CASP14 NMR-Guided Prediction

Description of NMR restraint data and formats

For each CASP-NMR Targets, the following data are provide

1. The CASP14 target sequence file (e.g. N1088_seq.txt)
2. **Ambiguous Contact Table** based on real NOESY data (e.g. N1088_AmbiR.txt)
3 Backbone dihedral angle ranges determined from chemical shift data using the program TALOS_N (e.g.N1088_DIHE.txt)

Predictors are encouraged to combine these NMR data with their modeling protocols and or predicted contacts [see for example 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)] to generate Sparse-NMR Models.

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). In some cases, these upper-bound distance estimates are sometimes made more precisely. The NOE data provided for CASP14 uses calibrated upper-bound distances (e.g. < 3.6 Å), but it is not uncommon for modelers to set all of these upper-bound distance values to 5.0 Å. Each spectrum contains 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 (larger than about 200 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 (i.e. some H atoms are replaced biosynthetically with ^{2}H) 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	NOES cross peak identifier	Distance	First hydrogen atom	Second hydrogen atom
90	87	2	3.6	H	H
28	122	8	5.0	H	HB
28	90	8	5.0	H	HB
28	224	8	5.0	H	HB
28	24	8	5.0	H	HB

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, i.e., Ala24, Ala90, Ala122, and Ala224). The latter is called an ambiguous distance restraint. For ambiguous restraints, at least one possible assignment should be true (e.g. at least one of the corresponding distances in the model should be < 5 Å); it is also possible that multiple assignments are satisfied for a single NOESY peak, 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, if the correct frequency of resonance of one of the two hydrogen atoms actually involved in the interaction is not known or incorrectly assigned, or if there is an error in the resonance assignments.

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

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

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, ^{13}C-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, as provided in the .seq file. The **Ambiguous Assignment Table** (above) exemplifies the format used to list NMR-based contacts in the CASP14 Ambiguous Contact File (e.g. N1088_AmbiR.txt).