CASP13 NMR-Guided Prediction

Description of NMR restraint data and formats

NMR-based contacts

A NMR resonance signal (aka NOESY *cross peak*) in a so-called multidimensional NMR NOESY spectrum corresponds to an interaction between a pair of hydrogen atoms that are close in 3D space (i.e. < 5 to 6 Å) within the protein structure. By matching the frequency coordinates of the cross peak to the resonance frequencies of the atoms in a list of NMR chemical shift assignments, it is possible to assign each NOESY cross peak to a specific interaction, thereby experimentally identifying the identities of the two atoms that are close in space. This constitutes a *distance restraint*, that can be used in protein structure modeling/determination. Typically, such a distance restraint is represented as an upper distance limit of 5 Å (in practice, this may differ depending on sample concentration and spectral quality). Each spectrum contains several 100's to a few 1000's of NOESY cross peaks.

To facilitate the correct identification of the proton pair to be restrained, the signals in multidimensional NMR NOESY spectra are separated based on the frequency of the resonance of nitrogen or carbon nuclei that are bound to the protons of interest (e.g. the amide nitrogen ^{15}N for backbone amide groups). In this way, two amide hydrogens that casually happen to have the same resonance frequency can be distinguished if their bound ^{15}N atoms resonate at different frequencies.

For larger proteins (> ~ 20 residues) efficient nuclear relaxation causes the resonances to broaden, eventually resulting in signal heights so small they cannot be detected. To overcome this problem, proteins can be prepared in which all carbon atoms are perdeuterated and thus not detected, providing much more narrow resonances (and higher signal) for the remaining proton sites. The amide nitrogen atoms can be re-protonated by back exchange from solvent water, and methyl sites can be protonated by biosynthetic methods. The resulting NOESY NMR spectra have NOESY cross peaks only between backbone amide (designated H), side chain amide (HD and HE), Ala Methyl (HB), and Ile, Leu and Val methyl (HG and HD) resonances.

Despite the above considerations, not all ambiguities in the assignment of NOESY signals can be resolved especially for larger proteins. Resonance overlap is also affected by the limits in experimental resolution that can be achieved in particular spectral dimensions. Therefore, in general, each signal in the NOESY spectrum can be assigned to **one or more** pairs of interacting hydrogen atoms, as exemplified below.

First Residue number	**Second residue number**	**First hydrogen atom**	**Second hydrogen atom**	**NOESY Cross Peak identifier**
90	87	H	H	2
28	122	H	HB	8
28	90	H	HB	8
28	224	H	HB	8
28	24	H	HB	8

In this example, there is only one possible assignment for NOESY cross peak #2 (interaction between the amide protons of residue 87 and 90) whereas there are four possible assignments for NOESY cross peak #8 (interaction between the amide proton of residue 28 and the methyl group of four different alanines). The latter is called an **ambiguous distance restraint**. For ambiguous restraints, at least one possible assignment should be true (i.e. correspond to a distance < 5 to 6 Å); it is also possible that multiple assignments are satisfied, if the observed signal is actually caused by the accidental overlap of multiple signals. Conversely, it may happen that **none** of the possible assignments is consistent with the real structure; this is the case if noise in the spectrum was mistakenly taken as a real signal, or the correct frequency of resonance of one of the two hydrogen atoms actually involved in the interaction is not known or incorrectly assigned.

Note that NMR does not distinguish between the three protons of a methyl group, which all have the same frequency of resonance; NMR-derived distances involving methyl groups correspond to a combination of the individual distances according to the following formula (Eqn. 1).

$$d_{eff}(H-HB) = \left(\sqrt[6]{\sum_{i=1}^{3} d^{-\frac{1}{6}}(H-HB_i)} \right)^{-1} \qquad Eqn.1$$

where d(H-HB_i) is the distance between the H atom and the i-th atom of the methyl group in the static structural model. Therefore, the following notation has been used for NOE-based contacts involving methyl groups: HB for the methyl of Ala, HG1 and HG2 for the methyls of Val, HD1 and HD2 for the methyls of Leu, and HD1 (the only methyl of Ile usually ^{13}C enriched) for the delta-methyl of Ile. When the two methyls of the same Val or Leu have the same resonance frequency, all six atoms are considered as a single group (designated QG or QD, respectively), and are included in the formula above, with *i* running up to 6.

In addition to backbone amide HN and sidechain methyl groups, NMR signals in perdeuterated, 13C-methyl labeled proteins may arise from HD21/HD22/HE21/HE22 atoms of the amide groups of the sidechains of Asn and Gln.

Residue numbering follows the same numbering as in the protein sequence. The **Ambiguous Assignment Table** (above) exemplifies the format used to list NMR-based contacts in the CASP13 Ambiguous Contact File (e.g. T0953s2_ambiR.txt).

Residual dipolar coupling

Recent reviews[1-4] outline the theory and demonstrate the utility of Residual Dipolar Coupling (RDC) data in a broad spectrum of applications to macromolecules. RDCs arise from the dipolar interaction between two magnetically active nuclei in the presence of the external magnetic field of an NMR instrument. This interaction is normally averaged to zero by the isotropic tumbling of molecules in their aqueous environment. The introduction of partial order and transient molecular alignment reintroduces dipolar interactions by minutely limiting isotropic tumbling. This partial order can be introduced in numerous ways, including molecular orientation by the inherent magnetic anisotropy susceptibility of molecules, incorporation of artificial tags (such as lanthanides) that exhibit magnetic anisotropy, stretched gels, or in a liquid crystal aqueous solution (as illustrated in Figure 1).

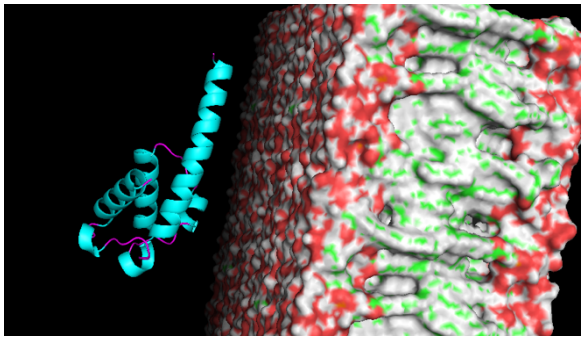

Figure 1. Bicelle crystalline solution is one method of inducing partial alignment.

The RDC interaction phenomenon has been formulated in different ways[5-8]. Equation 2 can be used as the simplest method of calculating RDC data from a given structure, with the parameters D_a and R (which for CASP13 are provided for each orientation in the RDC data file) defining the transient molecular orientation tensor. In this equation the value of r_{ij} corresponds to the physical separation of the two interacting nuclei (e.g. ^{15}N and bound ^{1}H) in units of Ångstrom, and (θ,φ) are the spherical coordinates of the vector joining these two interacting nuclei. Please note that spherical coordinates of N-H vectors are relative to the coordinate system of the alignment frame[9,10], so that meaningful comparison of the computed RDCs requires that the molecular coordinate system is first optimally rotated into this alignment coordinate system. In relation to the bond length r, the common practice among some investigators is to use a constant value for all residues (values of 0.98 Å -1.02 Å have been reported in the literature). In the CASP13 simNMR simulations, we have optimized these N-H bond lengths for each residue, resulting in values ranging from 1.009 Å tro 1.011 Å.

$$D_{ij} = \frac{D_a}{r^3}\left[(3 \cdot cos^2(\theta) - 1) + \frac{3}{2} \cdot R \cdot sin^2(\theta)\cos(2\varphi)\right] \qquad Eqn\ 2$$

The fitness between computed and experimental RDC data can be measured using the RDC-rmsd[10,11] or Q-factor[10-12] metrics. In general, a value of Q-factor less than 0.2 indicates close structural fitness, while values around 0.3 indicates room for additional structural refinement. Note that in the case of N-H RDC data, the structural refinement may simply consist of optimizing the location of the amide protons.

Several approaches have been proposed[13-16] for obtaining the optimal molecular orientaton of a protein that will minimize the fitness between computed and experimental RDC data (measured in the RDC-rmsd or Q-factor metrics). One such approach guarantees optimization of the orientational search using a Singular Value Decomposition[11,14,17]. These variety of approaches have been incorporate in structure calculation software such as Xplor-NIH, CNS, and Cyana.

The programs REDCAT[10,11] and Pales[12] are designed specifically to assess the agreement between a set of coordinates and RDC data. For the specific objectives of CASP13, the REDCAT software package may be used for analysis of RDCs and evaluation of the fitness of a structure to the RDC data. The REDCAT software package can be downloaded from the following URL: https://ifestos.cse.sc.edu/?q=softwares#redcat.

Contact Prediction

For larger proteins and complexes, NMR data can be complemented by residue-residue contact predictions (ECs) based on evolutionary co-variance analysis[18]. Several methods are available to predict ECs[19-24]. As a standard set of predicted residue-residue contacts, we have provided the results of METAPSICOV[21,24] submissions to CASP13 (e.g. T0957s1_ECs.txt). Predictors are encouraged to also explore other methods for contact prediction for these NMR-guided prediction targets.

simNMR data

In this exercise, chemical shift lists and NOESY peak were simulated for CASP Free Modeling (FM) targets. These simulated NMR data were then analyzed with the NOESY assignment program ASDP (also called AutoAssign)[25,26] to generate an Ambiguous Contact Table for each protein. In addition, ranges of backbone dihedral angles for well-ordered residues in regular secondary structures have been estimated using ranges analogous to those provided by the standard method used to define dihedral angle restraints from backbone chemical shift data[27]. Backbone ^{15}N-^{1}H RDCs were also simulated for two different alignments using the program Redcat[10,11], and provided along with the corresponding alignment tensor parameter D_a and R. The uncertainty in each of these RDC values is ± 0.5 Hz.

CASP13 Predictors are provided:

1. The CASP13 target sequence file (T0953s2_seq.txt)
2. Ambiguous Contact Table based on simulated NOESY data (e.g. T0953s2_ambiR.txt)
3. Simulated backbone ^{15}N-^{1}H RDCs and alignment tensor parameters (e.g. T0953s2_RDCs.txt)
4. Predicted Contacts (or Evolutionary Co-variances, ECs) from the METAPSICOV submission to CASP13 (e.g. T0953s2_ECs.txt)
5. Backbone dihedral angle ranges as would be determined from chemical shift data (e.g. T0953s2_dihed.txt)

Predictors are encouraged to combine these NMR data with predicted contacts (see Tang et al, 2015[18]), either as provided with the simNMR data, or as predicted by other methods.

Real NMR data will also be provided for one or two CASP targets.

References

1 Prestegard, J. H., Al-Hashimi, H. M. & Tolman, J. R. NMR structures of biomolecules using field oriented media and residual dipolar couplings. *Q Rev Biophys* **33**, 371-424 (2000).

2 Bax, A., Kontaxis, G. & Tjandra, N. Dipolar couplings in macromolecular structure determination. *Methods Enzymol* **339**, 127-174 (2001).

3 Prestegard, J. H., Bougault, C. M. & Kishore, A. I. Residual dipolar couplings in structure determination of biomolecules. *Chem Rev* **104**, 3519-3540 (2004).

4 Blackledge, M. Recent progress in the study of biomolecular structure and dynamics in solution from residual dipolar couplings. *Prog Nucl Mag Res Sp* **46**, 23-61, doi:10.1016/j.pnmrs.2004.11.002 (2005).

5 Tolman, J. R., Flanagan, J. M., Kennedy, M. A. & Prestegard, J. H. Nuclear magnetic dipole interactions in field-oriented proteins - Information for structure determination in solution. *P Natl Acad Sci USA* **92**, 9279-9283 (1995).

6 Prestegard, J. H., Mayer, K. L., Valafar, H. & Benison, G. C. Determination of protein backbone structures from residual dipolar couplings. *Methods Enzymol* **394**, 175-209, doi:10.1016/S0076-6879(05)94007-X (2005).

7 Bryson, M., Tian, F., Prestegard, J. H. & Valafar, H. REDCRAFT: a tool for simultaneous characterization of protein backbone structure and motion from RDC data. *J Magn Reson* **191**, 322-334, doi:10.1016/j.jmr.2008.01.007 (2008).

8 Shealy, P., Liu, Y., Simin, M. & Valafar, H. Backbone resonance assignment and order tensor estimation using residual dipolar couplings. *J Biomol NMR* **50**, 357-369, doi:10.1007/s10858-011-9521-5 (2011).
9 Losonczi, J. A., Andrec, M., Fischer, M. W. & Prestegard, J. H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson* **138**, 334-342, doi:10.1006/jmre.1999.1754 (1999).
10 Valafar, H. & Prestegard, J. H. REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* **167**, 228-241, doi:10.1016/j.jmr.2003.12.012 S1090780703004361 [pii] (2004).
11 Schmidt, C., Irausquin, S. J. & Valafar, H. Advances in the REDCAT software package. *BMC Bioinformatics* **14**, 302, doi:10.1186/1471-2105-14-302 (2013).
12 Zweckstetter, M. NMR: prediction of molecular alignment from structure using the PALES software. *Nat Protoc* **3**, 679-690 (2008).
13 Warren, J. J. & Moore, P. B. A maximum likelihood method for determining D(a)(PQ) and R for sets of dipolar coupling data. *J Magn Reson* **149**, 271-275, doi:10.1006/jmre.2001.2307 (2001).
14 Valafar, H. & Prestegard, J. H. REDCAT: a residual dipolar coupling analysis tool. *J Magn Reson* **167**, 228-241, doi:10.1016/j.jmr.2003.12.012 (2004).
15 Miao, X., Mukhopadhyay, R. & Valafar, H. Estimation of relative order tensors, and reconstruction of vectors in space using unassigned RDC data and its application. *J Magn Reson* **194**, 202-211, doi:10.1016/j.jmr.2008.07.005 (2008).
16 Mukhopadhyay, R., Miao, X., Shealy, P. & Valafar, H. Efficient and accurate estimation of relative order tensors from lambda-maps. *J Magn Reson* **198**, 236-247, doi:10.1016/j.jmr.2009.02.014 (2009).
17 Losonczi, J. a., Andrec, M., Fischer, M. W. & Prestegard, J. H. Order matrix analysis of residual dipolar couplings using singular value decomposition. *J Magn Reson* **138**, 334-342, doi:10.1006/jmre.1999.1754 (1999).
18 Tang, Y. *et al.* Protein structure determination by combining sparse NMR data with evolutionary couplings. *Nature Methods* **12**, 751-754, doi:10.1038/nmeth.3455 (2015).
19 Morcos, F. *et al.* Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences of the United States of America* **108**, E1293-1301, doi:10.1073/pnas.1111471108 (2011).
20 Marks, D. S. *et al.* Protein 3D structure computed from evolutionary sequence variation. *PLoS One* **6**, e28766, doi:10.1371/journal.pone.0028766 (2011).
21 Jones, D. T., Buchan, D. W., Cozzetto, D. & Pontil, M. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics* **28**, 184-190, doi:10.1093/bioinformatics/btr638 (2012).
22 Kamisetty, H., Ovchinnikov, S. & Baker, D. Assessing the utility of coevolution-based residue-residue contact predictions in a sequence- and structure-rich era. *Proc. Natl. Acad. Sci. USA* **110**, 15674-15679, doi:10.1073/pnas.1314045110 (2013).
23 Michel, M. *et al.* PconsFold: improved contact predictions improve protein models. *Bioinformatics* **30**, i482-i488, doi:10.1093/bioinformatics/btu458 (2014).
24 Jones, D. T., Singh, T., Kosciolek, T. & Tetchner, S. MetaPSICOV: combining coevolution methods for accurate prediction of contacts and long range hydrogen bonding in proteins. *Bioinformatics* **31**, 999-1006, doi:10.1093/bioinformatics/btu791 (2015).
25 Huang, Y. J., Tejero, R., Powers, R. & Montelione, G. T. A topology-constrained distance network algorithm for protein structure determination from NOESY data. *Proteins* **62**, 587-603, doi:10.1002/prot.20820 (2006).

26 Huang, Y. J., Mao, B., Xu, F. & Montelione, G. T. Guiding automated NMR structure determination using a global optimization metric, the NMR DP score. *J Biomol NMR* **62**, 439-451, doi:10.1007/s10858-015-9955-2 (2015).

27 Shen, Y. & Bax, A. Protein structural information derived from NMR chemical shift with the neural network program TALOS-N. *Methods Mol Biol* **1260**, 17-32, doi:10.1007/978-1-4939-2239-0_2 (2015).