Analysis of the performance of the CHESHIRE and YAPP methods at CASD-NMR round 3

Andrea Cavalli[1,2] and Michele Vendruscolo[1]

[1]*Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, UK*
[2]*Institute for Research in Biomedicine (IRB), 6500 Bellinzona, Switzerland*

Correspondence to: andrea.cavalli@irb.usi.ch, mv245@cam.ac.uk

Abstract

We present an analysis of the results obtained at CASD-NMR round 3 by the CHESHIRE and the YAPP methods. To determine protein structures, the CHESHIRE method uses solely information provided by NMR chemical shifts, while the YAPP method uses an automated assignment of NOESY spectra. Of the ten targets of CASD-NMR round 3, nine CHESHIRE predictions and eight YAPP ones were submitted. The eight YAPP predictions ranged from 0.7 to 1.9 Å Cα accuracy, with an average of 1.3 Å. The nine CHESHIRE predictions ranged from 0.8 to 2.6 Å Cα accuracy for the ordered regions of the proteins, with an average of 1.6 Å. Taken together, these results illustrate how the NOESY based YAPP method and the chemical shift based CHESHIRE method can provide structures of comparable quality.

Introduction

The CASD NMR initiative is aimed at assessing the performance of different protein structure determination methods that use information obtained through nuclear magnetic resonance (NMR) spectroscopy [1, 2]. The assessment is blind to the participants, as they are required to submit the results of their calculations without knowing in advance the correct structures of the targets. This type of assessment was inspired by the CASP exercise [3], which has been extremely helpful in the last 20 years in driving the development of structure prediction methods.

Although in principle NMR structure determination can be carried out using a variety of different NMR parameters, the CASD NMR exercise is focussed on two main approaches, as they are the most commonly applied ones. In the first, unassigned NOESY spectra are used as the source of structural information, and in the second assigned chemical shifts are instead employed. In this article we describe the results obtained using the YAPP method, which uses the first type of approach, and the CHESHIRE method [4, 5], which uses the second.

In the first two rounds of CASD NMR [1, 2] the NOESY peak lists provided were refined against the initial structural models during the determination of the reference structures and were, therefore, almost devoid of artifacts. This procedure simplifies the task for NOESY based methods or methods based on chemical shifts that use in some way the NOEs. It does, hoverer, not affect the performance of chemical shifts based approaches such as CHESHIRE. To further test the ability of automated procedures, in the third round of the CASD experiment it was decided to add an additional step. In a first phase unrefined peak lists where released together with chemical shift assignment, while the refined lists where hold back for 6 weeks. Structure calculations were thus performed with chemical shifts and/or unrefined peak lists, and with chemical shift and/or refined peak lists.

Methods

CHESHIRE: 'CHEmical SHIft REstraints'. The CHESHIRE method consists of a three-phase computational procedure [4, 5]. In the first phase, the chemical shifts and the intrinsic secondary structure propensities of amino acid triplets are used to predict the secondary structure of the protein under investigation. In the second phase, the secondary structure predictions and the chemical shifts are used to predict the backbone torsion angles. These angles are screened against a database to create a library of trial conformations of three- and nine-residue fragments spanning the sequence of the protein. In the third phase, a molecular fragment replacement strategy is used to assemble low-resolution structural models. The information provided by chemical shifts is used in this phase to

guide the assembly of the fragments. The resulting structures are refined with a hybrid molecular dynamics and Monte Carlo conformational search using a scoring function defined by: (1) the agreement between experimental and calculated chemical shifts, and (2) the energy of a molecular mechanics force field. This scoring function ensures that a structure is associated with a low CHESHIRE score only if it has a low value of the molecular mechanics energy and is highly consistent with experimental chemical shifts. In the calculations described in this article, typically 50,000 structures are generated for each target, and the best scoring one was submitted.

YAPP: 'Yet Another Peak Processor'. YAPP is an automated iterative procedure for simultaneous NOE assignment and structure calculation (**Figure 1**). YAPP uses an efficient implementation of torsion angular dynamics (TAD) [6, 7] to perform simulated annealing. YAPP uses chemical shifts information solely to assign atoms to potential peaks. A hydrogen atom (or a hydrogen atom covalently bound to a 'heavy' atom) is assigned to the corresponding NOESY peak dimension if its chemical shift value (or the chemical shift value of the bonded atom pairs) is within a given tolerance from the peak position. The default tolerance used in these calculations was set to 0.05 ppm for ^{1}H and 0.5 ppm for ^{13}C and ^{15}N. Peak intensities are converted into upper distances restraints using a $1/r^6$ relationship. After this initial chemical shift based assignment, the YAPP procedure continues with a step of assignment selection and seven cycles of assignment refinement. During the assignment selection step, YAPP performs structure calculations in an extended configuration space (X, λ), where X are the protein coordinates and $\lambda = (\lambda_p)_1^N, 0 < \lambda_p < 1$ is a vector of 'assignment' coordinates that is used to switch-off assignments that are systematically violated (see below).

The energy function used during this phase is the linear combination of a soft-core van der Waals term and extended restraints term

$$E_{YAPP} = E_{vdw} + E_{rest}.$$

The extended restrains term is

$$E_{rest} = \sum_{peak\ p} \begin{cases} \lambda_p (d_p - d_p^0)^2, d_p > d_p^0 \\ 0, if\ d_p < d_p^0 \end{cases} + D^2(1 - \lambda_p^2)$$

where $d_p = \left(\sum_a d_{i^a j^a}{}^{-6}\right)^{-1/6}$ is 'ambiguous' distance computed from all the assignment to a peak p, d_p^0 the upper distance restraint and D an additional parameter that controls the acceptable degree of violation.

From the previous equation, it can be seen that it is energetically more convenient for the whole system to turn off a restraint (λ_p =0) if it is violated by a distance larger than the user-selected value D, which in the cases discussed here was set to 2 Å. This strategy enables to identify assignments that can be satisfied from those that are systematically violated. The calculation of an extended structure, i.e. a pair (X, λ), is repeated 240 times with a simulated annealing protocol of 25,000 steps of torsion angle dynamics in 'protein space' coupled with a plain Metropolis Monte Carlo in the λ space. The 20 structures with the lowest energy are then used to remove assignments that cannot be satisfied and have λ_p =0.

After the selection of subset of restraints that can be satisfied up to a violation of about $D = 2$ Å, the assignment is refined for 7 additional cycles. The refinement cycles are carried out only in the protein space, i.e. the values λ_p are fixed to 1. In each cycle 240 structures are calculated using a simulated annealing protocol that consist of 25,000 steps of torsion angular dynamics. After each cycle ($n = 1, \cdots, 6$), all assignments that violate the upper distance limit in at least 5 of the 20 lowest energy structures by more that a given cutoff in are removed from the list used in the next iteration. The cutoff is progressively reduced from 0.9 to 0.1 from cycle 1 to cycle 6. Finally, a bundle of 240 structures is calculated using the final set of restraints. The structure with the lowest energy, together with the corresponding assignment, is kept as result. The whole procedure (selection, refinement and final calculation) is repeated 50-100 times, generating, therefore, 50-100 independent structure/assignment pairs.

Consistency checks

In CASD NMR round 3 we applied the CHESIRE and the YAPP methods using a completely automated procedure. Once a structure or a bundle of structures is generated, an assessment is made about its reliability. This step is carried out by performing a series of quality control tests on the final structures. The criteria used to assess the calculations are described below.

CHESHIRE consistency check. One of the critical steps for the accurate calculation of a structure from chemical shifts with CHESHIRE is the identification of a continuous 'structured' sub-sequence to use in the assessment. This step is considered because the proteins under investigation have often tags or unstructured regions at one of the termini. The removal of these unstructured regions is important in the CHESHIRE procedure because they increase the noise in the scoring function and thus they can make the sampling of the correct structure more difficult. The sequences were, therefore, trimmed using an automated procedure. This procedure removes segments at the beginning and the end if: (1) they are predicted to be 'coil' in terms of secondary structure, or (2) it is not possible to find enough fragments for these regions. A structure resulting from the CHESHIRE calculations passes the

consistency check if the calculations themselves generate a bundle of structures whose pairwise root mean square distance (RMSD) is below a certain cutoff (**Table 1**).

YAPP consistency check. To assess whether or not a YAPP bundle of structures is reliable, in the YAPP approach we first calculate the RMSD over the more ordered subset of amino acids, requiring this value to be less that 2.5 Å. Then we require the coverage of the ordered amino acids to be more that 80%; unstructured tails are not taken into account. To define the set of ordered amino acids we use the following method. First, for each pair of structures in the bundle we compute the largest set of amino acids that can be aligned. This is done using the MAXSUB algorithm [8]. The MAXSUB is fast algorithm that used linear optimization and some heuristics to find the maximal (largest) subset of amino acids that have an RMSD smaller that a given cutoff; here we use 4 Å. Then we define the set of ordered amino acids as those that can be aligned by MAXSUB in at least 80% of the pairwise comparisons. By applying these criteria we found that the two structures computed using the datasets provided to CASD NMR round 3 by the Arrowsmith group did not pass the consistency check. The results on calculation on the remaining structures are presented below. Inaccurate calculations identified by these criteria are shown in red in **Table 2**.

More insidious failures are those that are not detected by these consistency checks. These failures will be described separately below in the Results section.

Results

We participated to CASD-NMR round 3 with two methods developed in our group, the CHESHIRE method [5, 9] and the YAPP method. Both methods are implemented in the software package ALMOST [4]. CHESHIRE uses only information provided by backbone chemical shifts, while the YAPP method performs automated assignment of NOESY spectra. YAPP calculations were carried out with unrefined and refined peak list.

Analysis of the Cheshire results

From the 10 targets in CASD NMR round 3, 9 were used with CHESHIRE (**Table 1**). In all cases but one the structures computed by CHESHIRE are in good agreement with the reference conformations in the structured regions (**Figure 2**). The average accuracy of the CHESHIRE calculations is 1.67 Å,

with coverage of more than 80% of the structure; the coverage was determined using the MAXSUB algorithm with a cutoff of 4 Å [8]. For example, the accuracy in the calculation of the 99-residue protein HR6430A (PDB ID 2LA6) is 1.23 Å over 86 residues and 0.87 Å over 77 residues. These results show that although only backbone chemical shifts are used in the calculation of the final structure, the CHESHIRE procedure is able to generate structures that are of a quality comparable to those determined by methods that use considerably more information, such as NOESY-derived distances and is some cases RDCs.

In the remaining case, the target HR8254A (PDB ID 2M2E), the CHESHIRE structure was found to have a 4.61 Å RMSD from the reference structure. This large RMSD value is mainly due to the fact that in the reference structure of the C-terminal α-helix is completely extended, while in the structure generated by CHESIRE it is slightly bent (**Figure 2)**. Although without additional experiments it is difficult to decide which of the two structures is more correct, the comparison of the two structures could reveal which structural parameters could be most effective in solving this issue. It could be that the refinement procedure in CHESHIRE has a too strong bias towards compact structures and therefore bends this α-helix to reduce the radius of gyration. But it could also be that the lack of NOEs between α-helix 3 and the rest of the protein and the use of angular restraint obtained by TALOS [10] biases this portion of the protein towards a too idealized α-helix.

This type of problems was already noted in CASD NMR round 2. In that assessment chemical shift based methods were reported to have a poor performance with respect to NOESY based methods for one of the targets, namely the 97-residue protein AR3436A (PDB ID 2KJ6). For this target, CHESHIRE, ROSETTA and ROSETTA-DP produced structures with RMSD values from the reference structure larger that 3.3 Å (**Figure 3**). A subsequent re-analysis performed on the raw data, however, revealed that, although the chemical shifts assignment of the backbone was nearly complete and correct, the side-chain chemical shifts were incomplete and had misassignments [11]. After correcting these issues, structures computed with NOESY based methods matched the one obtained with chemical shifts alone. The RMSD of the CHESHIRE with the new, corrected reference structure is in fact 1.84 Å, and not 3.3 Å as previously reported [2] (**Figure 3**). These results show that, although chemical shift based structure calculations use less data than NOESY based ones, their accuracy can be equally accurate, and can help identify misassignments.

Analysis of the YAPP results

The YAPP protocol was applied to all the 10 targets of CASD NMR round 3 (**Table 2**). In 8 of 10 cases it provided structures in good agreement with the reference ones (**Figure 4**). In two cases the results of the calculations did not pass the internal consistency checks and were not submitted.

Interestingly, the 8 targets for which a submission was made were from the NESG centre, while the 2 targets for which a calculation was not completed were from the Arrowsmith group. Although a more accurate analysis will be needed, we suggest that these results may be a consequence of the different ways of selecting peaks between these two groups.

The average precision (bundle width) in calculations with raw peak lists is 1.20 Å, while in the case of structures computed with refined peak lists the average precision is 0.96 Å (**Table 2**). These results show that manual pruning of peaks has a small but noticeable effect on the precision of the structures determined with the YAPP method. The main effect is, however, that, at least on our hands, in one case the determination with raw peaks lists was not reliable, indicating that on average the effect of manual editing of the peak lists is small, but might improve the success rate of automated assignment protocols.

We also note that YAPP differs from other automated assignment protocols like CYANA or ARIA [12, 13] in one important aspect. The YAPP protocol is not aimed at finding a unique NOE assignment, but treats the assignments as variables. In this view, all structures in a YAPP bundle are the result of an independent assignment calculation and are, therefore, computed with slightly different NOESY-derived restraints. Each of the structures/assignment pair represents therefore a possible NOE assignment. It is therefore interesting to see how different the assignments are among them selves. **Figure 5** illustrates this aspect of the YAPP calculations in the case of the target HR2876C. As expected, most of the assignments are present in all 50 calculations done in this case. Interestingly, however, a non-negligible fraction of them are present in less that 5% of the structures. These results suggest that the assignment of NOEs to a structure may not be unique. To better quantify this type of variability we performed a consensus analysis, which is summarized in **Table 3**. On average all assignments are of same size, but only a small fraction (~55% for raw lists, ~77% for refined lists) are present in all the assignments. Structures computed only this subset of restraints in both cases very well defined (1.05 Å, 0.57 Å), but have a large RMSD to the reference PDB (2.26 Å, 1.32 Å). These results indicate that NOE lists may contain conflicting data that derive from slightly different structures and that the reference one represents them an average.

Conclusions

We have presented an analysis of the performance of the CHESHIRE and YAPP methods at the CASD NMR round 3. The results that we have described indicate that chemical shift based methods and NOESY based methods enable the determination of structures of comparable accuracy and that

they provide complementary information that can be used to correct possible artifacts in the calculations.

Compliance with Ethical Standards

The authors declare that they have no conflict of interest. This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Rosato, A., et al., *CASD-NMR: critical assessment of automated structure determination by NMR.* Nat Methods, 2009. **6**(9): 625-6.
2. Rosato, A., et al., *Blind testing of routine, fully automated determination of protein structures from NMR data.* Structure, 2012. **20**(2): 227-36.
3. Moult, J., et al., *A large-scale experiment to assess protein structure prediction methods.* Proteins, 1995. **23**(3): ii-v.
4. Fu, B., et al., *ALMOST: an all atom molecular simulation toolkit for protein structure determination.* J Comput Chem, 2014. **35**(14): 1101-5.
5. Cavalli, A., et al., *Protein structure determination from NMR chemical shifts.* Proc Natl Acad Sci U S A, 2007. **104**(23): 9615-20.
6. Guntert, P., C. Mumenthaler, and K. Wuthrich, *Torsion angle dynamics for NMR structure calculation with the new program DYANA.* J Mol Biol, 1997. **273**(1): 283-98.
7. Jain, A., N. Vaidehi, and G. Rodriguez, *A Fast Recursive Algorithm for Molecular-Dynamics Simulation.* J Comp Phys, 1993. **106**(2): 258-268.
8. Siew, N., et al., *MaxSub: an automated measure for the assessment of protein structure prediction quality.* Bioinformatics, 2000. **16**(9): 776-85.
9. Cavalli, A., R.W. Montalvao, and M. Vendruscolo, *Using chemical shifts to determine structural changes in proteins upon complex formation.* J Phys Chem B, 2011. **115**(30): 9491-4.
10. Shen, Y., et al., *TALOS+: a hybrid method for predicting protein backbone torsion angles from NMR chemical shifts.* J Biomol NMR, 2009. **44**(4): 213-23.
11. Zhang, Z., et al., *Robust and highly accurate automatic NOESY assignment and structure determination with Rosetta.* J Biomol NMR, 2014. **59**(3): 135-45.
12. Herrmann, T., P. Guntert, and K. Wuthrich, *Protein NMR structure determination with automated NOE assignment using the new software CANDID and the torsion angle dynamics algorithm DYANA.* J Mol Biol, 2002. **319**(1): 209-27.
13. Habeck, M., et al., *NOE assignment with ARIA 2.0: the nuts and bolts.* Meth Mol Biol, 2004. **278**: 379-402.

Target	Accuracy cutoff 4 Å		Accuracy cutoff 2 Å	
HR2876C	1.91	79/97	0.95	58/97
HR8254A	4.61	68/71	1.52	39/71
YR313A	2.16	102/119	1.74	89/119
HR2876B	2.36	97/107	1.22	65/107
StT322	1.26	26/63	0.95	22/63
OR135	1.44	74/83	1.11	64/83
OR36	2.75	128/134	2.60	126/134
HR6430A	1.24	86/99	0.86	77/99
HR6470A	0.76	48/69	0.76	48/69

Table 1. Assessment of the quality of the structures calculated using the CHESHIRE method. Columns 2 and 4: Cα-RMSD (in Å) between CHESHIRE and reference PDB computed with MAXSUB with a cutoff of 4 Å and 2 Å, respectively. Columns 3 and 5: fraction of the amino acids used in the alignment.

Target	Raw		Final		RANGE	From
	Bundle	C α	Bundle	C α		
HR2876B	0.82	1.54	0.79	0.93	11-107	GM
HR2876C	1.25	2.20	0.80	1.30	16-93	GM
HR5460A	4.13	4.88	0.88	1.16	11-28:31-159	GM
HR6430A	0.89	1.16	0.71	1.02	12-99	GM
HR6470A	1.08	1.44	0.82	0.68	10-59	GM
OR135	1.61	2.63	1.48	1.76	3-76	GM
OR36	1.27	2.21	1.29	1.90	1-48:51-128	GM
YR313A	1.49	2.12	0.96	1.05106	16-36:38-43:45-112:114-116	GM

Table 2. Assessment of the quality of the structures calculated using the YAPP method. Columns 2 and 4: average pairvise backbone RMSD (in Å) between all models in the bundle. Columns 3 and 5: average backbone RMSD between structures in the YAPP bundle and reference PDB structure. Column 6: range of amino acids used in the aligment. Column 7: source of the NMR data used in the calculations (GM: Montelione lab).

	Raw	Final
# NOEs**	2589.88 ± 21.33	3455.84 ±13.689
# NOEs cons	1429	2677
Bundle	1.05	0.57
RMSD	2.26	1.33
Bundle NOE + TALOS	0.82	0.43
RMSD NOE + TALOS	1.09	0.89

Table 3. Consensus analysis of target HR2876C.

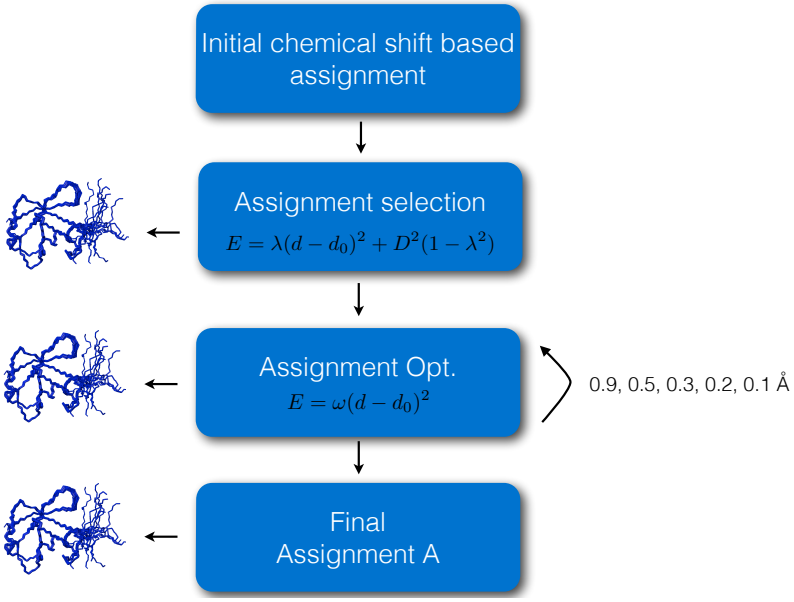

Figure 1. Flow diagram of the YAPP method.

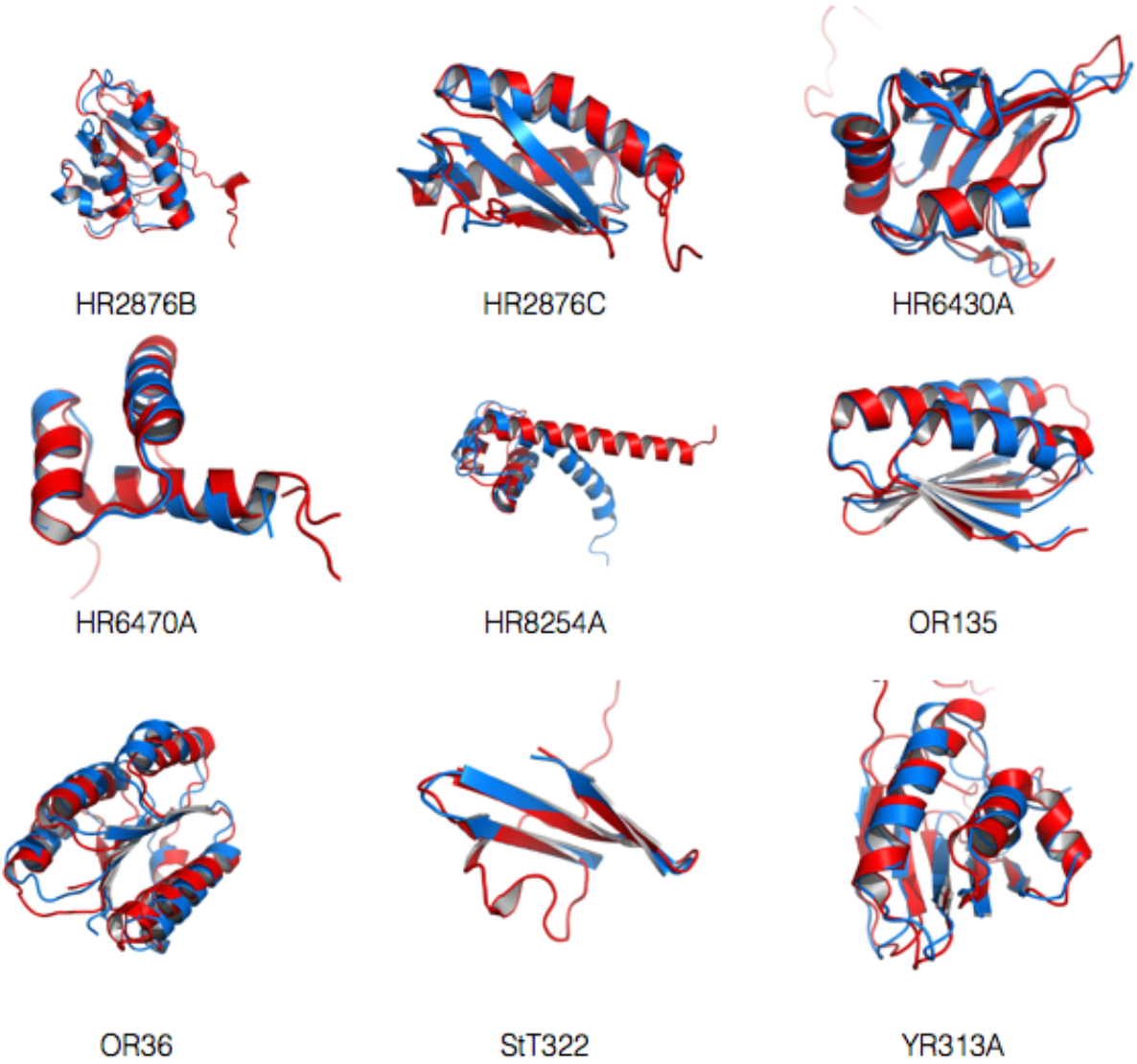

Figure 2. List of structures obtained using the CHESHIRE method.

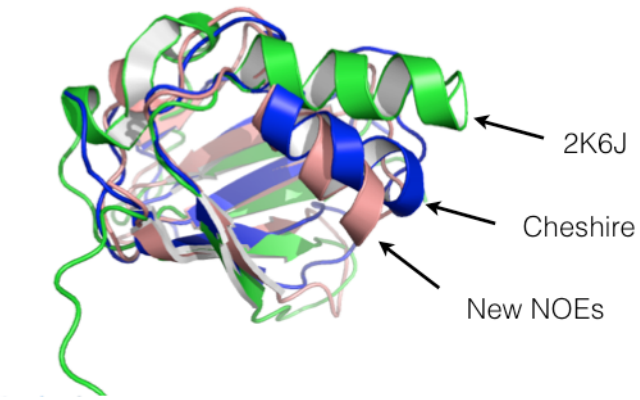

Figure 3. Comparison of the structures 2K6J, the new NOESY-based structure and the CHESHIRE structure from CASD NMR round 2.

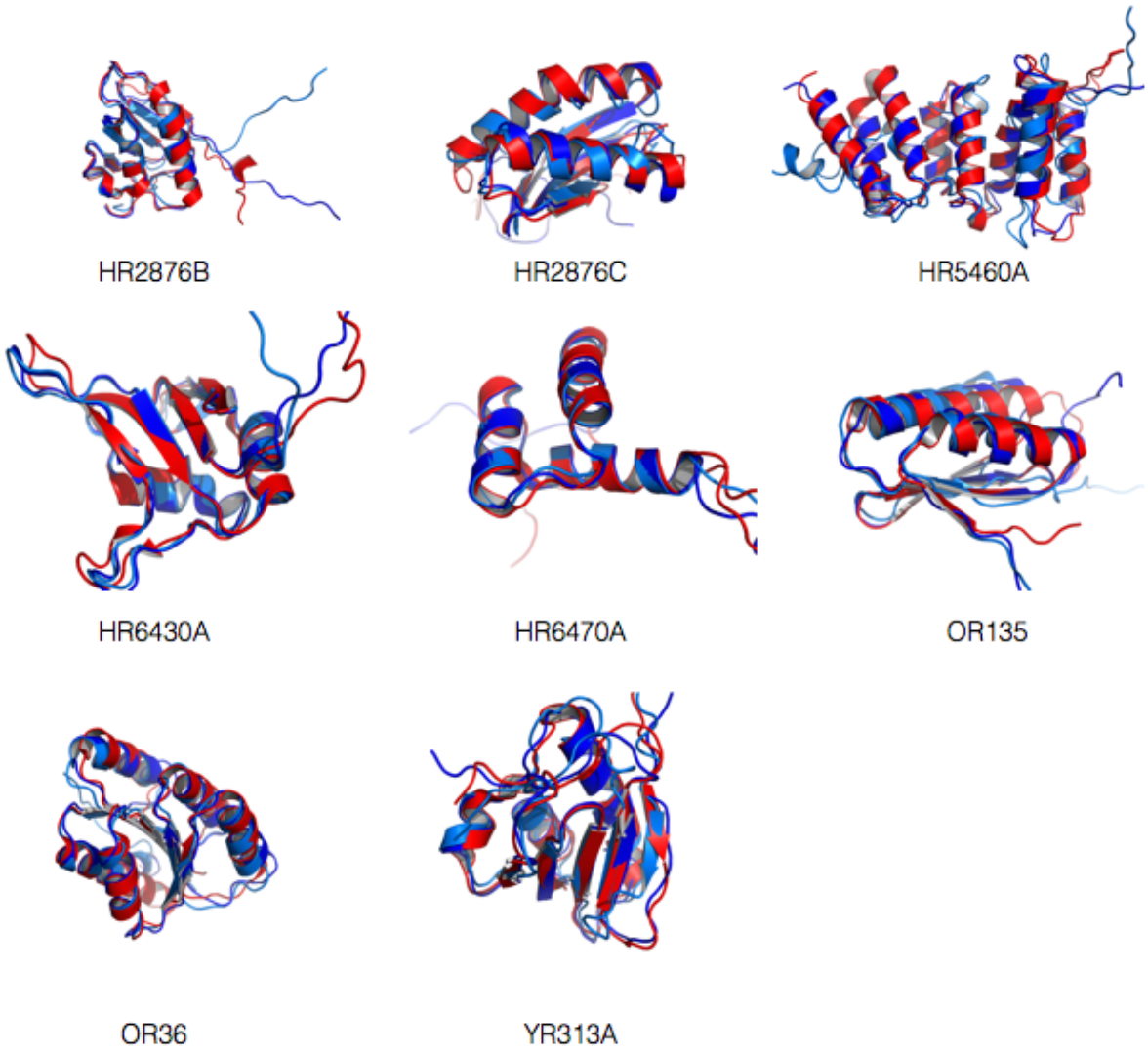

Figure 4. List of structures obtained using the YAPP method.

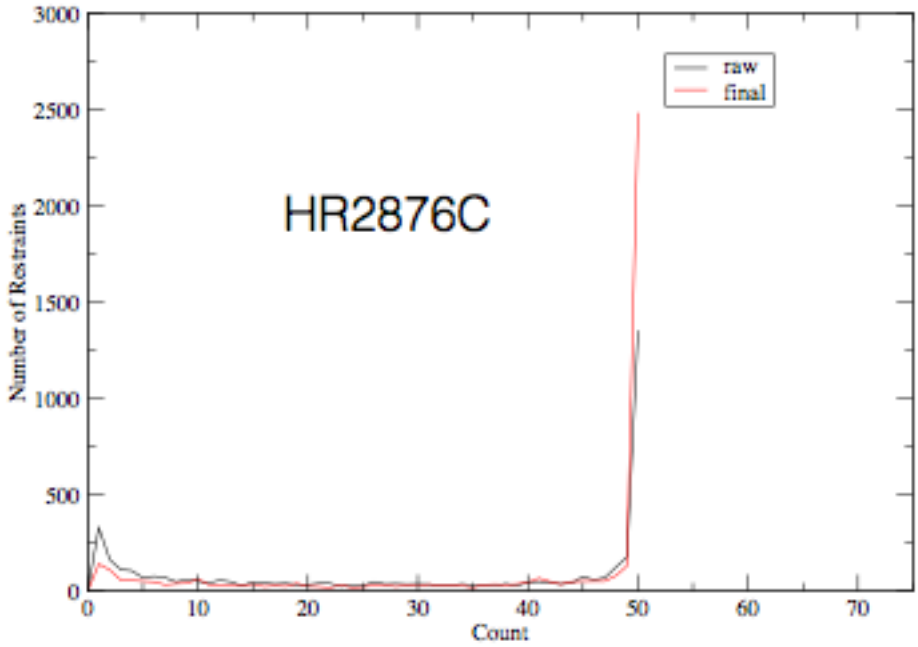

Figure 5. Analysis of the different assignments obtained with the YAPP method for the target HR2876C.