignments by both manual and semi-automated methods or they can be directly used for further studies, e.g. on ligand binding, protein dynamics, or post-translational modification, such as phosphorylation.

1. Introduction

Protein resonance assignment is a prerequisite for the biophysical investigation of proteins by NMR spectroscopy, such as studies of protein–protein and ligand-binding interactions, dynamics, or post-translational chemical modifications. Despite of many years of progress in the development of resonance assignment protocols, it often is still a time-consuming step, which typically requires the manual analysis of multiple 2D and 3D NMR experiments. For many proteins of interest their 3D structures have already been determined by X-ray crystallography and deposited in the Protein Databank (PDB) [1]. For these protein systems, studies of their biophysical and biological properties by NMR can provide useful information about their function. In many such cases, the reliable assignment of a subset of suitably positioned residues can prove valuable. It is then desirable to identify such assignments by keeping the number of multidimensional NMR assignment experiments at a minimum allowing one to shift available resources toward applications.

The most common NMR assignment strategies of proteins are based on the sequential assignment method [2,3]. This strategy uses the primary sequence information of a protein to derive assignments without requiring any 3D structural information. Automated versions of this assignment strategy have been implemented in a number of software packages [4–9]. This strategy is the method of choice when the objective of the study is the determination of the 3D protein structure itself. During the structure determination process it is common to assign more and more proton resonances whose distance constraints, in turn, help further improve the 3D structure [8,10] or, if available, use 3D structural information from homologous proteins [11].

For proteins whose 3D structure is already known from X-ray crystallography, the use of high-resolution 3D structural information can considerably facilitate the resonance assignment process. A general strategy proposed for this task used residual dipolar couplings (RDCs) in a single alignment medium along with database-derived average chemical shifts of the Cα and Cβ nuclei of the different amino acid types [12]. The combinatorial assignment problem, whose complexity grows with $N!$ where N is the number of residues, was represented in terms of a complete bipartite graph and the optimal global solution was obtained by using the Hungarian optimization method, which is efficient also for larger proteins. Other structure-based assignment strategies were developed specifically for paramagnetic proteins [13] or by primarily using ^{1}H–^{1}H NOESY-derived distance information [14–19]. Because protons of sequential residues are in close spatial vicinity,

* Corresponding author at: Department of Chemistry & Biochemistry, The Ohio State University, Columbus, OH 43210, United States.
E-mail address: bruschweiler.1@osu.edu (R. Brüschweiler).

http://dx.doi.org/10.1016/j.jmr.2015.02.013
1090-7807/

NOESY experiments provide local connectivity information, in addition to long-range distance constraints. Recent work demonstrated for small to medium–large proteins (<160 amino acids), the power of a single 3D ^{13}C- or ^{15}N-edited NOESY experiment for simultaneous resonance assignment and structure determination [8,20].

Here, we present a method that produces reliable assignments for a subset of residues if the protein structure is known. The method finds the optimal global assignment based on predicted and experimental chemical shifts and it identifies residues whose assignments are most robust. The continuous growth of both the protein data bank (PDB) [1] and the BMRB chemical shift database [21] has led to the recent development of methods for the increasingly accurate empirical prediction of chemical shifts of backbone nuclei of proteins with known 3D structures [22–28]. This assumes that the available X-ray crystallographic structure is representative for the NMR conditions, which can be in solution or in the solid state. The method presented here uses two complementary heteronuclear 3D experiments of a uniformly ^{15}N, ^{13}C-labeled protein that connect backbone resonances of sequential residues as input and it produces assignments based on global matching between the chemical shifts determined experimentally and ones predicted from the 3D structure. We demonstrate the method, which is termed PASSPORT for "Protein assignment strategy using chemical shift predictions for proteins with known structure", using just two triple-resonance 3D experiments as input.

2. Methods and materials

Because the currently available chemical shift predictors have on average an accuracy that is lower than the chemical shift differences between NMR cross-peaks that are next to each other in crowded regions of a protein NMR spectrum, such as a 2D ^{15}N–^{1}H HSQC spectrum, the reliable assignment based on a single spectrum is not feasible without the use of additional information. Here, we use local sequential connectivity information obtained from 3D CBCA(CO)NH [29] and 3D HNCA experiments [29–31] to generate clusters of resonances that belong to stretches of neighboring amino acids of variable lengths. These chemical shift clusters have a higher specificity than individual chemical shifts and thereby allow the better discrimination between correct and incorrect assignments.

In the 3D CBCA(CO)NH experiment a strip along the carbon dimension corresponding to the cross-section at the $^{1}H^{N}$ and ^{15}N chemical shifts of a given residue $i+1$ contains two carbon cross-peaks, which belong to the ^{13}Cα and ^{13}Cβ nuclei of the preceding residue i. For the 3D HNCA experiment the analogous NH($i+1$) strip contains two peaks along the carbon dimension with one corresponding to the ^{13}Cα($i+1$) resonance of residue $i+1$ and the other belonging to the ^{13}Cα(i) resonance of the preceding residue. In principle, superposition of the two spectra permits identification of the Cα chemical shift of a NH group and the Cα, Cβ chemical shifts of the preceding residue. The resulting five chemical shifts N($i+1$), HN($i+1$), Cα($i+1$), Cα(i), Cβ(i) form a group, which is called "sequential resonance cluster". From a known 3D structure, e.g. taken from the PDB, the chemical shifts of all sequential resonance clusters can be predicted and the root-mean-square difference (RMSD) of the chemical shifts (in units of ppm) between each predicted and each experimental sequential resonance cluster can be determined as follows:

$$C_{ij}=\sqrt{\frac{w_C\left(\delta^{\mathrm{pred}}_{C\alpha,i}-\delta^{\mathrm{exp}}_{C\alpha,j}\right)^2+w_C\left(\delta^{\mathrm{pred}}_{C\beta,i}-\delta^{\mathrm{exp}}_{C\beta,j}\right)^2+w_C\left(\delta^{\mathrm{pred}}_{C\alpha,i+1}-\delta^{\mathrm{exp}}_{C\alpha,j+1}\right)^2+w_N\left(\delta^{\mathrm{pred}}_{N,i+1}-\delta^{\mathrm{exp}}_{N,j+1}\right)^2+w_H\left(\delta^{\mathrm{pred}}_{H,i+1}-\delta^{\mathrm{exp}}_{H,j+1}\right)^2}{5}} \quad (1)$$

where δ^{pred} and δ^{exp} are the predicted and experimental chemical shifts, respectively. C_{ij} is a measure of the match between an experimental resonance cluster i and a putative cluster j. The lower C_{ij} the better, and thus more likely, is the match. Different types of nuclei are given different relative weights w_i in order to account for differential intrinsic chemical shift variations, whereby w_c, w_N and w_H were set to 1.0, 0.16, and 4.0 respectively. The assignment results are insensitive to small changes of the weights. They were initially defined by the average prediction error for each type of nucleus and they were subsequently adjusted to optimize the assignment results. For most resonance clusters, all five terms are included. In those cases where some of the terms are not present, C_{ij} is rescaled correspondingly (by dividing in Eq. (1) by the actual number of terms). For example, because there is no $\delta_{C\beta}$ term for GLY residues, C_{ij} is rescaled by a factor 1.25. In case there is an obvious inconsistency between experimental and predicted chemical shifts, such as when GLY is compared with non-GLY residues, C_{ij} will be set to a very large value reflecting a very poor match and making this match prohibitively expensive, i.e. effectively forbidden, during weighted matching.

For well-resolved resonances, the HNCA experiment can provide sequential connectivity information for stretches of more than 2 residues. Whenever possible, we use sequential residue pairs that contain up to 2 additional residues in each direction and hence to the RMSD goes over resonance clusters that contain resonances from up to 5 residues. This further increases the specificity and accuracy of assignments. Since Cα resonances tend to be crowded, there are often multiple possibilities to select the preceding and the following amino acid from the HNCA spectrum. In such cases, the combination is selected that gives the minimal RMSD with respect to a predicted resonance cluster. This implies that the combination selected for a given experimental cluster depends on the predicted cluster it is compared with.

Once the cost matrix **C** has been determined according to Eq. (1), the selection of the best assignment is based on weighted matching, which is a global optimization algorithm [32]. It identifies a unique one-to-one map between computed and experimental resonance clusters that minimizes the sum of all resonance cluster RMSDs. The global minimum is found efficiently by means of the Hungarian algorithm [33]. All chemical shifts were predicted using the PPM_One, Sparta+ [23], and Shiftx2 [27] chemical shift prediction programs. For protein parts that are missing in the X-ray crystal structure, we use the corresponding random coil chemical shifts, because these protein parts often adopt random coil conformations in solution. The random coil chemical shifts used are the chemical shift averages taken from the BMRB database [21].

Results of our method show that it produces for proteins with 100–200 residues 50% correct assignments. Clearly, the proposed assignments are only of limited use, unless false positive assignments can be identified with high confidence. To identify and eliminate false positive assignments, we add random Gaussian distributed noise with a standard deviation of 0.6 ppm to all elements of the cost matrix of Eq. (1) before solving the weighted matching problem. This procedure is repeated 30 times and the resulting assignments are compared with each other. A confidence level is determined for each assignment according to the number of identical assignments obtained during the 30 iterations. For example, if a resonance cluster is assigned to a particular protein position 24 times out of the 30 times of tries, the confidence level of this assignment is 80%.