Analysis of MD trajectories in GROMACS

David van der Spoel

What does MD produce?

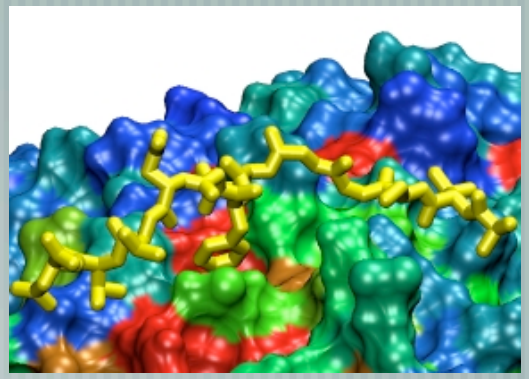
- Energy terms E(t)
- Coordinates x(t)
- Velocities v(t)
- Forces f(t)

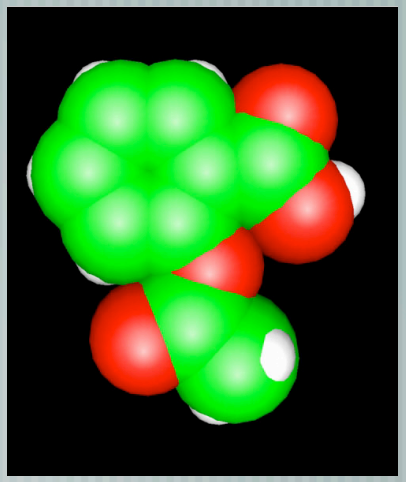
Managing your files

- **trjcat** - merging trajectories
  - concatenating
  - demultiplexing REMD
- **trjconv** - converting trajectories
  - scaling, translating, rotating
  - periodic boundary conditions
  - least squares superposition
- **eneconv** - merging and converting energy files

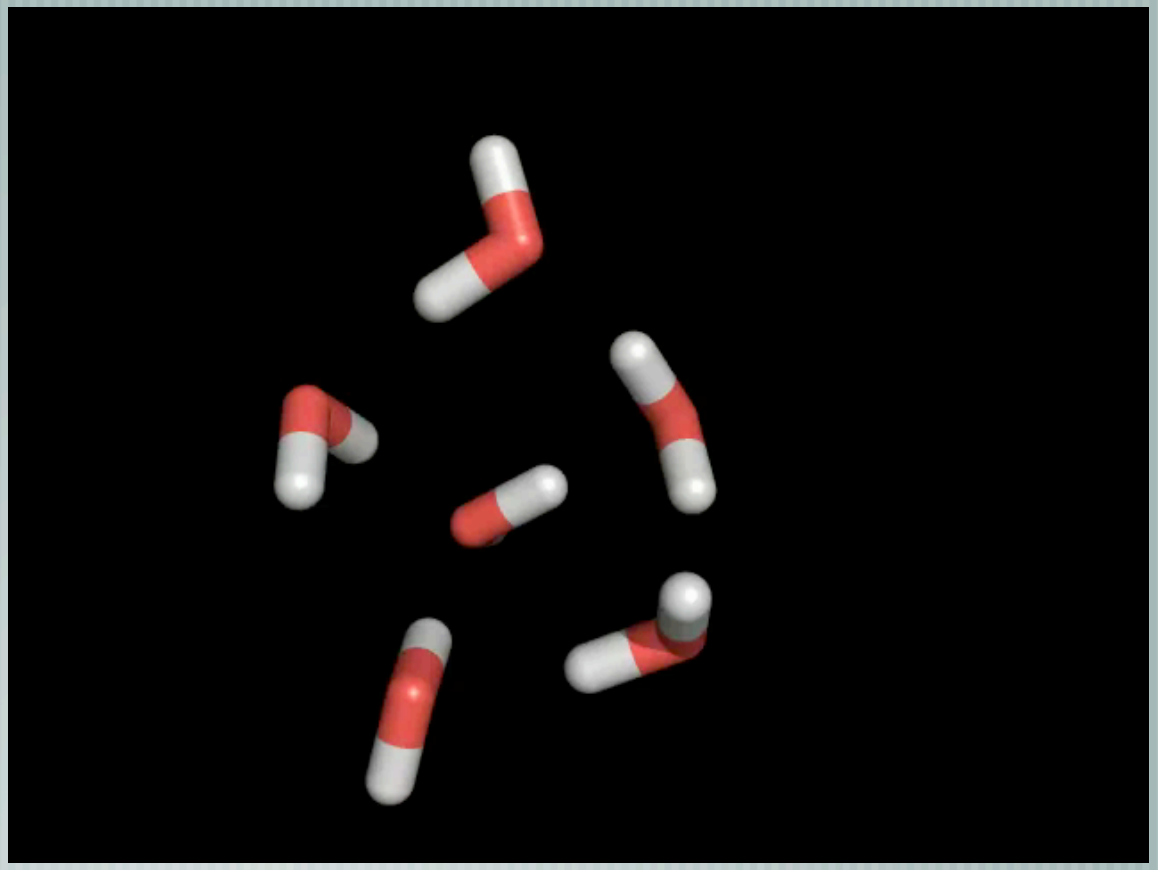
Viewing your trajectory

- **ngmx** - part of GROMACS
- VMD - http://www.ks.uiuc.edu
- gOpenMol - http://www.csc.fi/gopenmol
- pymol - http://pymol.sf.net



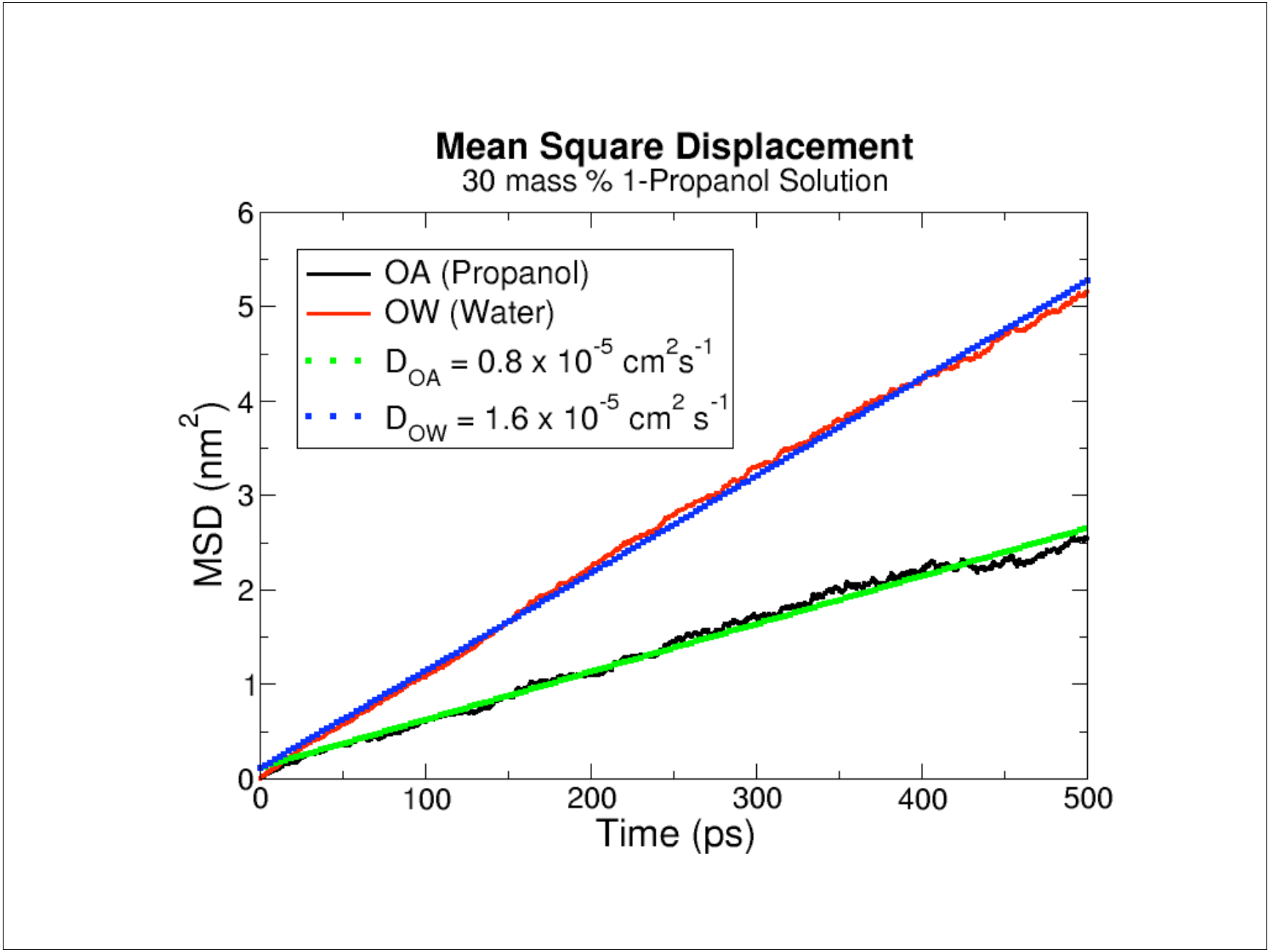


Dynamics = Movies



Quantifying Dynamics

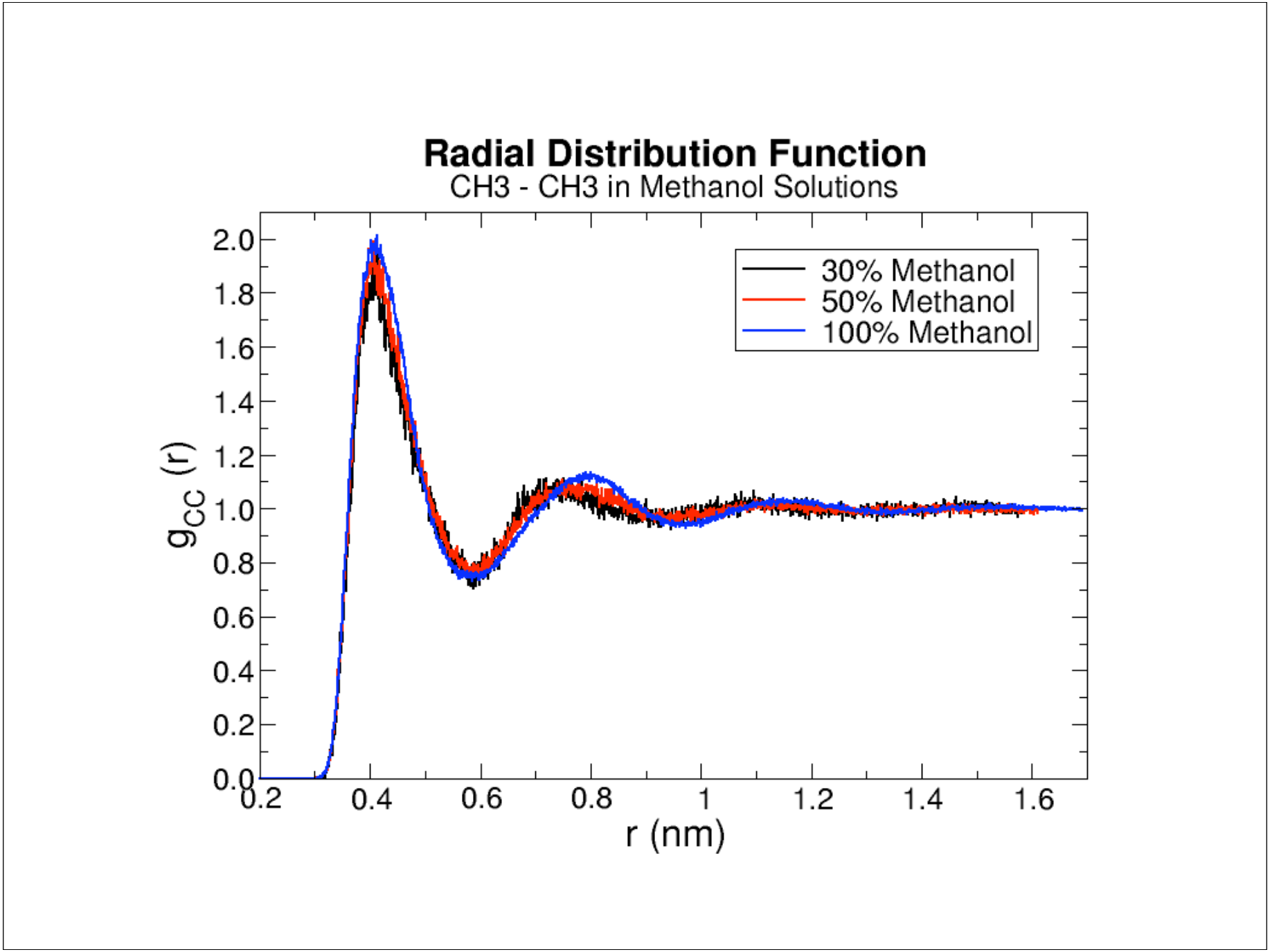
- Einstein relation: $6Dt = \lim_{t \to \infty} (\mathbf{r} - \mathbf{r}_0)^2$
- **g_msd** - Mean Square Displacement and diffusion constants
- Green-Kubo relation: $3D = \int_0^\infty \mathbf{v}(t) \cdot \mathbf{v}(0) dt$
- **g_velacc** - Velocity Autocorrelation Function


Mean Square Displacement
30 mass % 1-Propanol Solution
OA (Propanol)
OW (Water)
D_OA = 0.8 x 10^-5 cm^2 s^-1
D_OW = 1.6 x 10^-5 cm^2 s^-1
MSD (nm^2)
Time (ps)
0
1
2
3
4
5
6
0
100
200
300
400
500

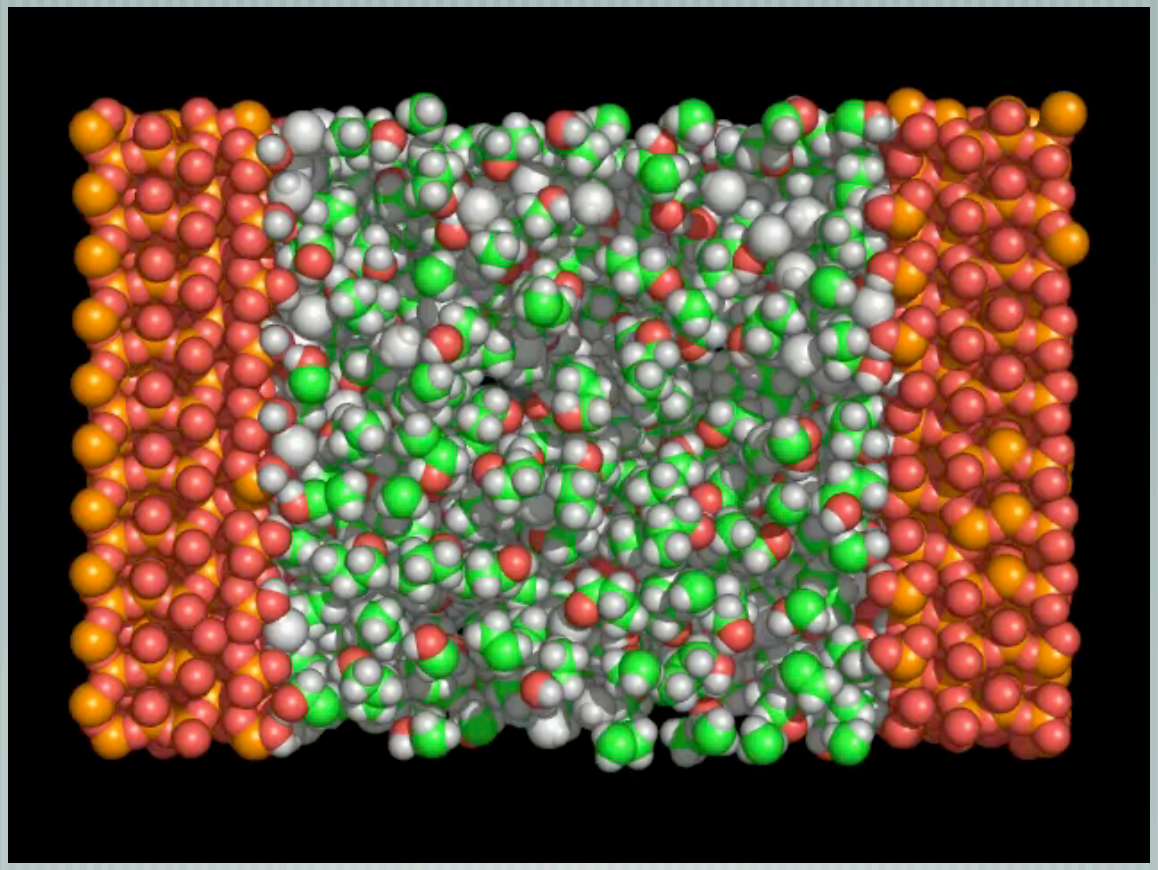
Liquid Structure

- g_rdf - Radial Distribution Function gives the local density of species B around species A relative to the average density

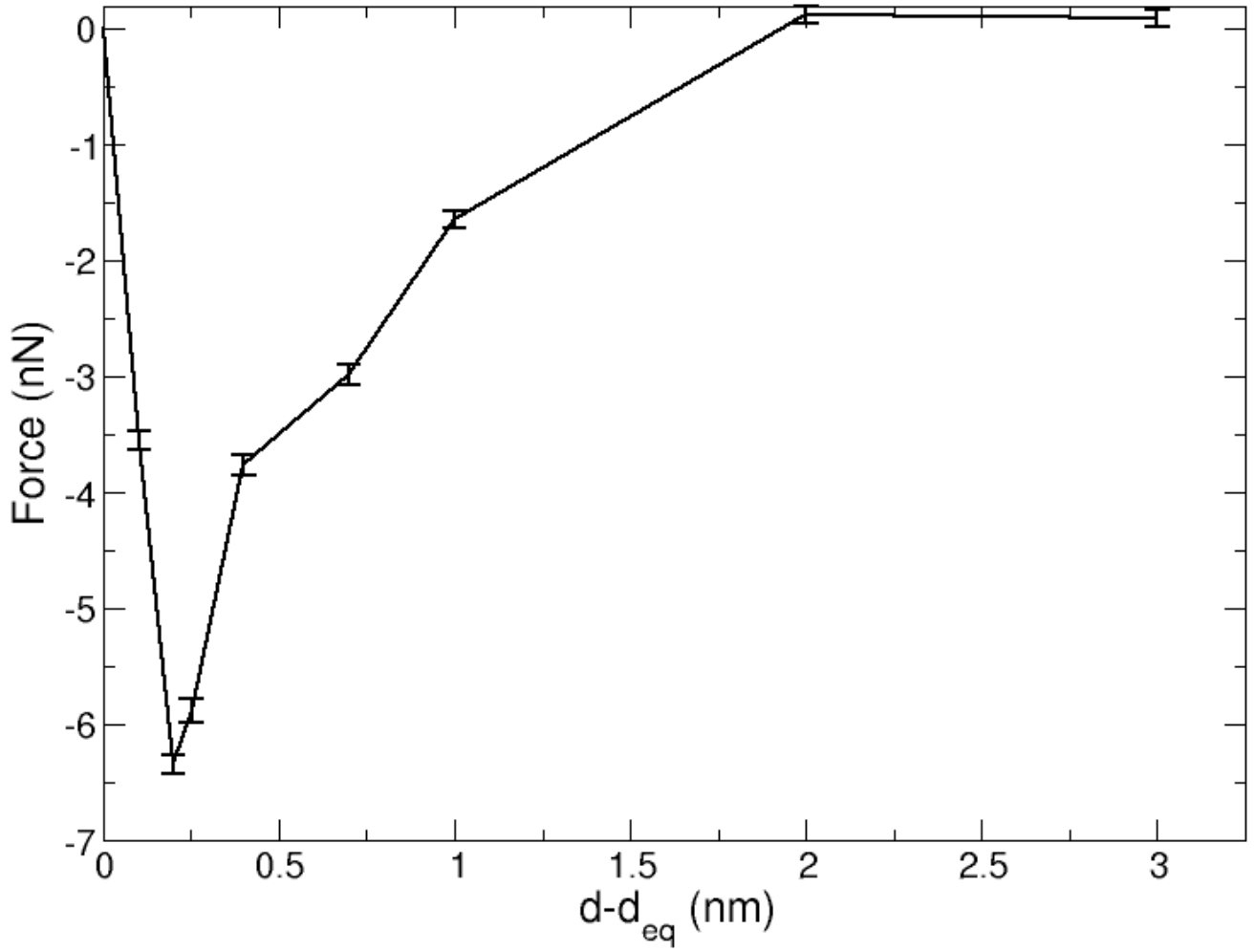
$$g_{AB}(r) = \frac{\rho_B(r)}{\langle \rho_B \rangle} = \frac{1}{N_A \langle \rho_B \rangle} \sum_{i \in A}^{N_A} \sum_{j \in B}^{N_B} \frac{\delta(r_{ij} - r)}{4\pi r^2}$$


Radial Distribution Function
CH3 - CH3 in Methanol Solutions
30% Methanol
50% Methanol
100% Methanol
g_{CC} (r)
r (nm)
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6

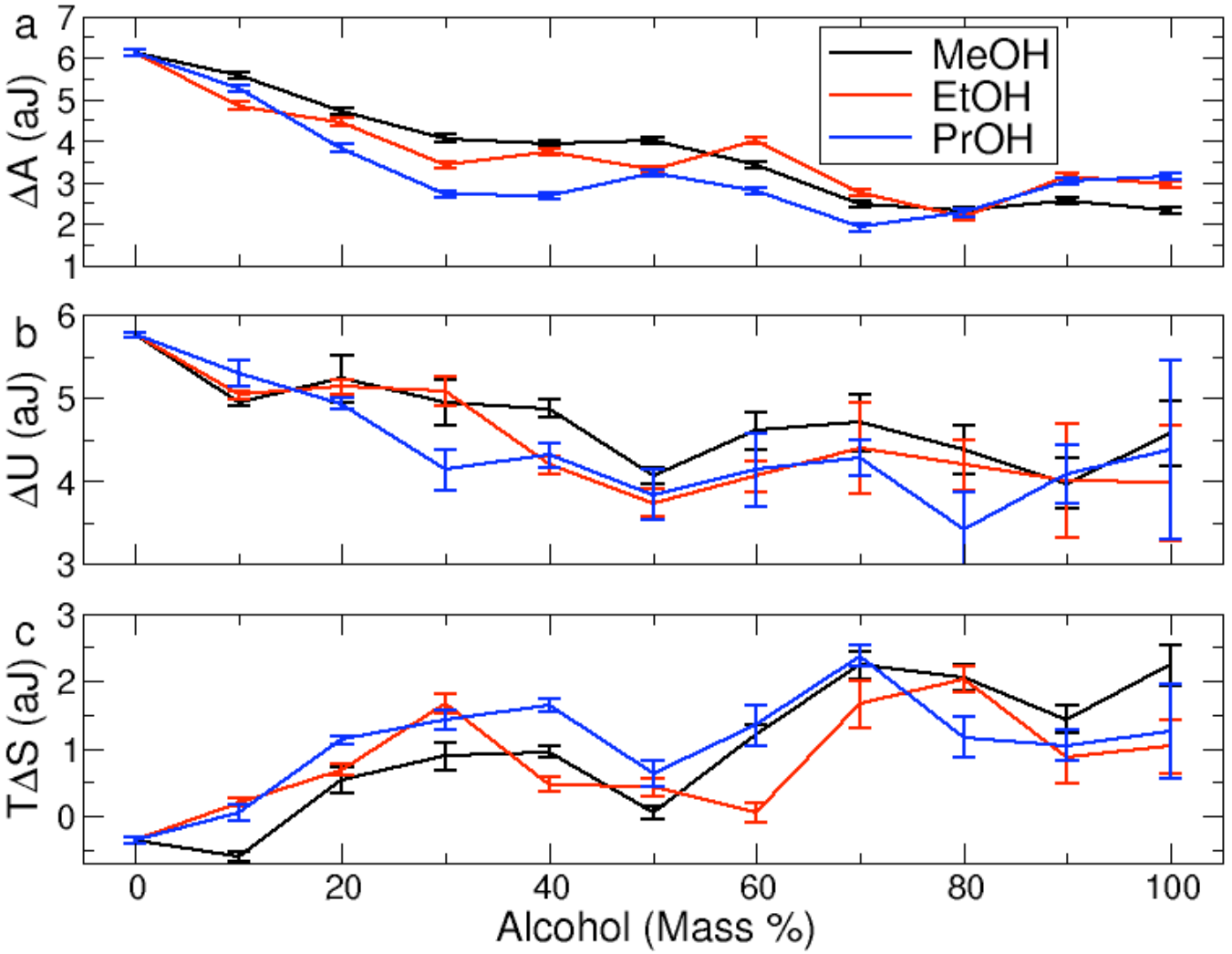
Forces between "particles"



30 Mass% Methanol solution between quartz particles



Integrated force until infinity yields Helmholtz energy



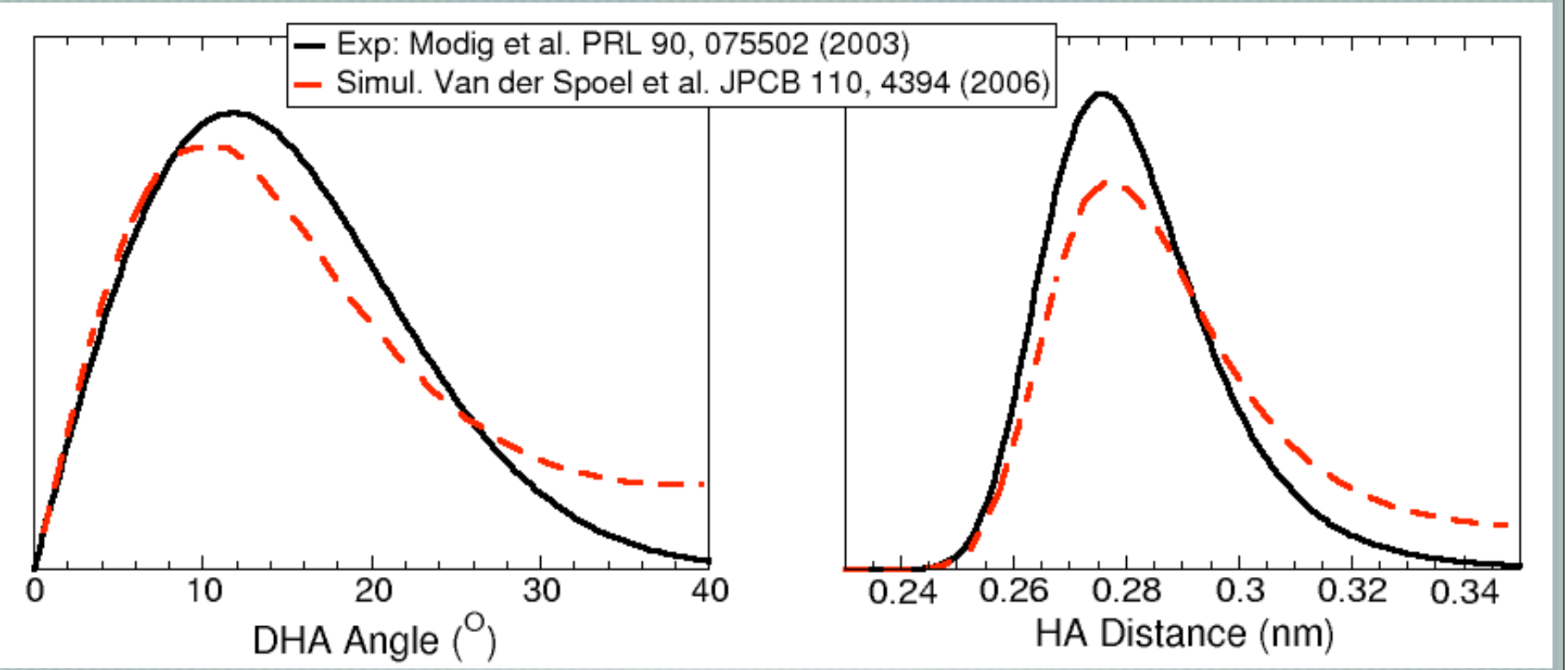
Using forces ctd.

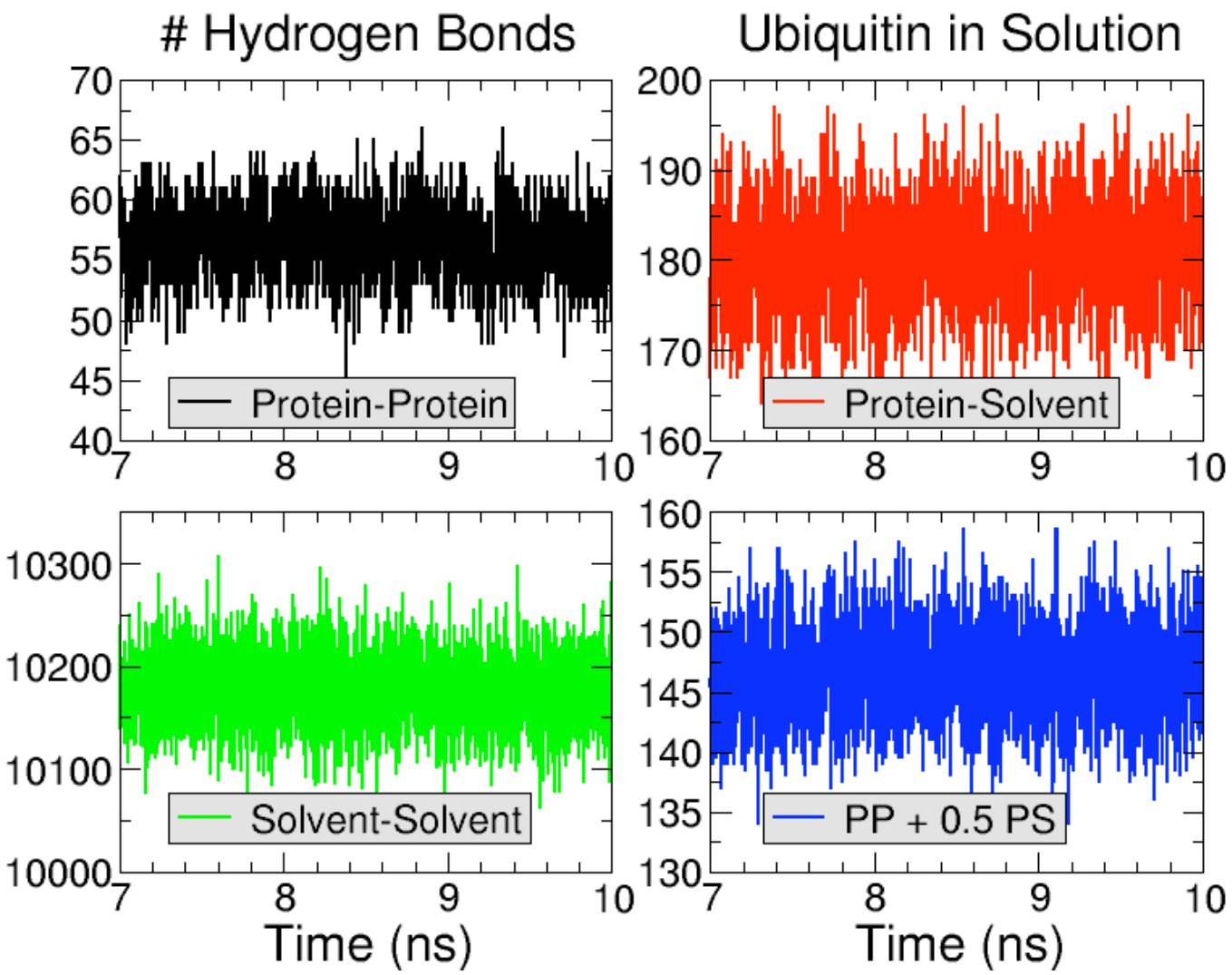
- Using the forces directly from MD is not very accurate due to large fluctuations
- Other methods to compute potential of mean force (PMF) are implemented in GROMACS:
  - Atomic Force Microscopy (AFM) Pulling, where you gently pull on a molecule
  - Umbrella sampling

Hydrogen Bonds

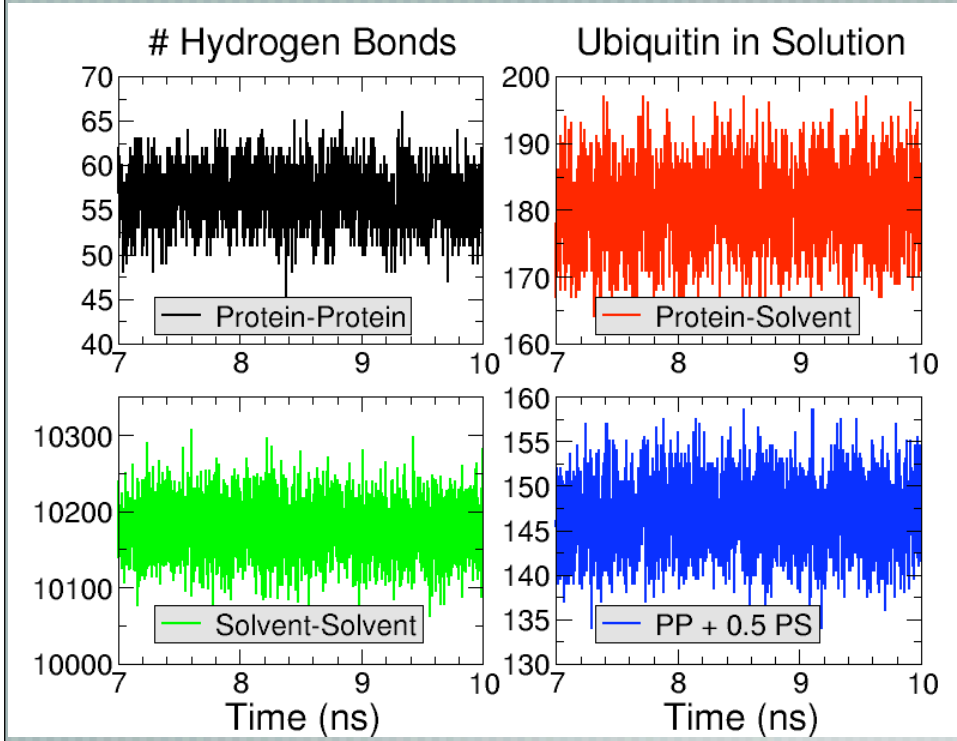
- Can be defined using either geometrical or energetic criterion
- g_hbond - geometric criterion only
  - $r < r_{HB}$ (typically 0.35 nm)
  - $alpha < alpha_{HB}$ (typically 30 degrees)

Hydrogen Bonds in Water




Hydrogen Bonds
Ubiquitin in Solution
Protein-Protein
Protein-Solvent
Solvent-Solvent
PP + 0.5 PS
Time (ns)
Time (ns)

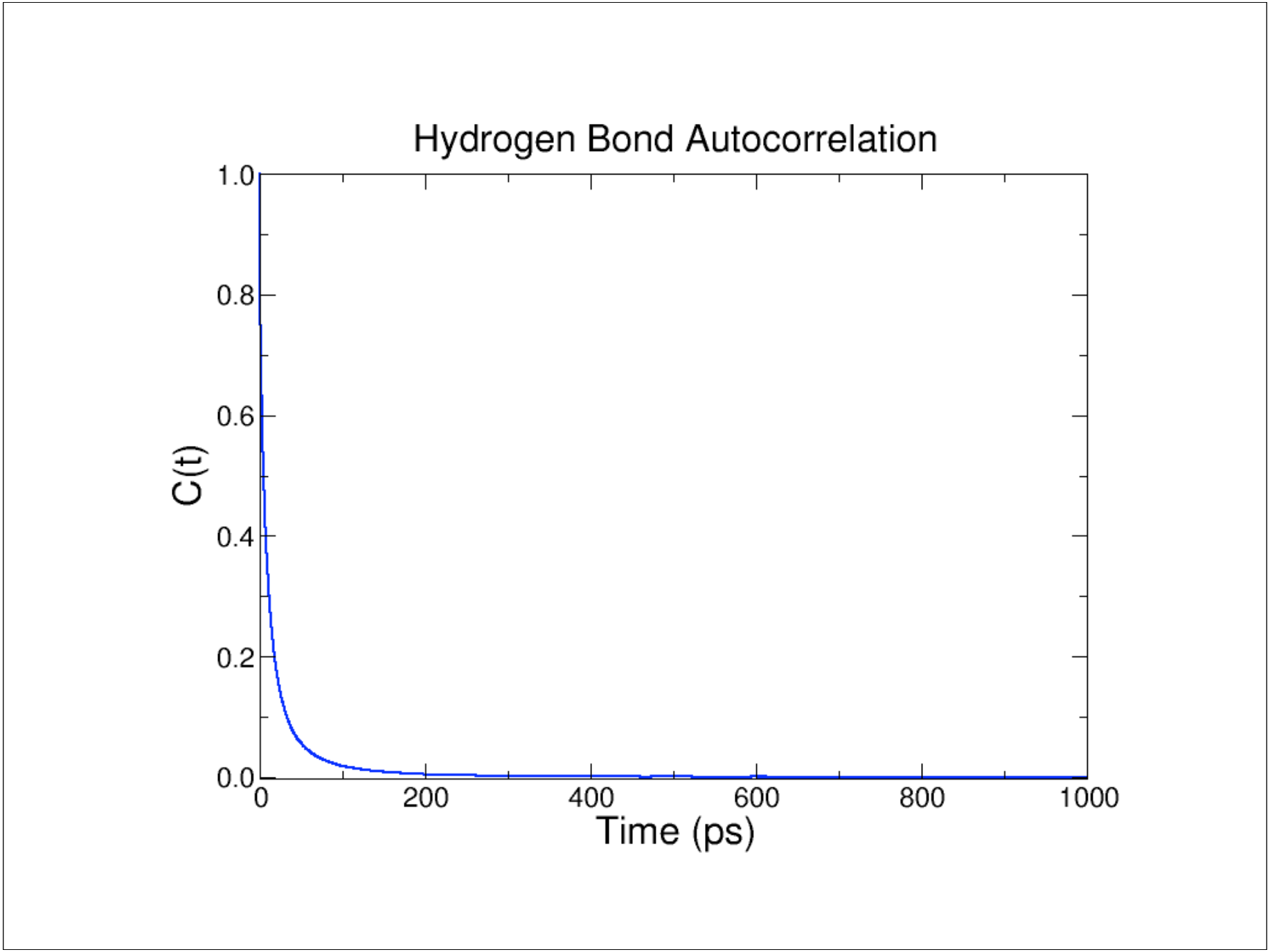
< # HB in Ubiquitin >



- PP: 56 (3)
- PS: 180 (6)
- SS: 10180 (30)
- PP+0.5PS: 147 (4)

HB Kinetics

- Define a binary function $H_i(t)$
  - $H_i(0) = 1$
  - $H_i(t) = 1$ when hydrogen bond i exists
  - $H_i(t) = 0$ when hydrogen bond i does not exist
- Compute the autocorrelation function $C_i(t) = \langle H_i(0)H_i(t) \rangle$


Hydrogen Bond Autocorrelation
C(t)
1.0
0.8
0.6
0.4
0.2
0.0
0
200
400
600
800
1000
Time (ps)

HB Kinetics

- Define rate equation: $\frac{dC(t)}{dt} = kC(t) - k'N(t)$
- Compute C(t), N(t) from simulation
- Numerically differentiate C(t)
- Numerically solve for k and k′
- Hydrogen bond life time $\tau_{HB} = 1/k$

HB Thermodynamics

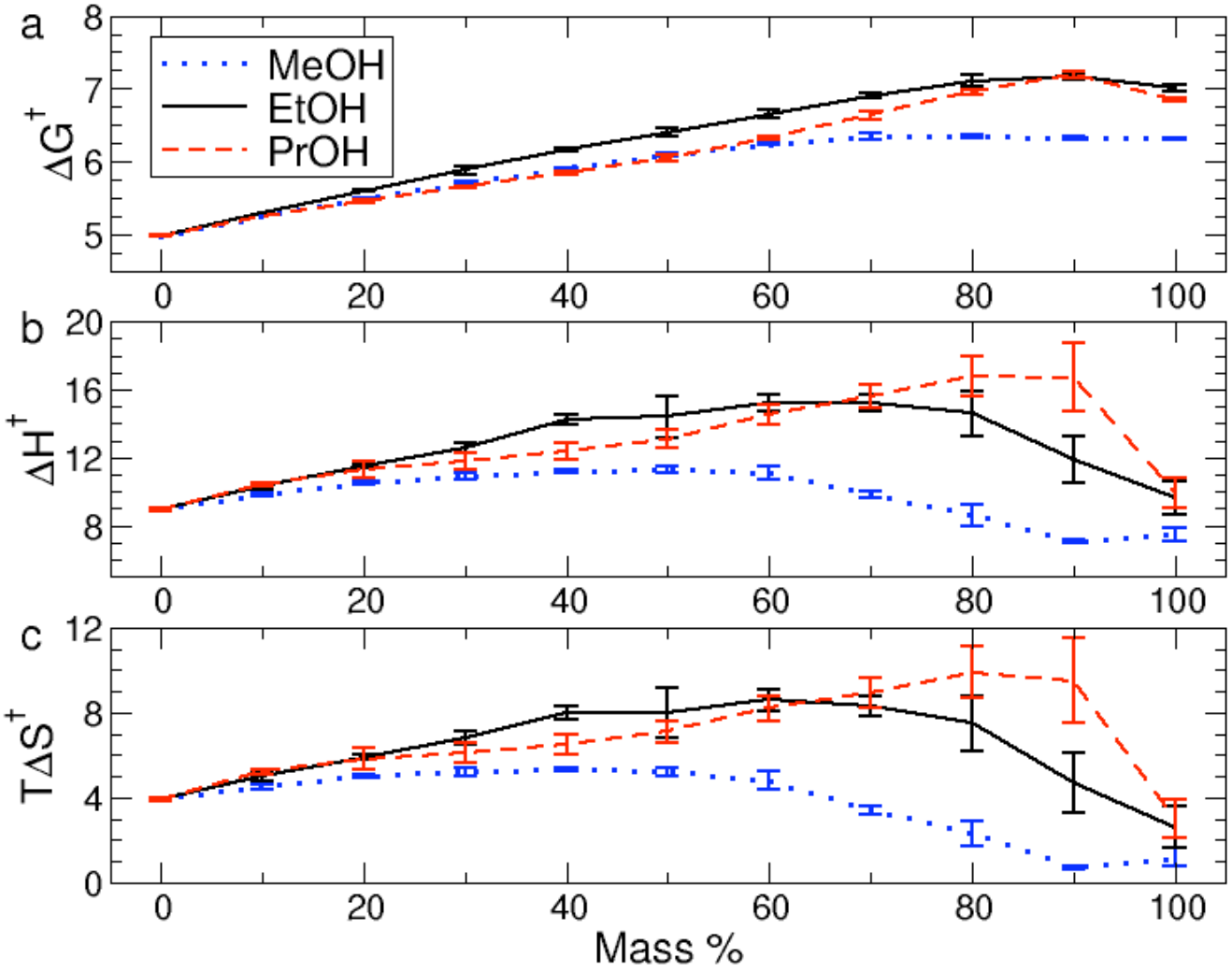
- Eyring process:

$$\tau_{HB} = \frac{h}{k_B T} \exp\left(\frac{\Delta G^{\ddagger}}{k_B T}\right)$$

- Compute DG at different temperatures
- Extract DH from Arrhenius equation:

$$k = A \exp^{-E_A / k_B T}$$

HB Thermodynamics



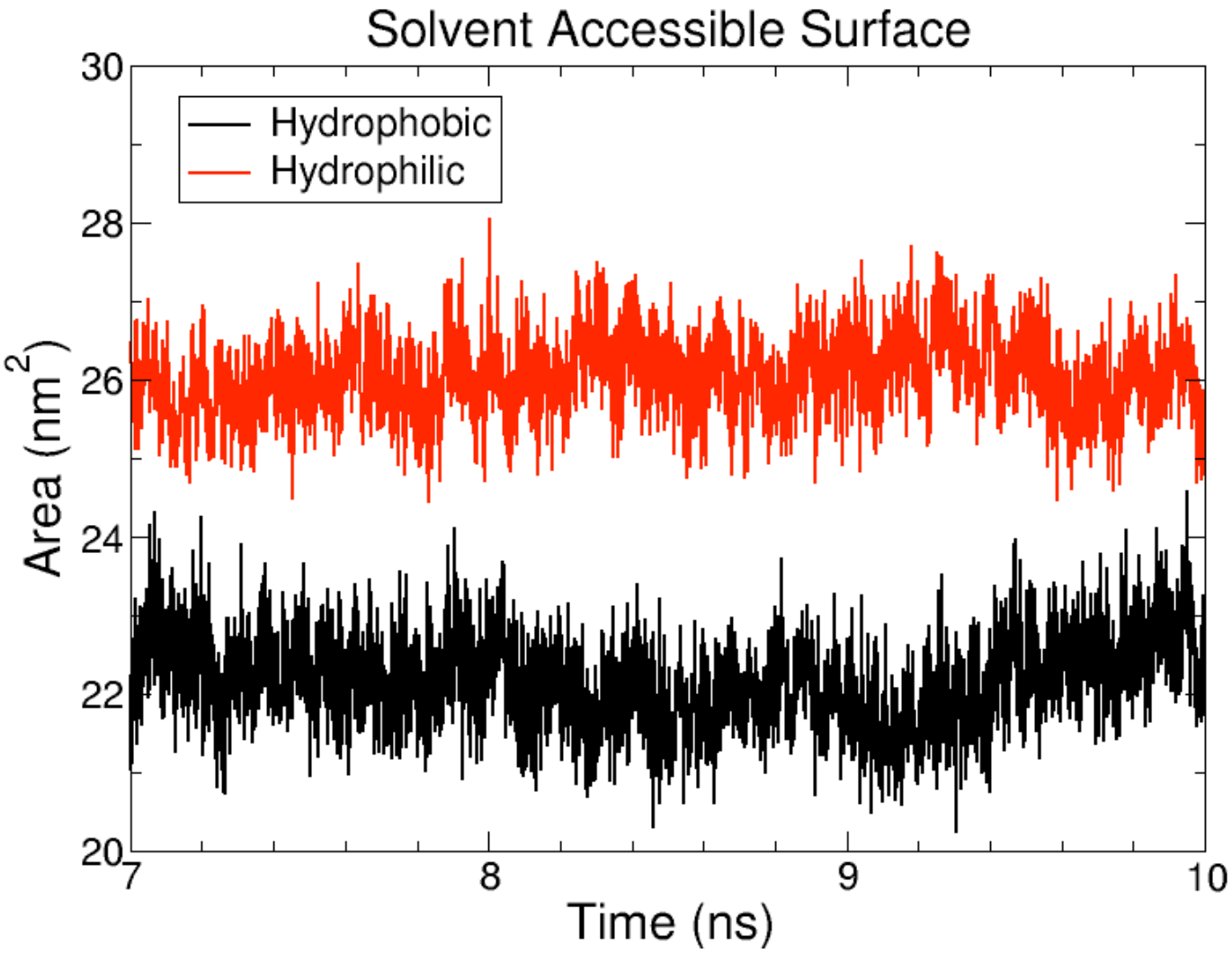
Surface Area

- A number of algorithms have been published
  - Connolly surface
  - MSMS (Scripps)
  - Double Cube Lattice Method (NSC, EMBL-Heidelberg)
  - more...
- **g_sas** uses NSC
- Solvent Accessible Surface Area (SASA) often used as descriptor for e.g. computing ΔG_{solv}

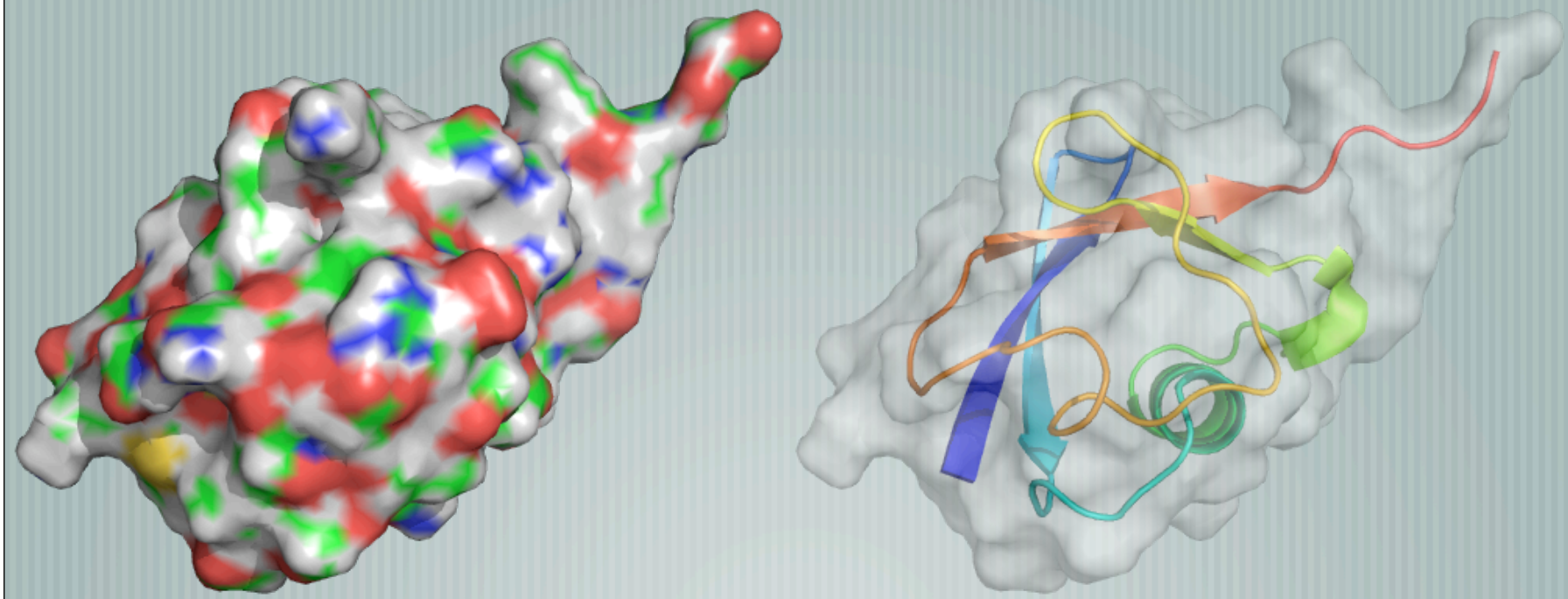
Double Cube Lattice Method

- Map triangles on the surface of an atom (sphere)
- Connect the triangles between different atoms
- Compute surface area by summing triangle areas

Eisenhaber et al. JCC **16**, p. 273 (1995)


Solvent Accessible Surface
Hydrophobic
Hydrophilic
Area (nm^2)
30
28
26
24
22
20
7
8
9
10
Time (ns)

Surface Representations

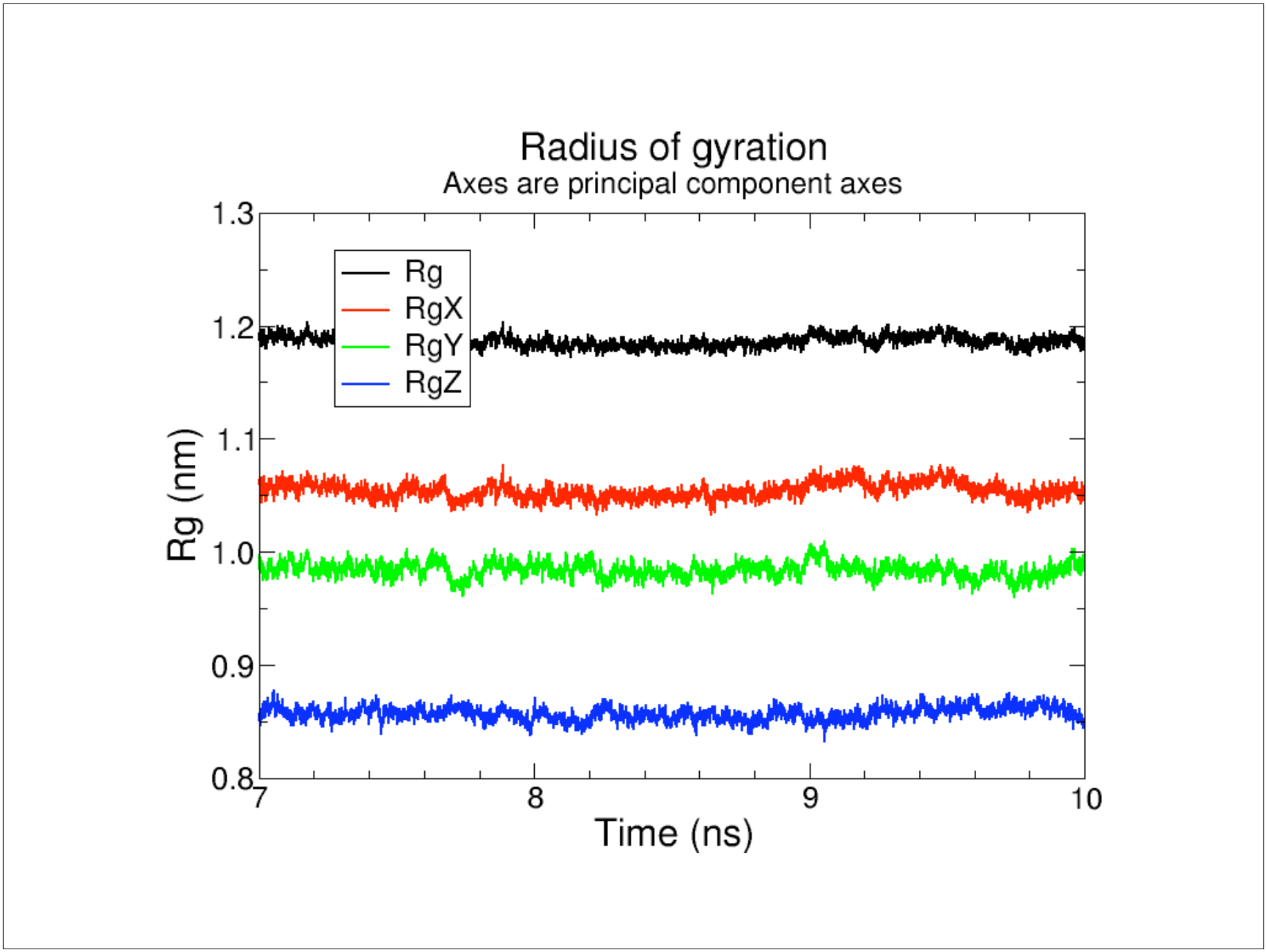


Ubiquitin surfaces created with PyMol

Radius of Gyration

- A measure of the size of a molecule
- Can in principle be determined experimentally, but not very accurately
- **g_gyrate** can align the molecule along its primary axes before computing the radius of gyration

$$R_g = \left(\frac{\sum_i m_i \left(\mathbf{r}_i - \mathbf{r}_{cm} \right)^2}{\sum_i m_i} \right)^{1/2}$$


Radius of gyration
Axes are principal component axes
Rg
RgX
RgY
RgZ
Rg (nm)
1.3
1.2
1.1
1.0
0.9
0.8
7
8
9
10
Time (ns)

Root Mean Square Deviation

- Measure of how related structures are
- RMSD: requires LSQ superposition of molecules

LSQ Superposition

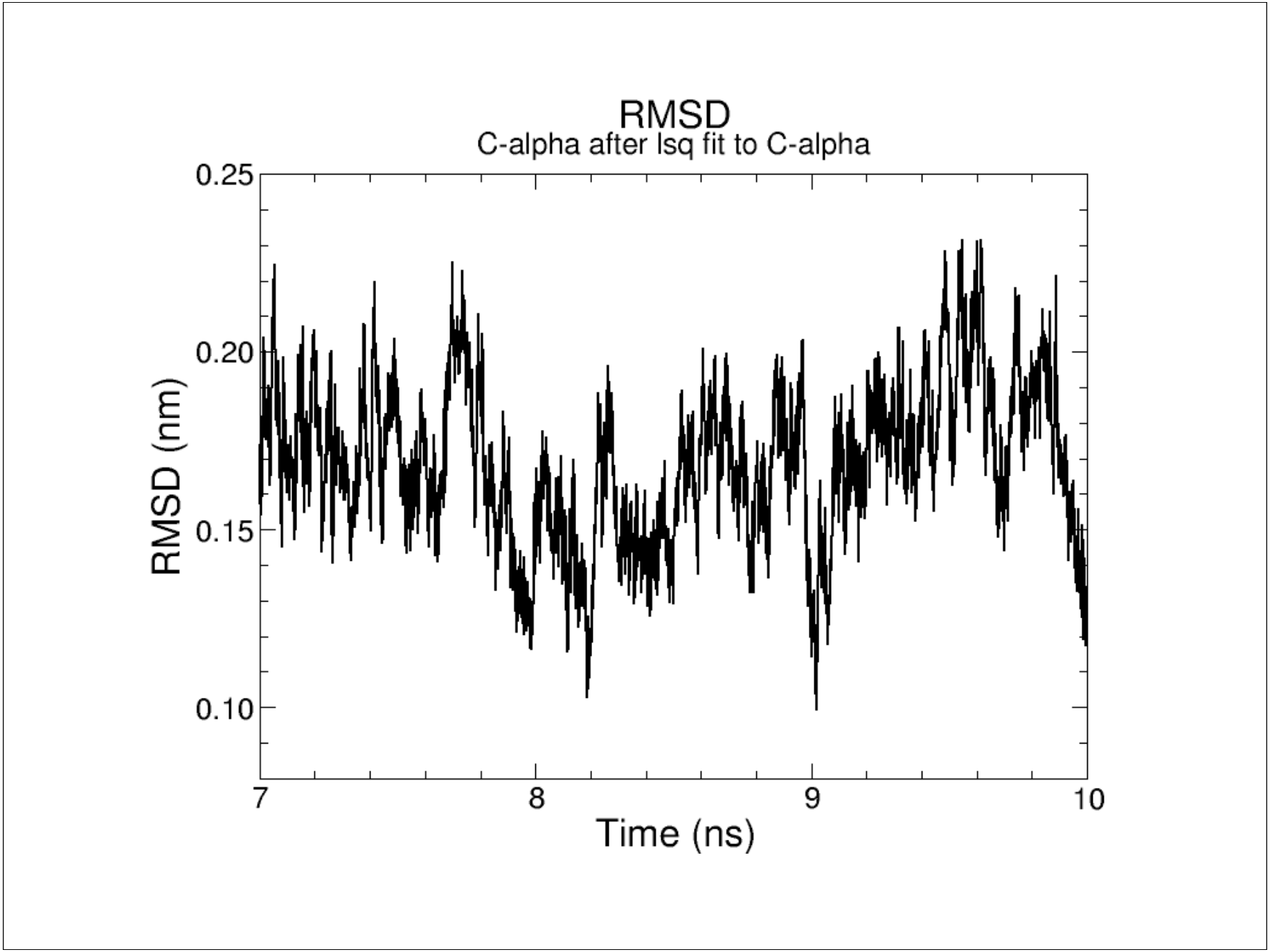
- Minimize $\chi^2 = \sum_i (\mathbf{r}_i^{ref} - \mathbf{M}\mathbf{r}_i)^2$
- Where M is a six-dimensional matrix describing rotation and translation
- Problematic when molecules are very different
- Advantage is that this is a well-known quantity

RMSD

$$RMSD = \left(\frac{\sum_i m_i (\mathbf{r}_i^{ref} - \mathbf{M}\mathbf{r}_i)^2}{\sum_i m_i} \right)^{1/2}$$

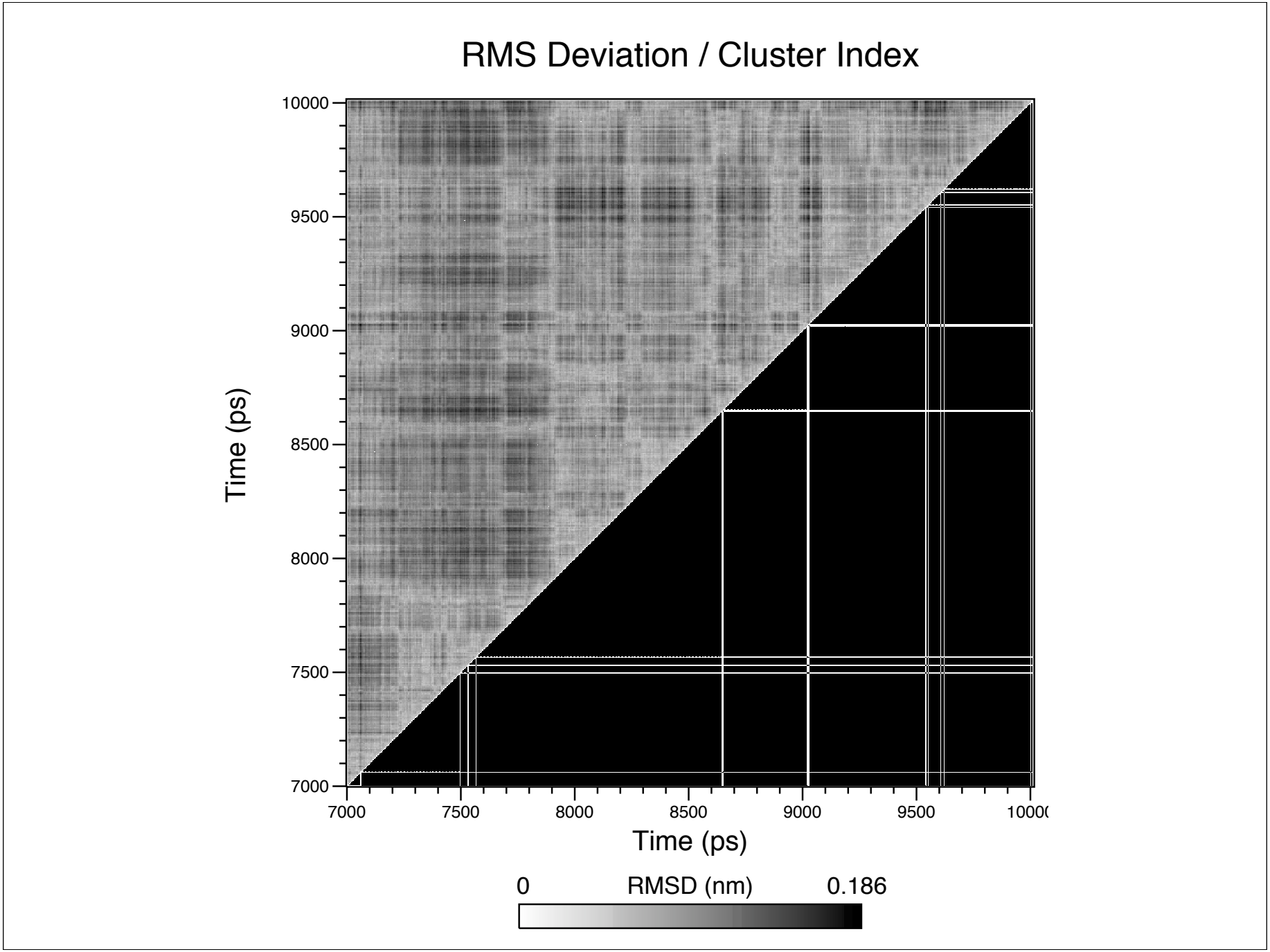
- Alternative: distance based RMSD

$$RMSD = \frac{1}{N} \left(\sum_i^N \sum_j^N (\mathbf{r}_{ij}^{ref} - \mathbf{r}_{ij})^2 \right)^{1/2}$$


RMSD
C-alpha after lsq fit to C-alpha
RMSD (nm)
0.25
0.20
0.15
0.10
7
8
9
10
Time (ns)

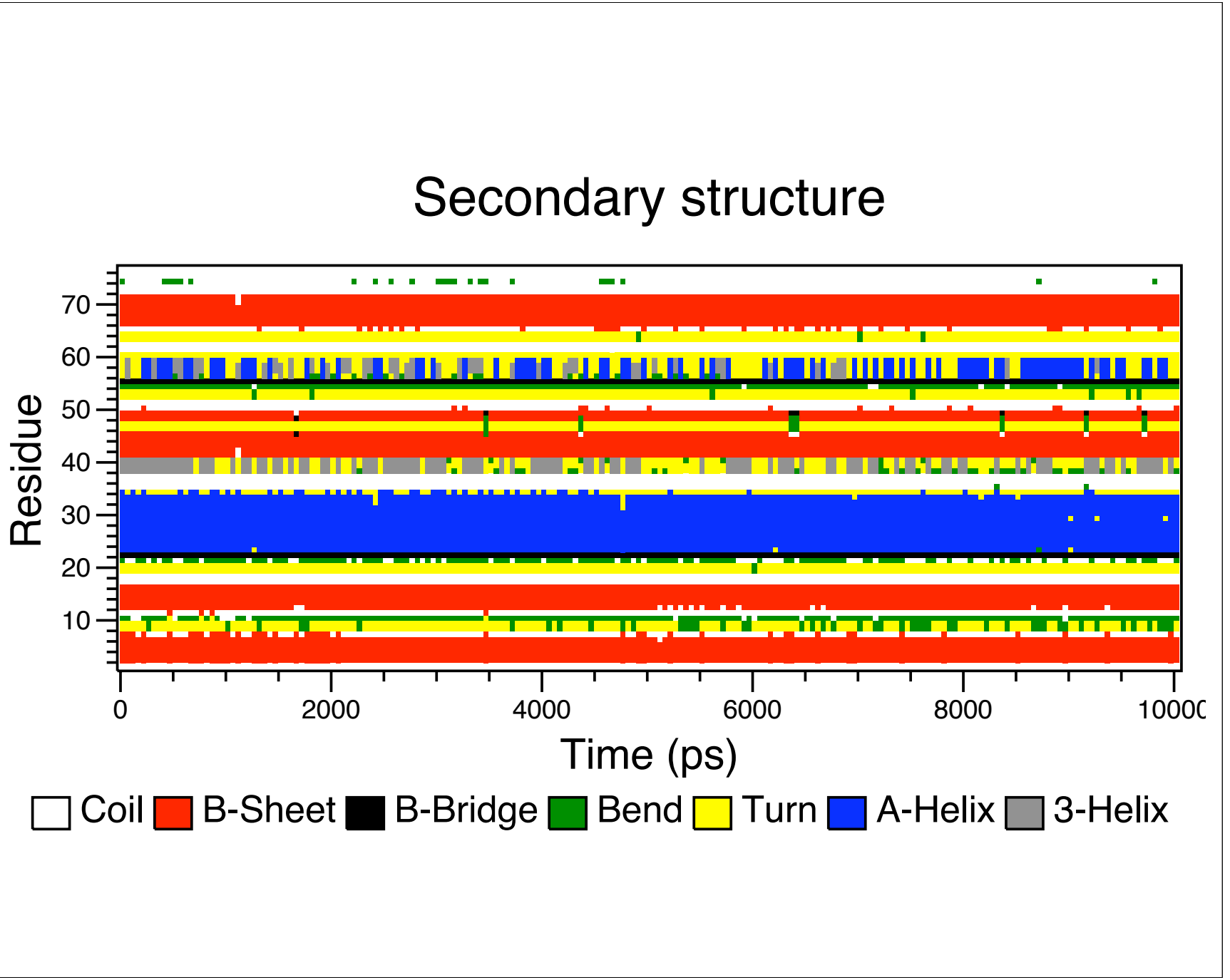
Clustering

- How many different conformations are there in a simulation trajectory?
- Clustering algorithms can give an answer
- g_cluster implements a number of popular algorithms, all of which have their own issues
  - Compare RMSD between pairs of structures
  - Group structures that have relatively low RMSD w.r.t. each other

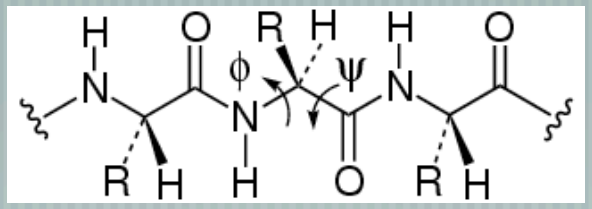

RMS Deviation / Cluster Index
10000
9500
9000
8500
8000
7500
7000
Time (ps)
7000
7500
8000
8500
9000
9500
10000
Time (ps)
0
RMSD (nm)
0.186

Secondary Structure

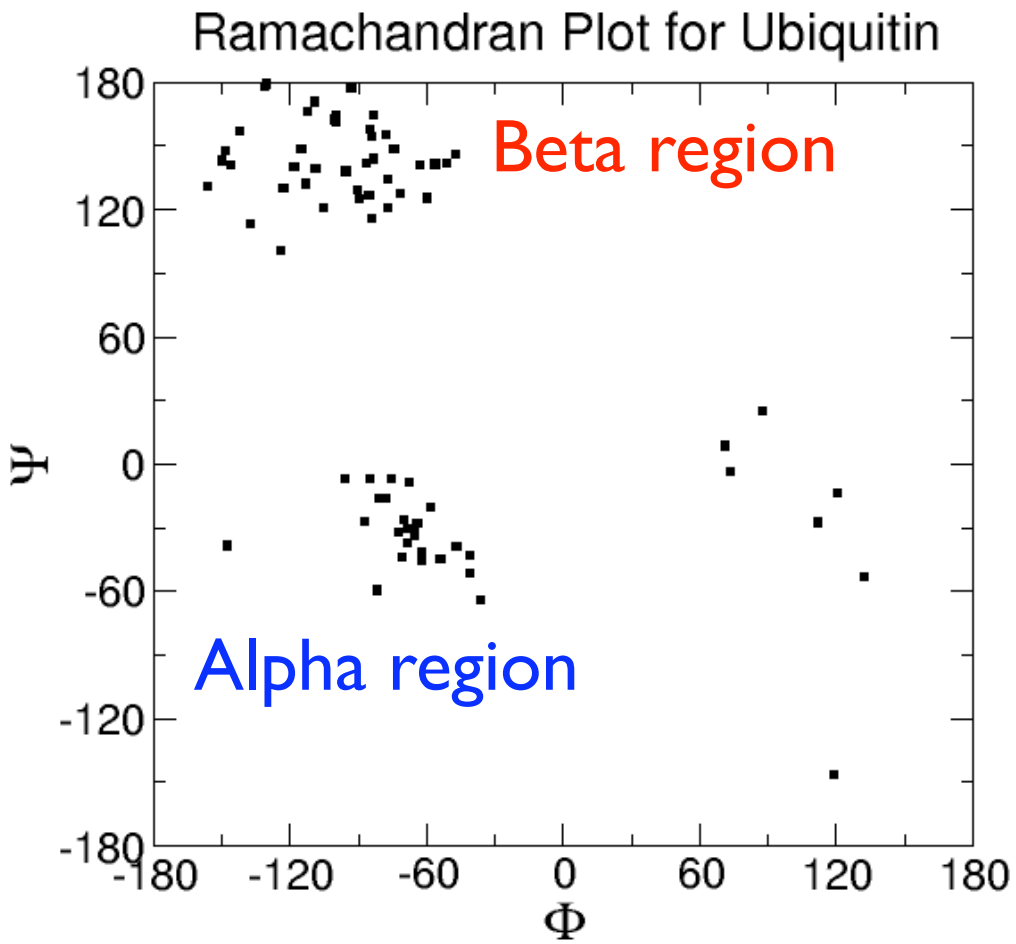
- Very important indicator of protein structure
- DSSP: dictionary of secondary structure in proteins (Kabsch & Sander, Biopolymers 22, p. 2577 (1983)), based on backbone torsion angles and hydrogen bonds
- do_dssp computes SS from trajectories by executing an external DSSP program (http://swift.cmbi.ru.nl/gv/dssp/)


Secondary structure
Residue
70
60
50
40
30
20
10
0
2000
4000
6000
8000
10000
Time (ps)
Coil
B-Sheet
B-Bridge
Bend
Turn
A-Helix
3-Helix

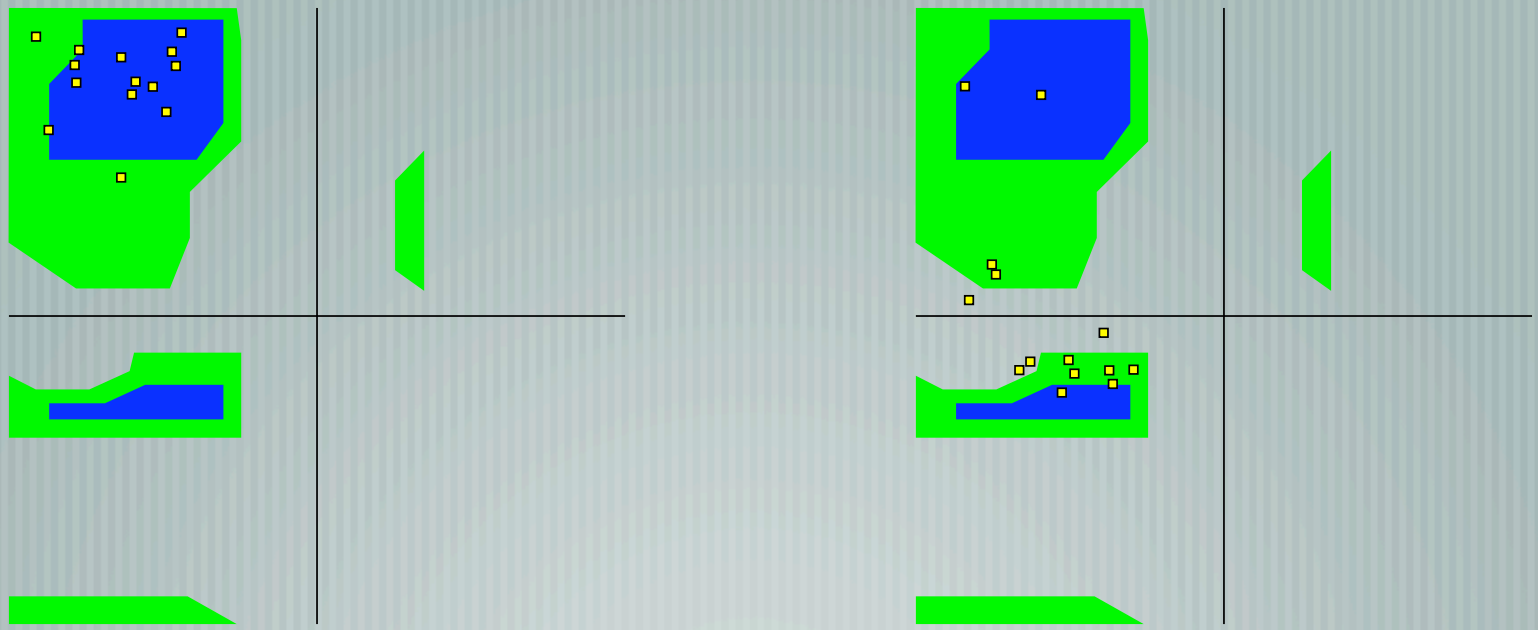
Ramachandran plot



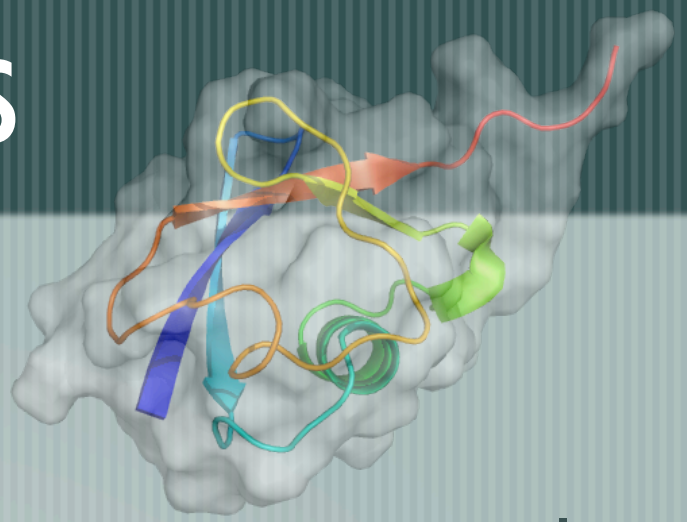
- Shows correlation of backbone angles Phi & Psi
- Certain areas are energetically unfavorable ("disallowed")
- Secondary structure (Alpha helix, Beta sheet) have common backbone angles
- **g_rama** computes them


Ramachandran Plot for Ubiquitin
Beta region
Alpha region
180
120
60
0
-60
-120
-180
-180
-120
-60
0
60
120
180
Ψ
Φ

Example Ramachandran



Protein Dynamics



- Consider a protein in solution
- The protein has an overall rigid motion: rotation+translation
- Within the protein secondary structure elements will be relatively rigid, but may fluctuate collectively
- Within secondary structure elements atoms will fluctuate mainly about their average positions

NMR Relaxation

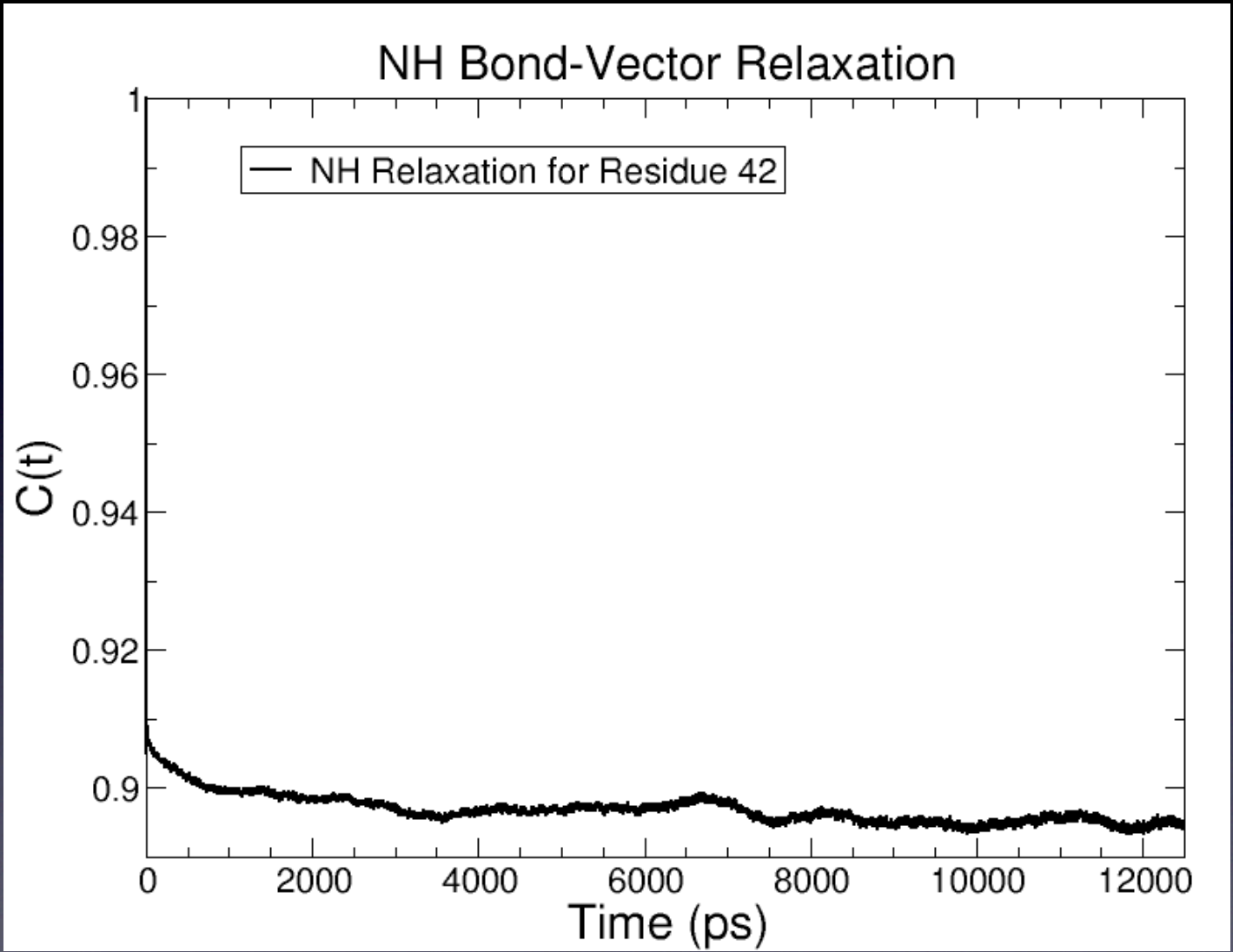
- NMR can be used to probe the intramolecular dynamics in e.g. proteins
- Result is an average over long time and many molecules
  - Q: how long is long and how many is many?
- Observables from NMR experiments can often be derived from MD simulations

NH Bond-Vector Relaxation

- If we disregard the overall motion of the protein we can write the motion of a distance vectors as:

$$C(t) = \left\langle \frac{P_2\left(\mathbf{r}_{ij}(0) \cdot \mathbf{r}_{ij}(t)\right)}{r_{ij}^3(0)\ r_{ij}^3(t)} \right\rangle$$

- Bond vectors in MD are usually rigid
- It is relatively easy to study NH bond vectors in NMR using ^{15}N labeled protein


NH Bond-Vector Relaxation
NH Relaxation for Residue 42
C(t)
1
0.98
0.96
0.94
0.92
0.9
0
2000
4000
6000
8000
10000
12000
Time (ps)

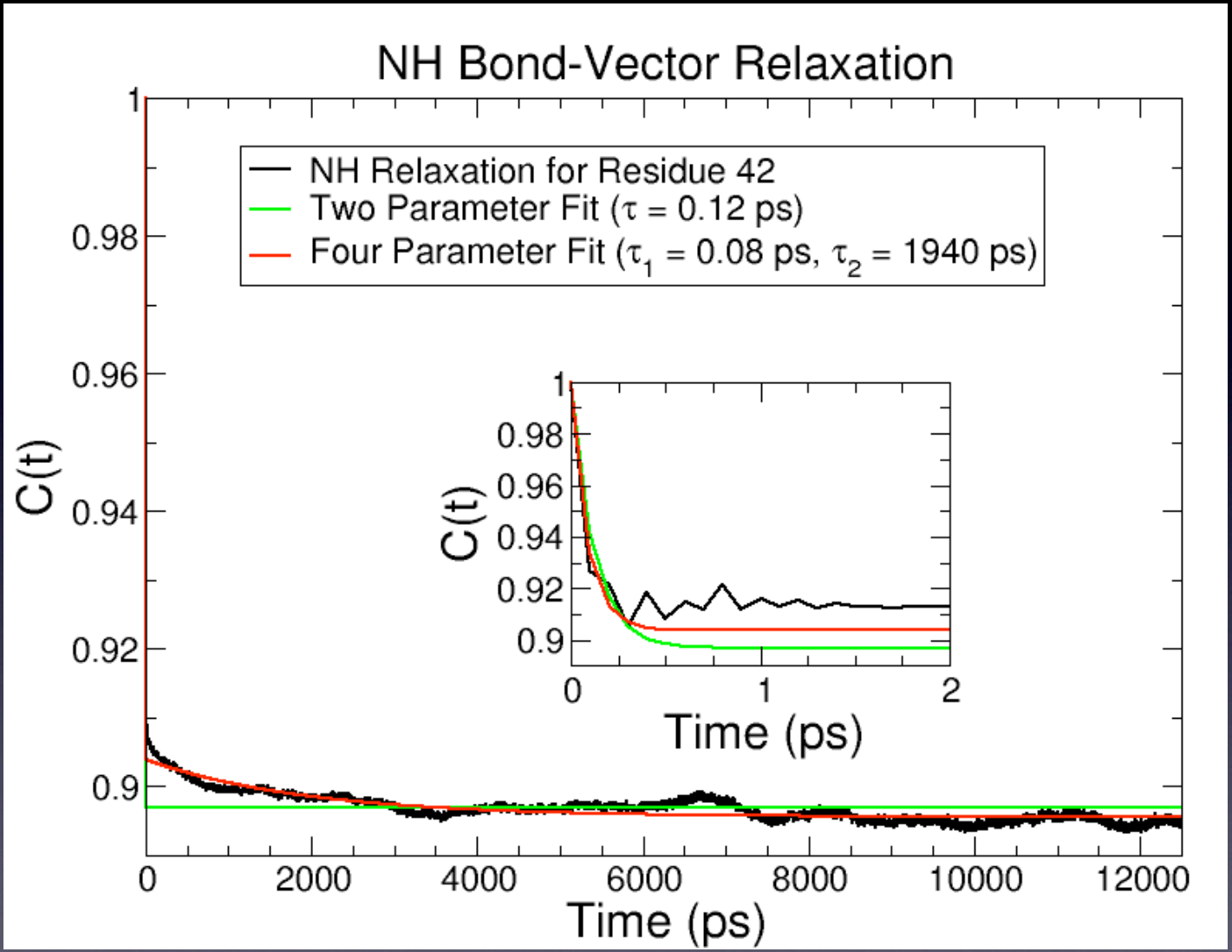
Lipari & Szabo Analysis

$$C(0) = \langle r_{ij}^{6} \rangle$$

$$C(\infty) = S^2 \langle r_{ij}^{-6} \rangle$$

- Motion at long time can be described by an order parameter
  - $S^2 = 0$: motion unrestricted
  - $S^2 > 0$: motion restricted
- Approximation:

$$C(t) = \langle r_{ij}^{-6} \rangle \sum_{i=0}^{\infty} a_i e^{-t/\tau_i} \qquad \tau_0 = \infty$$


NH Bond-Vector Relaxation
NH Relaxation for Residue 42
Two Parameter Fit (τ = 0.12 ps)
Four Parameter Fit (τ1 = 0.08 ps, τ2 = 1940 ps)
C(t)
Time (ps)
1
0.98
0.96
0.94
0.92
0.9
0
2000
4000
6000
8000
10000
12000
0
1
2

Summary I

- Many different tools available
- Protein analysis
- Liquid analysis
- Dynamics
- Electrostatic properties

Tips

- Think before you perform a simulation what kind of information you would like to get out
- Check availability of experimental structures and other data (e.g. energies, dynamic properties)

Acknowledgements

- Alexandra Patriksson, Erik Marklund for Ubiquitin simulations and analyses
- Csaba Hetenyi for Ramachandran plots and the idea for this lecture
- The GROMACS team for writing and correcting the GROMACS manual

but wait, there's more...

GROMACS

Groningen Machine for Chemical Simulations

USER MANUAL

Version 3.3