NMR for the non-spectroscopist

Even if the article does not explain very much in detail how NMR works, it points out the main principles of NMR applied to proteins. Most importantly:

- How to interpret 1D or 2D NMR spectra
- How to obtain the 3D structure of a protein in a solution
- How to determine the quality of a NMR structure
- Advantages and disadvantages of NMR

The main principle of spectroscopie is to induce a transition from one state to another state in a molecule. The energy of such transitions is well defined according to the principles of quantum chemistry.

- In infrared, we induce a transition in the vibration mode.
- In UV spectroscopy, we induce electronic transitions
- In NMR, we know that the different permitted orientations of a dipole in a magnetic field correspond to different energy states. In order to cause a transition, we need to provide the exact amount of energy in the form of an electromagnetic pulse.

What is NMR?

- NMR spectra arise from transitions made by atomic nuclei between different energy states.
- The nuclei of many isotopes carry magnetic dipoles, that take up different orientations in a magnetic field → in the magnet of an NMR spectrometer.
- According to quantum mechanics, transitions between states with certain energies are permitted.
- When we apply pulses of electromagnetic radiation at frequencies that precisely match those energy gaps, we are able to observe transitions that give rise to NMR signals.
- The frequency at which a given nucleus resonates depends on its chemical environment, that is, nuclei in different environments resonate at different frequencies. For example, a proton that is bound to a N will not have the same frequence as a proton bound to O.
- 1D NMR spectrum: plot with the chemical shift in the X axis and the intensity (number of nuclei existing in the same chemical environment). This is a 1D NMR plot.
- Raw frequencies (MHz) scale with the size of the NMR magnet. To correct for this, we report resonance frequencies as “chemical shifts” (ppm).
- Key property in NMR: excited states are relatively long-lived (nano-millisecond range) compared to other spectroscopy techniques→ we can have multiple transfer steps in a single experiment. This means: If we have two atoms H and C bound to each other in a given environment, and we apply a frequency that matches that of A atom, this atom will become excited, and since the excitation is long-lived, it will be able to transfer its excitation to the C atom and we will be able to see both things in the same experiment. Thus, we will obtain a signal that corresponds to these two frequences (that of H and that of C). This property enables us to do 2D NMR plots, in which we show the chemical shift of H against the chemical shift of C. For every H-C bound pair, we will have a signal in the plot.

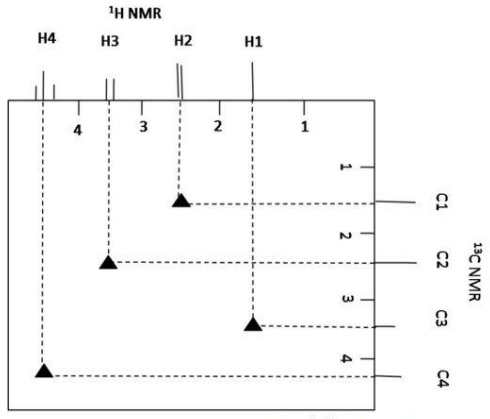

Fig 13. ^{1}H-^{13}C COSY spectrum

Since we can measure the frequencies of each of the nuclei through which excitation is passed, we can obtain signals that correlate the frequencies of many nuclei → multidimensional spectra: each transfer can be visualized as an independent nuclear frequency dimension (axis), and signals occurring at the intersections of two or more frequencies indicate a correlation between the corresponsing nuclei.

In a protein, if we plot a 2D N-H NMR, we would obtain a point for every amide of the peptide bond, that is, a point for every residue.

For sugars, the chemical environment is very similar. For DNA, there are not that many H-s so the signal is not that clear.

Types of NMR spectra:

- 1D H-NMR: shows signals for each H atom.
- 2D N-HSQC: a signal for each covalently bonded H-N group. Each signal has an intensity and two chemical shifts (one for the H and one for the N). In essence, one peak for each residue in the protein → excellent high-resolution fingerprint of the protein.
  - One peak for every backbone amide proton.
  - A pair of peaks for each Trp
  - A pair of peaks the side chain amide groups in Asn and Gln.
- C-HSQC: a signal for each covalently-bonded H-C pair. Not as well resolved as N-HSQC.

How much **sample** do I need?

- Some years ago big quantities were needed. Nowadays, 50 ug are enough, and they can be recovered and used for other purposes after the experiment.
- We can work with natural abundances of the isotopes, but the signal will not be that good. We usually work on recombinant proteins that have been overexpressed in minimal media containing nutrients enriched in those isotopes.
- The sample needs to be homogeneous, but we sometimes work with mixed samples.
- Instead of using water, we use D_2O, so that they wont resonate.
- Be careful with the pH: if it is too high, protons in the sample will exchange with the solvant and the signal will have poor quality.

Information that can be deduced from an NMR experiment:

- Is the protein **folded**?
  - Unfolded → all residues are exposed to solvent → poor signal dispersion.
  - Folded → beta sheets have a rich environment and will present a better dispersion than alpha helices.
  - Molten globule: we can have transitions in the ms range, which are be observable in an NMR spectrum. The peaks will not be well-defined.

- Is the protein **agreggated**?
  - Agreggation broadens signals Conformational exchange only broadens the signals of the nuclei whose environment has changed.

- Is the protein **dynamic**? We can know if the protein undergoes μs-ms conformational exchange. If we have a protein that has 120 residues and we do not see that many points in the spectrum, it might be because there are disordered regions in the protein. The time they need to get excites is too long for them to be observable in an NMR spectrum.

- How **stable** is the protein? Degradation is visible in NMR spectra.

Other important parameters:

- Strength of the magnetic field: higher strength → better signal
- Molecular weight: over 50kDa it becomes difficult.
- Temperature: higher T → better sigal (except for proteins that are not stable over some T).

- Buffer: No firm guidelines.
- Membrane proteins: most important → the choice of solubilizing detergents

How do we obtain a protein 3D structure from NMR data?

NMR does not directly give us a fixed image of the protein (as other forms of spectroscopy do). Moreover, in NMR we study proteins in solution rather than fixed (as in X-ray crystallography). Thus, we have to build a 3d structure from indirect evidence. We call these indirecte evidence "structural restraints". This also has advantages:

- We measure much more thana static image of a protein
- We can study proteins in their native solution

La description de ces trois restreintes est un peu technique, et je ne suis pas sûr que vous soyez questionné là-dessus, mais j'ai quand même fait le point.

Triple resonance experiments:

- First task: find the CS of every atom in the molecule → resonance assignment (H, N, C). We usually assign around 90% of the atoms.
- Now we gather structural restraints. Usually: proton-proton distances, dihedral angles and hydrogen bonds.
  - **H-H distance**: the dipolar interactions between protons, measured as a H, H nuclear Overhauser effect (NOE) in 2D NOESY experiments. This is one of the main sources of information in NMR, because it yields information about the distance between protons, which are very abundant in proteins. These restraints are so important in NMR that they can be used on their own. Protons that are close to each other will have a greater dipolar interaction.
  - **Angles**: scalar couplings (between pairs of nuclei that are close in the covalent structure of the molecule) → the intensity of this measurement is related to the backbone dihedral phi and psi angles.
  - **Hydrogen bonds**: we assume that structured proteins exchange fewer protons with the environment than disorder proteins do. If we analyse the $D_2O \leftarrow\rightarrow H_2O$ exchange rate, the exchange rate will be higher in hidrogen bonds. can also be inferred as structural restraints.

Once we have found the structural restraints, we can use them as restraints in a MD simulation. It is best to have many restraints.

MD simulations that we use to calculate NMR structures are not equivalent to the ones we use to simulate the dynamic behaviour of a biomolecular system. In NMR we only want to obtain a structure that fits a given set of restraints, the trajectories of the atoms are not important.

How can we interpret the structure that we have obtained?

We have performed the MD simulations several times and we have obtained an ensemble of structures that satisfy the experimental restraints. The ensemble that we obtain does not correspond to the different conformations of the protein! It represents the superimposition of different structures that the protein takes.

The regions that undergo rapid changes will not be represented in our structure. Some mobile parts will appear, yet we are usually interested in the conserved regions.

How to assess the quality of NMR structures?

We can calculate the rmsd of the ensemble of structures and take the mean structure, but we have no guarantee that it will be better than any of the individual structures in the ensemble, because NMR structures show averaged conformations of the protein.
We can use Procheck to evaluate the stereochemical features of the structures (for example, Ramachandran plot). We have to be careful not to take into account the highly mobile regions of the protein, because they limit resolucion. Note: resolution cannot be measured since NMR structures are not images. Number of restraints per residue.

NMR works better in small molecules than in big ones. Why? The correlation time increases with molecule size. If the transition time is too long when compared to the excitation time.

NMR structure validation

Some interesting notes:

- The more specific data included, the less general modeling is required.
- The very high precision with which atoms are described in silico can be misleading, as the uncertainties of the process are not immediately apparent.
- Validation can be knowledge-based or model vs data.
- The set of structures obtained in the end do NOT correspond to the conformers. The set will depend on the experimental data used and also on the modeling process.

Blind testing of routine, fully automated determination of protein structures from NMR data

Is it possible to generate NMR structures in a fully automated manner?
Why atuomated? Because it would be faster, easier and more reproducible.

They took 10 experimental data sets with unassigned nuclear Overhauser effect spectroscopy (NOESY) peak lists for various proteins of unknown structure. Their structures were computed using different fully automatic programs. The resulting structures were compared among them and also against manually solved reference structures (that were not available during the blind test).

Typical NMR protocol:
1) Resonance assignment: the chemical shifts (CS) observed in multidimensional spectra are assigned to their corresponding protein atoms.
2) NOESY assignment: thousands of dipolar-coupling effects are identified in multidimensional NOE spectra and assigned and converted into interatomic distance restraints.
3) Additional conformational restraints from RDCs, scalar couplings, CS data.
4) Structure generation: software programs are used generate a set of protein conformations (called a bundle of conformers) that should satisfy these restraints.
5) Structure refinement: the bundle of conformers is energetically refined through restrained MD simulations.

2-4 are performed iteratively in an integrated manner to maximize the number of conformational restraints obtained while guaranteeing self-consistency. We want them to converge to a self-consistent set of conformational restraints from which the final bundle of low pseudoenergy conformers is calculated.

Data: assigned CS lists and unassigned NOESY peak lists.

Three groups of automated methods:

1) Only NOESY data to obtain distance restraints for structure calculations: CYANA, UNIO, ASDP, ARIA.

2) CS data augmented by NOESY data:
- NOESY information is used to re-rank CS-based results: CS-DP-Rosetta
- Uses CS-generated structures to perform NOESY assignments and extract distance restraints: Chesire-YAPP.

3) Only CS data as experimental information: Chesire and CS-Rosetta.

They compared two of the structures with X-ray crystallography structures of close homologs.

How do we quantify convergence? Rmsd among the automatically generates structures.

How do we measure accuracy? Rmsd agains the reference structure. Also, Global Distance Test total score (GDT_TS) (thee two meaures are anticorrelated).

Results:
- Three of the NOESY-based programs automatically and consistently generated acceptable structures.
- CS-based methos augmented with NOESY data.
- Pure CS-methods had the poorest performance. Also, they are computationally more demanding.
- Geometric and stereochemical quality of the structures is not a good indicator of their accuracy, because it is mainly determined by the force field used in the refinement step. Even structures with a wrong fold can give good stereochemical results.
- DP score: quantifies the agreement between the structures and unassigned NOESY data.