Computation of solution structures of peptides and proteins from NMR data

Orlando Crescenzi

Dept. of Chemistry, University Federico II of Naples

From the perspective of computational chemistrty

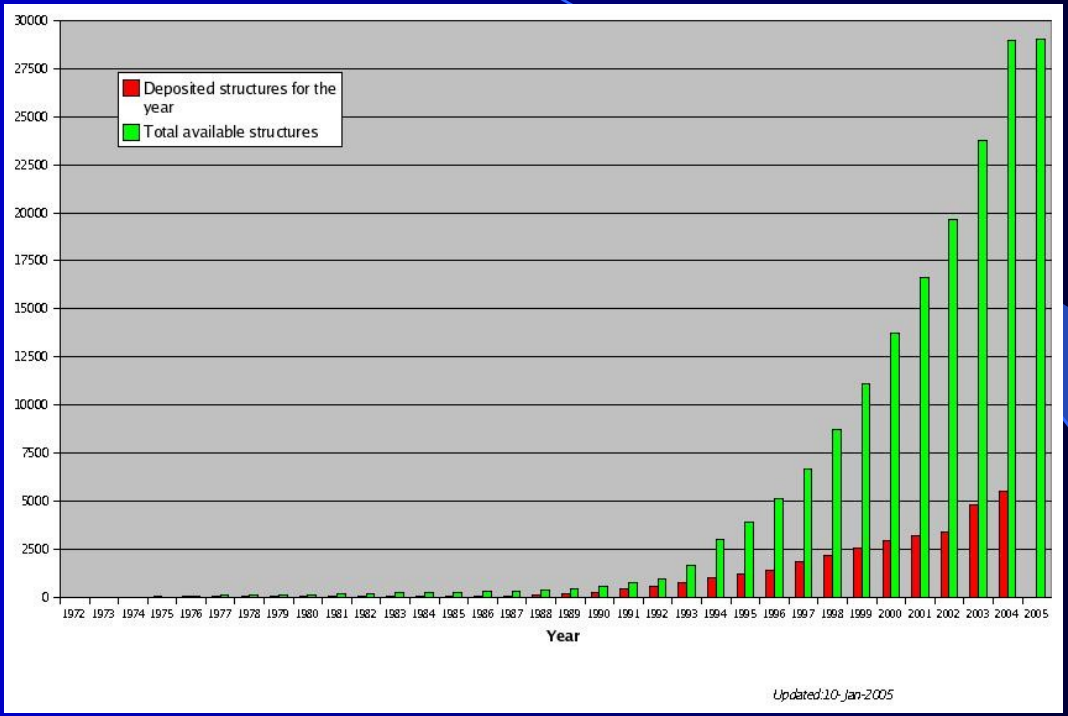
- Protocols for NMR-based structure determination represent a typical overlap area between spectrocopy and computational approaches
- A feeling of the reliability of the computed structures can only be acquired if one knows how they are computed
- A knowledge of the current procedures represents a background to discuss the additional information that can be gained by more elaborate computational methods (e.g. ab initio computation of NMR parameters)

Which kind of molecules ?

- Moleculed of known connectivity
- With a well-defined three-dimensional "structure" in solution (i.e. limited conformational flexibility)
- With a sufficiently dense network of protons
- With a limited size and a reasonably isotropic tumbling behaviour

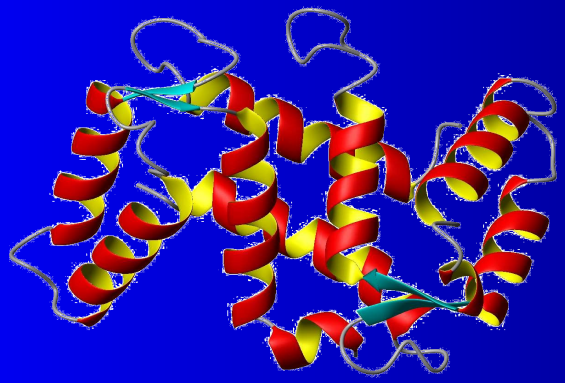
- Apart from proteins and well-structured peptides, also nucleic acids and possibly polysaccharides

The present of structural biology:
protein data bank holdings

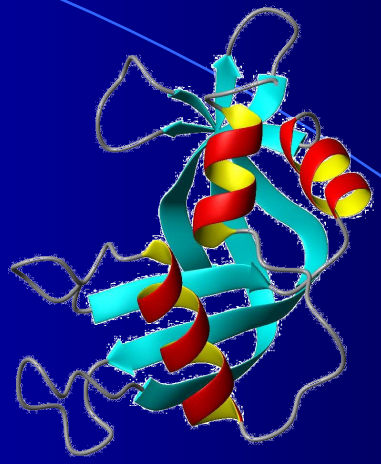


	Proteins, Peptides, and Viruses	Protein/Nucleic Acid Complexes	Nucleic Acids	Carbohydrates	Total
X-ray diffraction and other	23183	1122	770	11	25086
NMR	3599	105	637	2	4343
Total	26782	1227	1407	13	29429

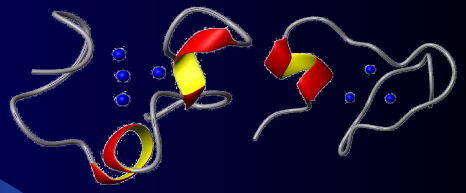
Examples of NMR protein structures



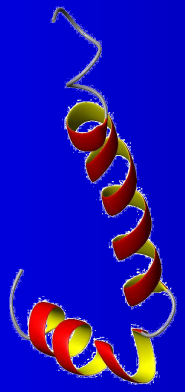
Calcyclin dimer
Calcium-binding EF hand



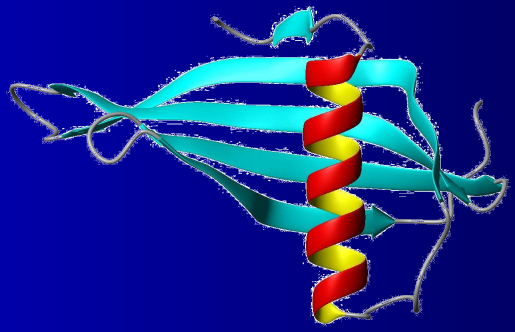
BS-Rnase monomer
Ribonuclease with special activities



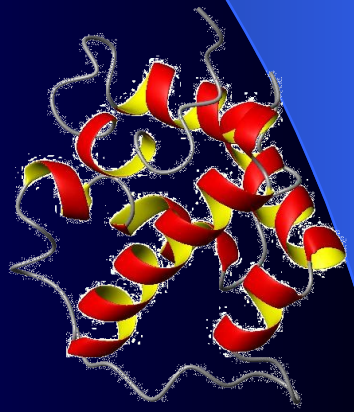
Fish metallothionein domains
Divalent ions metabolism



Aβ 1-42
Amyloid peptide



Monellin
Plant sweet protein

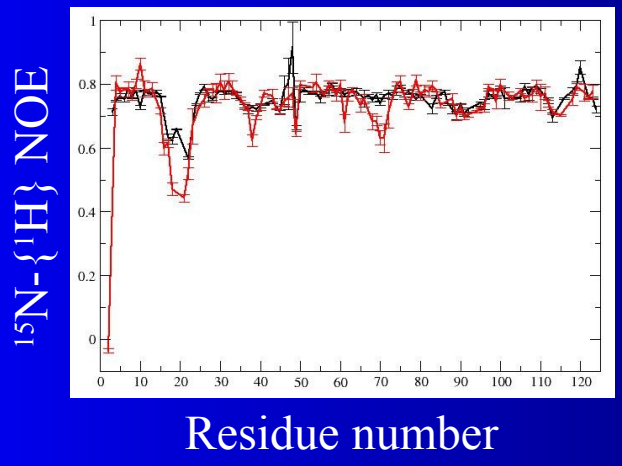


Csp
Insect chemosensory protein

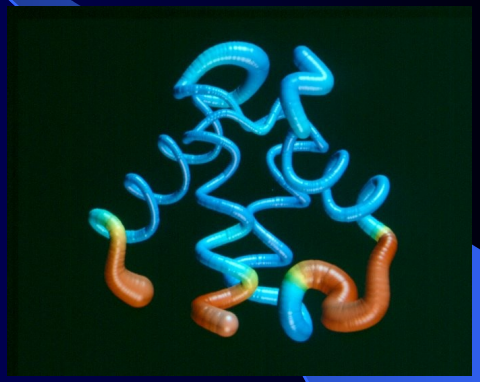
NMR versus X-ray diffraction for protein structure determination

- X-rays:
  - Uses a much larger number of experimental data; hence a correspondingly higher structural quality
  - Provided a suitable crystal can be obtained, the procedure for X-ray structure determination is faster
  - Even very large systems can be studied by X-ray diffraction
- NMR:
  - The protein is studied in solution (or in micelles, membranes...), i.e. closer to the biological environment
  - Structural distorsions related to crystal contacts are absent (especially important for floppy molecules like peptides)
  - Unfolded and partially folded states can be examined
  - Sensitive to dynamical processes on a variety of time scales

NMR to probe dynamical phenomena



Experimental dynamical informations (e.g. heteronuclear NOEs) only report on selected time-scales



- Experimental structures are static snapshots of highly dynamic molecular systems
- Biological processes (recognition, interaction, folding) require molecular motions (from femtoseconds to seconds): active development in this area
- Some preliminary knowledge of the system dynamics is also useful to optimize experimental and computational strategies

Size limitation in NMR

- For a relatively rigid system, linewidths increase with the rotational correlation time τ_c, i.e. with system size; with standard approaches, a practical limit is reached at around 30 kDa
- Recent developments (TROSY, partial deuteration) are extending this limit: however, many biochemical systems are way too large for NMR structure determination
- In favourable cases, the problem can factorized by analyzing independent domains
- NMR can provide important informations even in the absence of a complete structure determination: e.g. with chemical shift assignment, mapping of binding sites, relaxation studies, etc.

NMR parameters and molecular structure

- The magnetic hamiltonian which undelies the NMR spectrum:

$$\hat{H} = \sum_C \hbar\gamma_C \boldsymbol{B_0}\cdot(\boldsymbol{1}-\sigma_C)\cdot\hat{\boldsymbol{I}}_C + \frac{\hbar^2}{2}\sum_C\sum_{D\neq C}\gamma_C\gamma_D\hat{\boldsymbol{I}}_C\cdot(\boldsymbol{D}_{CD}+\boldsymbol{J}_{CD})\cdot\hat{\boldsymbol{I}}_D$$

B_0: external field
I_C: nuclear spins
σ_C: nuclear shielding tensors
D_{CD}: direct dipolar interaction between spins I_C and I_D
J_{CD}: scalar coupling between spins I_C and I_D

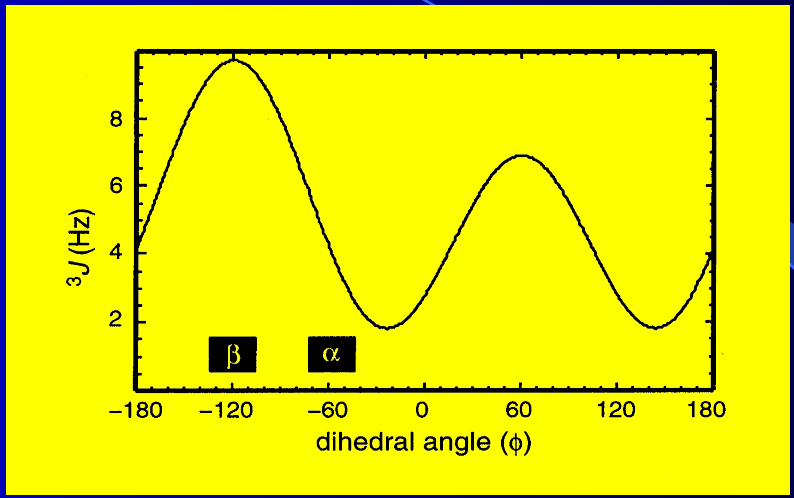
- From NMR experiments, we extract (averaged) parameters related to the tensorial quantities in the equation; these convey the indirect structural information we use;
- Conversely, is we know the molecular structure, the tensors can be estimated by suitable (quantum mechanical) methods.

NMR parameters useful for structure determination: the traditional set...

- ^{1}H-^{1}H NOEs: primary source of structural information
  - particularly direct, albeit approximate, dependence on structure: NOE ~ r_{HH}^{-6}
  - strongly non linear, therefore flexibility has a large impact
  - NOEs provide a large number of constraints
  - long-range contacts strongly restrain the folding of the chain; conversely, many short-range NOEs are irrelevant

- *J*-couplings: an additional source of structural information
  - provide a small set of restraints
  - very local in character (e.g. ^{3}J most affected by internuclear dihedrals)
  - dependence on structural parameters described by approximate (Karplus type) relationships
  - sometimes ambiguous (i.e. more than one dihedral can give rise to the same ^{3}J)

- E.g. $^{3}J_{\mathrm{HN\text{-}H\alpha}}$



- Amide H/D exchange rates, temperature coefficients $\Delta\delta/\Delta T$
  - help identify amide groups involved in strong, stable H-bonds
  - H-bond acceptor must be determined from other evidence

and some new entries

- Residual dipolar couplings
  - arise when molecular tumbling is slightly anisotropic; partial alignment is either spontaneous (at very high field), or can be induced by appropriate techniques (bicelles, strained gels)
  - provide an indication of the orientation of NH bonds w.r.t. a common frame: this allows for the relative positioning of secondary structure elements

- Trans H-bond scalar couplings
  - in favourable cases, measurable scalar couplings through a H-bond: $^{3h}J_{NC'}$ (^{15}N–H...O=$^{13}C'$)
  - ^{15}N,^{13}C labelling required
  - coupling is small (< ~ 1 Hz), and decreases rapidly with H-bond length
  - both H-bond partners identified → direct information on secondary structure

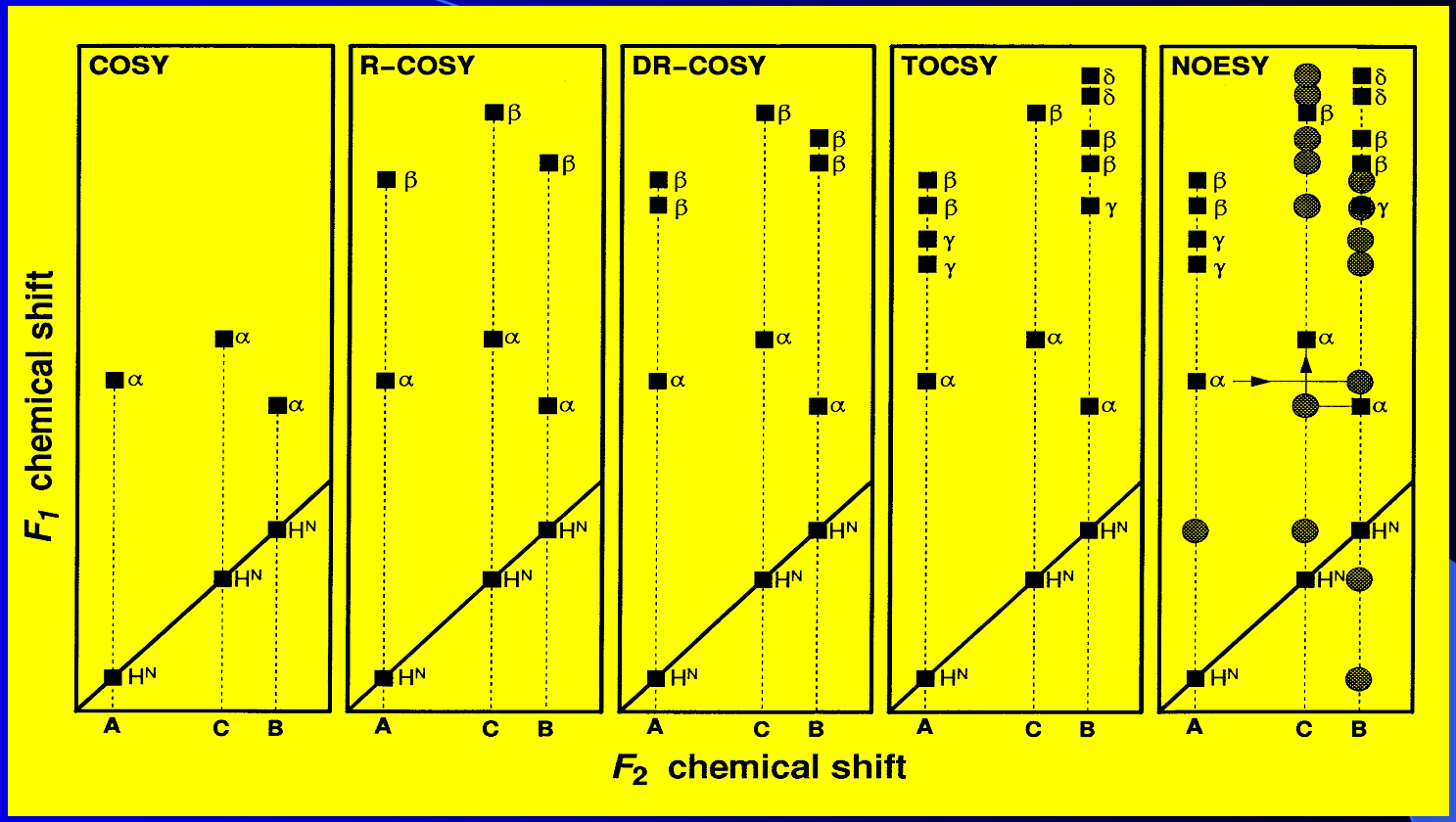
- Chemical shifts
  - assigned at the very beginning
  - traditionally only interpreted to produce a chemical shift index
  - potentially, lots of structural information, but...
  - dependence on structure very complex: a sophisticated model is required (QM)

Some experimental aspects

- Protein spectra are collected in H_2O
  - otherwise NH protons would be lost !
  - just enough D_2O to allow for locking (on HDO signal)
  - can add buffers, salts, and even use mixed solvents etc.
  - huge water signal can be suppressed by a variety of pulse schemes
- A D_2O sample is also useful
  - analysis on non-exchangeable protons is simplified
  - better quality spectra, since water suppression is not needed
- Sample must be pure, and of relatively high concentration
  - NMR is intrinsically insensitive
  - 0.5 – 2 mM concentrations are typical
  - for lower concentrations, cryoprobes
- Sample must be stable for weeks, and not prone to aggregation
  - Can check for aggregation by NMR diffusion measurements

The homonuclear approach

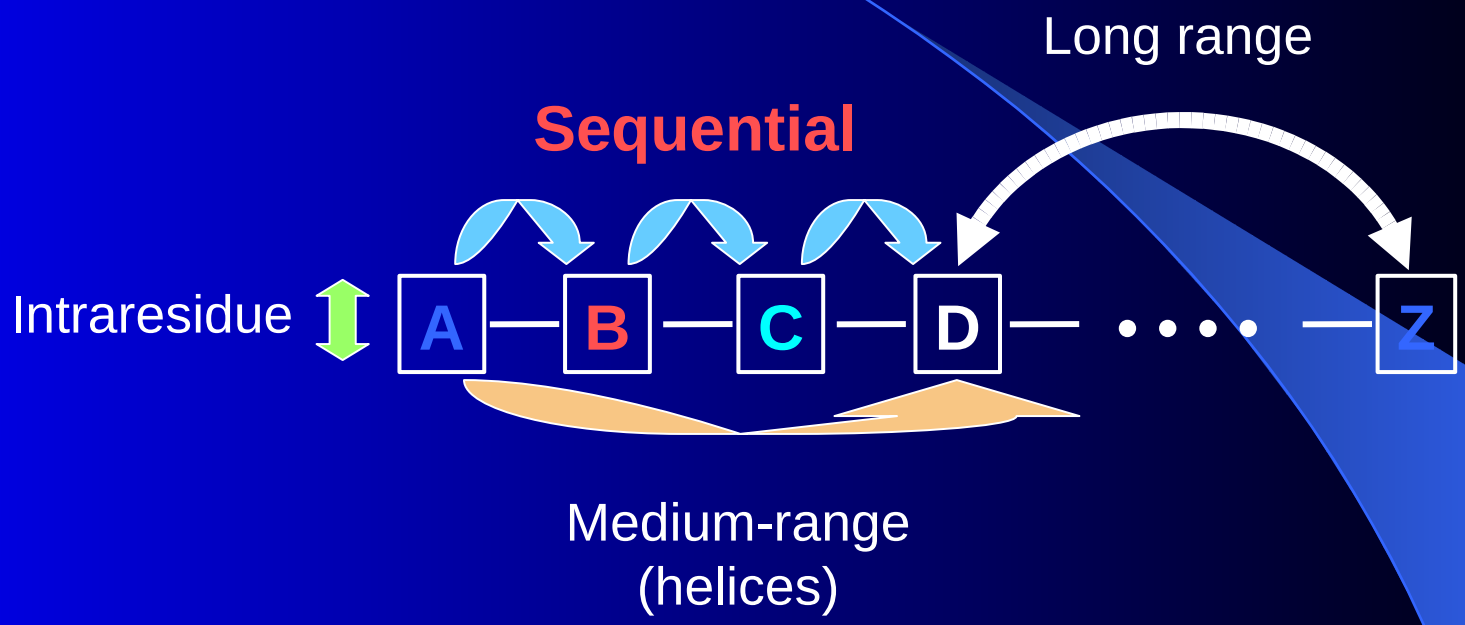
- Only suitable for small (< 10 kDa) proteins and peptides
- An unlabelled sample is used
  - still widely used for peptides (chemical synthesis of labelled peptides is very expensive)
  - if protein is expressed, can as well go directly to ^{15}N labelling
- A standard set of 2D spectra is acquired
  - COSY and TOCSY to identify the spin systems (sets of protons with resolved scalar couplings), starting from the amide protons
  - spin systems are characteristic for aminoacid classes
  - ^{1}H-^{1}H couplings do not extend from one residue to another: hence...
  - the required additional information is extracted from the NOESY spectrum
  - depending on the secondary structures, characteristic patterns of interresidue contacts are observed: e.g. H^{N}_{i}- H^{N}_{i+1}; H^{α}_{i}- H^{N}_{i+1}; H^{β}_{i}- H^{N}_{i+1}; etc.
  - this allows to juxtapose the spin systems (sequence specific assignment)



place into sequence

identify spin systems

- At this stage, only a subset of NOEs is examined
  - intraresidue, to complete spin system identification
  - sequential and characteristic medium range, to juxtapose spin systems



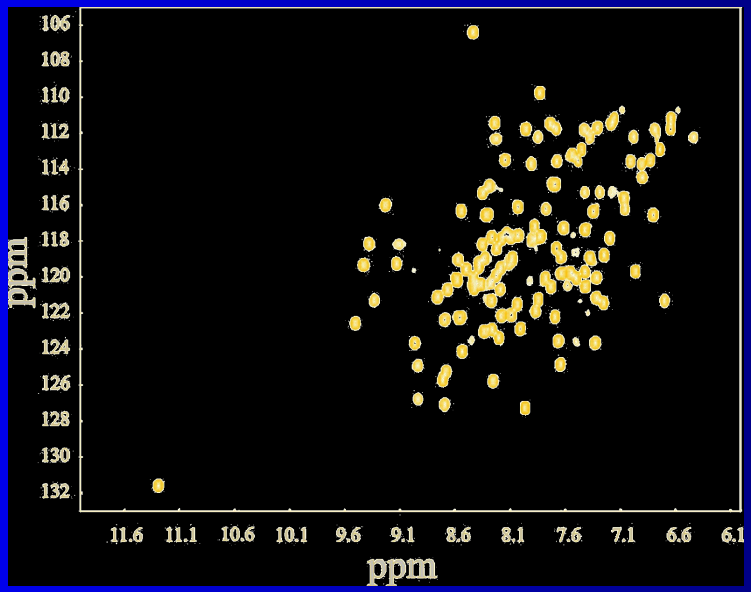
- At the end of this phase, assignment of proton chemical shifts is complete
  - analysis of the NOESY spectrum is still partial, but secondary structure information is already available

The heteronuclear approach

- Labelled samples are required: either uniformly labelled with ^{15}N...
  - reasonably cheap
  - NMR experiments are simple
  - can get along with this for medium size proteins
- Or with both ^{15}N and ^{13}C
  - much more expensive
  - pulse sequences are tricky
  - necessary for large proteins
  - allows for identification of more constraints, i.e. better final structures
- Labelling is achieved by expression in microorganisms grown on labelled nutrients
  - *E. coli* by far the most common choice
  - yests and insect cells also possible
  - in perspective, maybe cell free expression

Advantages from labelling

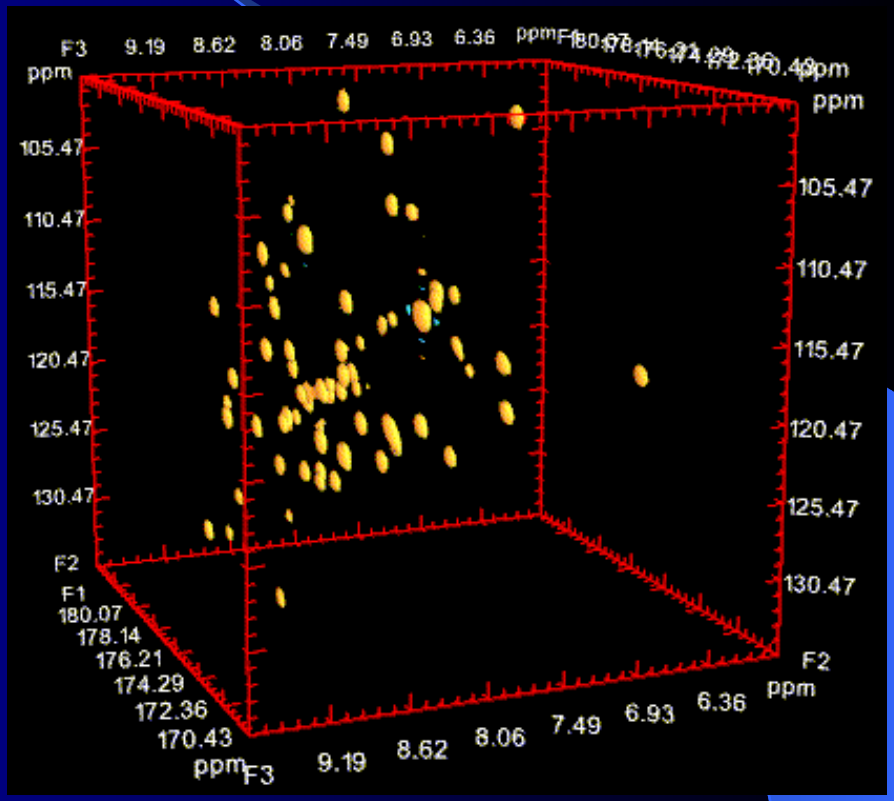
- Both in singly and in doubly labelled samples: overlapping proton resonances resolved by the large dispersion of heteronuclei



^{15}N shifts

^{1}H shifts

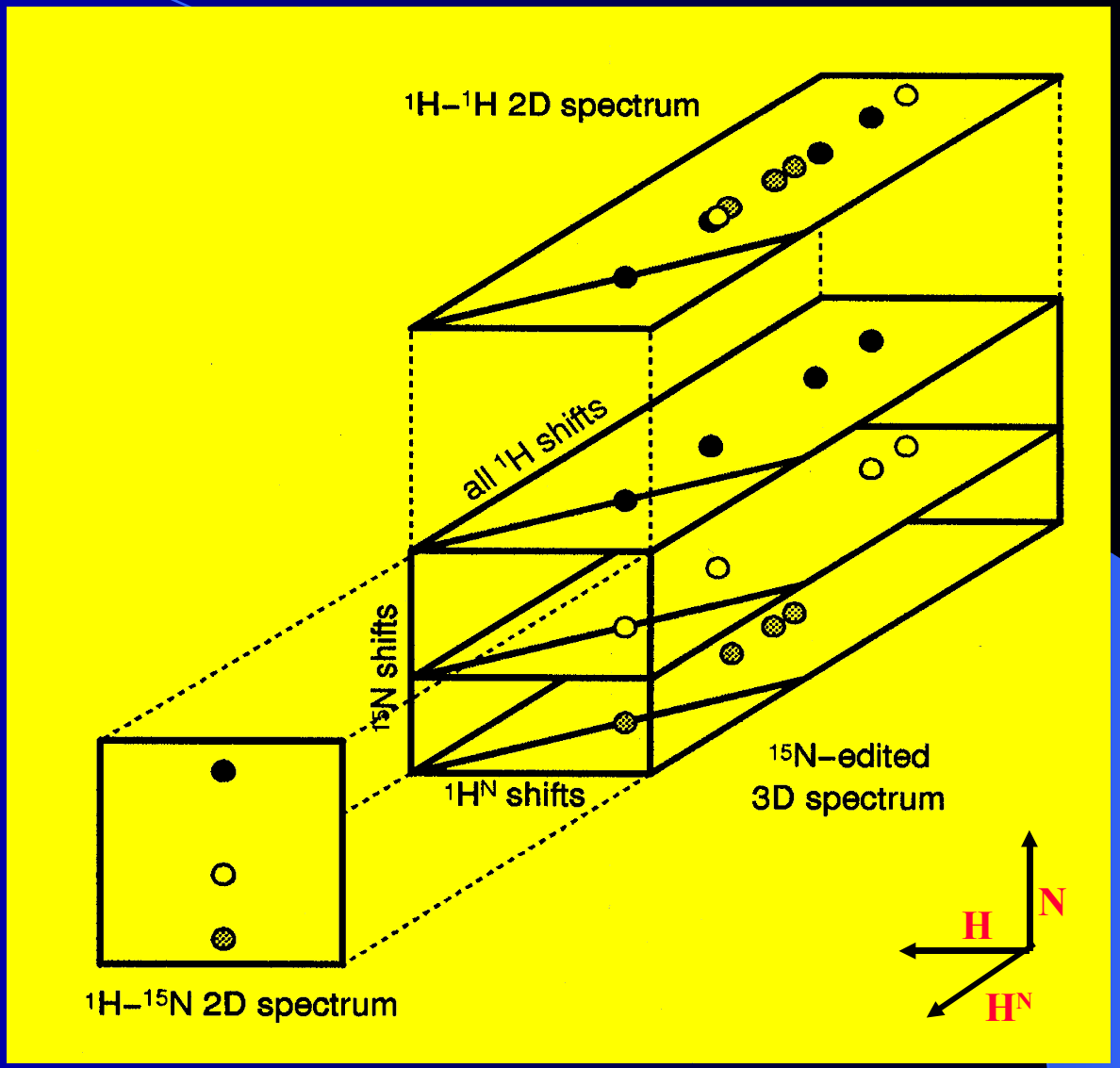
HSQC of odorant binding protein: a double-resonance 2D spectrum



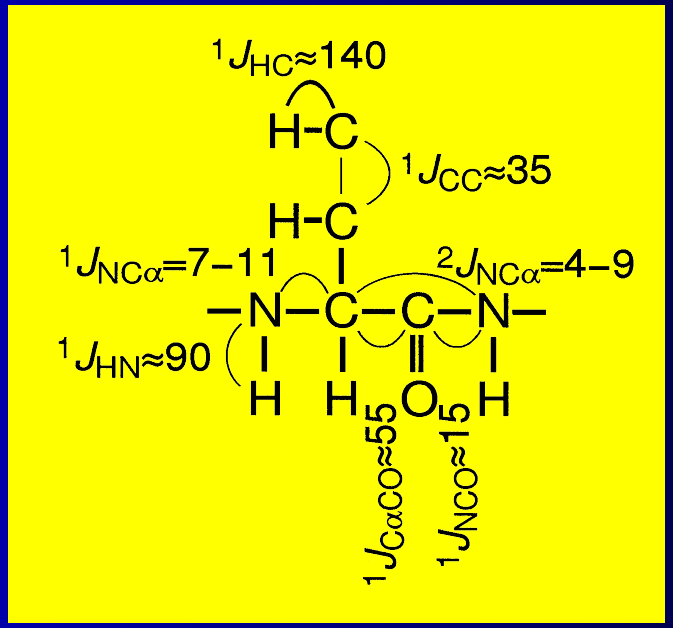
HNCO of ubiquitin: a triple-resonance 3D spectrum

TOCSY HSQC

^{1}H ^{1}H ^{15}N



- In doubly labelled samples: heteronuclear *J*-couplings extend across residues
  - no more NOESY-based sequential assignments
  - protocols rely exclusively on scalar correlations
  - much faster, more complete, even fully automated chemical shift assignments
  - heteronuclear *J*-couplings also encode structural information

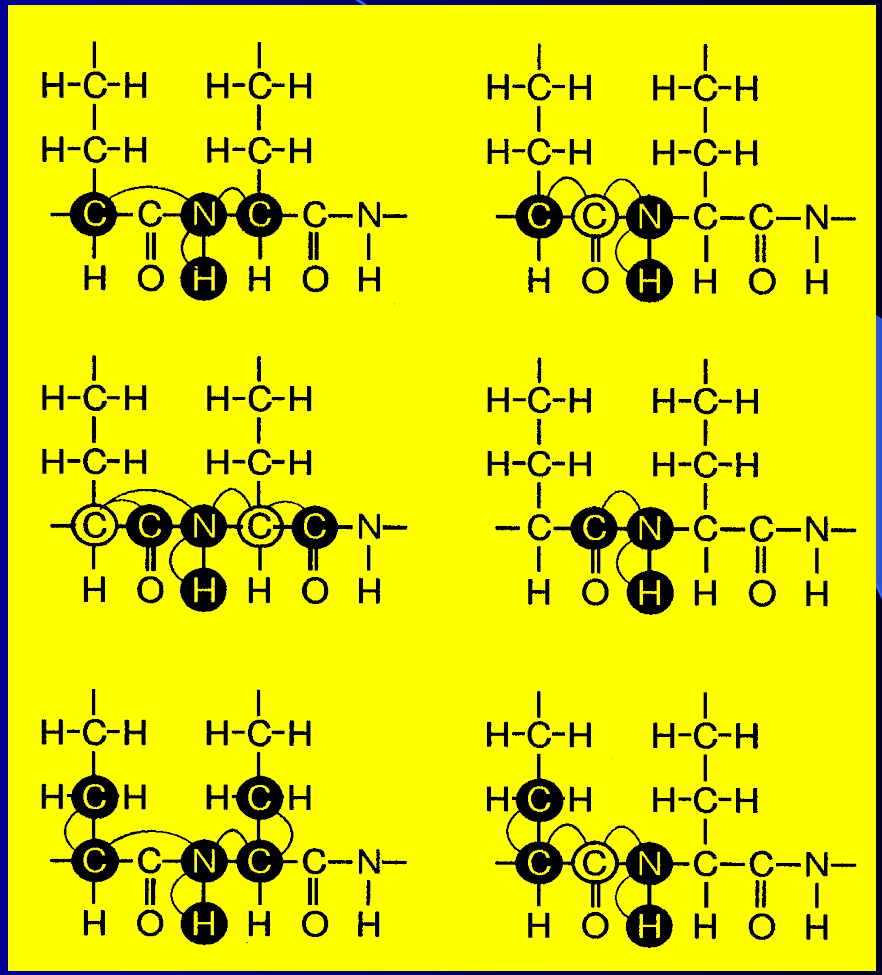


- Lots of new experiments, both for C/N assignment...

HNCA / HN(CO)CA

HN(CA)CO / HNCO

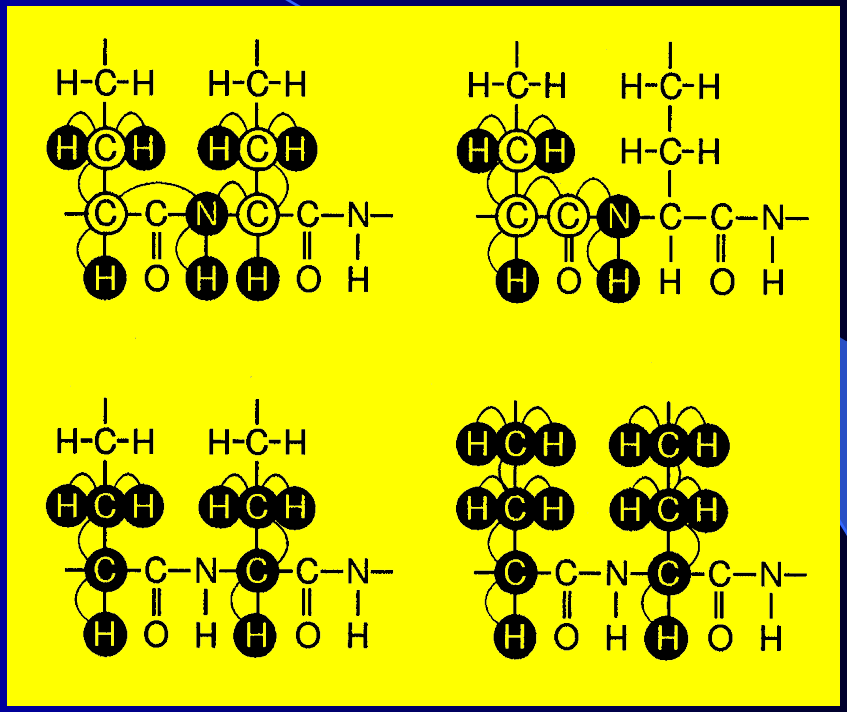
HNCACB / HN(CO)CACB



- And for proton assignment

HBHA(CBCA)NH /
HBHA(CBCACO)NH

HCCH-COSY /
HCCH-TOCSY



- At the end of this phase, assignment of proton, nitrogen and carbon chemical shifts is complete
  - NOESY spectrum not yet analyzed

NOE assignment

- Assuming that chemical shift assignment has been achieved
  - chemical shift assignment should be essentially complete
  - diastereotopic pairs (e.g. side chain HB2/HB3) usually not (yet) assigned

- Next stage is based on NOEs
  - prepare peak lists (peak picking)
  - can be done automatically, however usually need to check
  - one list for each NOESY experiment (different mixing times, heteronuclear editing, H_2O versus D_2O...)
  - integrate NOEs (or else use intensities)

- Assignment criteria
  - for each NOESY cross peak: search chemical shift lists for matches in all dimensions (within some tolerance)
  - in 2D NOESY: no restriction on proton assignments
    ✂ → most NOEs are ambiguous at this stage
  - in edited experiments: definite relationship between protons and corresponding heteronuclear dimensions (e.g. 3D: $H_i - H^N_j - N_j$; 4D: $C_i - H_i - H^N_j - N_j$)
    ✂ → many more unambiguous NOEs

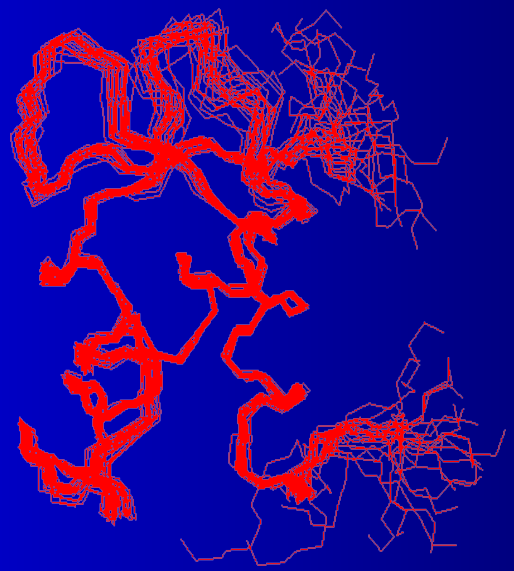
Prepare restraints for structure calculation

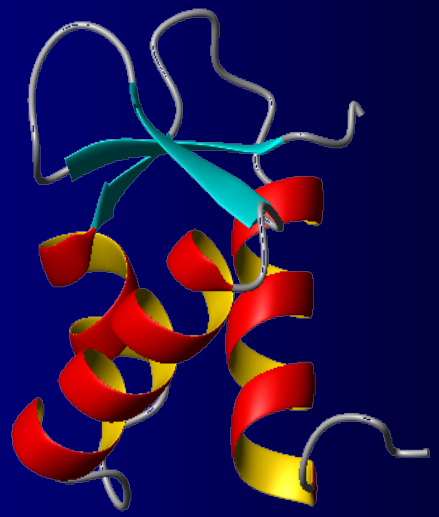
- Calibrate NOEs
  - separate calibration for each individual peak list
  - can use NOE = A * r_{HH}^{-6} dependence for all types
  - or else individual empirical dependences (NOE = B * $r_{HH}^{-\gamma}$, $\gamma = 4 - 6$) for different classes, like backbone, side-chain, methyls
  - free parameters (A, B...) are fitted to bring a subset of distance restraints to the statistically expected value
- These distances are applied as upper bounds (upl)
  - many factors can bleach a NOE contact
  - "anti-NOEs" can be checked individually and applied in specific cases
- Correct for non-stereospecifically assigned diastereotopic pairs
  - several technicalities
  - in essence, modify / loosen the constraints to allow for both possible assignments

Structure calculation

- At present, almost only simulated annealing protocols
  - potential energy from force field complemented with restraint violation penalties to produce an overall "potential energy"
  - high temperature molecular dynamics to sample large portions of PES
  - slow cooling to end up in favourable minima

- Repeat many times (e.g. 50 – 200 runs)...
  - and select final structures with lowest "potential energy"
  - ensemble of structures compatible with experimental restraints → NMR bundle





Simplified force fields

- Conformational freedom mostly related to dihedral angles
  - get rid of fast stretching and bending motions, and increase time step
  - once algorithmic problems have been solved, torsion angle dynamics (TAD) much faster than traditional MD

- In a sense, NMR structures are actually models
  - bond lengths and valence angles are fized at their reference values
  - even after refinement, no direct experimental indication on these parameters

Once you have a first bundle, use it to improve NOE assignments...

- Assignment criteria
  - same criteria as in first run, and in addition...
  - assignments should correspond to short interproton distances

  - e.g. for a given ambiguous cross peak of a 2D NOESY
    - with possible assignments in D1: 12.HN, 18.HN
    - and possible assignments in D2: 48.HA, 64.HA

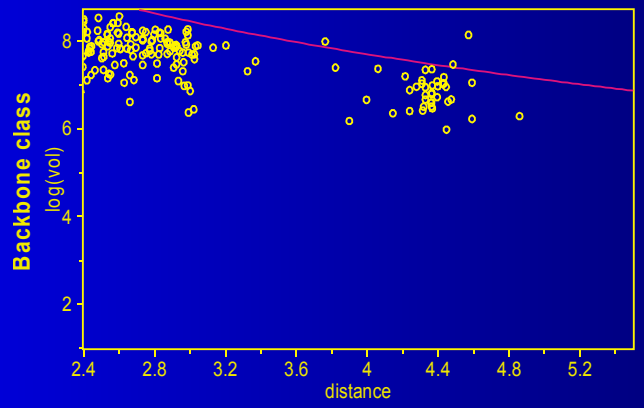
    - check all distances in the NMR bundle:

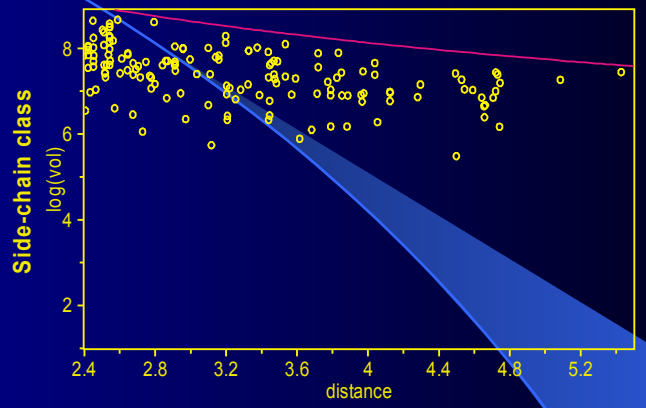
12.HN – 48.HA	2.6 +/- 0.2 Å
12.HN – 64.HA	6.9 +/- 0.6 Å
18.HN – 48.HA	7.7 +/- 0.8 Å
18.HN – 64.HA	12.2 +/- 1.1 Å

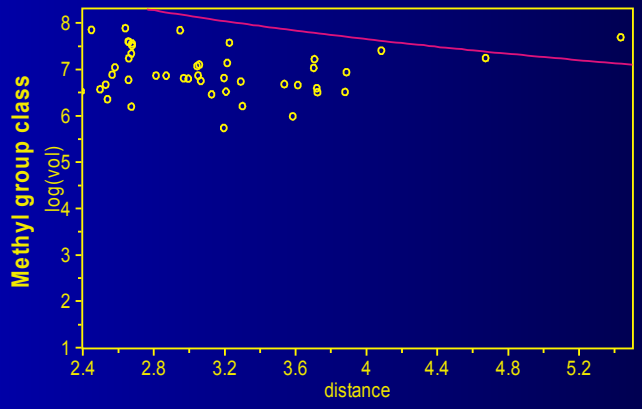
    - since only one assignment corresponds to a (reproducibly) short distance, assign as 12.HN – 48.HA

and NOE calibration...

- Plot current upper bounds versus average distance in the bundle

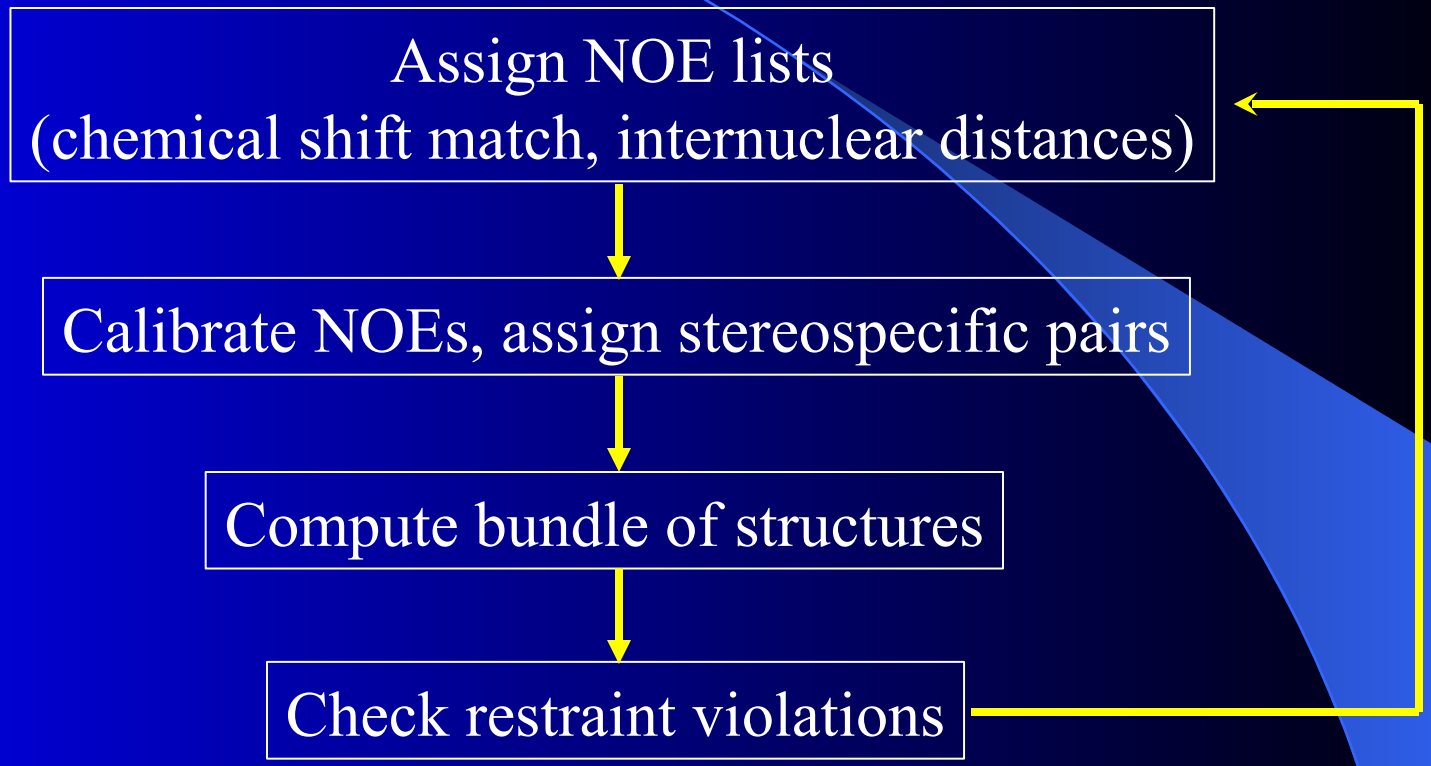






- If necessary, change calibration parameters for optimal match
- Improve diastereospecific assignments (check both possibilities)

up to convergence



- Convergence is reached when no more NOEs can be assigned

Structure refinement

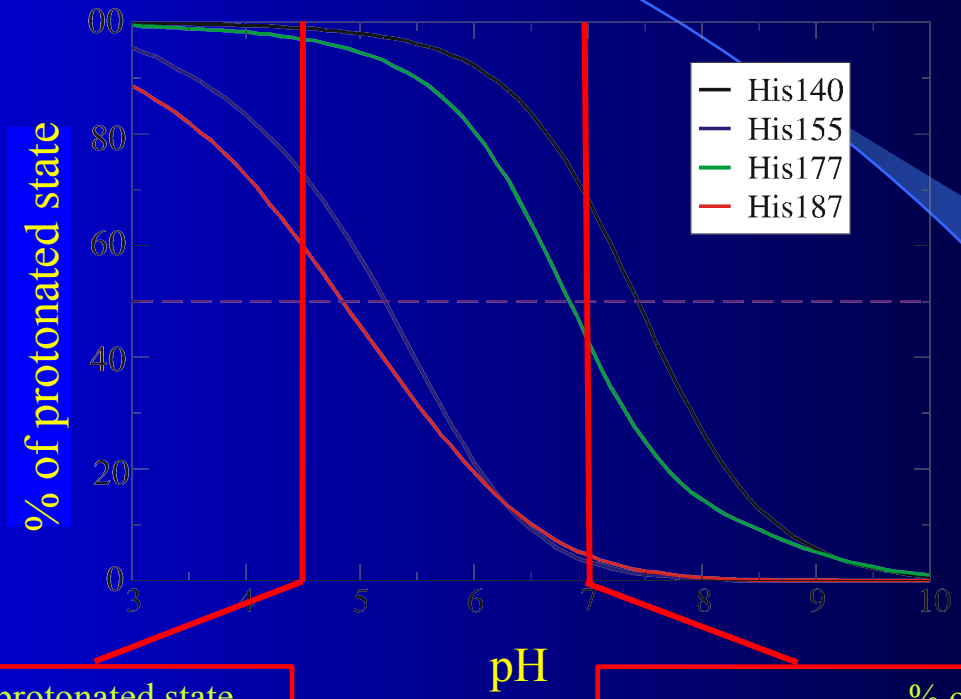
- Use a more realistic biomolecular force field (no more fixed bond angles; maybe bond distances still frozen)
- Lower relative weight of restraint violation term
- Take into account the solvent
  - [old way: distance-dependent dielectric constant]
  - implicit solvent model
  - box of explicit waters
- Perform energy minimization, or better molecular dynamics (necessary with explicit solvent), on all bundle models
- Check for restraint violations (and if necessary go back to TAD....)

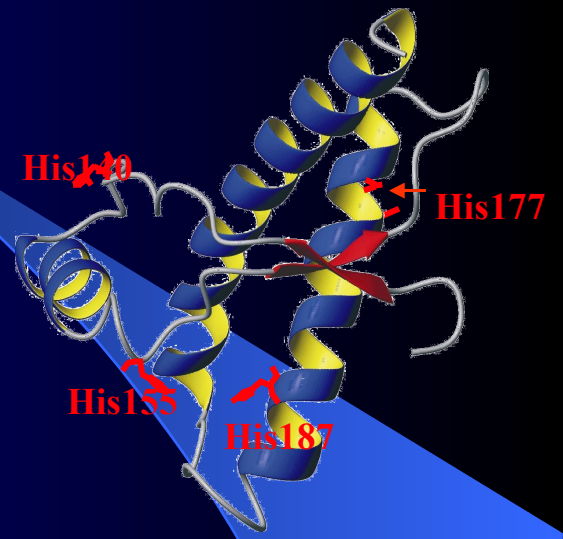
At this stage, side chain protonations states need to be defined

- Can guess based on random coil pK_a's and solution pH...
- But pK_a's in proteins are often quite unusual
- A better guess from protein electrostatics

Theoretical estimation of histidine pK_a

Solving the Poisson-Boltzmann equation for protein electrostatics





	% of protonated state mean values at pH 4.5
His140	99 ± 2
His155	73 ± 26
His177	97 ± 2
His187	60 ± 25

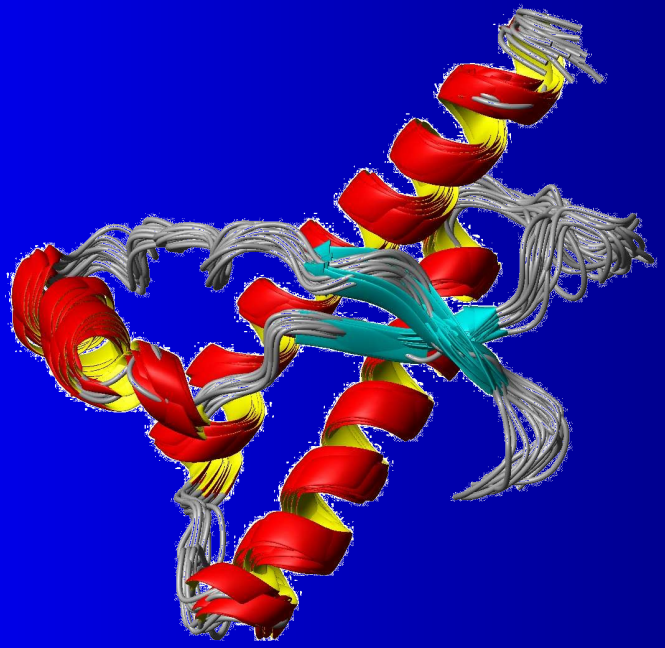
His140 and His177 are protonated

	% of protonated state mean values at pH 7.0
His140	68 ± 19
His155	3 ± 0.3
His177	43 ± 22
His187	5 ± 0.7

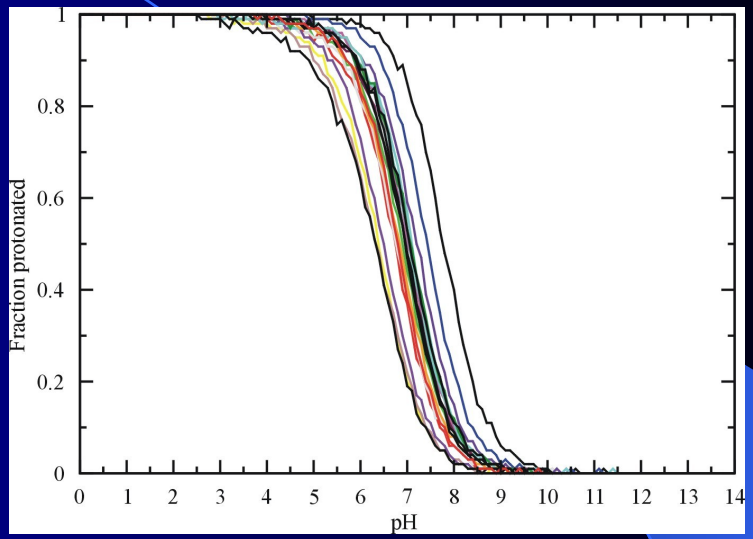
His155 and His187 are neutral

Taking into account dynamics

Averages over the whole bundle of NMR structures



Bundle of 20 NMR structures

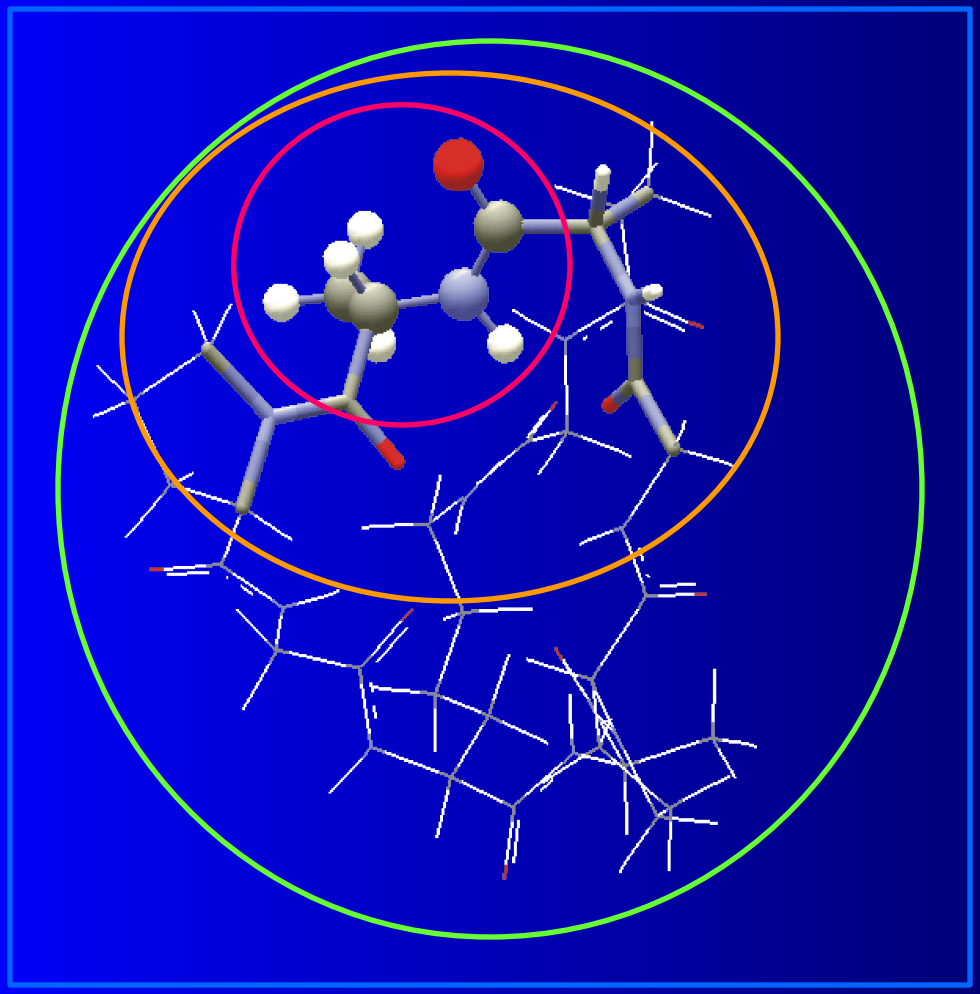


Computed His177 titration curves (for each individual structure)

Averaging on experimental structures and/or on MD frames can be used in the computation of other parameters as well

Ab-initio chemical shift calculation:

All H^N protons of a peptide



BCP2, Calcium-binding bicyclic peptide

Three-layer model,
PCM to describe bulk solvent

The innermost layer:
PBE0 / 6-311+G(2d,2p)

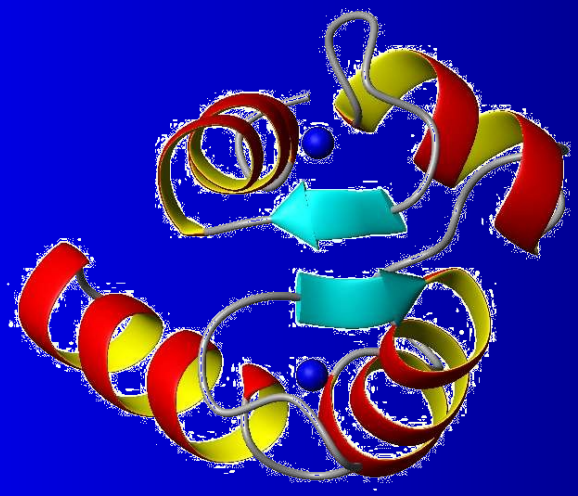
The second layer:
HF / 6-31G(d)

The rest of the molecule:
point charges (Amber)

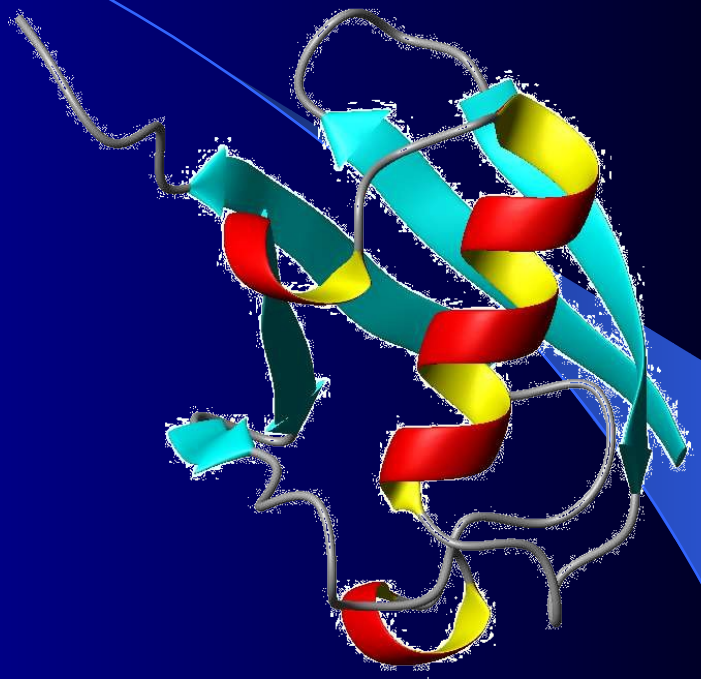
Bulk solvent:
PCM

Our current targets:

All H^N, N, C^α, C' of a protein



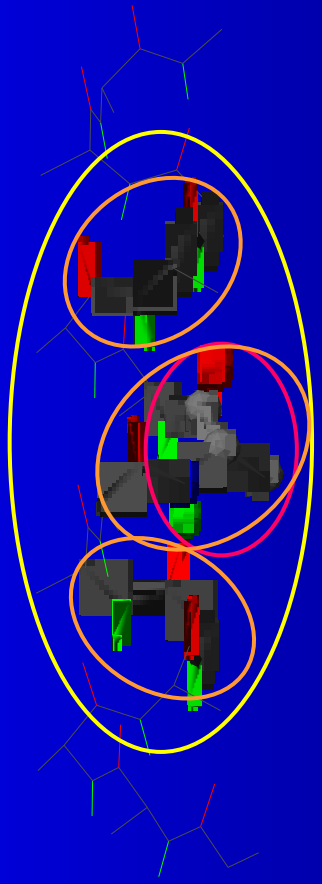
Calbindin D9k
Calcium-binding EF hand



Ubiquitin
Protein degradation signal

Main issues: proper partitioning scheme, long-range effects

^{1}H, ^{13}C and ^{15}N isotropic chemical shifts: alanine tetradecapeptide analog



Octapeptide analog: HF/6-31G*

Dipeptide analog and H-bonded residues: PBE0/6-311+G(d,p)

Computed chemical shifts: PBE0/6-311+G(d,p)

Foreseeable development areas for structural biology

Computation of spectroscopic parameters

↓

Ab-initio methods required
(mixed QM/QM/MM + environment)

↓

Other parameters
(e.g. chemical shifts)

↓

Structure validation,
positioning of metal ions

Biomolecular dynamics

↓

Semiempirical methods possible
(extended simulations times)

↓

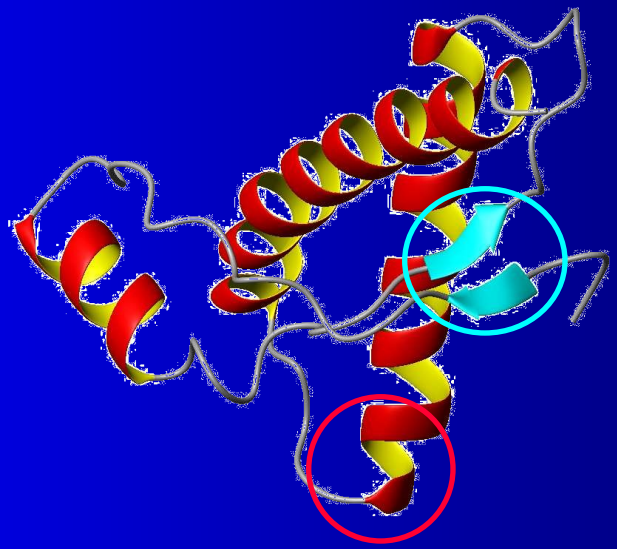
Conformational transitions
(e.g. protein folding / unfolding)

↓

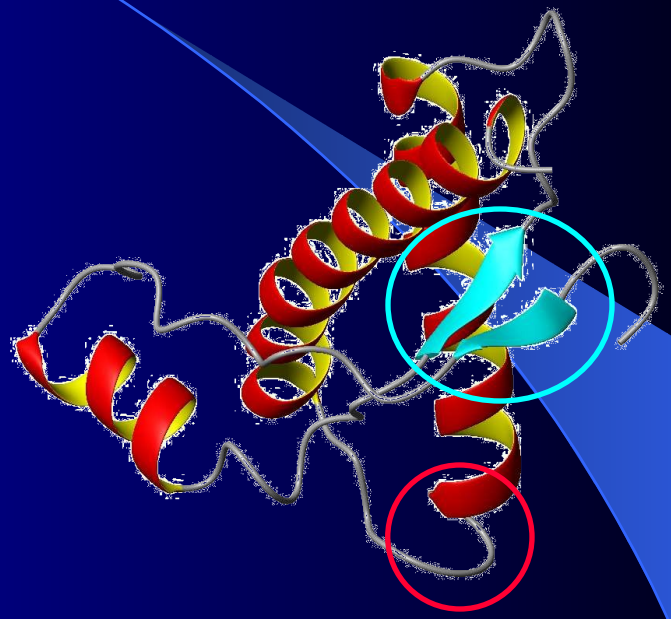
Reactivity,
averaging of parameters

Heavy computational requirements

Molecular dynamics over long periods:
pH-Driven rearrangement in human prion protein



Simulation at neutral pH
(histidines are neutral)



Simulation at acidic pH
(histidines are protonated)

β-Sheet elongation and α-helix destructuration occur during the 10-ns simulation at acidic pH