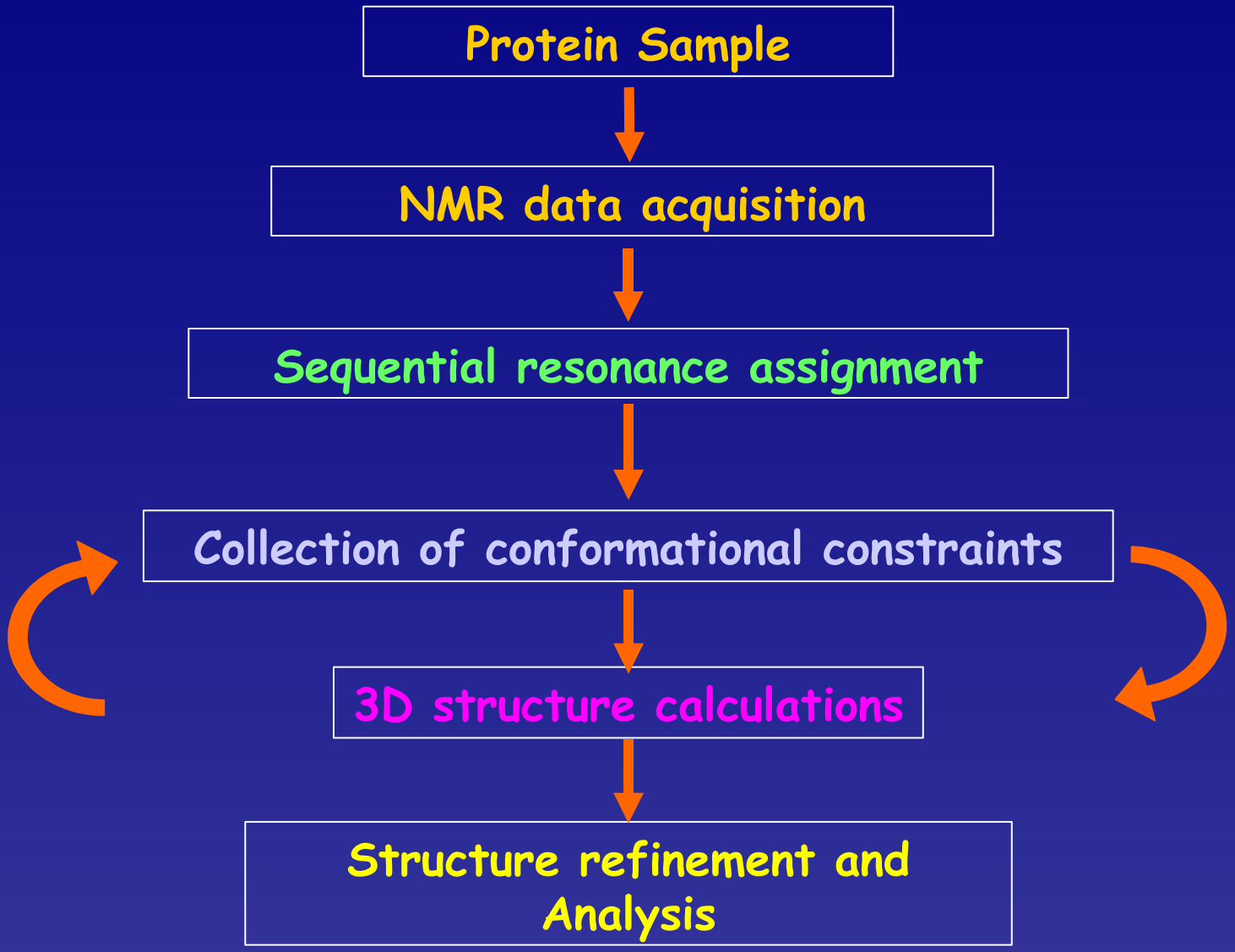
Structure determination through NMR



Acylphosphatase
SMALLEST known enzyme: 103 residues
ferredoxin-like βαββαββ sandwich domain

total number of atoms: 1661

H: 830

N: 135

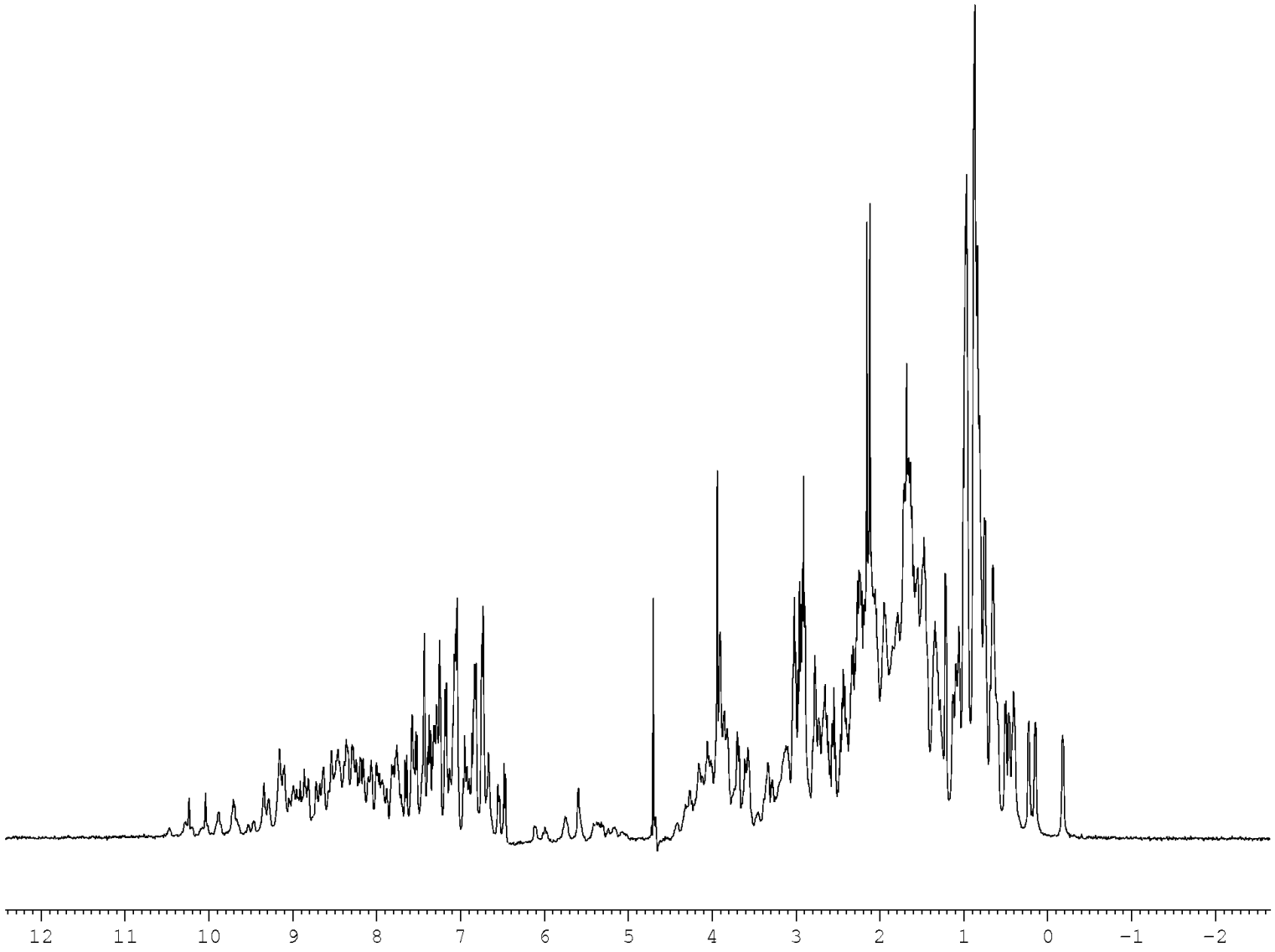
C: 537

O: 157

S: 2

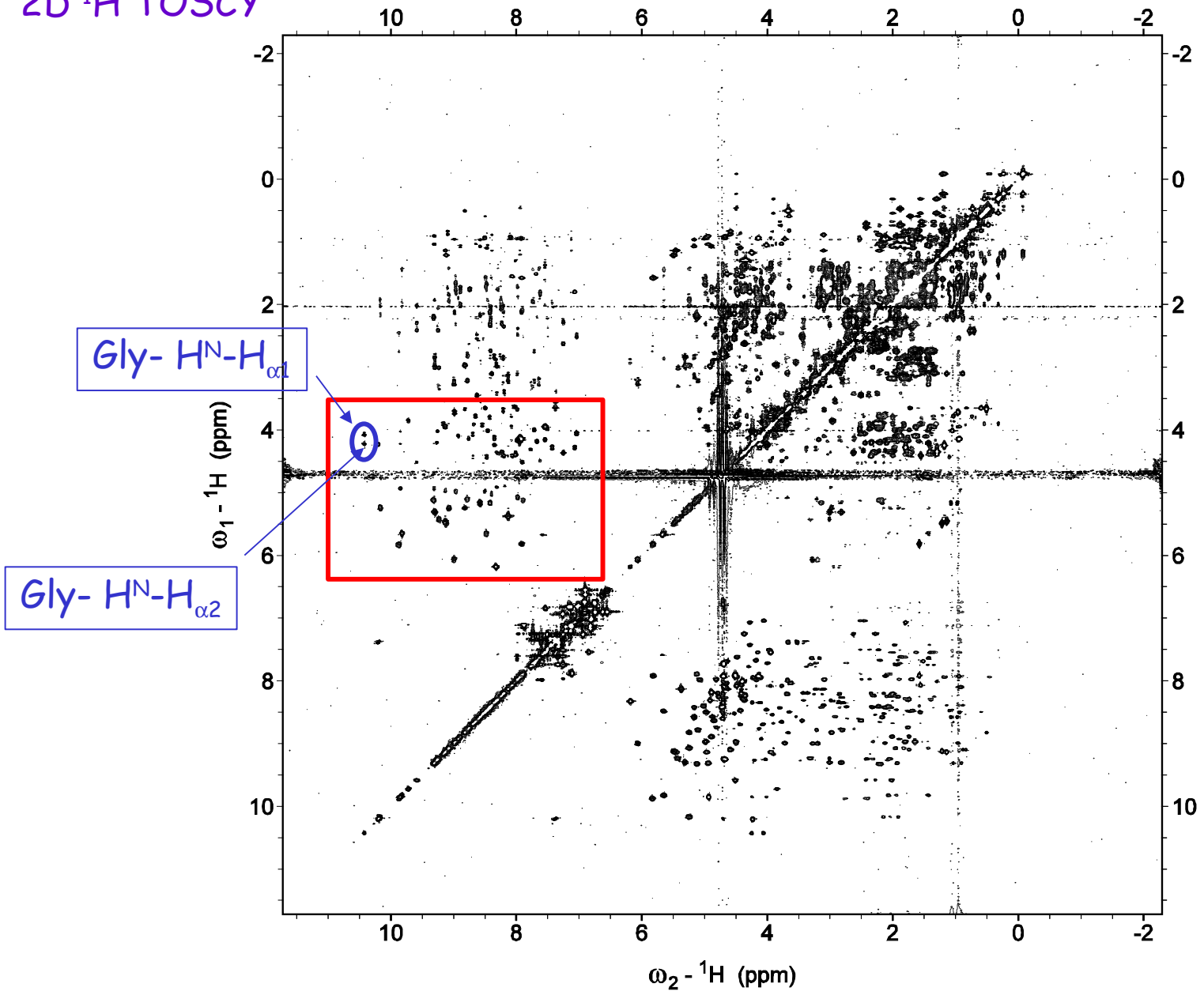
Corazza et al. Proteins (2006)
Pagano et al. J. Biomol NMR (2006)

^{1}H 1D spectrum of Acylphosphatase

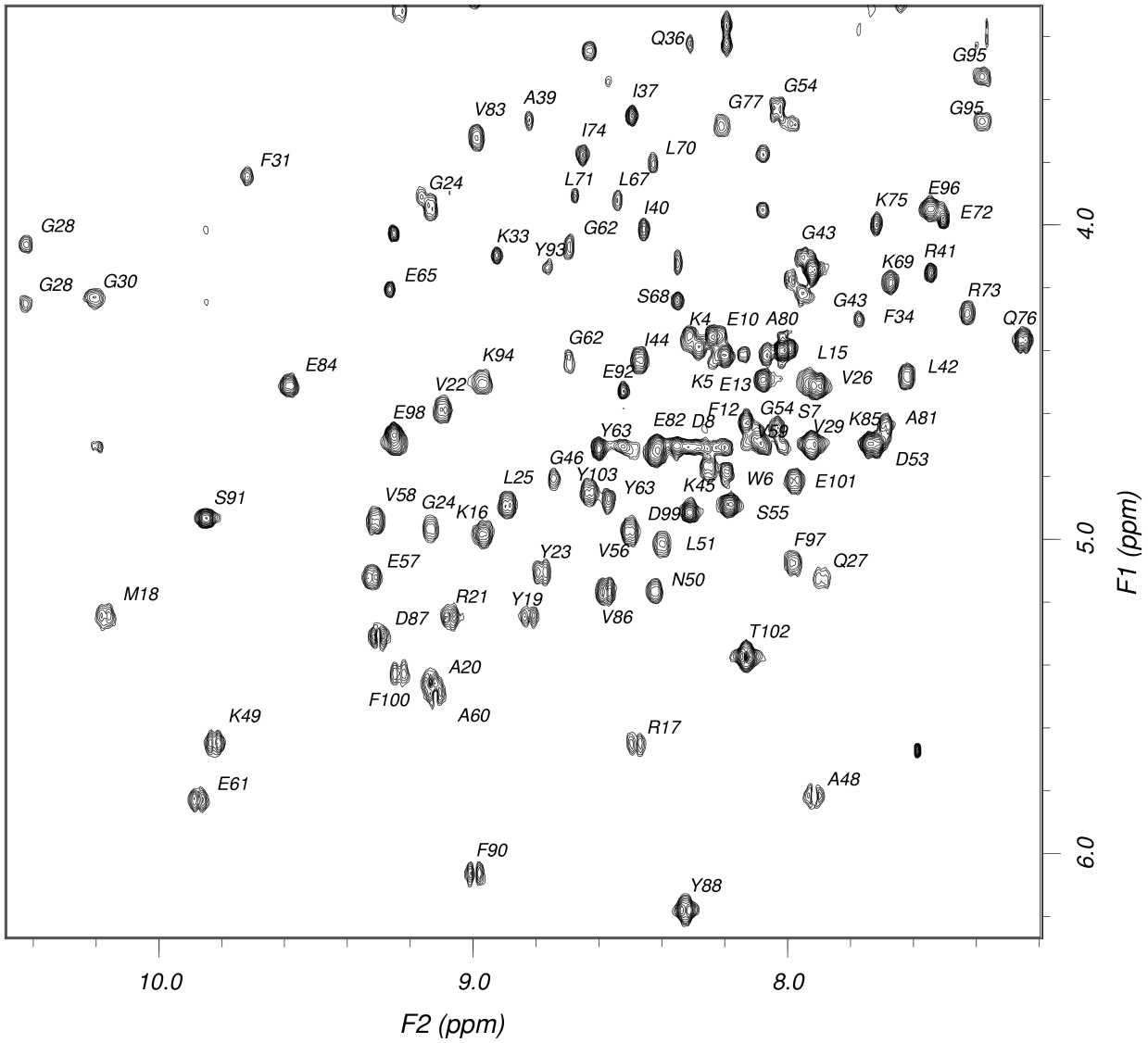


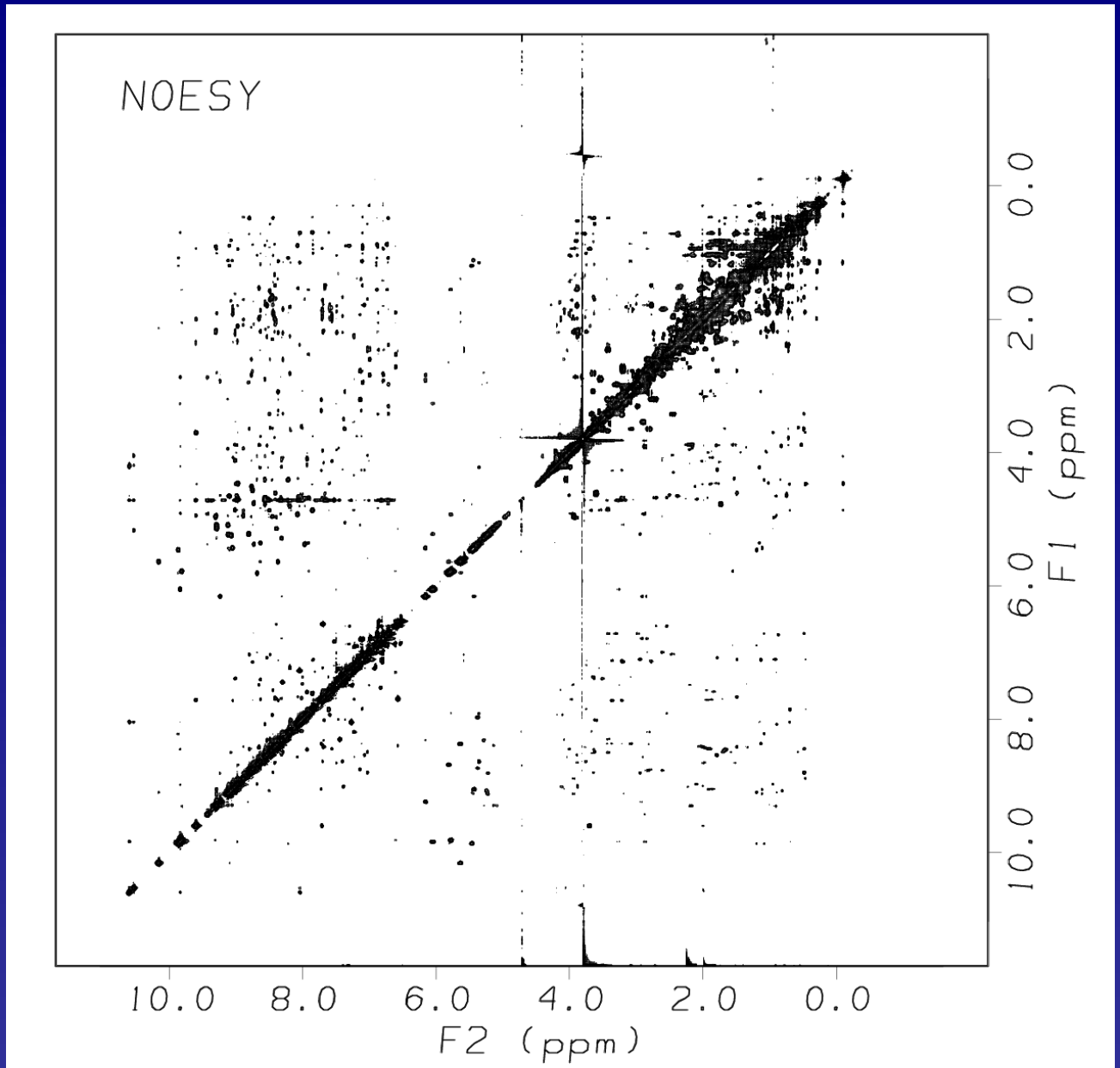
^{1}H (ppm)

2D ^{1}H TOSCY

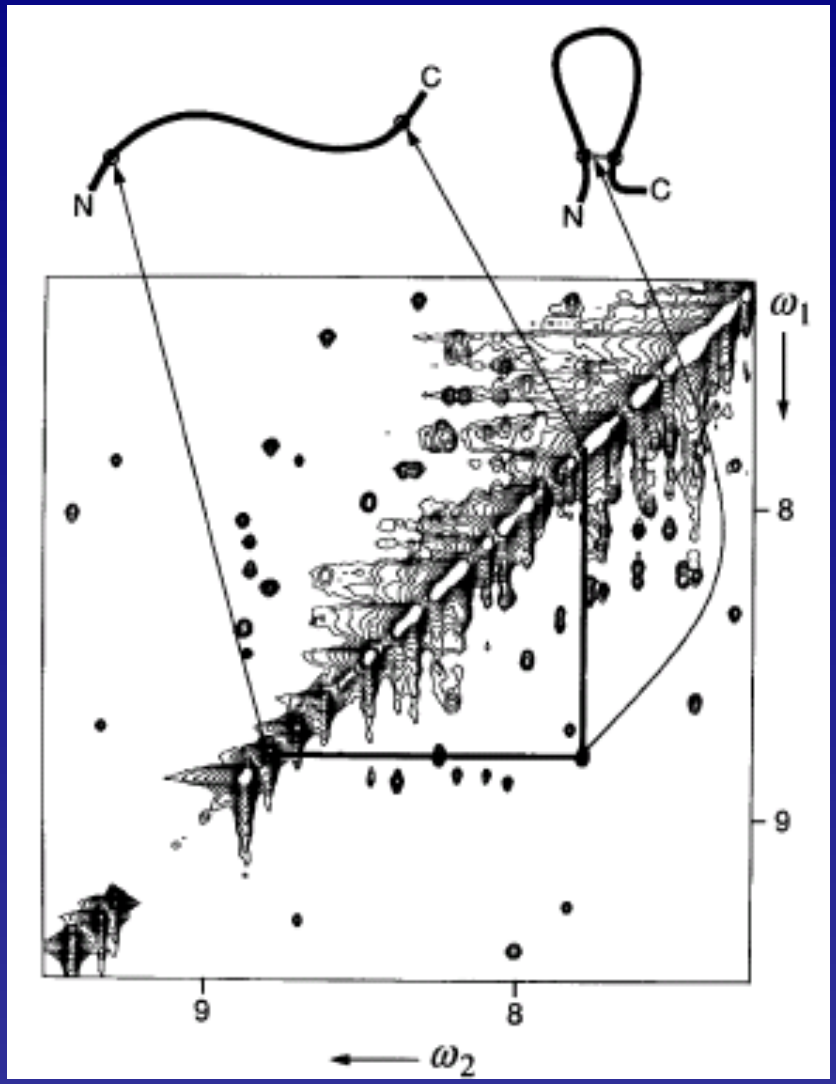

Gly- H^N-$H_{\alpha 1}$
Gly- H^N-$H_{\alpha 2}$
ω_1 - ^{1}H (ppm)
ω_2 - ^{1}H (ppm)
10
8
6
4
2
0
-2

Assigned fingerprint

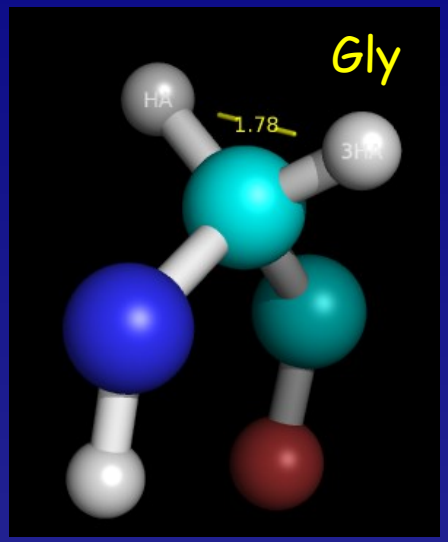


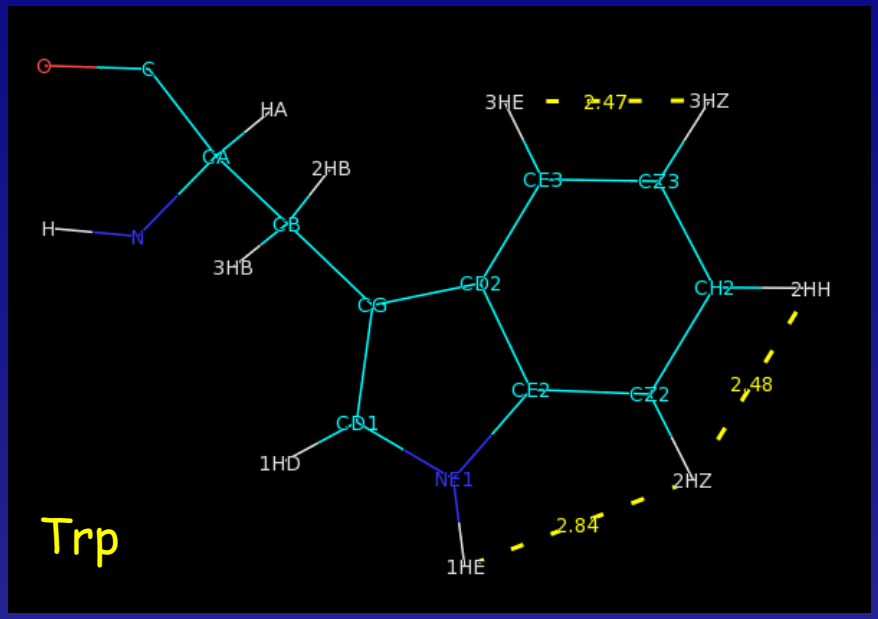

NOESY
10.0
8.0
6.0
4.0
2.0
0.0
F2 (ppm)
0.0
2.0
4.0
6.0
8.0
10.0
F1 (ppm)

NOESY—A Powerful Technique to Study Spatial Structure



- The NOESY cross peaks are integrated
- A reference cross peak belonging to a chemically "fixed" distance is chosen





- The volumes are translated into distances according to:

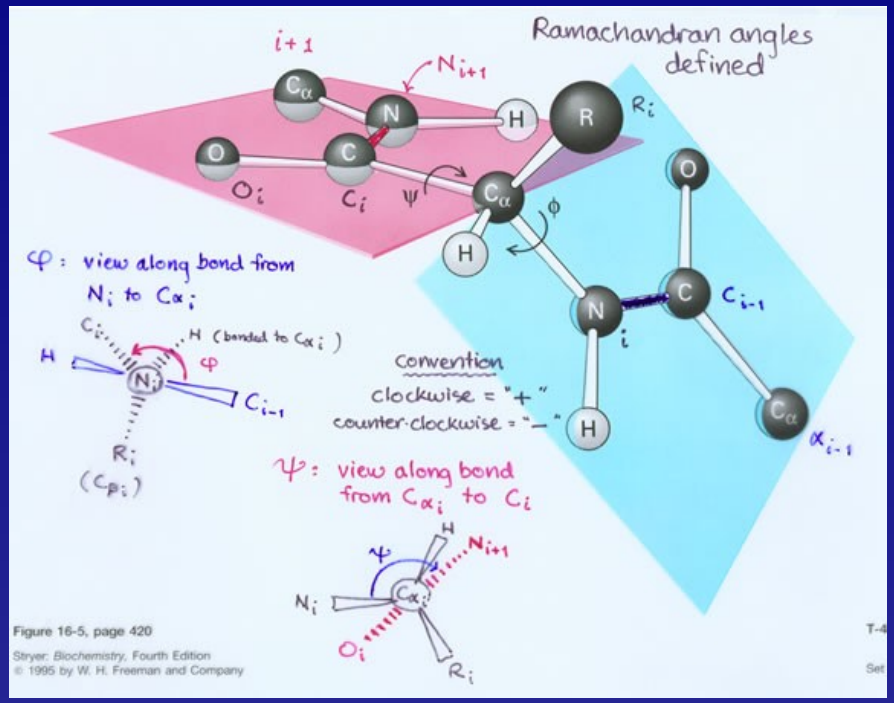
$$r_{ij} = r_{ref} (V_{ref}/V_{ij})^{1/6}$$

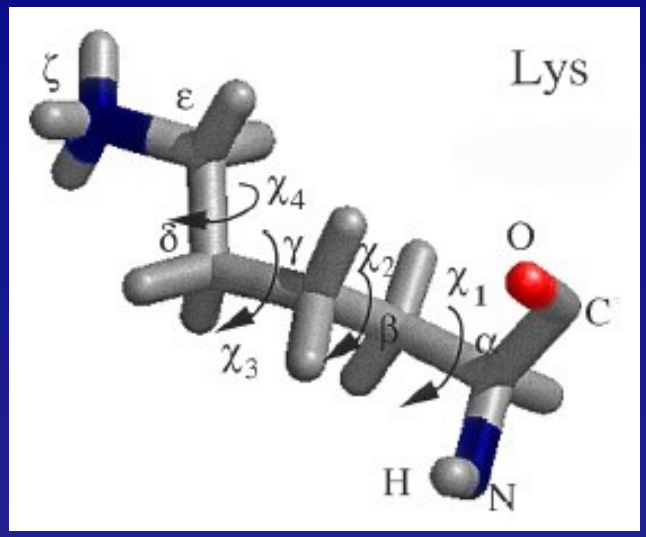
Classes of constraints:

1. intra-residue (i=j)
2. sequential (|i-j|=1)
3. medium range (1<|i-j|≤5)
4. long range (|i-j|> 5)

φ and χ^1 angles are obtained from

- $^3J_{H^N\text{-}H^\alpha}$ and $^3J_{H^\alpha\text{-}H^\beta}$ coupling constants measurements





φ and ψ angles are obtained from

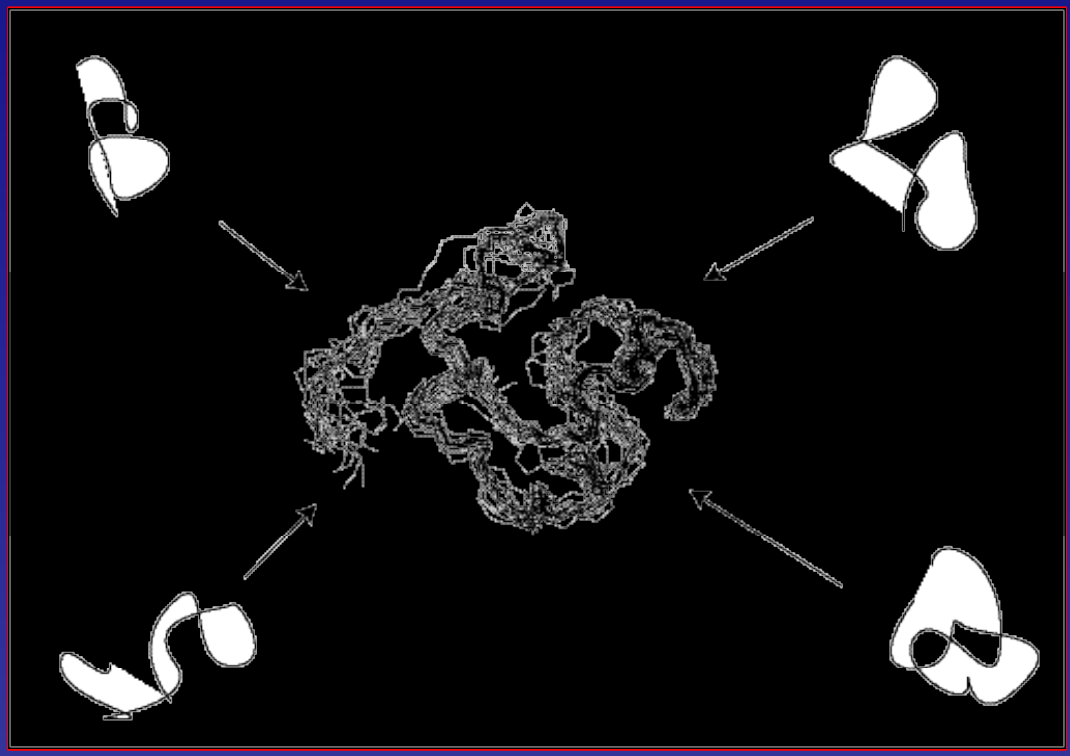
- Chemical shifts values of H^α, N^H, C^α, C', C^β Using TALOS approach

Conformational restraints

NOEs	→	Proton-proton distances
Coupling constants	→	Torsion angles
Chemical shifts	→	Torsion angles
H - bond	→	Proton-proton distances
RDCs	→	Bond orientations

Structure calculation methods based on conformational constraints

- Distance geometry
- Variable target function
- restrained molecular dynamics + simulated annealing



Distance Geometry: background

- One way to describe the conformation of a molecule other than by Cartesian or internal coordinates is in term of distances between all atom pairs.
- **Given the exact values for all distances among a set of points in the Euclidean space it is possible to determine the Cartesian coordinates for these points.**
- The distances can be represented by a symmetric NxN matrix where the elements (i,j) are $D_{ij} = |r_j - r_i|$. The diagonal elements are all zero.
- The metric matrix G can be calculated as

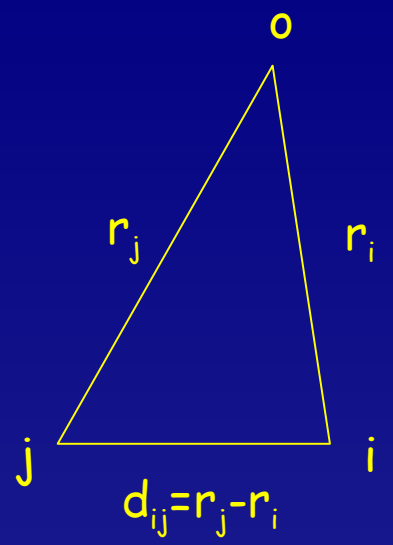
$$G_{ij} = r_i \cdot r_j = \begin{cases} \frac{1}{N}\sum_{k=1}^{N} D_{ik}^2 - \frac{1}{2N^2}\sum_{k,l=1}^{N} D_{kl}^2, & i = j \\ \frac{G_{ii} + G_{jj} - D_{ij}^2}{2}, & i \neq j \leq \end{cases}$$

  - G is related to the Cartesian coordinates $r_1, \ldots, r_N$ according:

$$r_i^{\alpha} = \sqrt{\lambda^{\alpha}}\, e_i^{\alpha} \qquad (\alpha = 1,2,3)$$

where λ^{α} are the eigenvalues of G and e_i^{α} are the n-dimensional eigenvectors.

embedding procedure



Cosine rule:

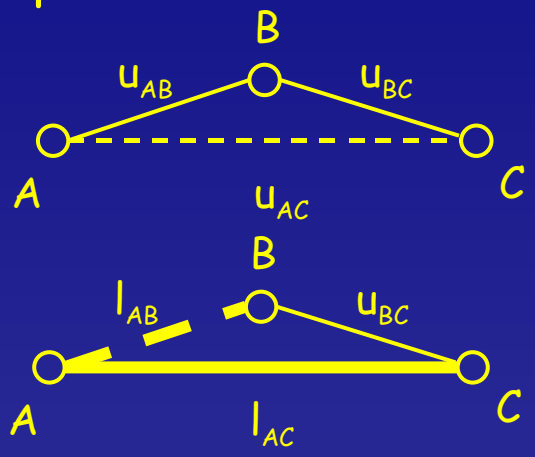
$$d_{ij}^2 = r_i^2 + r_i^2 - 2r_i \cdot r_j$$

$$r_i \cdot r_j = \frac{r_i^2 + r_i^2 - d_{ij}^2}{2}$$

Distance Geometry: application to NMR

- What is known in our case are **upper limits** derived from experimental constraints, **lower limits** due to van der Waals repulsion, and some exact distances from known bond length and angles. We do not have a complete set of distances. So our matrix is made by upper and lower bounds.

- Second we optimize this matrix by triangle inequalities by smoothing it.



$$u_{AC} \leq u_{AB} + u_{BC}$$

$$l_{AC} \geq l_{AB} - u_{BC}$$

- Basically, we randomize the distances between the atoms in the peptide, in the permitted interval between lower and upper bounds. These include normal bonds and NMR constraints. Then the embedding procedure is used to obtained the coordinates.

- From the procedure previously described it possible to obtain a set of Cartesian coordinates.

- What it is usually obtained are quite loose structures showing the correct fold, but with many inaccuracy in the geometry. Usually they have to be refined, either by MD followed by minimization or by straight minimization.

Structures calculated from distance geometry will produce the correct overall fold but usually have poor local geometry (e.g. improper bond angles, distances). Moreover it is not possible to introduce directly torsional constraints that have to be translated into distances.

Hence distance geometry must be combined with some extensive energy minimization method to generate physically reasonable structures.

It was the first method used to solve NMR structures.

Variable target function method

The basic idea is to minimize a target function that includes terms for experimental and steric restraints.

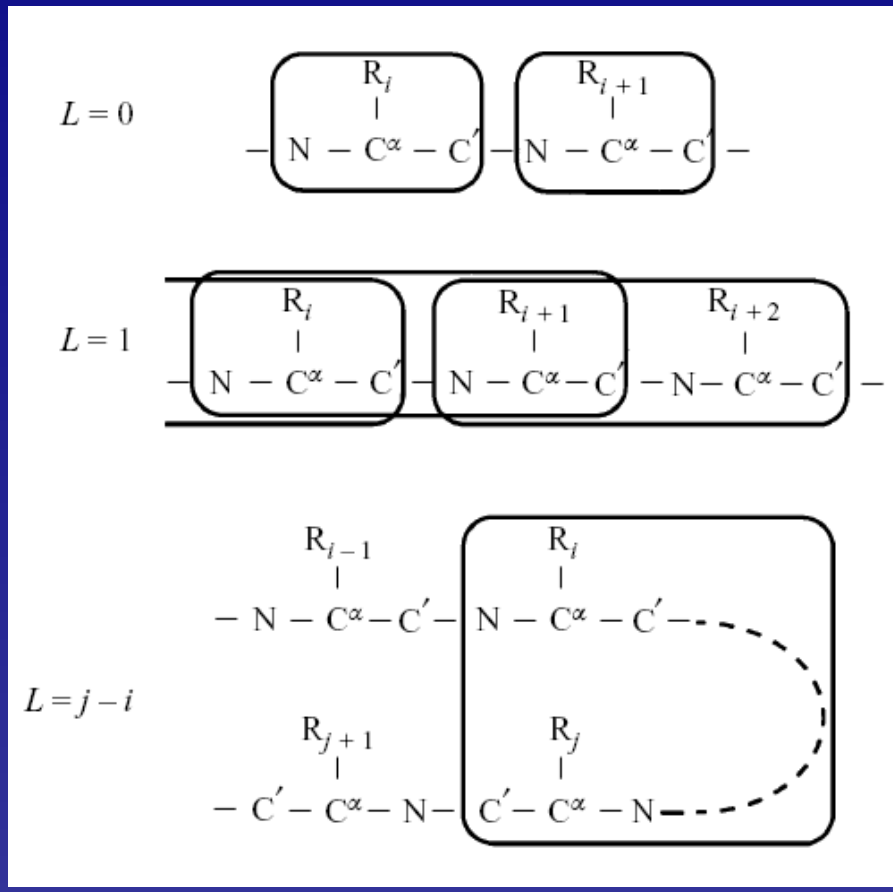
In order to avoid the problem of local minima the initially starting randomized structure is restrained by using in the order:

1. Intraresidual constraints (L=0)
2. Sequential constraints (L=1)
3. More distant constraints (L=j-i)

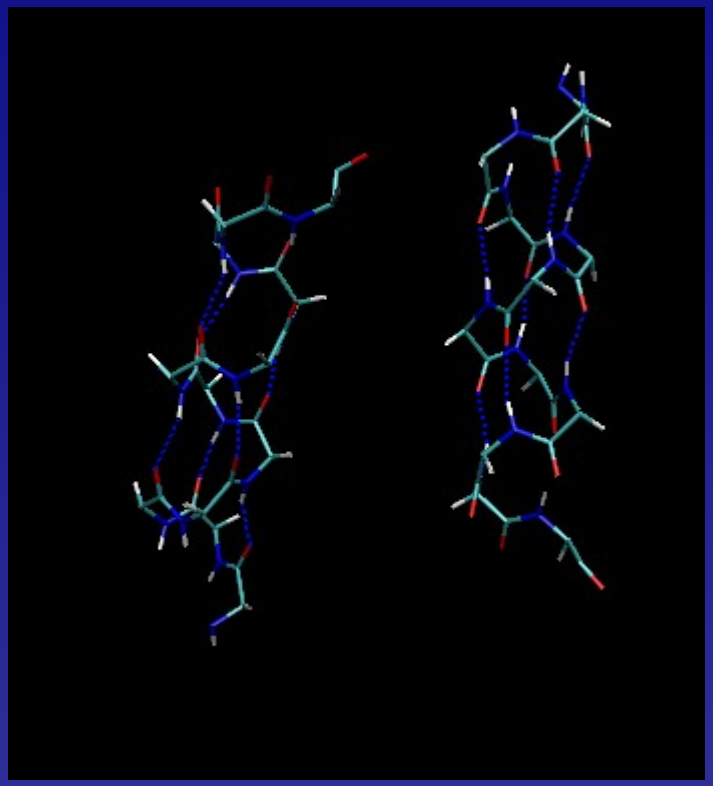
It is a conceptually simple method and works in the torsional angle space preserving the geometry during the calculation. DIANA (Güntert et al 1991)

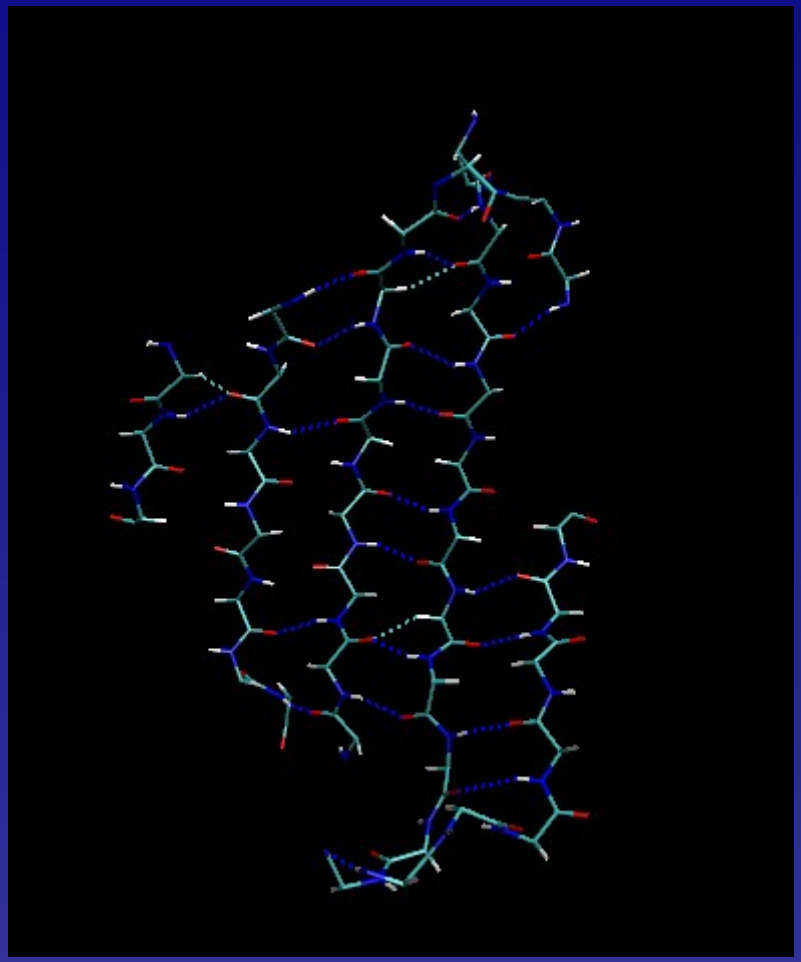
In DIANA the minimization is obtained using a simple conjugated gradient method.

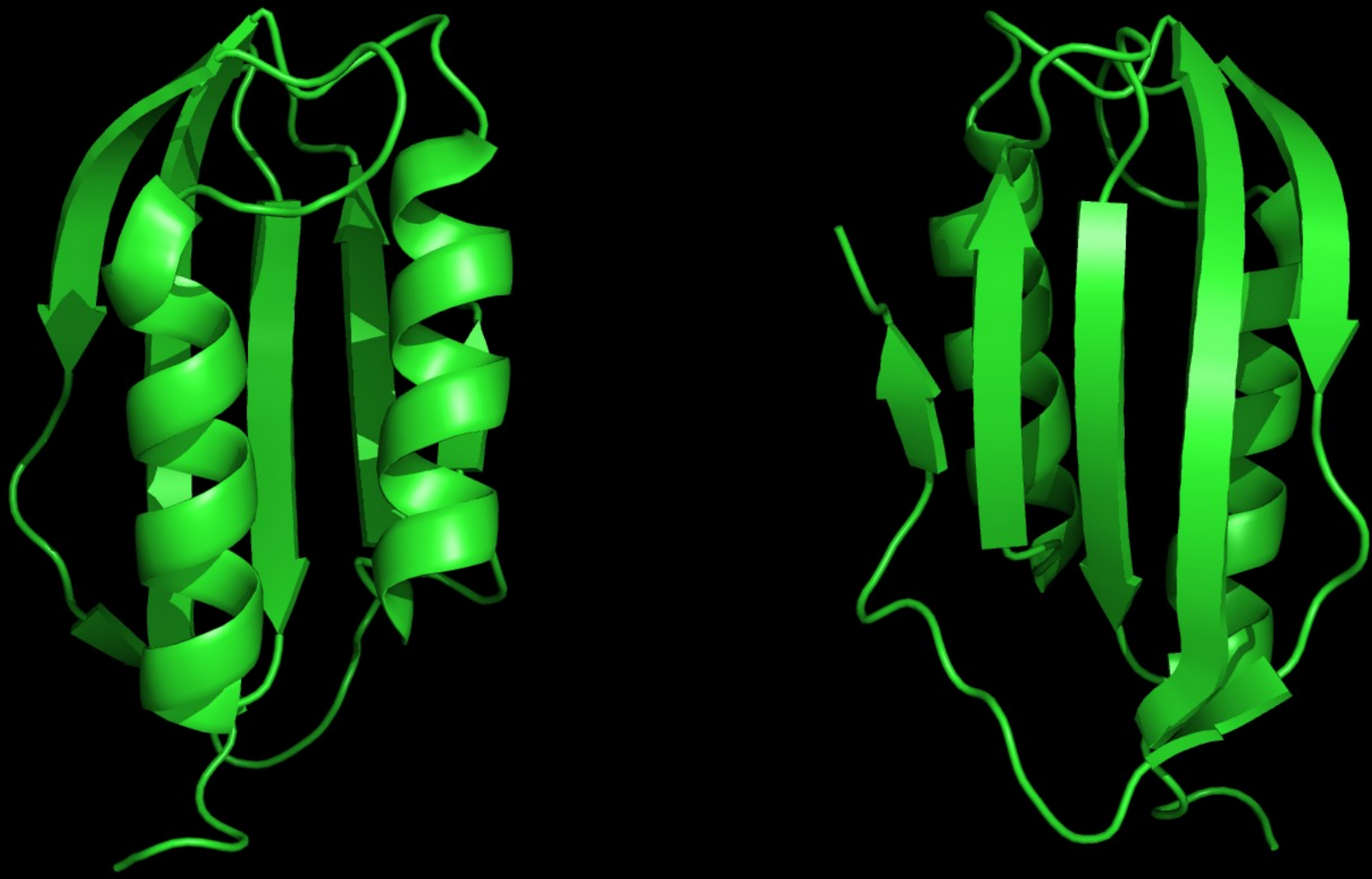
The yield of structures that converge is small.



Due to the importance given to local constraints the α-helical structures were solved more efficiently. Instead locally minimized conformations could be incompatible with long range constraints leading to β-sheets that are taken into account later.







rMD--Restrained molecular dynamics

Molecular dynamics involves computing the Newton equation of motion:

$$m_i \frac{d^2 r_i}{dt^2} = F_i, \quad i = 1, \ldots, n$$

$$F_i = -\nabla V_i$$

where V is the potential energy with respect to the atomic coordinates. Usually this is defined as the sum of a number of terms:

$$V_{total} = V_{bond} + V_{angle} + V_{dihedr} + V_{vdW} + V_{coulomb} + V_{NMR}$$

The first five terms here are "real" energy terms corresponding to such forces as van der Waals and electrostatic repulsions and attractions, cost of deforming

The NMR restraints are incorporated into the V_{NMR} term, which is a "pseudopotential" term included to represent the cost of violating the restraints, e.g the NOEs

$$V_{NOE} = \begin{cases} w_{NOE}(r_{ij} - u_{ij})^2, & r_{ij} > u_{ij} \\ 0, & l_{ij} < r_{ij} < u_{ij} \\ w_{NOE}(r_{ij} - l_{ij})^2, & r_{ij} < l_{ij} \end{cases}$$

where l_{ij} and u_{ij} are the lower and upper bounds of our distance restraint, and w_{NOE} is some chosen force constant, typically ~ 250 kcal mol^{-1} nm^{-2}

So it's somewhat permissible to violate restraints but it raises *V*. Often a simplified force field is used and the electrostatic is not taken into account.

XPLOR (Brünger, 1992) is one of the most used programs to solve NMR structures using rMD in Cartesian space.

The force filed used is:

$$V = \sum_{\text{bonds}} k_b(r - r_0)^2 + \sum_{\text{angles}} k_\theta(\theta - \theta_0)^2 + \sum_{\text{dihedral}} k_\phi(1 + \cos(n\phi + \alpha))$$

$$+ \sum_{\text{impropers}} k_\phi(\phi - \alpha)^2 + \sum_{\substack{\text{nonbonded}\\ \text{pairs}}} k_{\text{repel}}(\max(0, (sR_{min})^2 - R^2))^2$$

$$+ \sum_{\substack{\text{distance}\\ \text{restraints}}} k_d \Delta_d^2 + \sum_{\substack{\text{angle}\\ \text{restraints}}} k_a \Delta_a^2$$

k_i, s are force constants and weight; r_0, θ_0 are reference distance and angles. R_{min} is the distance at which the van der Waals potential has a minimum.

The equations of motion are numerically integrated using the leap-frog algorithm (an improvement of Verlet algorithm) according to the scheme:

$$v_i(t + \Delta t / 2) = v_i(t - \Delta t / 2) + \frac{F_i(t)}{m_i} \Delta t + O(\Delta t^3)$$

$$r_i(t + \Delta t) = r_i(t) + v_i(t + \Delta t / 2) \Delta t + O(\Delta t^3)$$

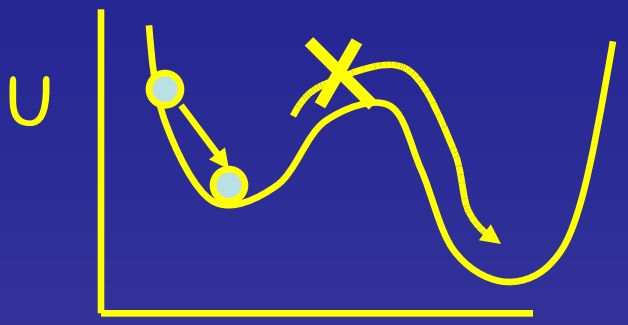
The time step Δt has to be of 10^{-15} s to take into account the fastest motion (bond oscillation). To increase the time steps it is possible to consider the bond length fixed (SHAKE methods).

Simulated Annealing

In MD usually the system is 'heated' to a physically reasonable temperature around 300 K. The amount of energy per mol at this temperature is ~ k_BT, were k_B is the Boltzmann constant. That is ~ 2 Kcal/mol.

This energy may be enough to overcome most local energy barriers but some may have a sufficiently low local energy minima that MD can not overcome. In these cases, use a more drastic search method called **simulated annealing** (because it simulates the cooling of glass).
Atoms are given kinetic energy by coupling to a "temperature bath" (typically "heat" to 1000-3000 K) and allow to slowly cool.

Repeatedly solve Newton's equations of motion for the ensemble of atoms.



The MD + SA procedure can be performed in the standard cartesian space or in the torsion angle domain.

NMR protein structure calculation

→ Conformational restraints from NMR measurements

→ Simulated Annealing + Molecular Dynamics

(Minimization of a hybrid energy function (Target function))

MD + SA can be performed both in Cartesian space and in torsion angle space.

Available programs:

- Xplor-NIH (CNS, XPLOR) (both cartesian space and TAD)
- **DYANA (CYANA) (only TAD)**

Structural calculations

Several (100-400) random structures are generated

TAD

The folded structures with the best agreement to the experimental constraints are taken (family of structures)

Experimental constraints

Torsion Angle Dynamics (TAD)

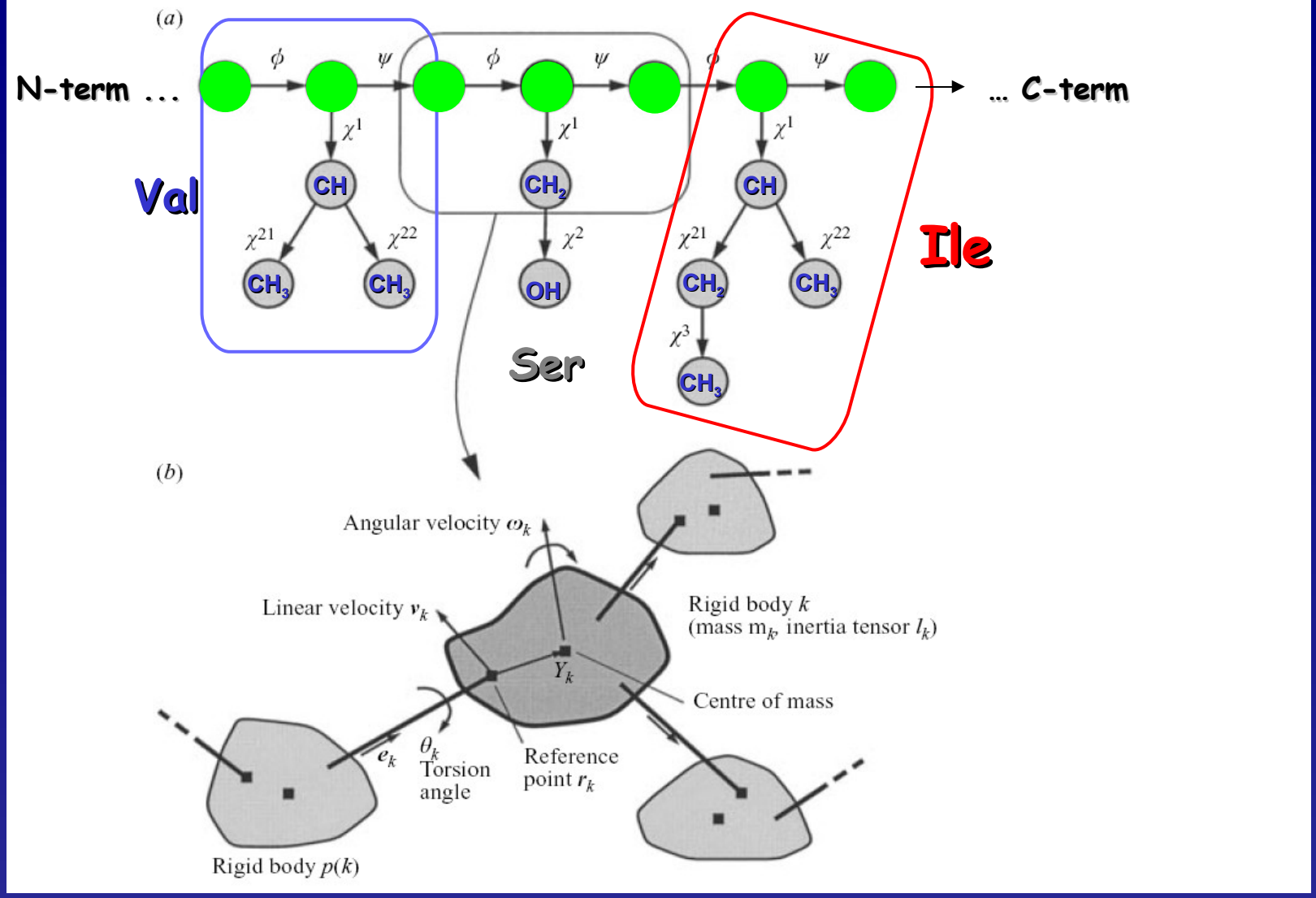
- Torsion angle dynamics = molecular dynamics (MD) in torsion angle space
- Classical mechanical equations of motion are solved in a system with N torsion angles as the only degrees of freedom
- About 10 times less degrees of freedom than in conventional Cartesian space MD
- Fixed bond lengths and bond angles:
  - no high frequency motions
  - longer integration time-steps, higher annealing temperatures

Generalized coordinates: $q_1 \ldots\ldots q_m$

$L = E_{kin} - E_{pot}$

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

Lagrange equation of motions



PETER GUNTERT *Quarterly Reviews of Biophysics* **31**, (1998), 145-237

The only degree of freedom are the torsion angles, that is rotation around a single bond.

Torsion Angle Dynamics (TAD)

- Newton's equation in generalized coordinates, $\theta_1, \ldots, \theta_n$

Quantity	Cartesian coordinates	Torsion angle space
Degrees of freedom	3N coordinates $x_1, \ldots, x_N$	n = number of torsion angles $\theta_1, \ldots, \theta_n$
Equation of motion	Newton's equations: $m_i \ddot{x}_i = -\nabla V$	Lagrange equations: $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_k}\right) - \frac{\partial L}{\partial q_k} = 0$ $L = E_{kin} - E_{pot}$
Computational complexity	Proportional to N	~ n^3 (linear equations) ~ n (tree structure)

Exploiting the tree structure of proteins, the computational cost for TAD is proportional to the system size.

The program DYANA

The target function represents the potential energy of the system

$$V = \sum_{c=u,l,v} w_c \sum_{(\alpha,\beta)\in I_c} (d_{\alpha\beta} - b_{\alpha\beta})^2 + w_a \sum_{i\in I_a} \left[1 - \frac{1}{2}\left(\frac{\Delta_i}{\Gamma_i} \right)^2 \right] \Delta_i^2$$

- Upper and lower bound restraints
- Van der Waals term
- Torsion angle restraints terms

Güntert P., Mumenthaler C., Wüthrich K., *J.Mol. Biol.*, 1997

DYANA steps

- Generation of random conformers (50 – 300).
- Short minimization to reduce high energy interaction (no hydrogen included).

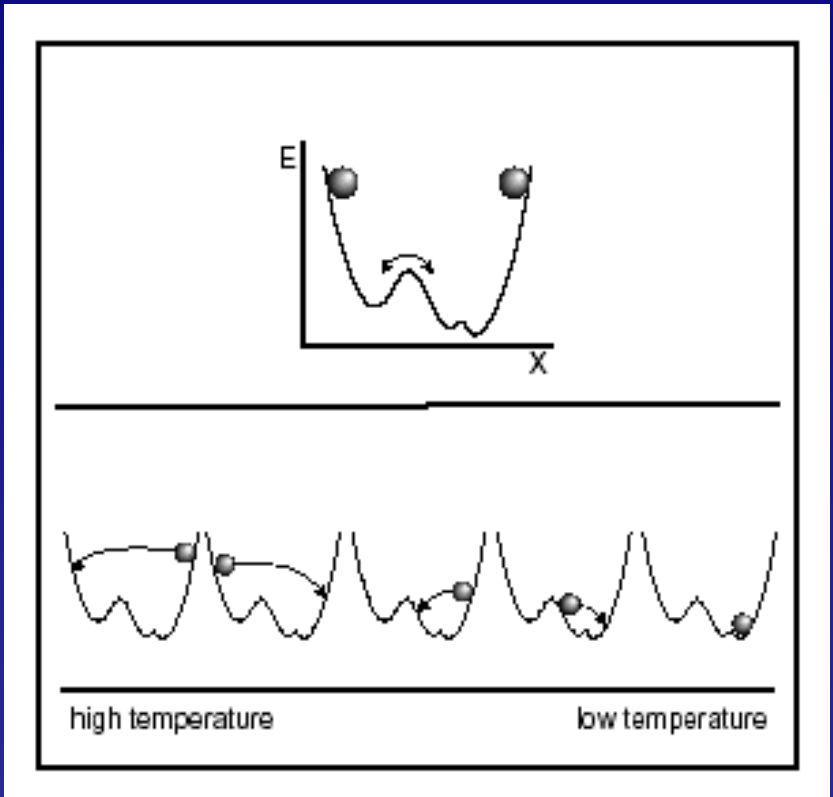
  a. 100 conjugate gradient steps – only restraints of neighbor residues.

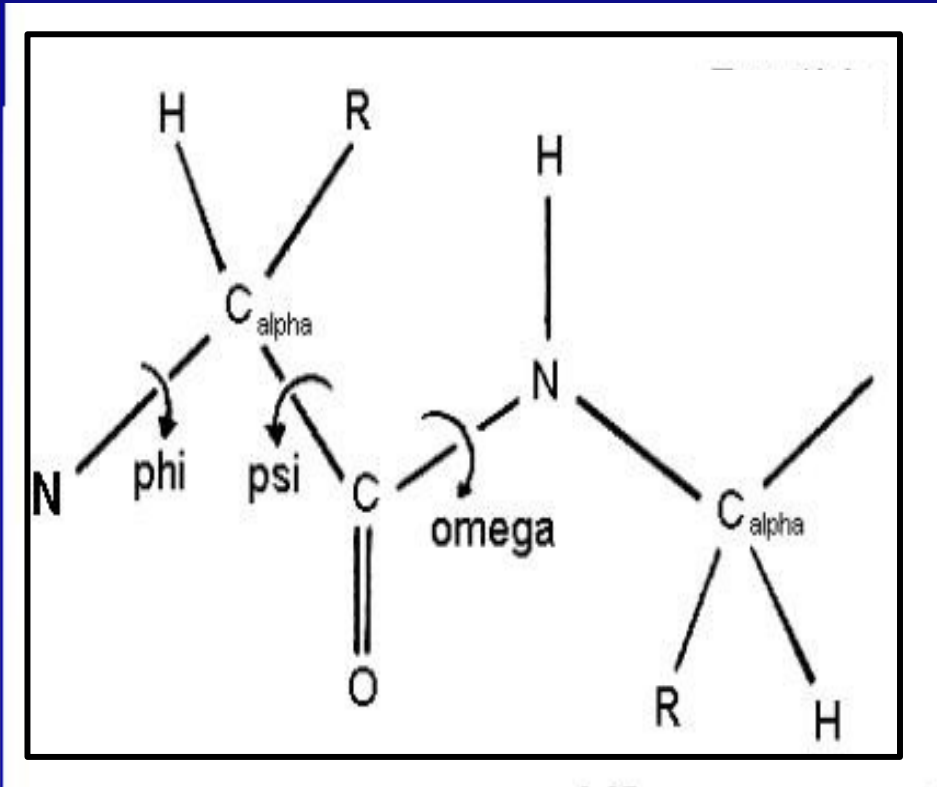
  b. 100 conjugate gradient steps – all restraints.

- Torsion angle dynamics calculation at high temperature. T_{high} = 10000 K. Δt = 2fs.
- Slow cooling TAD. Longer Δt. (100 fs)
- Incorporation of all hydrogens. Check of steric overlap. Conjugate gradient minimization is performed.
- Final 1000 steps of minimization.

NMR protein structure calculation

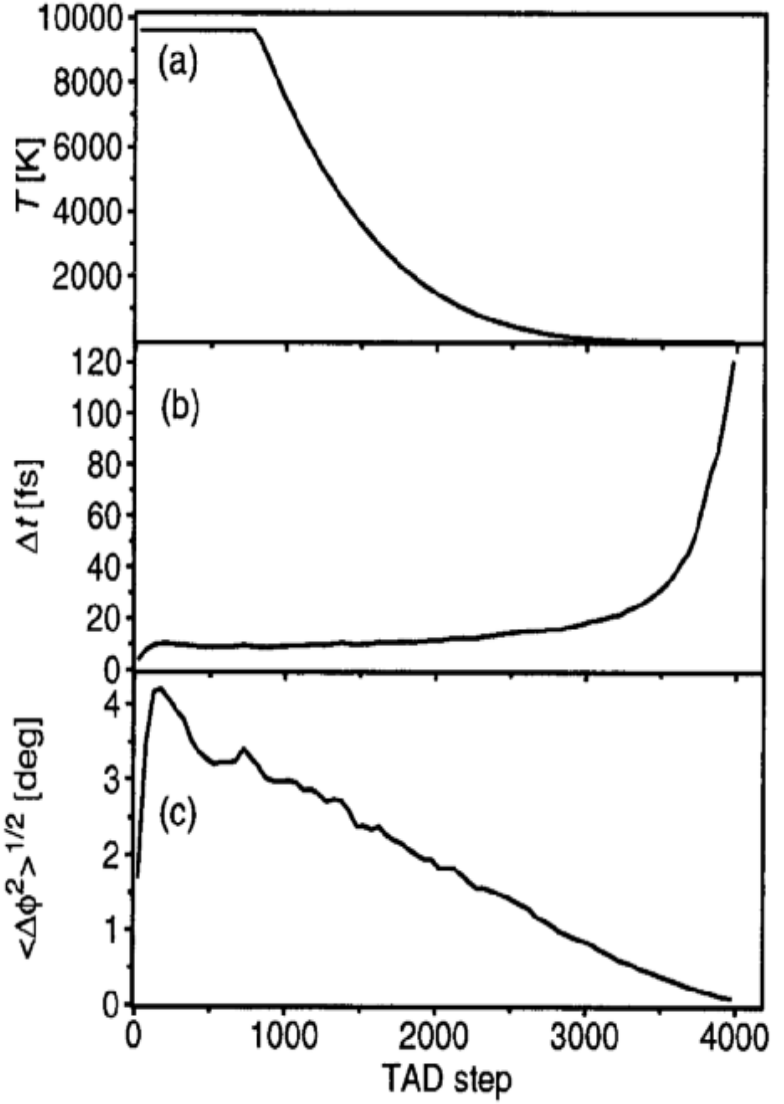
Simulated annealing with torsion angle dynamics





A starting structure is heated to a high temperature
During many discrete cooling steps the starting structure can evolve towards the energetically favourable final structure under the influence of a force field derived from the constraints.

The program DYANA



(DYnamics Algorithm for Nmr Applications)

Temperature of the heat bath to which the system is weakly coupled (default value for initial temperature T=9600K)

Integration time-step length, Δt, depending on the accuracy of energy conservation (short Δt at the outset and an increase above 100fs toward the end of the calculations)

Rms deviation on torsion angles along the TAD simulation

50-300 random conformers are annealed

The best 20 structures with the lowest target function are selected to constitute the representative structure family

Acceptance Criteria

Typically generate 50 or more trial structures, but not all will converge to a final structure that is physically reasonable or consistent with the experimentally derived NMR restraints. We want to throw such structures away rather than include them in our reported ensemble.

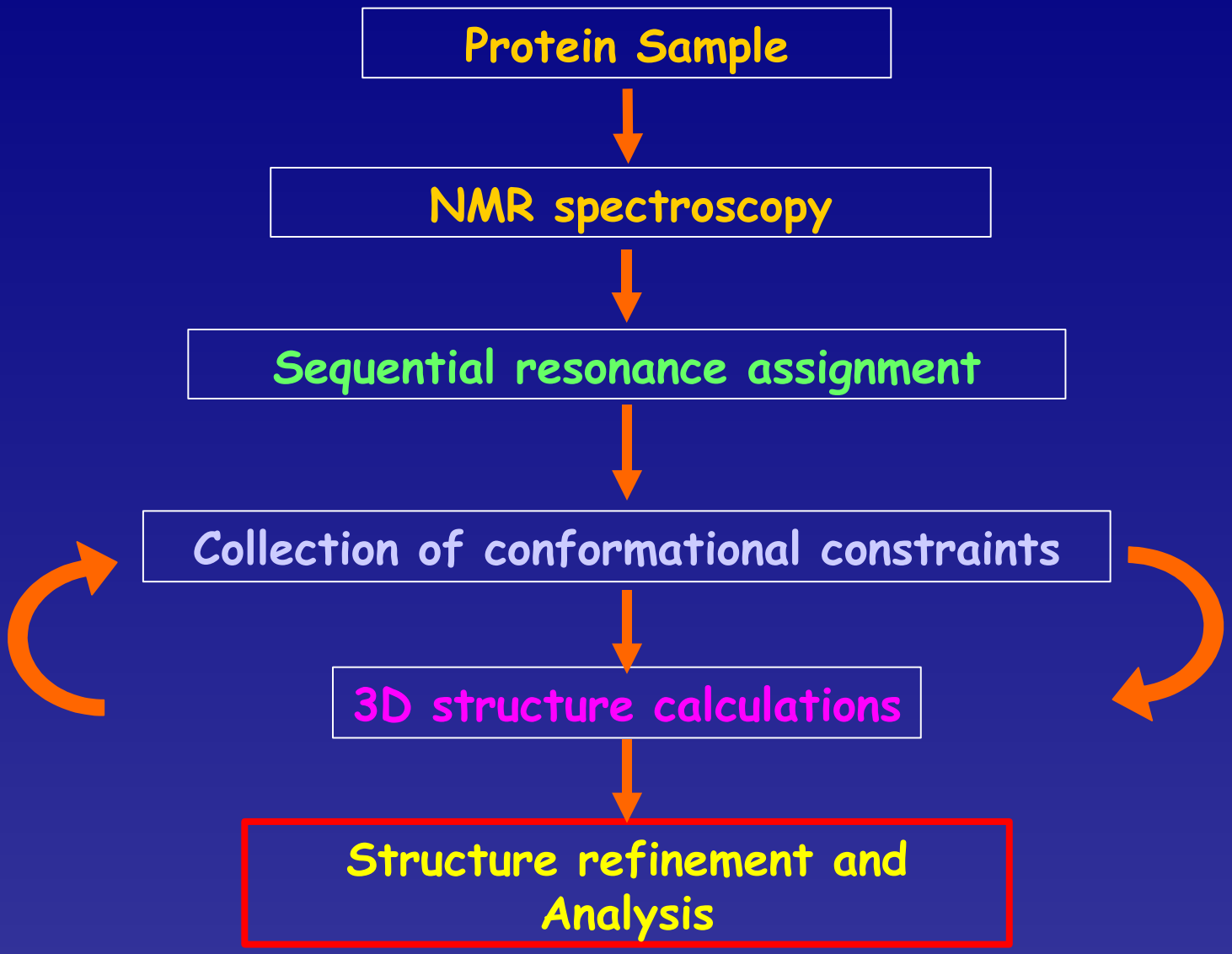
These are typical acceptance criteria for including calculated structures in the ensemble:

- no more than 1 NOE distance restraint violation greater than 0.4 Å
- no dihedral angle restraint violations greater than 5 degrees
- no gross violations of reasonable molecular geometry

Sometimes structures are rejected on other grounds as well:

- too many residues with backbone angles in disfavored regions of Ramachandran space
- too high a final potential energy in the rMD calculation

Structure determination through NMR



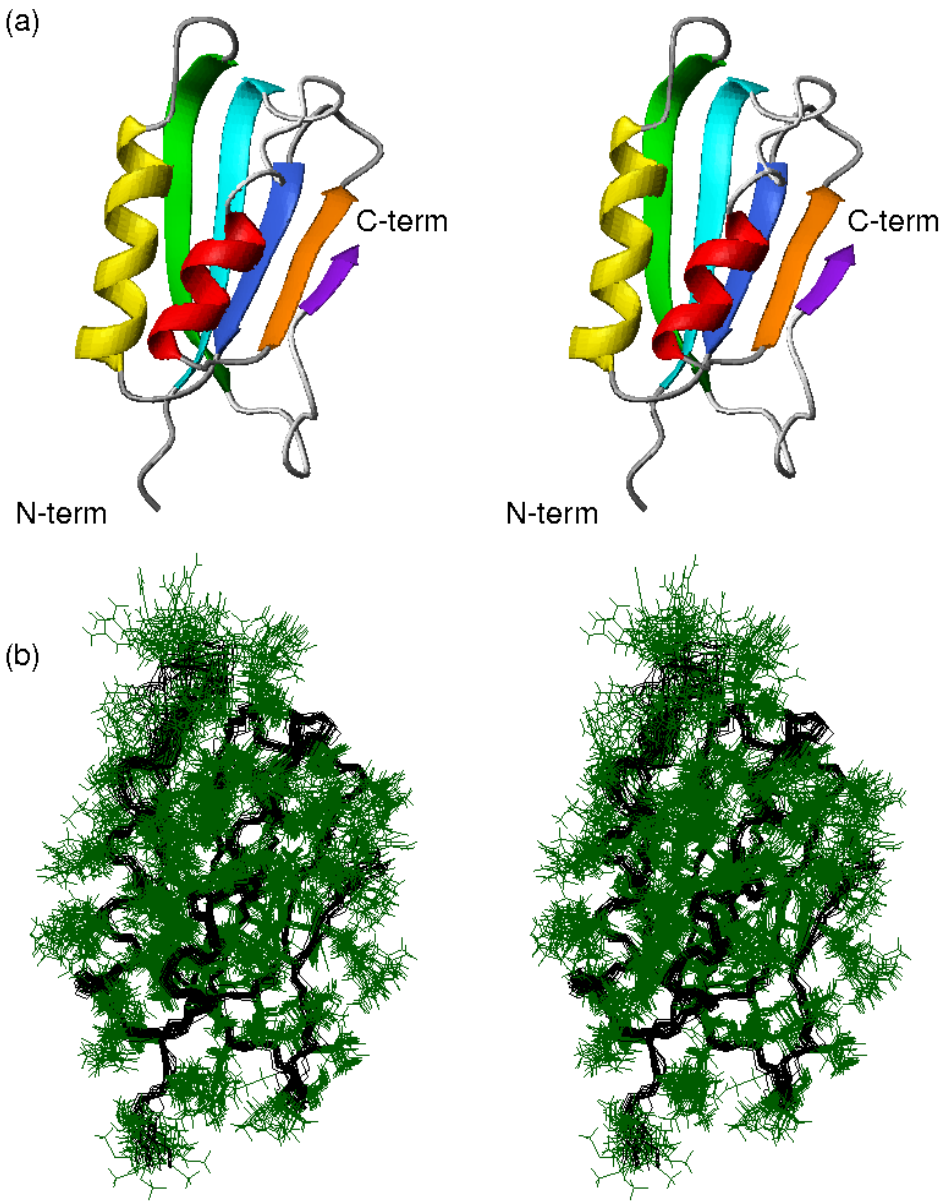
Structure refinement through REM and RMD

Restrained Energy Minimization and Restrained Molecular Dynamics

The force field used for the refinement is a complete one and better results are obtained for hydrated systems. During the refinement protocol both vdW and electrostatic are introduced. The calculation is always done in the cartesian space.

$$V_{tot}(r) = V_{stretching} + V_{bending} + V_{torsion} + V_{non-bonded}$$

$$V_{non-bonded} = \sum_{i,j} \left[\frac{A}{r_{ij}^{12}} - \frac{B}{r_{ij}^{6}} + \frac{1}{4\pi\varepsilon_0} \frac{q_i q_j}{r^2} \right]$$



Ensemble of 20 structures

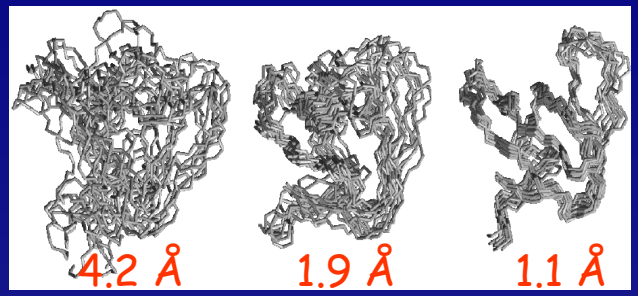
Quality of the structure

- Number of constraints > 15 per residue

- Precision

$$RMSD = \sqrt{\frac{1}{n}\sum |r_i - Rq_i|^2}$$

R matrix of rotation



Too low RMSD values are meaningless in solution at room temperature

- Accuracy $E_{pot} = \sum U_i\left(d_i - d_i^0\right)^2 + \sum W_i\left(1 + \cos\left(\vartheta_i - \vartheta_i^0\right)\right)^2 + VdW$

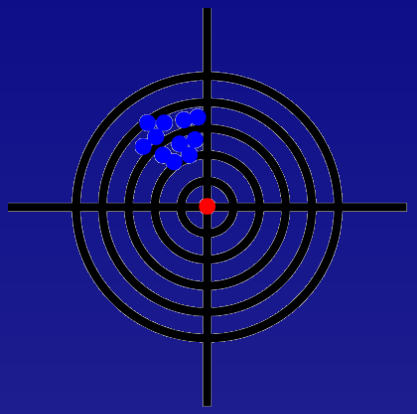
+ other constraint contributions

- **Procheck statistics** expected for a good quality structure:

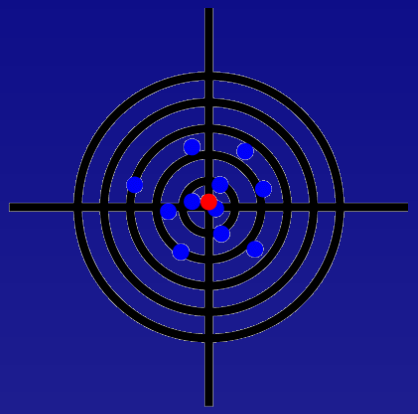
< 10 bad contacts per 100 residues
Average hydrogen bond energy in the range of 2.5-4.0 $Jmol^{-1}$
Overall G-factor > –0.5

Precision versus Accuracy

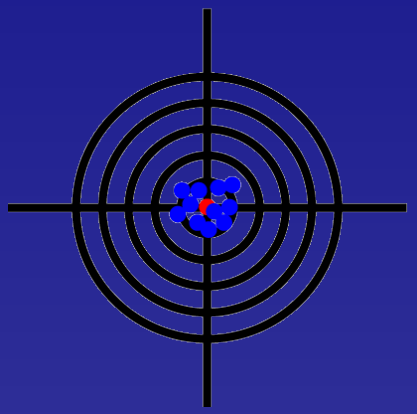
Precise,
not accurate



Accurate,
not precise



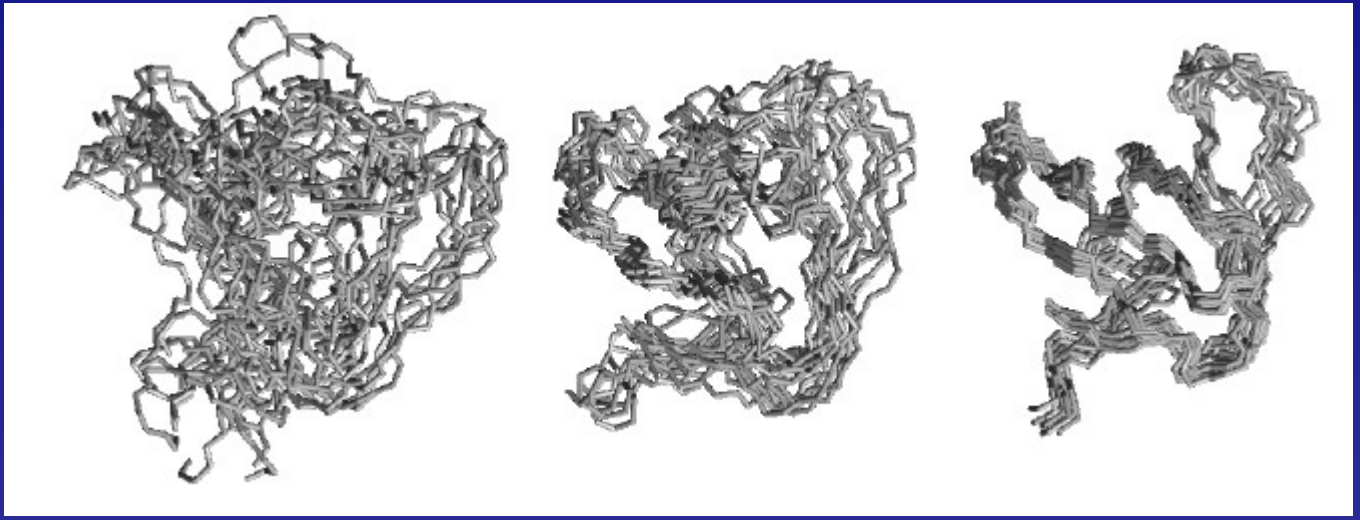
Precise
and accurate



Not accurate
and not precise

Improving the Quality of NMR Structures

- Stereospecific Assignments
  - Making stereospecific assignments increase the relative number of distance constraints while also tightening the upper bounds of the constratins
  - There is a direct correlation between the quality of the NMR structure and the number of distance constraints
    - more constraints → higher the precision of the structure



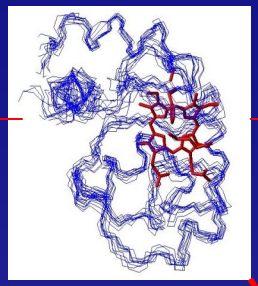
Increasing Number of NOE Based Constraints

Validation criteria for protein structures

- Local geometry:
  - Bond lengths, bond angles, chirality, omega angles, side chain planarity
- Overall quality:
  - Ramachandran plot, rotameric states, packing quality, backbone conformation
- Others:
  - Inter-atomic bumps, buried hydrogen-bonds, electrostatics

AQUA and PROCHECK-NMR

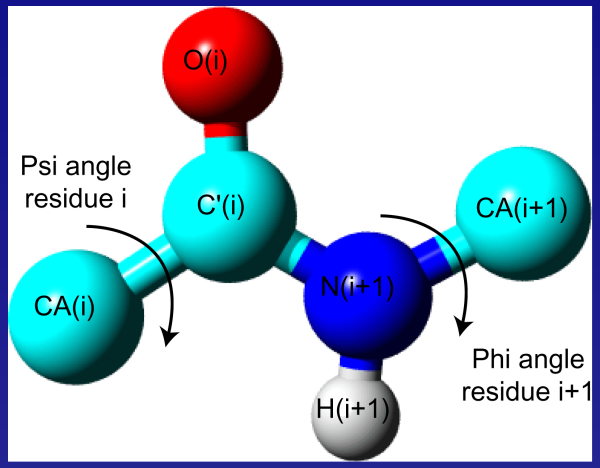
Analysis of an ensemble of NMR protein structures



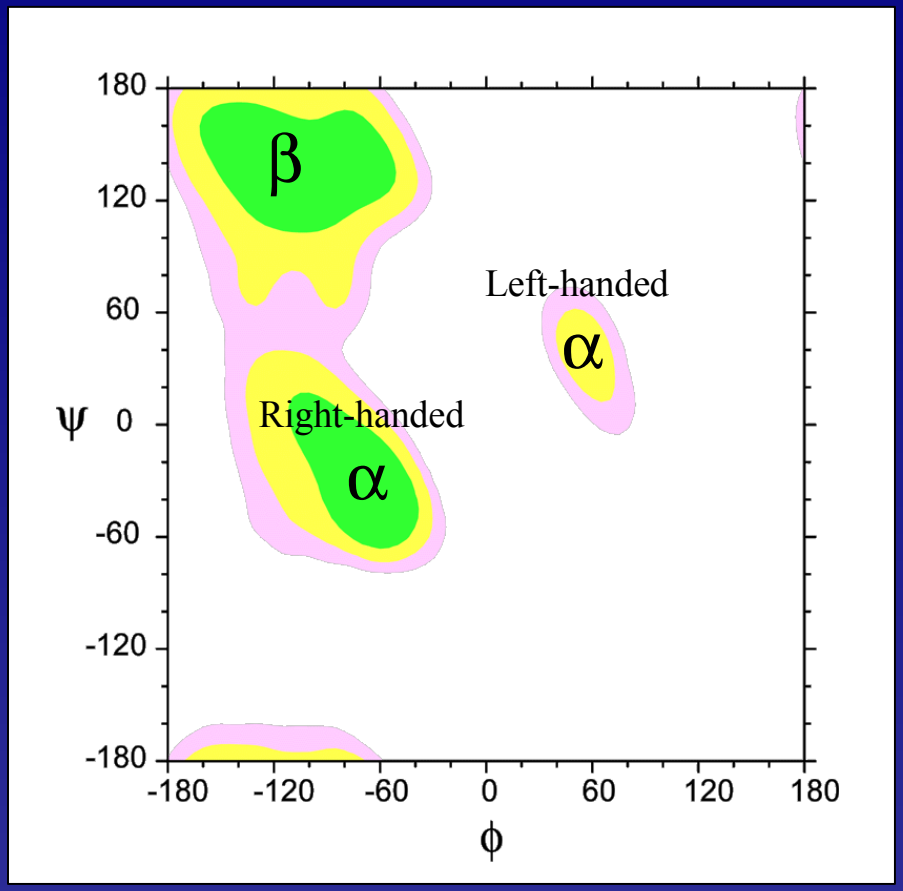
Degree of agreement of the structure with the experimental data

Quality of the geometrical properties of the model structures

Ramachandran Plot

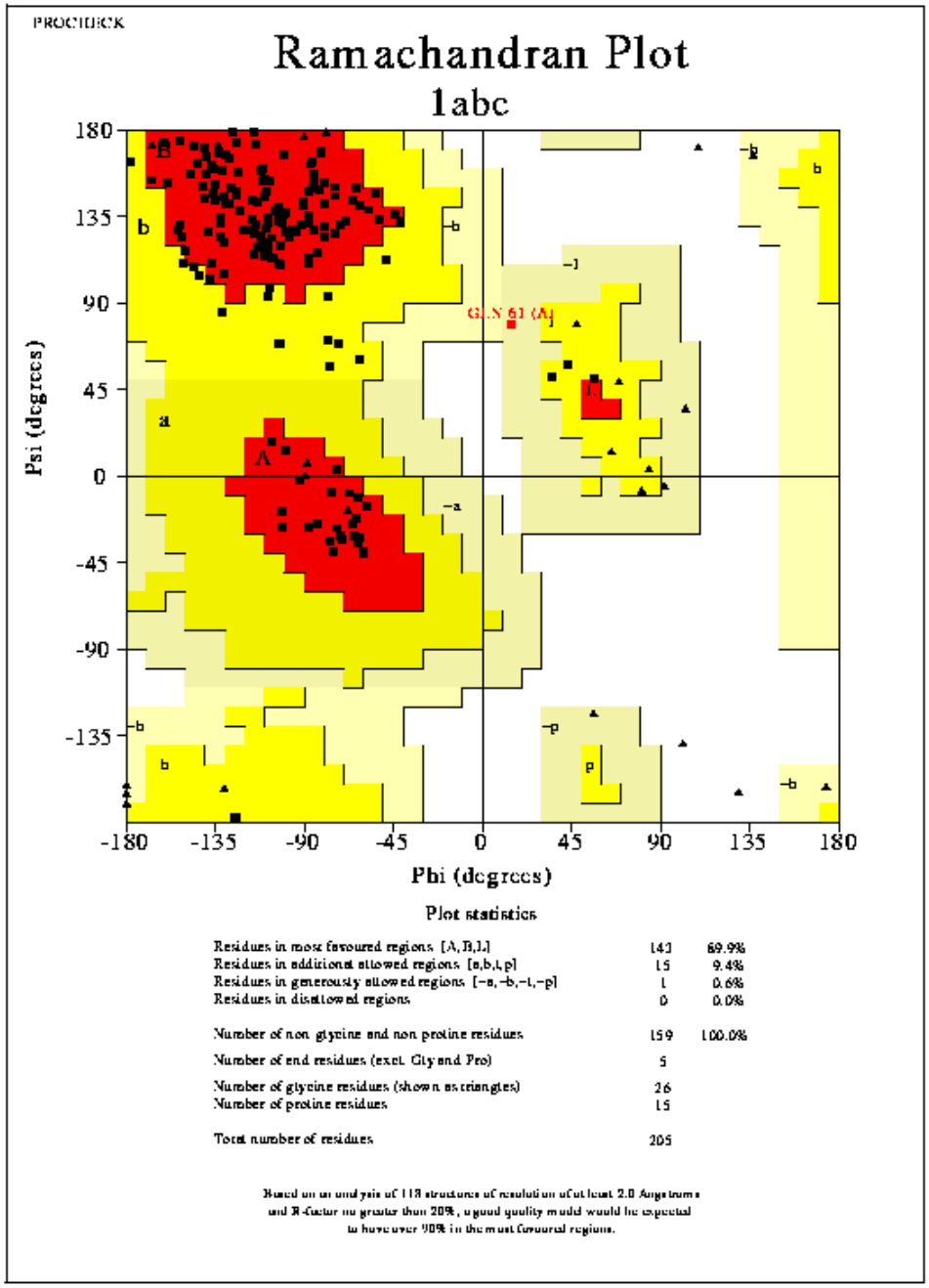


Phi and Psi angles



Ramachandran plot

Ideally, one would hope to have over 90% of the residues in these "core" regions



Residues in the most favored region (*A*, *B*, *C*) : 69.9 %
Residues in the add. allowed region (*A*, *B*, *C*) : 9.4 %
Residues in the gener. allowed region (*A*, *B*, *C*) : 0.6 %
Residues in the disallowed region (*A*, *B*, *C*) : 0 %

High Resolution NMR Structures

Usually ~15-20 NOE distance restraints per residue, but the total # is not as important as how many long-range restraints you have, meaning long-range in the sequence: $|i-j| > 5$, where i and j are the two residues involved

Good NMR structures usually have $\geq$ ~ 3.5 long-range distance restraints per residue in the structured regions

High-resolution structure will have backbone RMSD $\leq$ ~0.8 Å, heavy atom RMSD $\leq$ ~1.5 Å

Low RMS deviation from restraints (good agreement w/restraints) and good stereochemical quality:

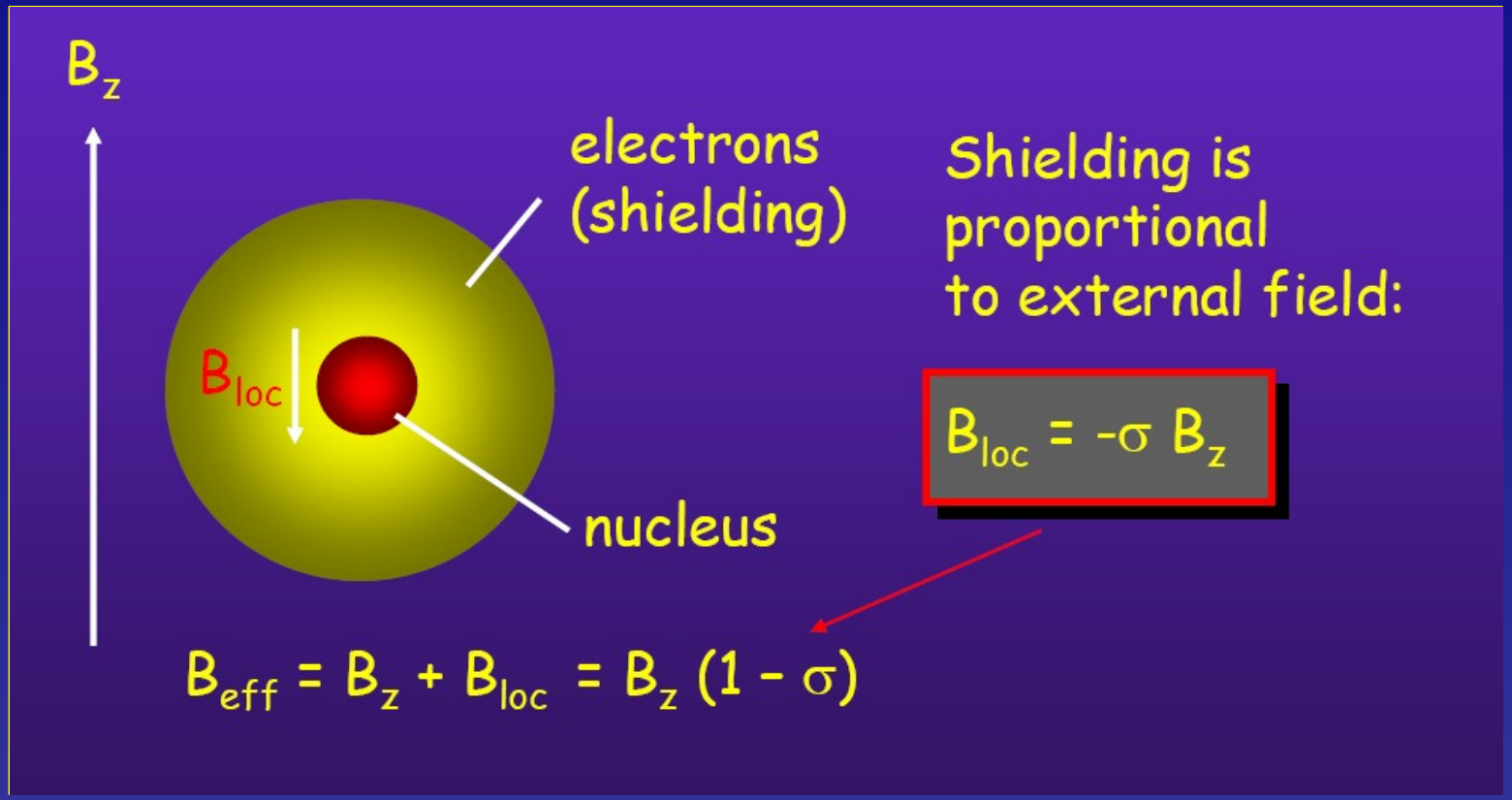
- -ideally >90% of residues in core (most favorable) regions of Ramachandran plot
- -very few "unusual" side chain angles and rotamers (as judged by those commonly found in crystal structures)
- -low deviations from idealized covalent geometry.

Chemical shift origin

The precise frequency absorbed by a nucleus in a sample depends on the chemical environment

or

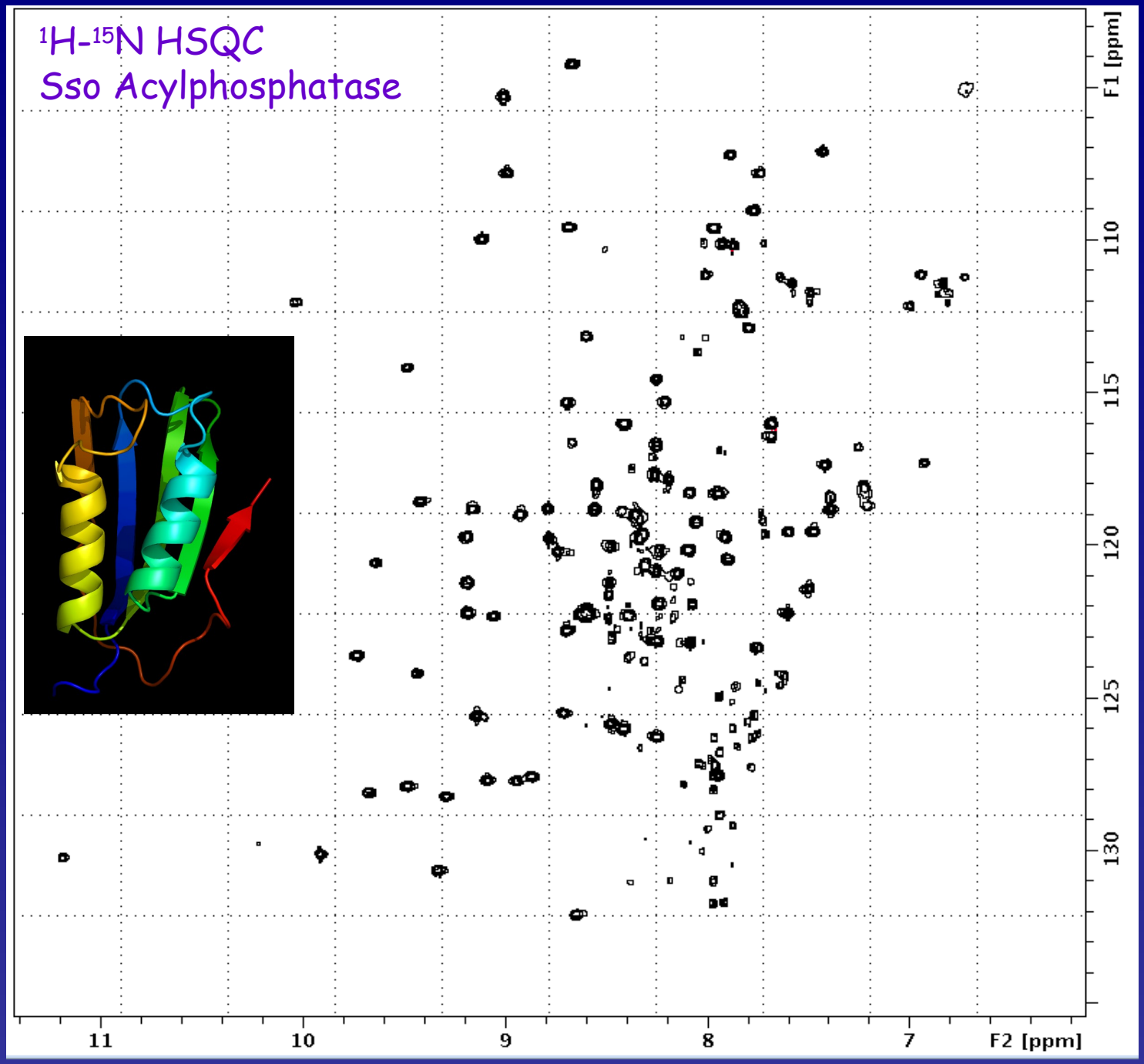
the **chemical shift** describes the dependence of nuclear magnetic energy levels on the electronic environment in a molecule.

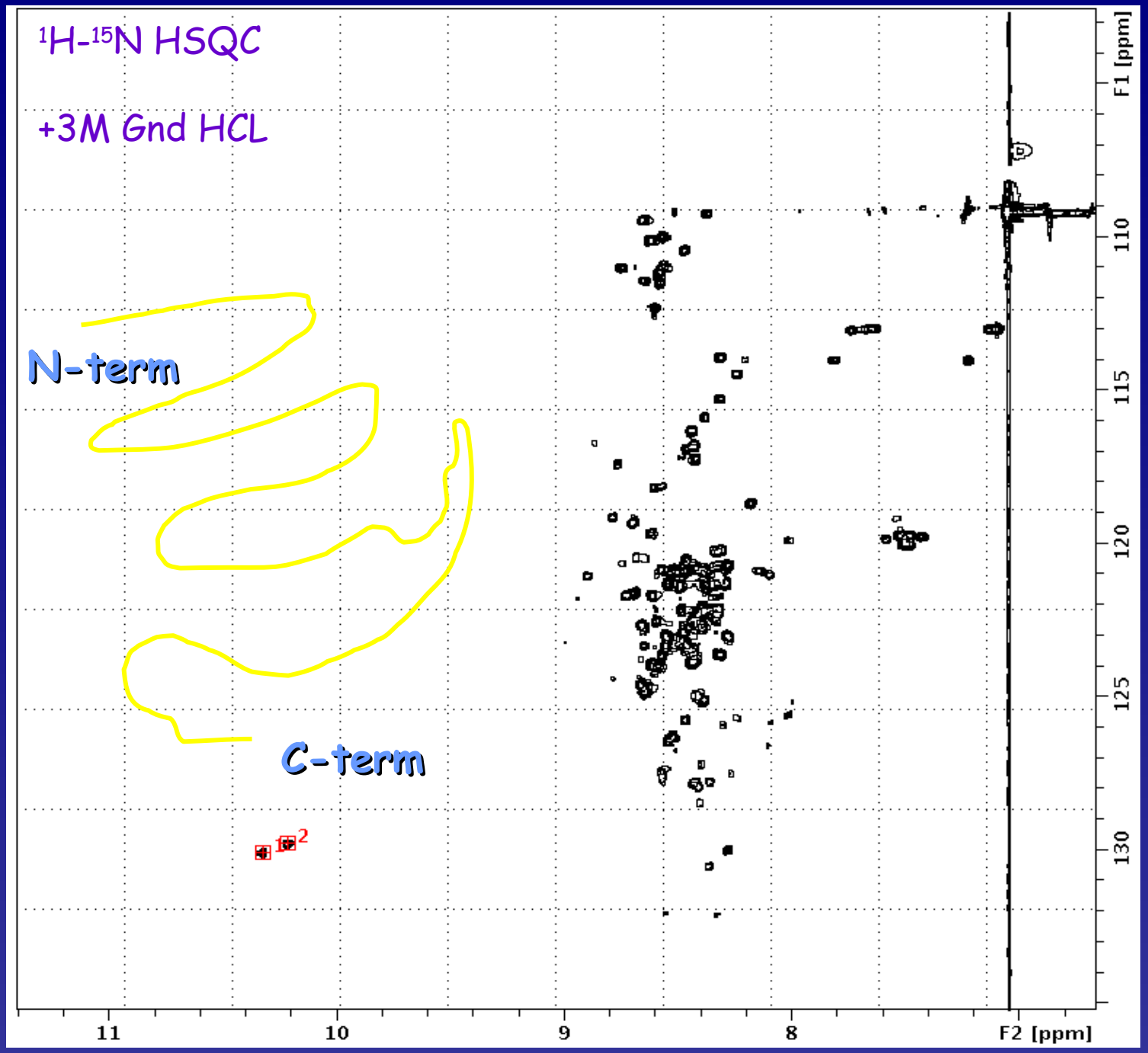


Chemical shift origin

Factors influencing the chemical shift:

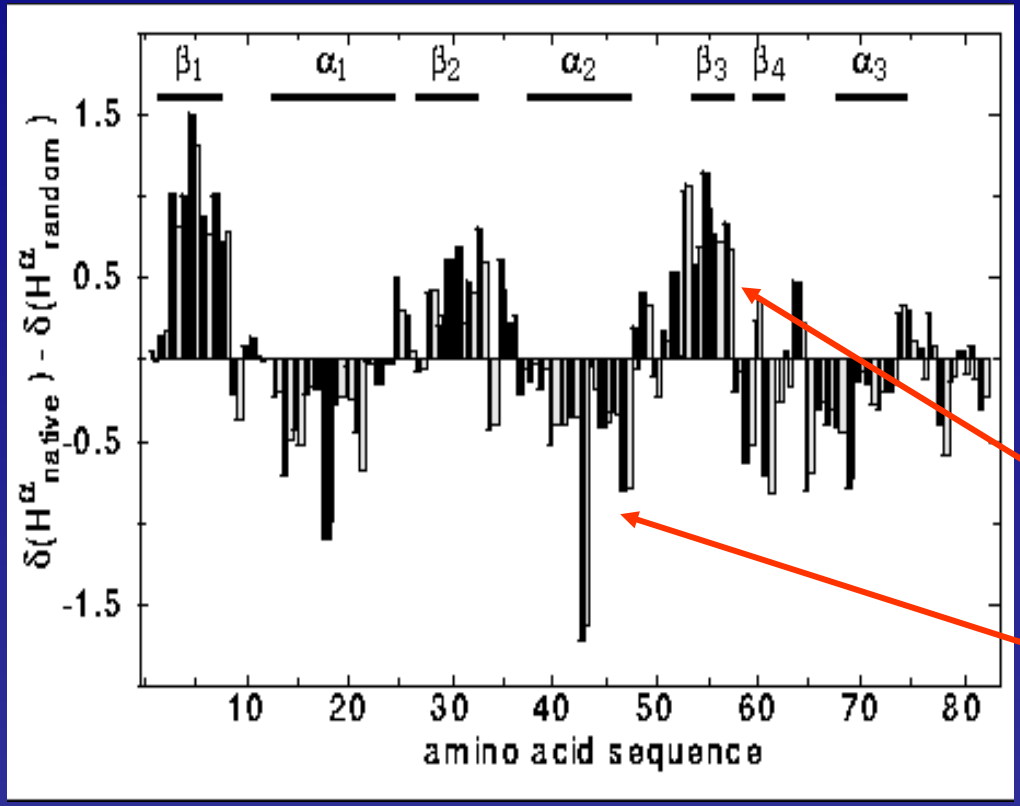
- nucleus shielding (electronegativity of the bound nuclei)
- presence of paramagnetic nuclei
- ring current effect (aromatic groups)
- chemical shift anisotropy (mediated in liquids)
- local electrostatic fields
- solvent


^{1}H-^{15}N HSQC
Sso Acylphosphatase
F1 [ppm]
110
115
120
125
130
^{15}N (ppm)
11
10
9
8
7
F2 [ppm]
^{1}H (ppm)


^{1}H-^{15}N HSQC
+3M Gnd HCL
N-term
C-term
F1 [ppm]
110
115
120
125
130
^{15}N (ppm)
11
10
9
8
F2 [ppm]
^{1}H (ppm)

Chemical shift index

As chemical shifts depend on the nucleus environment, it also contains structural information. Correlations between chemical shifts of Cα, Cβ ,CO, Hα and secondary structures have been identified.



Secondary chemical shift

Hα shift [measured – random coil] (Δ):

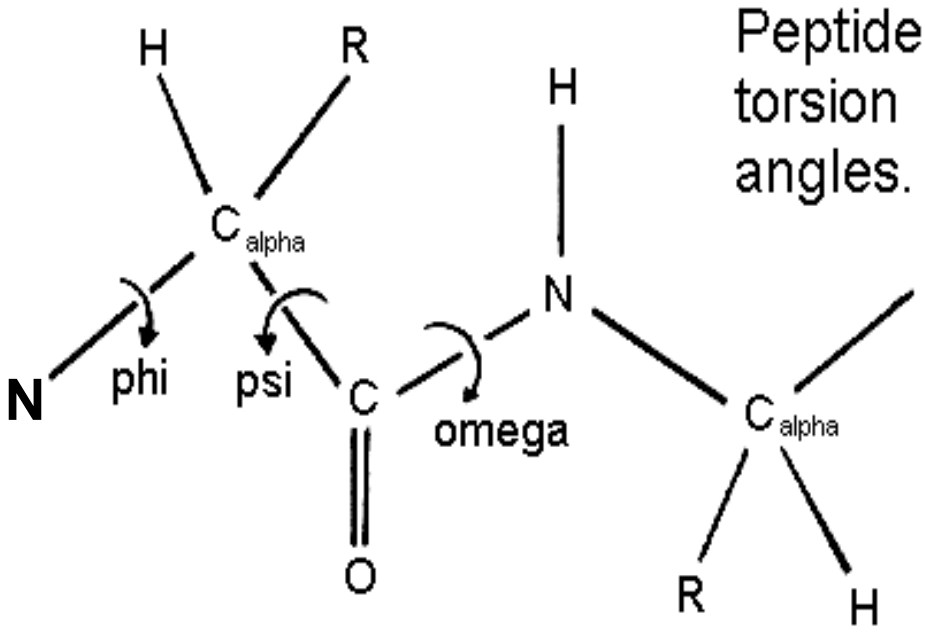
- > 0.7 ppm ⇒ CSI= 1
- - 0.7 < Δ < 0.7 ⇒ CSI= 0
- < - 0.7 ppm ⇒ CSI= -1

At least 4 consecutive residues with **CSI +1** ⇒ **β strand**.

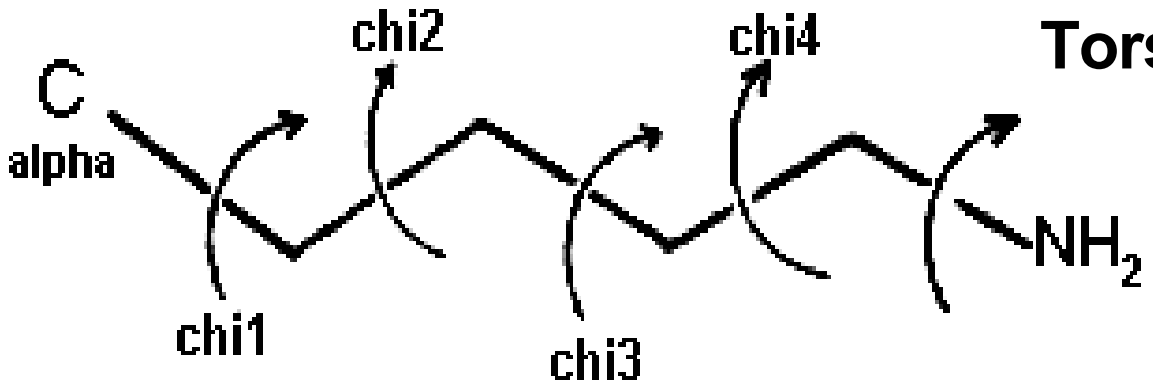
At least 4 consecutive residues with **CSI -1** ⇒ **α helix**.

All other regions are designated as coil

Protein structure and dihedral angles



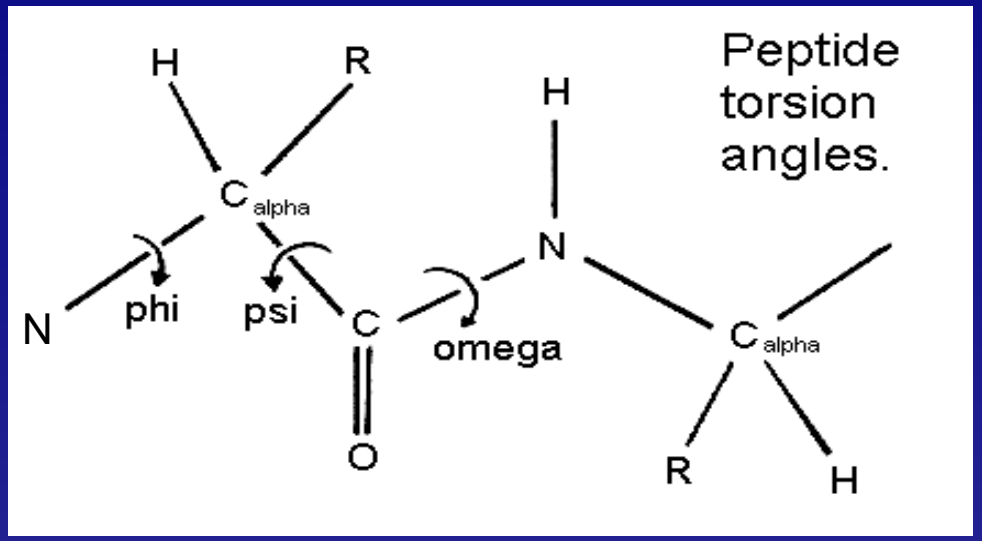
Peptide torsion angles.



Side chain Torsion angles.

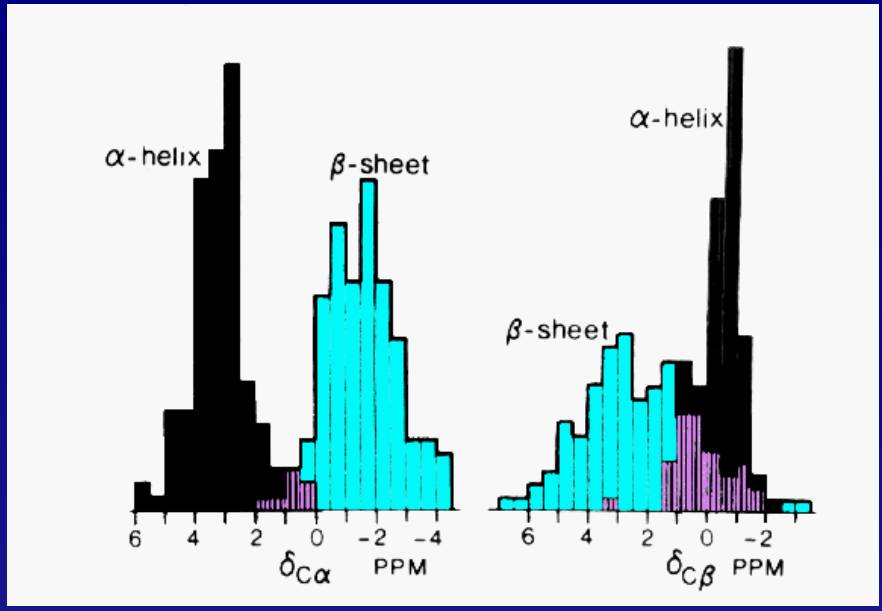
Protein Secondary Structure and backbone Chemical Shifts

- TALOS (http://spin.niddk.nih.gov/NMRPipe/talos/)
  - Given the Hα, Cα, Cβ, C', N chemical shift assignments and primary sequence



  - Compares the **secondary chemical shifts** against database of chemical shifts and associated high-resolution structure
    - comparison based on "triplet" of amino acid sequences present in database structures with similar chemical shifts and secondary structure
  - Provides potential ϕ, ψ backbone torsion constraints

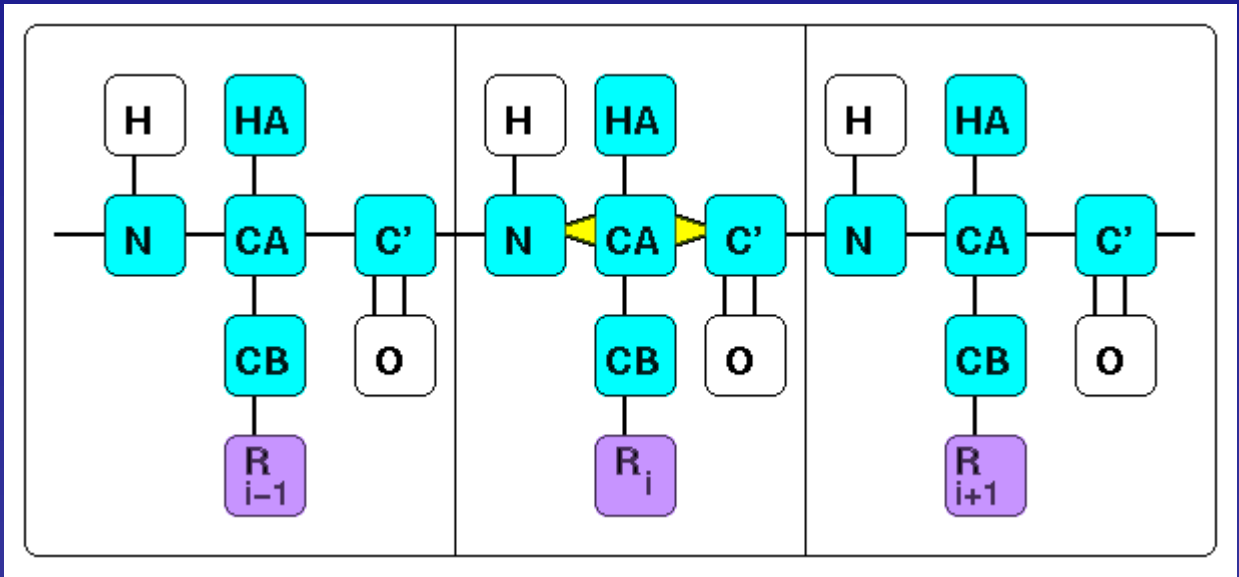
186 proteins in TALOS database



TALOS is a database system for empirical prediction of phi and psi backbone torsion angles using a combination of five kinds (HA, CA, CB, CO, N) of chemical shift assignments for a given protein sequence. The TALOS approach is an extension of the well-known observation that many kinds of secondary chemical shifts (i.e. differences between chemical shifts and their corresponding random coil values) are highly correlated with aspects of protein secondary structure. The goal of TALOS is to use secondary shift and sequence information in order to make quantitative predictions for the protein backbone angles phi and psi, and to provide a measure of the uncertainties in these predictions.

TALOS uses the secondary shifts of a given residue to predict phi and psi angles for that residue. TALOS also includes the information from the next and previous residues when making predictions for a given residue. So, in practice, TALOS uses data for three consecutive residues simultaneously (i.e. 15 total secondary shifts and 3 residue types) to make predictions for the central residue in a triplet.

The idea behind TALOS is that if one can find some triplet of residues in a protein of known structure with similar secondary shifts and sequence to a triplet in a target protein, then the phi and psi angles in the known structure will be useful predictors for the angles in the target.

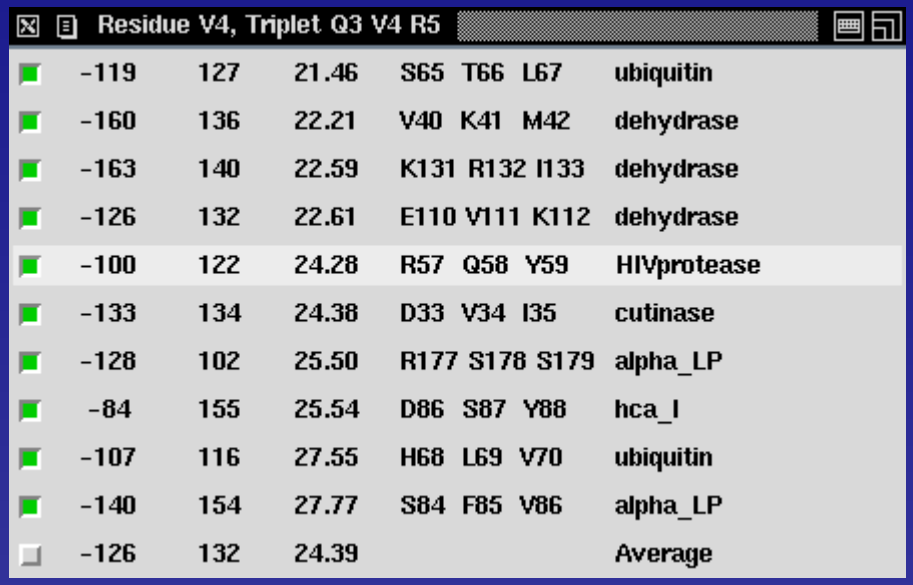


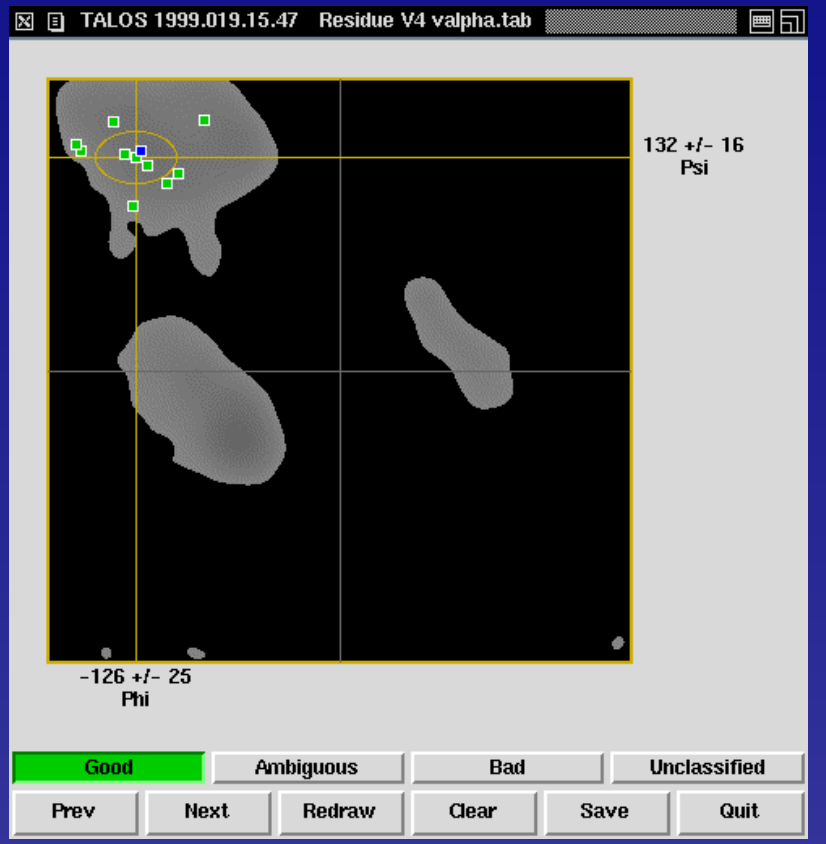
Green Good prediction (at most one outlier)

Yellow Ambiguous; no prediction

Red Bad prediction relative to a known structure

Gray No classification yet





TALOS reliability was tested by a cross-validation procedure

According to the tests:

- no predictions for 20% to 45% of the residues in a protein.
- predictions for about 72% of the residues on average.

In 45 out of 186 proteins studied, the TALOS results included no bad predictions ("bad" meaning substantially different from the crystal structure).

(IMPORTANT!) Over all 186 proteins, about 1.8% of the predictions made by TALOS were incorrect relative to the corresponding crystal structure.

Average uncertainty as reported by TALOS:

13.5 (12.9) degrees for phi, and 12.2 (12.4) degrees for psi.

(actual RMSD)

SPARTA Shen, Bax J. Biomol NMR (2007)

INVERSE PROBLEM:

protein structure is known ⟶ prediction of chemical shifts

SPARTA: empirical prediction of backbone chemical shifts (N, HN, HA, CA, CB, CO) from a given protein with known PDB coordinates.

The idea is that if one can find some triplet of residues in a protein of known structure with similar structure and sequence to a triplet in a target protein, then the backbone secondary chemical shifts for this protein will be useful predictors for the backbone secondary chemical shifts in the target.

How is the similarity measured? The similarity is measured as a score S(i,j) for a res i of the query protein and res j of the database:

$$S(i,j) = \sum_{n=-1}^{1} \left[k^{H}_{n,r} \Delta^{2}_{ResType} + k^{\phi}_{n} \left(\phi_{i+n} - \phi_{j+n} \right)^{2} + k^{\psi}_{n} \left(\psi_{i+n} - \psi_{j+n} \right)^{2} + k^{\chi^1}_{n} \Delta \chi^{1}_{i+n,j+n} \right]$$

Residue similariry

In practice, SPARTA searches a database for the 20 best matches to a given triplet in the target protein. The weighted averages chemical shifts of the central residues of these 20 matches are used as a prediction for the secondary shift of the central residue.

The SPARTA database was constructed using the most well-defined parts of high resolution (2.4 Angstroms or better) X-ray crystal structures to define the phi, psi and chi1 angles, as well as other structural information, such as hydrogen bonding and ring current shifts, which would be used to quantitatively correct the raw predicted shifts from database searching. This database currently includes data from 200 proteins, representing 24,166 triplets.

SPARTA reliability was tested by a cross-validation procedure

The RMS deviations in ppm:

N	HN	HA	CA	CB	CO
2.36	0.46	0.25	0.88	0.97	1.01

Ring current shifts and hydrogen bonding, is also considered.

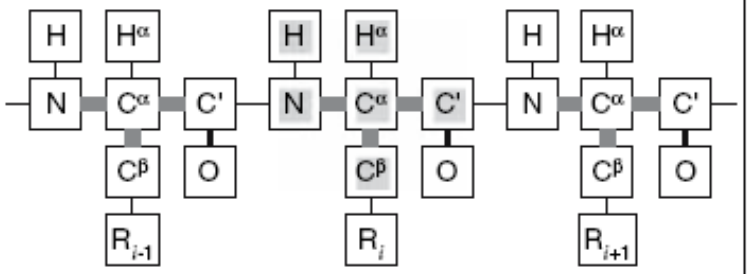
The secondary shifts in the SPARTA database are actually the **corrected shifts** using the calculated ring current shifts from PDB coordinates. The SPARTA predicted shifts for target protein are also corrected by adding the calculated ring current shifts from target protein. For HA and HN, the SPARTA-predicted secondary shifts are also corrected considering hydrogen bonds.

Protein backbone chemical shifts are extremely sensitive to the local conformation; therefore, SPARTA results for the residues in the flexible region or the with very large ring current shifts contribution may be less reliable.

SPARTA FLOWCHART

RCSB COORDINATES

Tripeptide (i–1,i,i+1)



200 PROTEINS

NMR chemical shifts from BMRB
X-Ray structures from RCSB (≤2.4Å)

TRIPEPTIDE DATABASE
(Pre-defined)

24166 Tripeptides (i-1, i, i+1)

ϕ, ψ and χ^1 angles

Secondary chemical shifts $\Delta\delta_r$
($\Delta\delta_r = \delta_r - \delta_{r,coil} - \delta_{r,prev} - \delta_{r,next} - \delta_{r,ring}$)

SPARTA PROGRAM

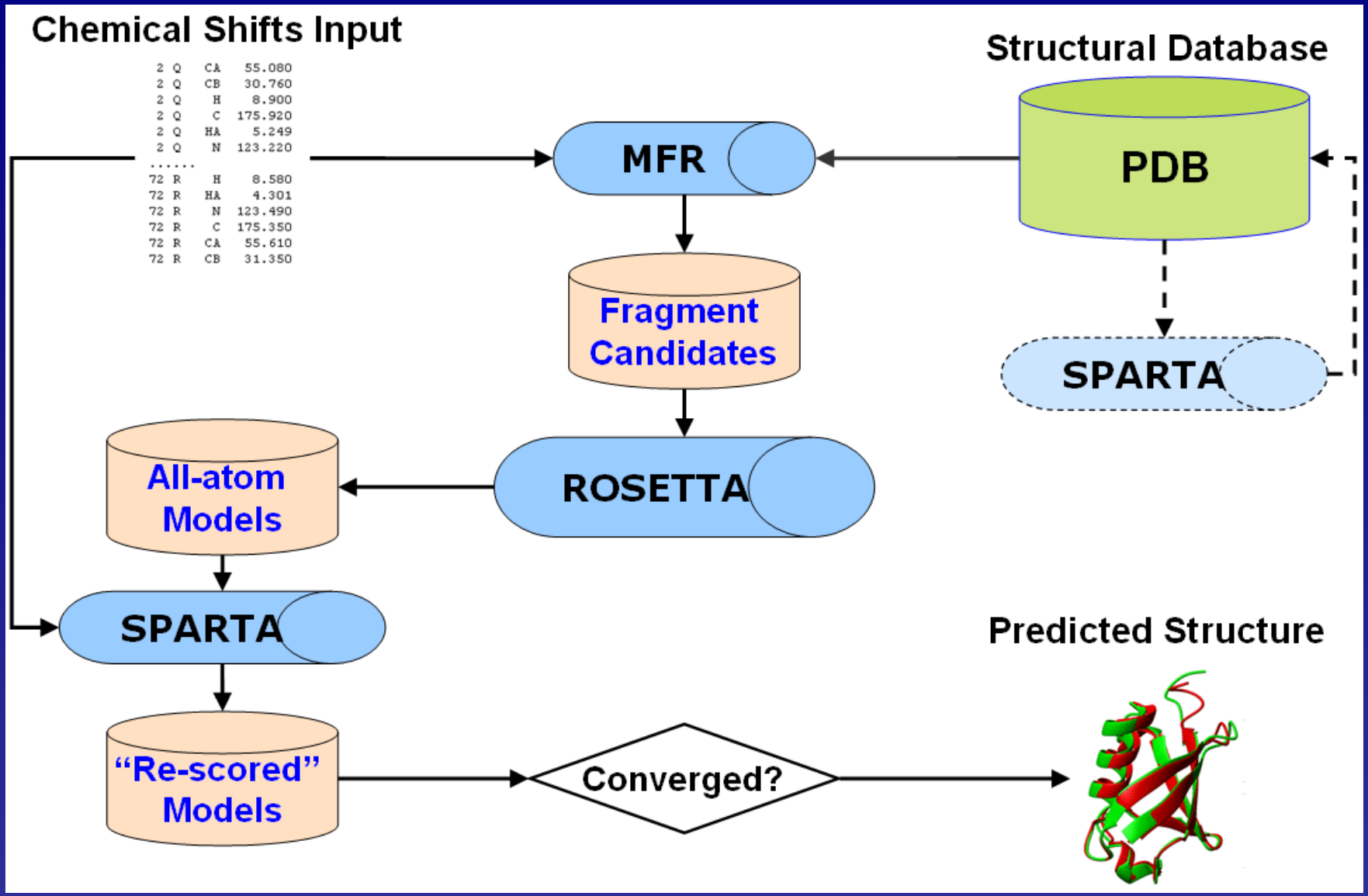
1. Calculate $\phi/\psi/\chi^1$ angles for (i–1, i, i+1), and ring current shifts $\delta_{r,i,ring}$ and H-bond for i
2. Calculate differences in ϕ/ψ and residue-type for residues (i-1,j-1), (i,j), (i+1,j+1), and overall similarity score, $S(i,j,r)$
3. Calculate average secondary shifts of residue j of the 20 lowest scores, $\langle\Delta\delta_{r,j}\rangle$
4. Add calculated ring current shifts $\delta_{r,i,ring}$, and $\delta_{r,i,coil}$, $\delta_{r,i,prev}$, $\delta_{r,i,next}$, to $\langle\Delta\delta_{r,j}\rangle$
5. Apply H-bond correction to $^1H^N$ and $^1H^\alpha$

Predicted chemical shifts $\delta_{r,i}$ for residue i
(r = ^{15}N, $^1H^N$, $^1H^\alpha$, $^{13}C^\alpha$, $^{13}C^\beta$, $^{13}C'$)

CS-ROSETTA

Chemical-Shift-ROSETTA (CS-ROSETTA) is a program that using as a sole input NMR chemical shift (13Cα, 13Cβ, 13C' , 15N, 1Hα and 1HN) generate protein structure.
Once the protein is doubly labeled (^{13}C and ^{15}N) backbone chemical shifts are generally available at the early stage of the NMR structure determination procedure, prior to the collection and analysis of structural restraints.
CS-ROSETTA approach, utilizes SPARTA-based selection of protein fragments from the PDB, in conjunction with a regular ROSETTA Monte Carlo assembly and relaxation method.

16 proteins, from 56 to 129 residues yielded full atom models that have 0.7-1.8 angstrom root-mean-square deviations for the backbone atoms relative to the experimentally determined X-ray or NMR structures.
This protocol potentially provides a new direction for high-throughput NMR structure determination, in particular in structural genomics.

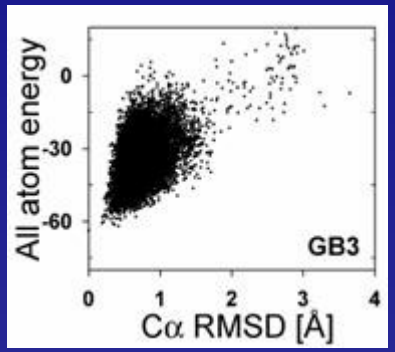

Chemical Shifts Input
2 Q CA 55.080
2 Q CB 30.760
2 Q H 8.900
2 Q C 175.920
2 Q HA 5.249
2 Q N 123.220
.....
72 R H 8.580
72 R HA 4.301
72 R N 123.490
72 R C 175.350
72 R CA 55.610
72 R CB 31.350
Structural Database
MFR
PDB
Fragment Candidates
SPARTA
All-atom Models
ROSETTA
SPARTA
Predicted Structure
"Re-scored" Models
Converged?

After finishing CS-ROSETTA structure generation, users have to decide whether the ROSETTA models are acceptable. For this purpose, it is convenient to plot the "landscape" of (re-scored) ROSETTA full-atom energies of all models with respect to their C_alpha RMSD values relative to the lowest-energy model.

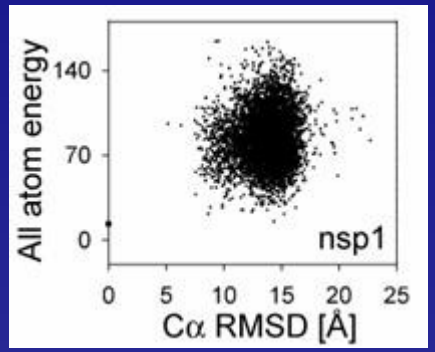
3. If the low energy models cluster within less than Cα RMSD of about 2 angstrom from the model with the lowest (re-scored) energy the structure prediction is successful and the 10 lowest energy models are accepted.

5. If no clustering around the low energy model is observed, the structure prediction has not converged and the low energy models can not be accepted.

1



2



By using the current method implemented in CS-ROSETTA package, 5,000 to 20,000 predicted CS-ROSETTA models are generally required to obtain the convergence. For small proteins (<= 90-100 amino acids), 1,000 to 5,000 predicted CS-ROSETTA models often sufficient. ROSETTA takes about 5-10 minutes to calculate one all-atom model on a single 2.4GHz CPU.

ROSETTA

Philosophy:

Try to mimic the relationship between local and global interactions in determining protein structure.

The final structure is obtained when fluctuation of local structures come together in a compact conformation (hydrophobic core, paired b strands, side chain interactions).

A library of fragments represent the range of possible local structures for all short sequences of the polypeptidic chain.

Strategy:

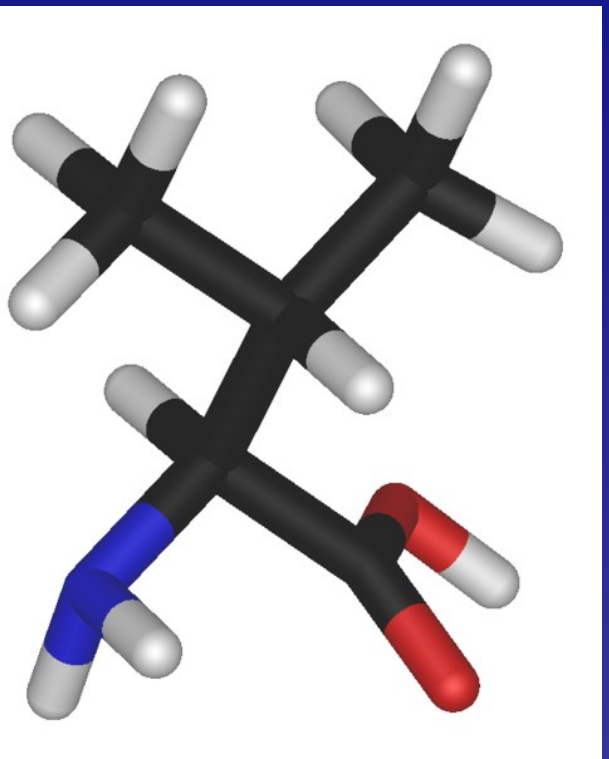
Selection, based on homology, of 200 fragments (9 residues long) and of 200 fragments (3 residues long) from a selected database

Compact structures are assembled by randomly combining the fragments using a Monte Carlo simulated annealing search. A scoring function that accounts for non local interactions (compactness, hydrophobic burial, strand pairing, …) is minimized.

Pseudo atom

- Degenerate pairs of methylene protons, QB
- Methyl groups QB (ala), QG1/QG2 (val), etc.
- Degenerate pairs of methyl groups QQG (Val), QQD (Leu)
- Phe/Tyr aromatic ring protons QD, QE

Pseudo Atom
Valine Example



QG1 QG2

CG1 CG2

QQG
CG

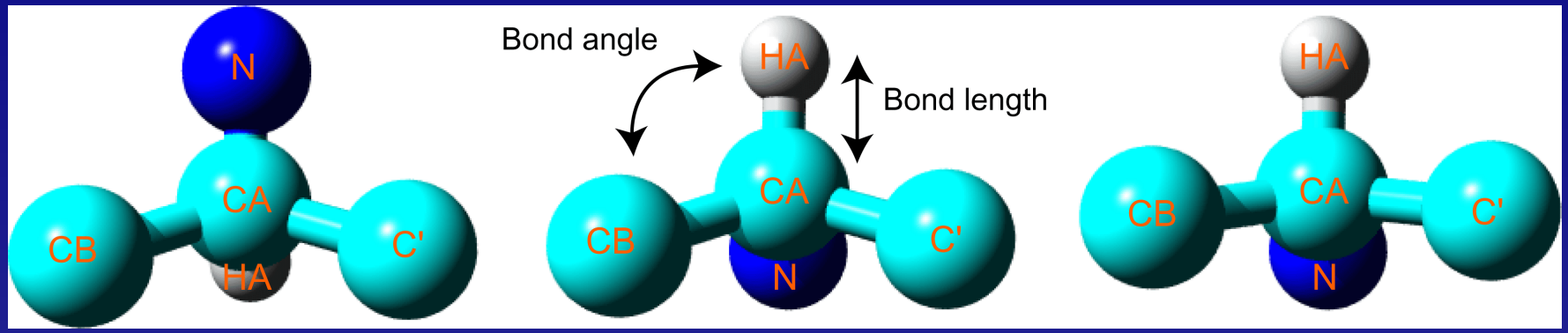
Structure quality through PROCHECK

- Covalent geometry
- Torsion angles
- Chirality
- Planarity
- Precision
- Restraint violations

Results are presented as plots suitable for publication

Laskowski R A, MacArthur M W, Moss D S & Thornton J M (1993). *J. Appl. Cryst.*, 26, 283-291.

Bonded geometry

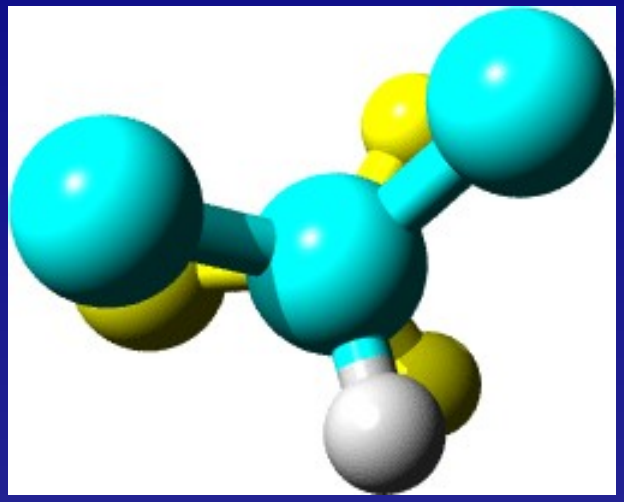


D-amino acid

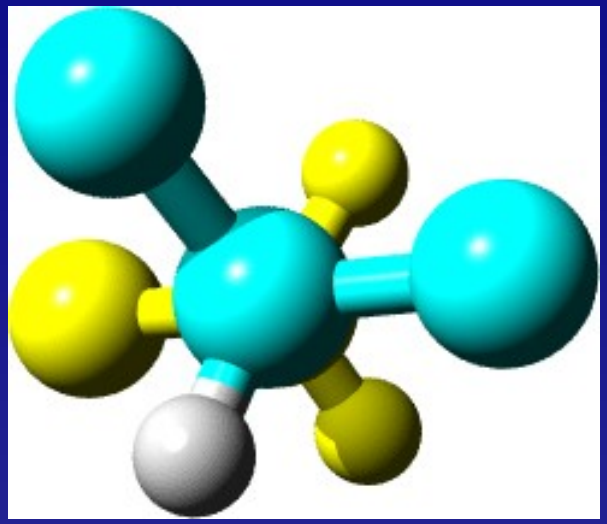
L-amino acid

Distorted Cα-chirality

Rotameric states

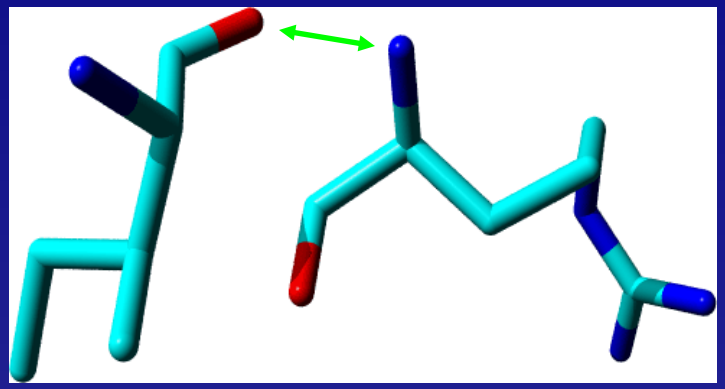


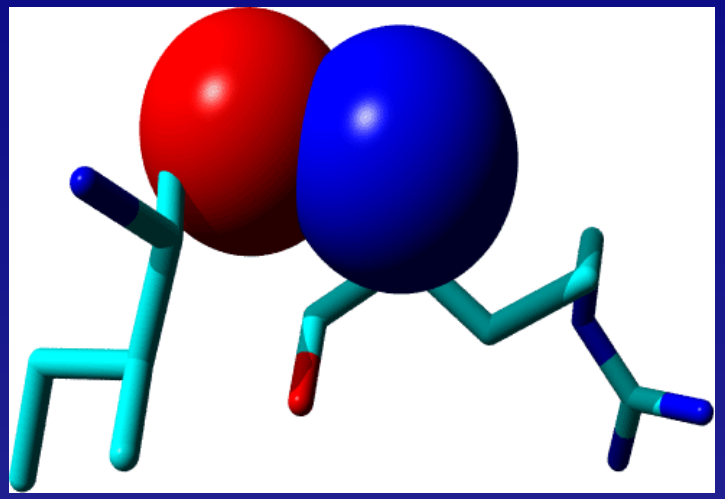
Eclipsed



Staggered

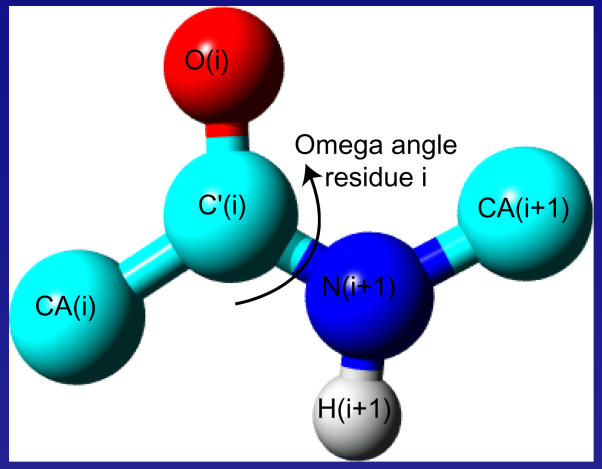
Inter-atomic bumps

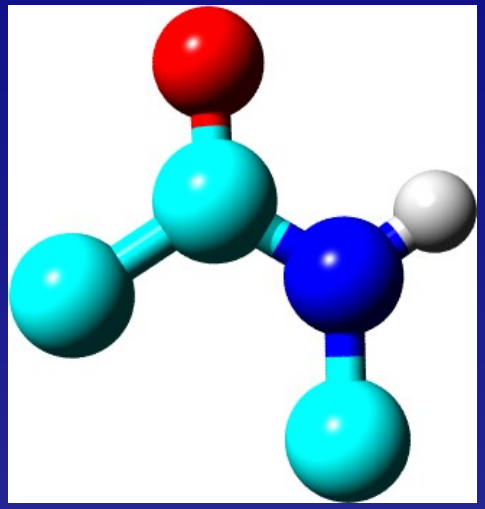




Overlap of two backbone atoms

Omega angles

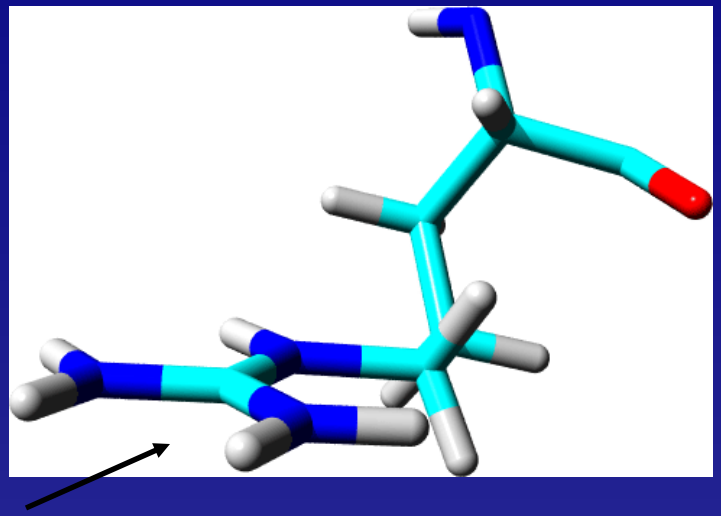




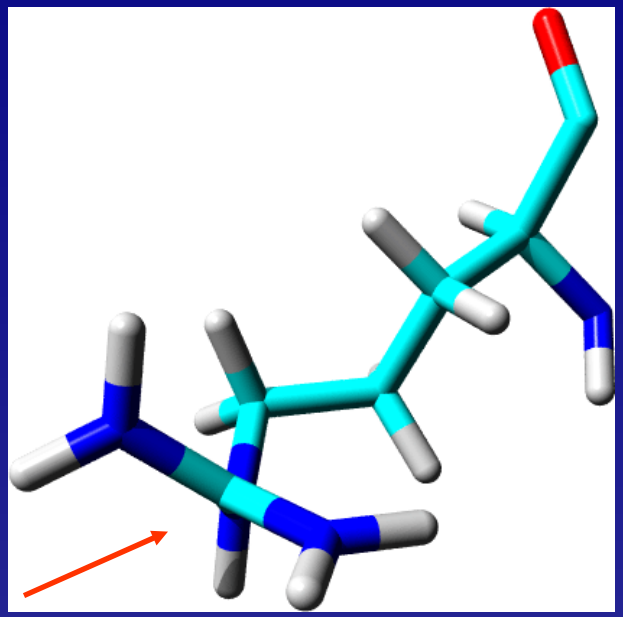
***Trans*-conformation**
(omega=180°)

***Cis*-conformation**
(omega=0°)

Side chain planarity

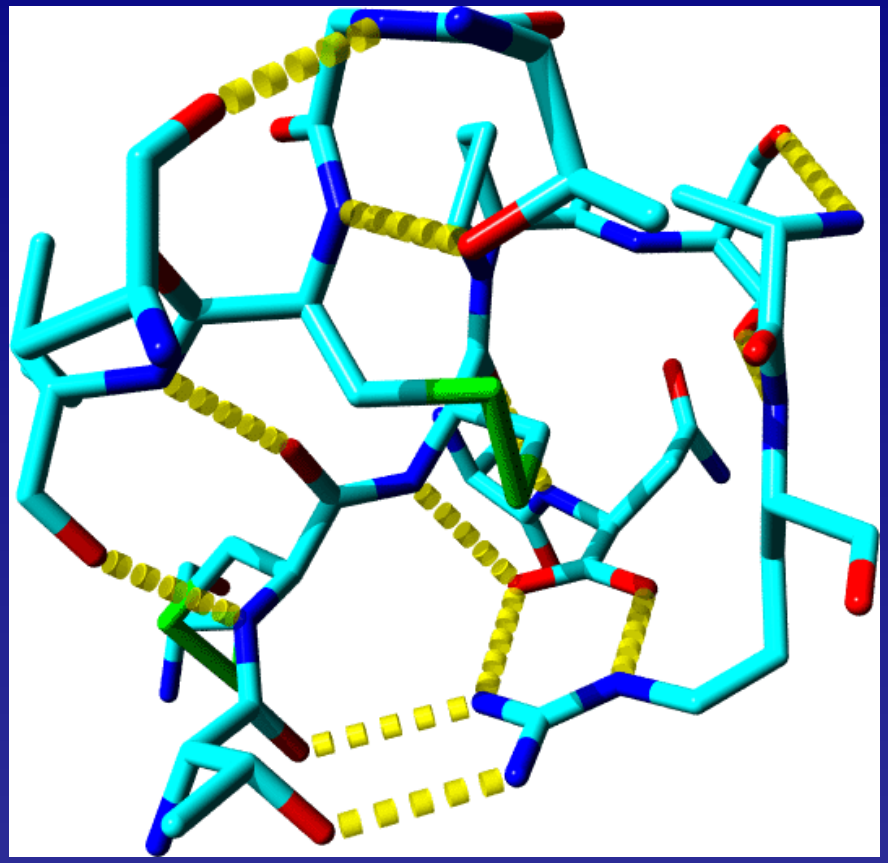


Planar ARG side-chain
(Good)



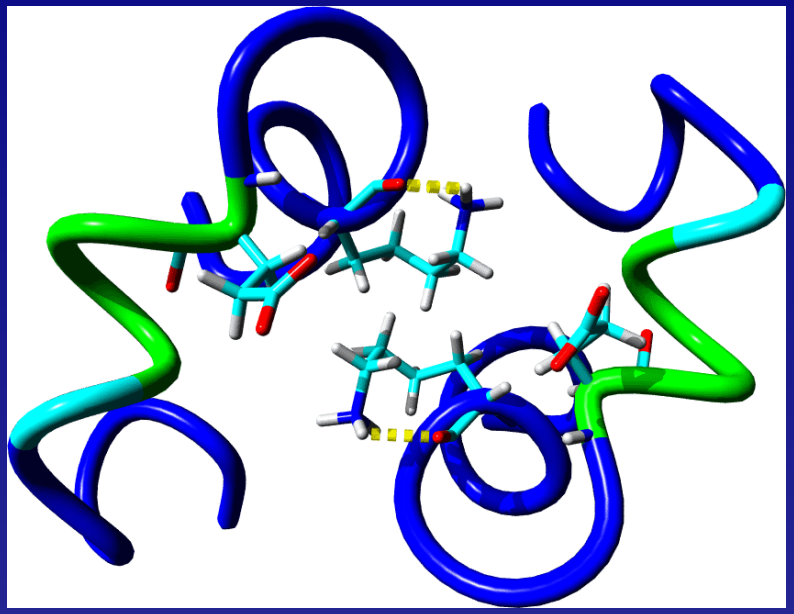
Non-planar ARG side-chain
(Bad)

Internal hydrogen bonding

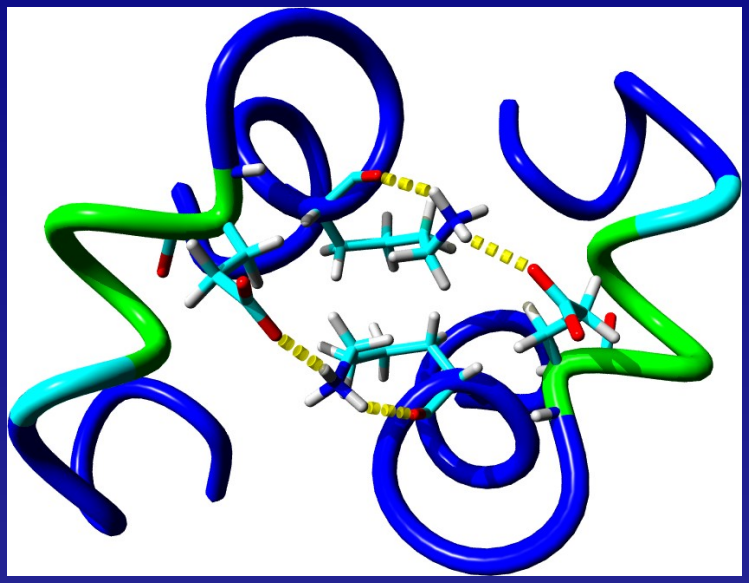


Internal hydrogen bonding in Crambin

Electrostatics

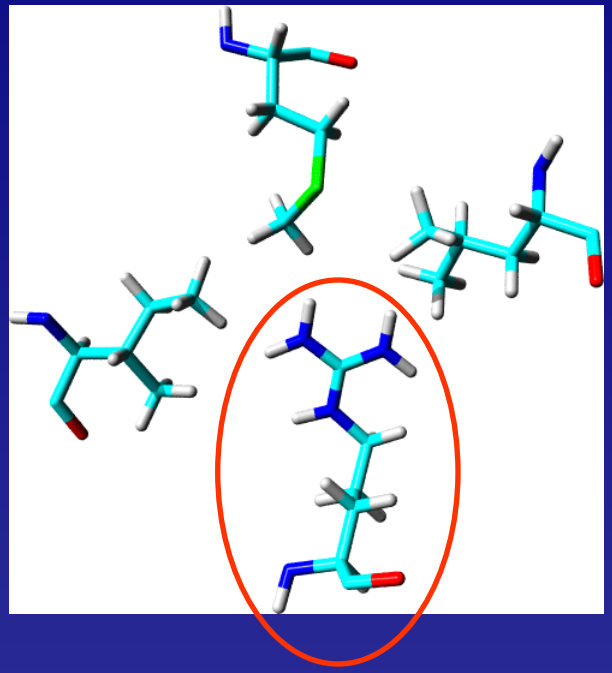


"Bad" electrostatics

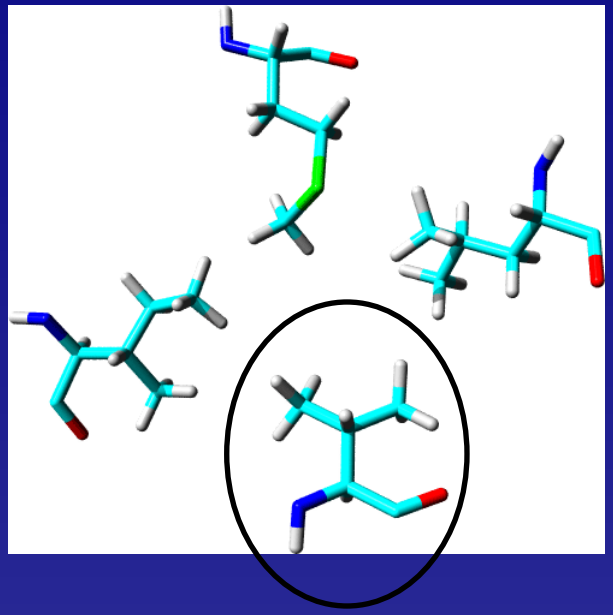


After energy minimization including electrostatics

Packing quality

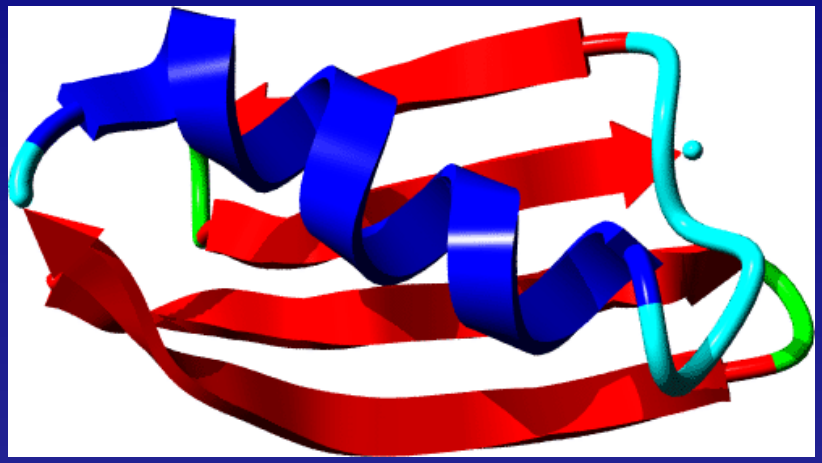


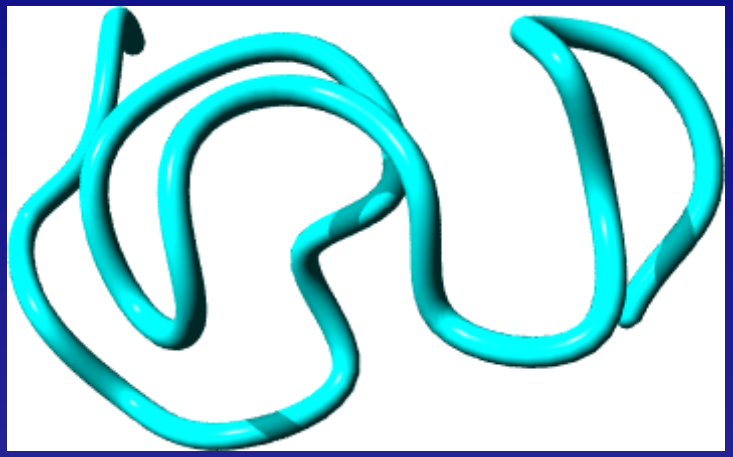
Bad packing



Good packing

Backbone Conformation





Very normal

Very unique