Vol.2018-BIO-56 No.6
2018/12/14

Validation of NMR protein structures using rigidity theory and chemical shifts

Kazuhito Nishiyama[†1,a)] Toshiki Saitoh[†1] Adnan Sljoka[†2]

Abstract: Most protein structures deposited in the Protein Data Bank (PDB) are solved with X-ray crystallography or Nuclear Magnetic Resonance (NMR) experiments. Unlike crystal structures, there is no reliable way to validate the accuracy of NMR structures, which is a big issue for users of PDB. To develop a highly dependable method for NMR validation, we compare two independent representations of flexibility for protein structures: Random Coil Index (RCI) which utilizes experimental NMR chemical shifts and method FIRST which is based on concepts in biophysics and mathematical rigidity theory. In initial results on a set of NMR structures, we show that the correlations between the two flexibility representations RCI and FIRST can accurately validate the quality of structures.

1. Introduction

The 3D structures of proteins are can be determined with experimental techniques. Solved structures are deposited into Protein Data BankPDB [2]. Most protein structures deposited in the PDB are solved with either X-ray crystallography or Nuclear Magnetic Resonance (NMR) experiments: X-ray crystals and NMR structures account for about 80% and 15% of PDB structures, respectively. The advantage of X-ray crystallography is that it can be used to measure structural information for proteins with large molecular weight, however it is difficult to elucidate realistic solution-behaved structures since the protein has to be crystallized and dynamical information is lost. On the other hand, NMR structure determination process is solution-based at more realistic temperatures, so it opens up possibilities to obtain dynamical structural information at physiological conditions. One difficult with NMR, is that structures cannot be large. 3D structures analyzed by X-ray crystallography can be validated using several measures (resolution, R-factor etc) [6], but there is no reliable validation of solved NMR structures. The goal of this work is to develop a method to validate the 3D structures obtained by NMR.

It is well known that the flexibility and dynamics are important measures in analysis of 3D structures and functions of proteins [3], [7]. We consider the validation of NMR structures by comparing flexibility data obtained from two different perspectives. One is a method which predicts the flexibility of proteins using Random Coil Index (RCI). RCI uses NMR chemical shifts which are readily available from Biological Magnetic Resonance Bank (BMRB) [8]. The other is a computational method for predicting rigidity of protein structures using 3D structural data, Floppy Inclusions and Rigid Substructure Topography(FIRST). In this paper, we calculated correlation coefficients of flexibility data from RCI and predictions from FIRST to validate the 3D structures of proteins.

[†1] Presently with Kyushu Institute of Technology
[†2] Presently with Kwansei Gakuin University
[a)] nishiyama.kazuhito361@mail.kyutech.jp

2. Methods

2.1 RCI from chemical shifts

The RCI is a method for predicting the flexibility of proteins by calculating an inverse weighted average of secondary chemical shifts and RMSDs for each residue from Molecular Dynamics simulation and NMR structural ensembles [1]. RCI is a structure-independent measure of backbone flexibility. In Fig. 1 we have shown an example of the output of RCI. The flexibility for each residue is quantified in the range of 0 to 1, where higher values indicate increased flexibility.

2.2 FIRST from three-dimensional structures

FIRST program gives fast computational prediction of flexible and rigid regions in a protein. FIRST starts with a 3D model of a protein and generates a molecular constraint multigraph which consists of hydrogen bonds, covalent bonds, salt bridges and hydrophobic interactions. Using a pebble game algorithm, which is based on combinatorial rigidity theory rigid clusters and flexible connections are obtained [4]. The strength of hydrogen bonds is calculated using an energy function which takes into an effect donor and acceptor geometry. FIRST computes rigid cluster at small energy increments (where weak hydrogen bonds are broken one by one). Fig. 2 shows a hydrogen bond dilution plot as an output of FIRST.

3. Results and Discussion

3.1 Correlation of flexibility by FIRST and RCI

We calculated correlation coefficients between FIRST and RCI, which are two independent measures of flexibility, in order to probe quality of the NMR ensemble. For each residue, we

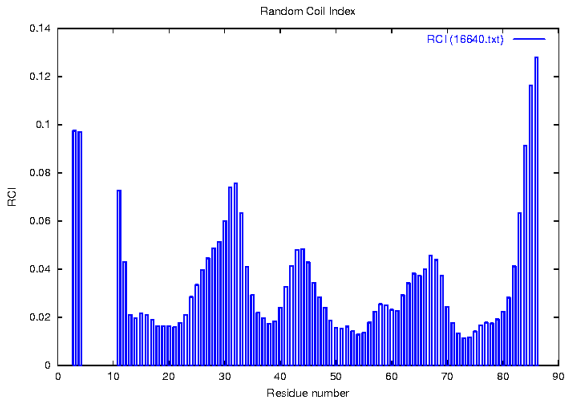

Fig. 1 An example of RCI for BMRB:16640

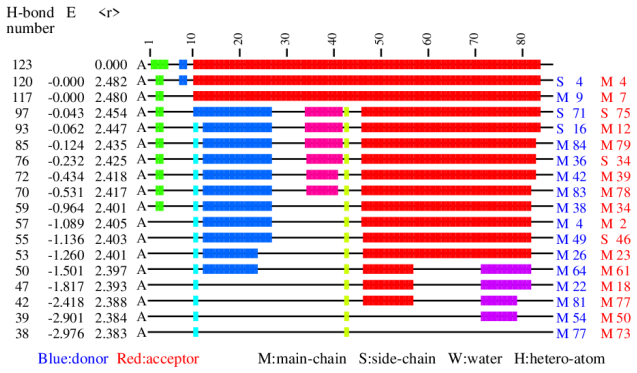

Fig. 2 An example of FIRST for PDBID:2krk model 1

compare the two values by RCI and FIRST: RCI gives a flexibility score of each residue and FIRST gives the energy score for each residue where the corresponding residue is not part of any large rigid cluster. We use the Spearmans rank correlation coefficient to calculate the correlation. Part of some residues in some proteins are missing from the BMRB and PDB. We omit such missing residues to calculate the coefficients.

We show the results with relatively high and low correlation coefficients in Table 1. We computed the coefficients for NMR structures given in [5]. There are typically 20 structural models in each NMR structure (ensemble). In Table 1, each column corresponds to a NMR structure and each entry is a correlation score between FIRST and RCI for particular model.

Table 1 Correlation coefficient

model	2krk	2jr2	2kpu	2kyi
1	0.802595	0.696875	0.411024	0.049442
2	0.757035	0.696271	-0.043820	0.210382
3	0.759777	0.669086	0.005229	0.302629
4	0.769155	0.714510	0.164312	0.098390
5	0.842937	0.726098	0.267498	0.352601
6	0.684400	0.649254	-0.122532	-0.324674
7	0.818045	0.754761	-0.200878	0.378246
8	0.852273	0.629765	0.252282	0.169703
9	0.821702	0.415430	0.161280	0.113094
10	0.829958	0.583535	0.130570	0.017155
11	0.757150	0.768129	0.064276	0.205364
12	0.763020	0.550566	0.360303	0.115026
13	0.781475	0.774403	0.117194	0.093904
14	0.808961	0.735259	0.152118	0.014693
15	0.814842	0.727847	0.138309	0.466381
16	0.858104	0.704040	0.190237	0.222528
17	0.827874	0.655879	0.315841	0.274563
18	0.755732	0.540068	0.114758	0.236790
19	0.799434	0.729901	0.202631	0.198602
20	0.741885	0.755533	0.295508	0.321201

3.2 Discussion

When correlation coefficients are high, we see that flexibility predicted from the 3D structures is consistent. In Table 1, correlations in 2krk and 2jr2 are about 0.8, 0.7 in most models, respectively. On the other hand, the correlations are low for 2kpu and 2kyi. For most proteins in our experiments, the correlation coefficients of all models are roughly the same since the values are in $\pm 0.1 \sim 0.15$ range from the average value of all models. For 2kyi, almost all models have correlation coefficients of approximately 0 to 0.4, but for model 6 it is -0.3.

Thus, for some PDBs, we found a correlation coefficient of a model that deviates from other models, suggesting there are likely errors in NMR structural ensemble members.

4. Conclusion and Future works

In this paper, we calculated correlation coefficients between flexibility measure of protein backbone using experimental chemical shifts as implemented in RCI and computational method FIRST using 3D structural data. If the correlation coefficient of a model is high, the 3D structure is consistent. Our proposed method on the reported test cases is a strong tool for validating NMR structure. Future work will further solidify this research.

References

[1] Mark V Berjanskii and David S Wishart. A simple method to predict protein flexibility using secondary chemical shifts. *Journal of the American Chemical Society*, 127(43):14970–14971, 2005.

[2] Helen M Berman, John Westbrook, Zukang Feng, Gary Gilliland, Talapady N Bhat, Helge Weissig, Ilya N Shindyalov, and Philip E Bourne. The protein data bank. *Nucleic acids research*, 28(1):235–242, 2000.

[3] Carl Ivar Branden et al. *Introduction to protein structure*. Garland Science, 1999.

[4] Donald J Jacobs, Andrew J Rader, Leslie A Kuhn, and Michael F Thorpe. Protein flexibility predictions using graph theory. *Proteins: Structure, Function, and Bioinformatics*, 44(2):150–165, 2001.

[5] Binchen Mao, Roberto Tejero, David Baker, and Gaetano T Montelione. Protein nmr structures refined with rosetta have higher accuracy relative to corresponding x-ray crystal structures. *Journal of the American Chemical Society*, 136(5):1893–1906, 2014.

[6] Randy J Read, Paul D Adams, W Bryan Arendall III, Axel T Brunger, Paul Emsley, Robbie P Joosten, Gerard J Kleywegt, Eugene B Krissinel, Thomas Lütteke, Zbyszek Otwinowski, et al. A new generation of crystallographic validation tools for the protein data bank. *Structure*, 19(10):1395–1412, 2011.

[7] William R Taylor and András Aszódi. *Protein geometry, classification, topology and symmetry: A computational analysis of structure*. CRC Press, 2004.

[8] Eldon L Ulrich, Hideo Akutsu, Jurgen F Doreleijers, Yoko Harano, Yannis E Ioannidis, Jundong Lin, Miron Livny, Steve Mading, Dimitri Maziuk, Zachary Miller, et al. Biomagresbank. *Nucleic acids research*, 36(suppl_1):D402–D408, 2007.