NMR Spectral Assignment and Structural Calculations

Lucia Banci
CERM – University of Florence

EMBO Global Exchange Course
Suwon, Korea 19-26 June, 2016
Structure determination through NMR

Protein Sample

NMR spectroscopy

Sequential resonance assignment

Collection of conformational constraints

3D structure calculations

Structure refinement and Analysis
Which experiments should I run to assess sample quality?

- Protein overexpression
- Purification
- \(^{15}\text{N}/^{13}\text{C}\) labelling

The protein in the NMR tube!

\(<25\text{ kDa about 240 AA}\) → \(^{13}\text{C},^{15}\text{N}\) labeling

\(>25\text{ kDa about 240 AA}\) → \(^{13}\text{C},^{15}\text{N}\) labeling + \(^2\text{H}\) labeling necessary!!
Is my sample OK for NMR?

$^1$H-$^{15}$N HSQC spectra give the protein fingerprint

**unfolded**

**folded**

Signals of unfolded proteins have little $^1$H dispersion, that means the $^1$H frequencies of all residues are very similar.

Folded proteins have larger dispersion

**Can I see all the peaks I expect?**

Count the peaks! 

Backbone NH (excluding prolines!)
Making resonance assignment

What does it mean to make sequence specific resonance assignment?

To associate each resonance frequency to each atom of the individual residues of the protein.
Assignment Strategy

The strategy for assignment is based on scalar couplings.
CBCA(CO)NH and CBCANH correlate amide groups \((\text{H and }^{15}\text{N})\) with \(C\alpha\) and \(C\beta\) resonances.

- \(1\text{H}_i^{-15}\text{N}_i^{-13}\text{C}_{\alpha_i}^{-13}\text{C}_{\beta_i}\)
- \(1\text{H}_i^{-15}\text{N}_i^{-13}\text{C}_{\alpha_{i-1}}^{-13}\text{C}_{\beta_{i-1}}\)

Experiments for backbone assignment

- HNCO
- HN(CA)CO
- HN(CO)CA
- HNCA

\(1\text{H}_{(i)}^{-15}\text{N}_{(i)}^{-13}\text{CO}_{(i-1)}\)
\(1\text{H}_{(i)}^{-15}\text{N}_{(i)}^{-13}\text{CO}_{(i)}\)
\(1\text{H}_{(i)}^{-15}\text{N}_{(i)}^{-13}\text{C}_{\alpha_{(i)}}\)
\(1\text{H}_{(i)}^{-15}\text{N}_{(i)}^{-13}\text{C}_{\alpha_{(i-1)}}\)
Experiments for backbone assignment

The chemical shifts of $\text{C}_\alpha$ and $\text{C}_\beta$ atoms can be used for a preliminary identification of the amino acid type.
The 'domino pattern' is used for the sequential assignment with triple resonance spectra.

Green boxes indicate sequential connectivities from each amino acid to the preceding one.
In H(C)CH-TOCSY, magnetization coherence is transferred, through $^1J$ couplings, from a proton to its carbon atom, to the neighboring carbon atoms and finally to their protons.
H(C)CH-TOCSY experiment

F2 (ppm) $^{13}$C

F3 (ppm) $^1$H

Isolucine

$^{13}$C

$^1$H
Conformational restraints

NMR experimental data

- NOEs
- Coupling constants
- Chemical shifts
- H-bonds
- RDCs
- Relaxation times
- PCSs
- Contact shifts

Structural restraints

- Proton-proton distances
- Torsion angles
- Proton-proton distances
- Bond orientations
- Metal-nucleus distances
- Orientation in the metal $\chi$ frame
- Torsion angles
NOE is based on a relaxation process due to dipolar coupling between two nuclear spins.

NOESY volumes are proportional to the inverse of the sixth power of the interproton distance (upon vector reorientational averaging)
The NOESY experiment:

All $^1\text{H}$ within 5-6 Å from a $^1\text{H}$ can produce a cross-peak in NOESY spectra whose volume provides $^1\text{H}-^1\text{H}$ distance restraints.
How are the distance constraints obtained from NOEs intensities?

CYANA NOEs calibration

The NOESY cross-peak intensities ($V$) are converted into upper distance limits ($r$) through the relation:

$$ V = \frac{K}{r^n} $$

where $K$ is a constant and $n$ can vary from 4 to 6. $K$ constant is initially determined from NOE’s between protons at fixed distance.

$$ \log V = \log K - n \cdot \log r $$

Classes of constraints

1. Backbone $V = A/d^6$
2. Sidechain $V = B/d^4$
3. Methyl $V = C/d^4$

Distances are given as value range.

Wuthrich, K. (1986) "NMR of Proteins and Nucleic Acids"
How are the distance constraints obtained from NOEs intensities?

Xplor-NIH Calibration of NOEs

The NOESY cross-peak intensities are converted into upper distance limits

<table>
<thead>
<tr>
<th>Classes of restraints</th>
<th>Distance ranges</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Very Weak</td>
<td>0 – 20%</td>
</tr>
<tr>
<td></td>
<td>1.8–6.0 Å</td>
</tr>
<tr>
<td>2. Weak</td>
<td>20 – 50%</td>
</tr>
<tr>
<td></td>
<td>1.8–5.0 Å</td>
</tr>
<tr>
<td>3. Medium</td>
<td>50 – 80%</td>
</tr>
<tr>
<td></td>
<td>1.8–3.3 Å</td>
</tr>
<tr>
<td>4. Strong</td>
<td>80 –100%</td>
</tr>
<tr>
<td></td>
<td>1.8–2.7 Å</td>
</tr>
</tbody>
</table>

0.5 Å are added to the upper bound of distances involving methyl groups in order to correct for the larger than expected intensity of methyl crosspeaks.

Dihedral angles

Backbone dihedral angles

Sidechain dihedral angles

Amide plane

ψ

α-Carbon

Peptide bond
Not free to rotate (nearly always 180)

ω

Amide plane

α-Carbon

ψ

φ

Not free to rotate (nearly always 180)
Dihedral angle restraints

\[ 3J(HN - H\alpha) = A \cos^2(\phi - 60^\circ) + B \cos(\phi - 60^\circ) + C \]

Karplus equation

\[ J_{HNH\alpha} > 8\text{Hz} \quad \text{–} 155^\circ < \phi < -85^\circ \]  
\[ J_{HNH\alpha} < 4.5\text{Hz} \quad \text{–} 70^\circ < \phi < -30^\circ \]  
\[ 4.5\text{Hz} < J_{HNH\alpha} < 8\text{Hz} \quad \text{coil} \]

\[ \phi, \psi \quad \text{are also determining the } J_{HNC} \text{ values} \]

\[ \beta \text{ strand conformation} \]

\[ \alpha \text{ helix} \]
As chemical shifts depend on the nucleus environment, they contain structural information. Correlations between chemical shifts of C$_\alpha$, C$_\beta$, CO, H$_\alpha$ and secondary structures have been identified.

### Chemical Shift Index

CSI’s are assigned as:

- C$_\alpha$ and carbonil atoms chemical shift difference with respect to reference random coil values:
  - $-0.7 \text{ ppm} < \Delta\delta < 0.7 \text{ ppm}$ 0
  - $\Delta\delta < -0.7 \text{ ppm}$ -1
  - $\Delta\delta > +0.7 \text{ ppm}$ +1

For C$_\beta$ the protocol is the same but with opposite sign than C$_\alpha$

Any “dense” grouping of four or more “-1’s”, uninterrupted by “1’s” is assigned as a helix, while any “dense” grouping of three or more “1’s”, uninterrupted by “-1’s”, is assigned as a $\beta$-strand. Other regions are assigned as “coil”.

A “dense” grouping means at least 70% nonzero CSI’s.
H-bonds as Structural restraints

Experimental Determination of H-Bonds:

- HNCO direct method
- H/D exchange indirect method

Distance and angle restraints

- Upper distance limit
- Lower distance limit

Distance between the donor and the acceptor atoms is in the range 2.7 - 3.2 Å

140° < N-H···O < 180°
Residual dipolar couplings

RDCs provide information on the orientation of (in principle each) bond-vector with respect to the molecular frame and its alignment in the magnetic field
Proteins dissolved in liquid, orienting medium

Some media (e.g. bicelles, filamentous phage, cellulose crystallites) induce to the solute some orientational order in a magnetic field

A small “residual dipolar coupling” results

Residual dipolar couplings

\[ \text{RDC}_{(IS)} \propto \Delta \chi \cdot f(\theta_i, \phi_i) \]

where \( \chi \) is the molecular alignment tensor with respect to the magnetic field and \( \theta_i, \phi_i \) are the angles between the bond vector and the tensor axes

Relative orientation of secondary structural elements can also be determined
How complete are the NMR structural restraints?

NMR mainly determines short range structural restraints but provides a complete network over the entire molecule.
3D structure calculations

Most Common Algorithms

- MD in cartesian coordinates/Simulated annealing
  - XPLOR-NIH

- MD in torsion angle space/Simulated annealing
  - XPLOR-NIH and CYANA

A random coil polypeptide chain is generated, which is folded through MD/SA calculations and applying experimental constraints
Molecular Dynamics (MD)

How the algorithms work:

• MD calculations **numerically solve the equation of motion** to obtain trajectories for the molecular system

• In Cartesian coordinates, the Newton’s equation of motion is:

\[
m_a \frac{d^2 \mathbf{r}_a}{dt^2} = \frac{\partial}{\partial \mathbf{r}_a} U[t | \sigma(\mathbf{r}_1, \ldots, \mathbf{r}_N)],
\]

• In torsion angle space the equations of motion (Lagrange equations) are solved in a system with N torsion angles as the only degrees of freedom. Conformation of the molecule is uniquely specified by the values of all torsion angles. **About 10 times less degrees of freedom than in Cartesian space**

\[
\frac{d}{dt} \left( \frac{\partial L}{\partial \dot{q}_k} \right) - \frac{\partial L}{\partial q_k} = 0
\]

\[
L = E_{\text{kin}} - E_{\text{pot}}
q = \text{generalized coordinate}
\]
How MD is used to find the lowest energy conformation?

- The potential energy landscape of a protein is very complex and studded with many local minima where a conformation can become “trapped” during MD calculations.

- A distinctive feature of MD simulations, when compared to the straightforward minimization of an energy function, is the presence of kinetic energy that allows the protein conformations to cross barriers of the potential surface.
Simulated annealing (SA)

- MD is combined with simulated annealing protocols
- The kinetic energy (provided in terms of temperature) defines the maximal height of energy barrier that can be overcome in MD simulations
- In protein structure calculations, temperature is varied along the MD simulation so as to sample a broad conformational space of the protein and to facilitate the search of the minimum of the hybrid energy function which combines energy terms with structural constraints
Simulated annealing (SA)

- Through SA, a molecule reaches its minimum energy configuration by slow cooling it after having sampled a broad conformation range at high temperatures.
- It is a general optimization method used to search for the minimum of very complex functions.
- Elaborated SA protocols have been developed to optimize the exploration of protein conformational space (e.g., several stages of heating and cooling, switching on/off atom-atom repulsion, etc.).
Knowledge about the topology of the system is needed:

- Experimental data are supplemented with information on the **covalent structure** of the protein (bond lengths, bond angles, planar groups...) and the **atomic radii** (i.e. each atom pair cannot be closer than the sum of their atomic radii)
How the algorithms work:

A sketch of what SA does

- A starting random structure is heated to very high temperature.
- During many cooling steps, the starting structure evolves towards (i.e., folds into) the energetically favorable final structure under the influence of the force field derived from the restraints.
A hybrid energy function is defined, that incorporates *a priori* information and NMR structural restraints as potential and pseudopotential energy terms, respectively.
How is the structure calculated:

Steps:
- a random coil conformation is generated
- an MD trajectory is calculated using the hybrid energy function as the potential energy
- during MD the temperature is gradually decreased to zero
- the end point of the trajectory is (close to) the minimum of the hybrid energy function

MD calculation with restraints

Lower hybrid energy
(target/ penalty function)
## CYANA and Xplor-NIH

<table>
<thead>
<tr>
<th>Feature</th>
<th>CYANA</th>
<th>Xplor-NIH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Covalent structure</td>
<td>Fixed</td>
<td>Restrained by potential energy terms</td>
</tr>
<tr>
<td>MD in Cartesian coordinates</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>MD in Torsion Angle Space (TAD)</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>SA protocol</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Structure refinement (in explicit water)</td>
<td>No</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Not just one time

- NMR structure calculations are always performed by computing, using the same restraints and algorithm, several different conformers, each starting from different initial random coil conformations.
- In general, some of the conformers will be good solutions (i.e. exhibit small restraint violations) whereas others might be trapped in local minima.
- The usual representation of an NMR structure is thus a bundle of conformers, each of which being an equally good fit to the data.
- Conformational uncertainty may originate from true flexibility of the molecule.
The NMR solution structure of a protein is hence represented by a bundle of equivalent conformers.

The backbone of a protein structure can be displayed as a cylindrical "sausage" of variable radius, which represents the global displacements among the conformers of the protein family:

- 2987 meaningful NOE
- 158 dihedral $\psi$ and 158 dihedral $\phi$ angle constraints
- RMSD to the mean structure is $1.25 \pm 0.23$ Å for the backbone and $1.75 \pm 0.14$ Å for all heavy atoms

Structure refinement

(Restrained) Energy Minimization (EM) and MD on the bundle of conformers

The calculated conformes are then refined applying the complete force field

- **EM**: the conformation with the local energy minimum is obtained. It will only locate the closest minimum. Cannot cross energy barriers
- **MD**: the conformational space is sampled through internal motions which depend on the potential generated by the atoms in the molecule and the kinetic energy, defined by the temperature.
- **(R)EM/(R)MD**: in addition to the classical force field, the structural restraints are also applied
- Performed in vacuum and/or in explicit solvent (water)
Analysis of the results

• How many conformers should be used to represent the solution structure?

  Around 10% (10-20) of calculated structures. It should be a number that is a reasonable compromise between statistics significance and data size with respect to their manageability in graphics and analysis programs.

• How should they be selected from the ensemble of conformers?

  The conformers with the lowest target/penalty function, i.e. with the best agreement with the experimental structural restraints are selected.

Accuracy of the Structure
For two sets of \( n \) atoms, RMSD is defined as the normalized sum of the root mean square deviations of the position of a given atom with that of the same atom in the second set (after superimposition of the structures of the bundle):

\[
RMSD = \sqrt{\frac{\sum (r_{ai} - r_{bi})^2}{n}}
\]

- two identical structures will have an rmsd of 0Å
- larger is the rmsd and more dissimilar are the structures
**Precision versus Accuracy**

- Precise, not accurate
- Accurate, not precise
- Precise and accurate
- Not accurate and not precise
Validation criteria

Protein Structures are assessed with respect to:

• Back-calculations of the experimental restraints

• Local geometry:
  – Bond lengths, bond angles, chirality, omega angles, side chain planarity

• Overall quality:
  – Ramachandran plot, rotameric states, packing quality, backbone conformation, side-chain planarity

• Others:
  – Inter-atomic bumps, buried hydrogen-bonds, electrostatics, packing quality
Validation of the NMR Structures

The most common programs used to evaluate the quality of the structures are:

• WHATIF (swift.cmbi.ru.nl)
• QUEEN
• CiNG (http://nmr.cmbi.ru.nl/icing) (WHATIF and PROCHECK-NMR)
• PSVS (http://psvs-1_4-dev.nesg.org/) (PROCHECK-NMR, MolProbity, Verify3D, Prosa II)

Structural Parameters

Ramachandran Plot

Phi and Psi angles

- Generously allowed
- Disallowed

Ideally, over 90% of the residues should be in the "core" regions
Thank you
Automated Spectral Assignment and Structure determination
Automated assignment programs

MARS

Used for automated backbone assignment (NH, CO, Cα, Cβ).
It requires manually pick-peaking of 3D spectra for backbone assignment, such as CBCANH, CBCACONH etc.

Input:

• Primary sequence
• Spectral data, i.e. chemical shifts of resonances grouped per residue and those of its preceding residue.
• Chemical shift tolerances
• Secondary structure prediction data (PSI-PRED)
AutoAssign

For automated backbone assignment (NH, CO, Cα, Cβ, Hβ and Hα). It requires manually pick-peaking of 3D spectra for backbone assignment, such as CBCANH, CBCACONH etc.

Input:

- peak list table of triple resonance spectra
- primary sequence
UNIO for protein structure determination

UNIO protocol operates directly on the NMR spectra.

http://perso.ens-lyon.fr/torsten.herrmann/Herrmann/Software.html

UNIO – Computational suite for fully/highly Automated NMR protein structure determination

- UNIO provides accurate and automated 3D protein structure determination.
- UNIO enables protein NMR structure determination within one week including the collection of NMR experiments.

UNIO for protein structure determination

<table>
<thead>
<tr>
<th>UNIO standard protocol</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Amino acid sequence of the protein</strong></td>
</tr>
<tr>
<td><strong>MATCH backbone assignment</strong></td>
</tr>
<tr>
<td><strong>Input</strong>: 4D and 5D APSY spectra or triple resonance spectra</td>
</tr>
<tr>
<td><strong>Output</strong>: backbone chemical shifts</td>
</tr>
<tr>
<td><strong>ATNOS/ASCAN side chain assignment</strong></td>
</tr>
<tr>
<td><strong>Input</strong>: 3D NOESY spectra</td>
</tr>
<tr>
<td><strong>Output</strong>: side-chain chemical shifts</td>
</tr>
<tr>
<td><strong>ATNOS/CANDID NOE assignment</strong></td>
</tr>
<tr>
<td><strong>Input</strong>: 3D NOESY spectra</td>
</tr>
<tr>
<td><strong>Output</strong>: assigned 3D NOESY peak lists and 3D protein structure with external program (XPLOR, CYANA, CNS etc)</td>
</tr>
</tbody>
</table>
Criteria for NOE assignment

for each cross-peak the initial possible assignments are weighted with respect to several criteria, and initial assignments with low overall score are then discarded.

- **Chemical shift agreement**
- **NOEs network-anchoring**
- **Compatibility with intermediate structure**

Automated NMR structure determination

- Automated NOESY spectral analysis using ATNOS-CANDID/CYANA
- The automated ATNOSCANDID algorithm assembled in UNIO proceeds in iterative cycles of ambiguous NOE assignment followed by structure calculation using torsion angle dynamics.

T. Herrmann, K. Wüthrich, and F. Fiorito
Does it always work??

Automatic

Manual

atx- like domain of hCCS protein (70 aa)

fHbp (274 aa)
CS ROSETTA generates 3D models of proteins, using only the $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$, $^{13}\text{C}'$, $^{15}\text{N}$, $^1\text{H}_\alpha$ and $^1\text{HN}$ NMR chemical shifts as input.

CS-ROSETTA involves two separate stages:

1. Polypeptide fragments are selected from a protein structural database, based on the combined use of $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$, $^{13}\text{C}'$, $^{15}\text{N}$, $^1\text{H}_\alpha$, and $^1\text{HN}$ chemical shifts and the amino acid sequence pattern.

2. These fragments are used for generate a structural model, using the standard ROSETTA Monte Carlo assembly and relaxation methods.