CHARMM Analysis Tools

Lennart Nilsson
Karolinska Institutet, Stockholm

MMTSB/CTBP

2006 Summer Workshop

Overview

- Basic analysis issues
- Main CHARMM analysis modules
- Single structure
- Trajectory
- Examples:
  Solvent, hydrogen bonding, clustering

Basic analysis issues

- What do I want to know from my simulation?
  How can this help with my scientific problem?
  HAVE TO BE ADDRESSED EARLY ! ! !
- A single number (energy value,…) – convergence?
- Properties of a single structure – representative?
- Properties of a population – how many?
- Time-dependence – time-scales?

- (almost) All analysis in CHARMM is done by post-processing a coordinate file or a trajectory

Main CHARMM analysis modules

- CORMAN: Single coordinate sets, averages. COOR xxxx (corman.doc)
- CORREL: Time-series data and correlation functions from trajectories (correl.doc)
- solvent properties: COOR ANAL, RDFSOL
- Energy, interaction_energy
- SCALar, gives access to internal data (charge, mass…)
- Quasi-harmonic modes (”essential dynamics”): VIBRAN (vibran.doc)

Single structure

- Geometry data:
  distance, angle, torsion angle, radius of gyration, least-squares-plane through set of atoms, sugar conformation (puckering), helix orientation
- Energy data:
  energy, interaction_energy, forces, contributions from each atom, components (SKIPE)
- Composite data:
  secondary structure (Kabsch&Sander/DSSP), hydrogen bonds, RMSD from other structure, solvent accessible surface area, volume, cavities

Geometric examples

quick 34 41 prints distance between atoms 34 and 41 (miscom.doc)

quick 34 41 52 prints angle between atoms 34, 41 and 52

radius of gyration:

coor rgyr mass select ires 3:28 end

minimum distance between two sets of atoms:

coor mindist sele segid prot end sele segid dna end

coordinate statistics (x,y,z min, max and average):

coor stat

dipole moment:

coor dipole

RMSD between main and comp coordinates, after optimal superposition:

coor orient rms

Solvent accessible surface area of each selected atom, disregarding the rest:

coor surface select segid prot .or. segid dna end

Most of these commands also set CHARMM variables (?xxx, subst.doc)

Energy examples

compute energy terms and forces, print energy and average force:

```
 energy
```

interaction energy between two sets of atoms;

energy terms only dependent on atoms in one set are not computed:

```
 update !set up necessary lists for energy calculation
 interaction select segid prot end sele segid lig end
```

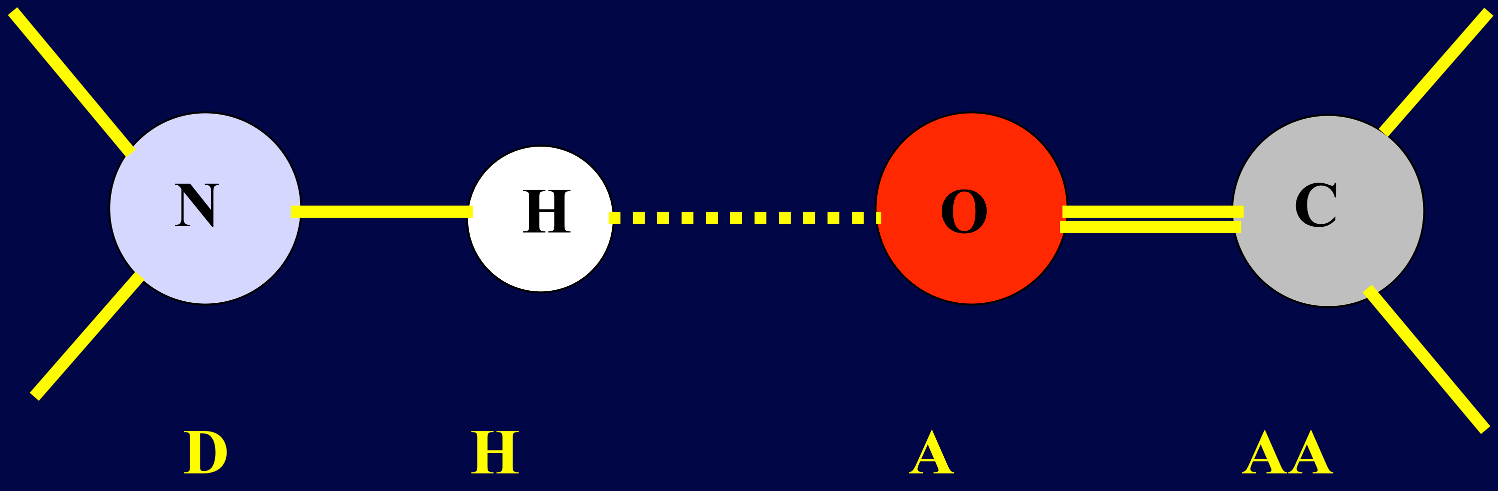
after an energy evaluation the force (kcal/mol/Å) on each atom is available:

```
 coor force comp
 scalar wcomp show sele type ca .and. ires 41:?NRES end
```

contributions from each atom, excluding bond energy terms (analys.doc):

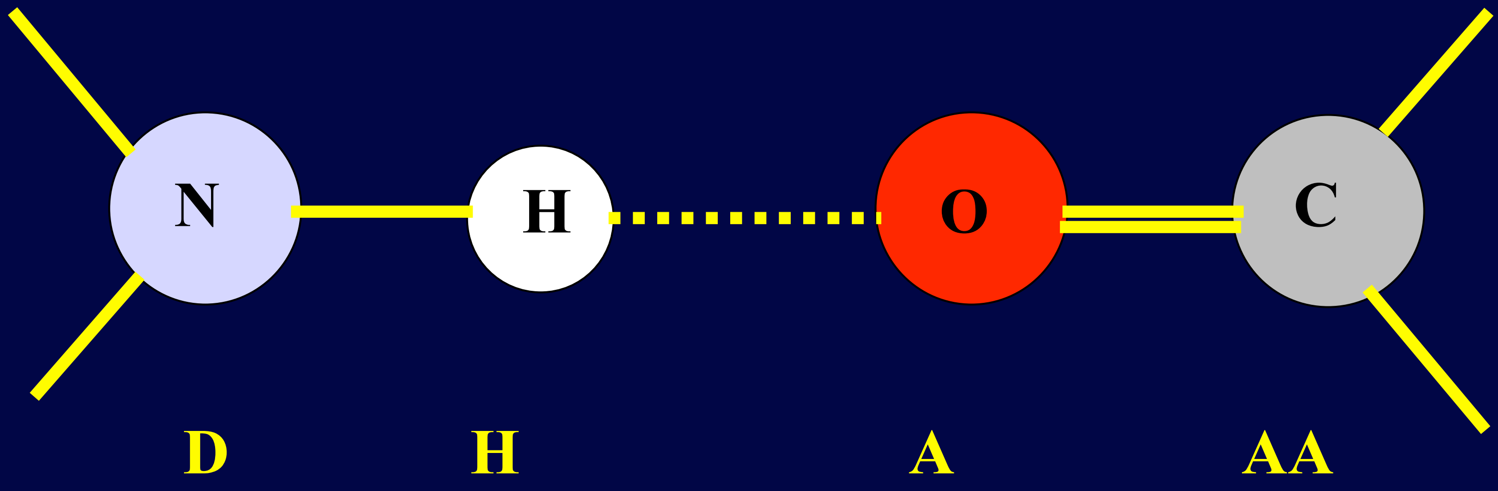
```
 analysis on
 skipe bond
 energy
 scalar econt show
 echo ?ener ?elec ?grms  !energy.doc
```

Hydrogen bonds



Two electronegative atoms can be closer than their vdW radii would normally allow if one of them (the **D**onor) is bonded to a hydrogen, which is partially shared by the other (the **A**cceptor)

Hydrogen bonds



Several parameters can be used to characterize the hydrogen bond:

r(D-A), r(H-A), ∧(D-H-A), ∧(H-A-AA), torsion(-A–AA-)

CHARMM has lists (in the PSF) of the Donors, Acceptors and Acceptor Antecendents, from DONOR and ACCEPTOR statements in the RTF (also available as commands)

Hydrogen bonds

- The fine details of hydrogen bonds have recently been analyzed using these geometric parameters (Morozov, A. V., Kortemme, T., Tsemekhman, K., and Baker, D. (2004) PNAS *101*, 6946-6951)

- When the hydrogen position is known (*ie,* all-atom force-fields) r(H-A) is a simple and useful indicator of hydrogen bonding; r(H-A)< 2.4Å (corresponds to r(D-A)<3.4Å) is a good criterion (De Loof, H., Nilsson, L., and Rigler, R. (1992) JACS *114*, 4028-4035)

Hydrogen bond criteria

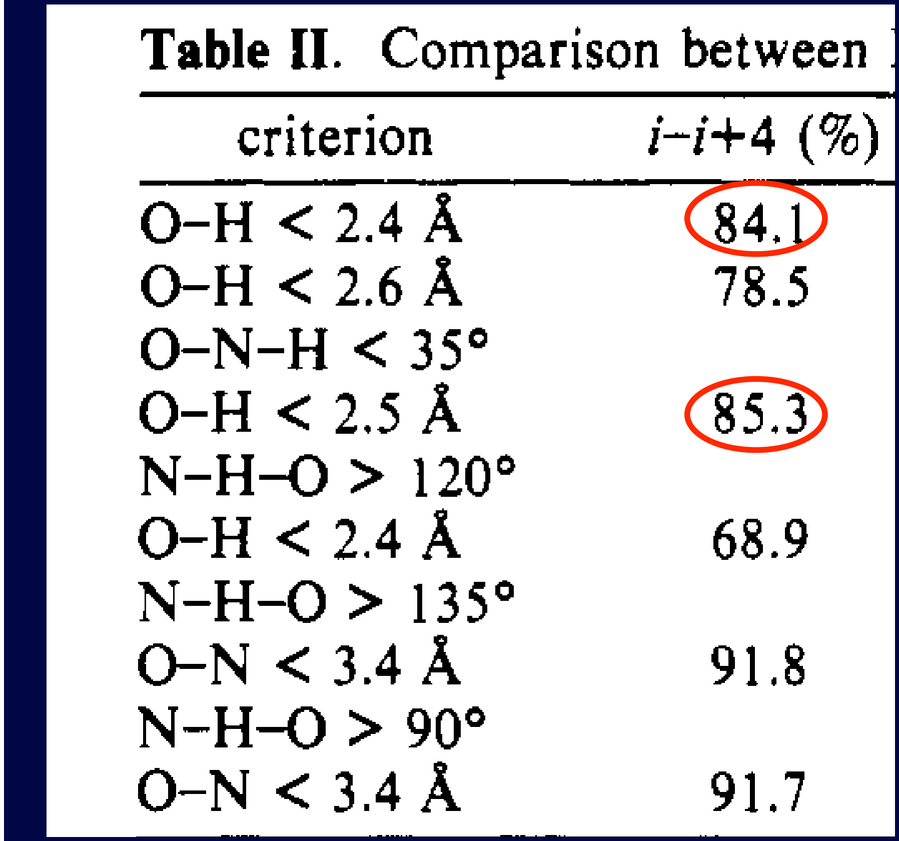
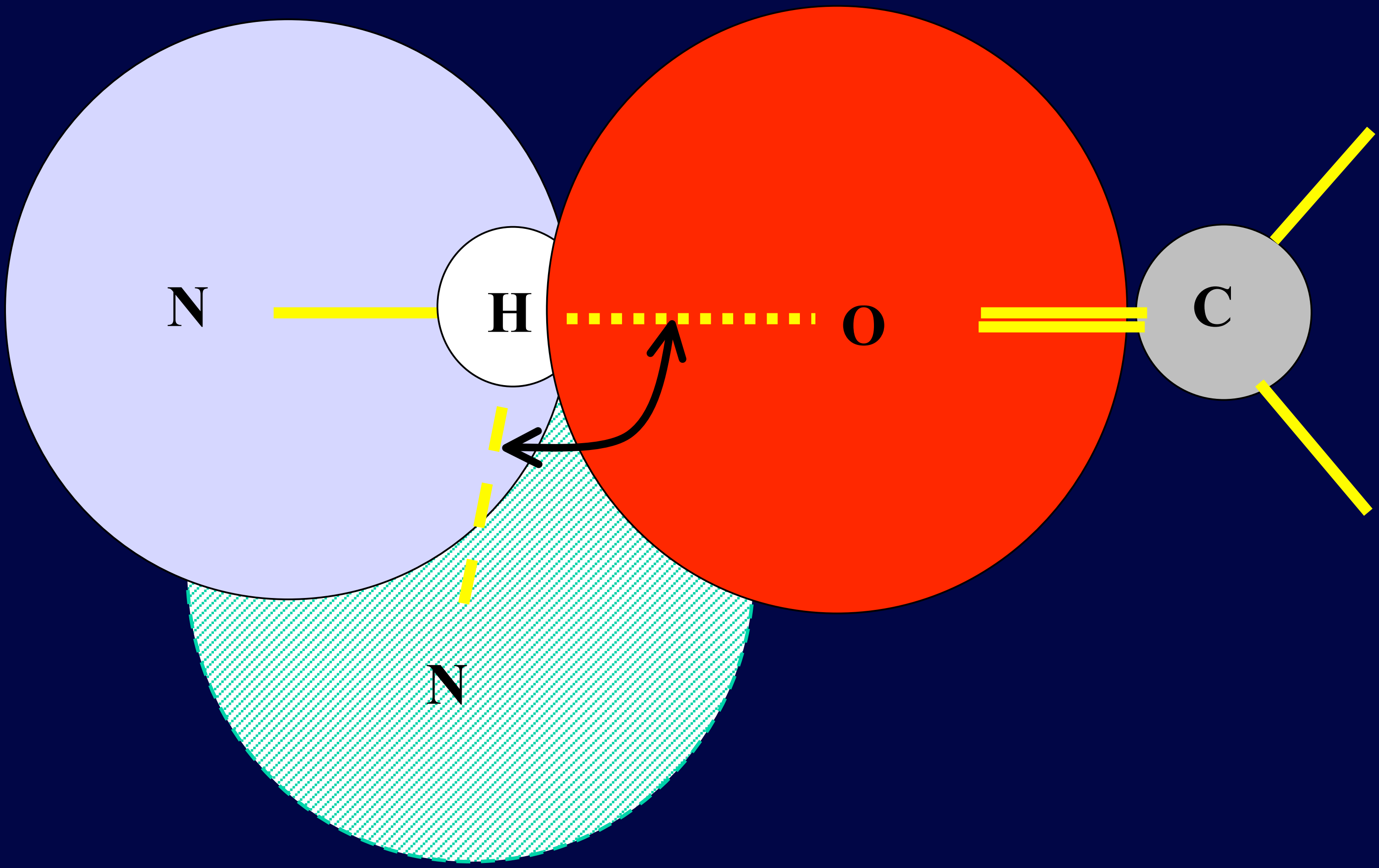


Table II. Comparison between

criterion	*i*–*i*+4 (%)
O–H < 2.4 Å	84.1
O–H < 2.6 Å	78.5
O–N–H < 35° O–H < 2.5 Å	85.3
N–H–O > 120° O–H < 2.4 Å	68.9
N–H–O > 135° O–N < 3.4 Å	91.8
N–H–O > 90° O–N < 3.4 Å	91.7

Fraction of backbone N-H…O hydrogen bonds found in a simulation using different criteria (DeLoof et al JACS).

The simple criterion r(H-A)<2.4Å is reasonable. A correlation between the r(H-A) and ∧(D-H-A) criteria was also observed, due to steric clash between A and D, explaining why the angle criterion adds little extra in this case.


N
H
O
C
N

Finding hydrogen bonds in a single structure with CHARMM

The COOR HBONd command (corman.doc) finds and lists all hydrogen bonds between two atom-selections based on r(H-A), and optionally ∧(D-H-A). The command uses acceptor/donor lists in the PSF to identify possible hydrogen bonding atoms. If DONO/ACCE statements are missing for some residue(s) in your RTF you have to add these and regenerate the PSF before using COOR HBONd (this may be the case for TIP3 residues).

COOR HBOND examples

```
coor hbond sele type hn end sele type O end
```

```
coor hbond sele segid prot end sele segid wat end verb
```

For each donor/acceptor in the first selection the number of hydrogen bonds to any acceptor/donor in the second selection is printed out.

Keyword VERBose gives a more detailed listing that includes the identity of the atoms involved in the second selection, and the actual geometry.

CHARMM substitution variables ?NHBOND and ?AVNOHB are set if VERBose is not used.

```
coor hbond sele segid prot end sele segid dna end -bridge tip3
```

This form finds and lists all instances where a residue with the name tip3 is hydrogen bonded to some atom in both selections, in this case water bridges between a protein and a DNA molecule.

PBC can be handled. The form COOR CONTact does not use acceptor/donor lists in the PSF.

Protein secondary structure

- Proteins usually have a high content of secondary structure elements, α-helices and β-sheets
- Several algorithms to **analyze** secondary from 3D structure
- CHARMM uses the Kabsch&Sander method (DSSP; Kabsch, W., and Sander, C. (1983). Biopolymers *22*, 2577-2637) which is based on patterns of backbone hydrogen bonds. The CHARMM implementation is very similar to DSSP. The default hydrogen bond criterion here is r(H-A)<2.6Å. CHARMM19 and CHARMM22 amide hydrogen names are recognized.

```
COOR SECS SELE SEGID PROT END SELE SEGID PROT END
```

Finds secondary structures in the first selection, within the context of the second selection, *eg*, a β -strand in the first selection will be recognized as such if it forms appropriate hydrogen bonds to residues in the second selection. Sets CHARMM variables, and returns flags in WMAIN array.

Analysis of trajectories

- To obtain averages, distributions, time-dependencies (correlation functions)
- Several modules/commands can directly access trajectories (coordinates or velocities), *eg* CORREL and NMR

Trajectory specification:

first *n* nunit *k* begin *m1* skip *m2* stop *m3*

Access *k* unformatted files on units *n* to *n+k-1*. Extract every *m2*:th coordinate set between *m1* and *m3. m1, m2, m3* are specified as **integration step numbers** from the start of the whole simulation.

Compute average structure and RMS fluctuations (in WMAIN):

```
open read unform unit 51 name myfile_4.trj
open read unform unit 52 name myfile_5.trj
coor dyna first 51 nunit 2 begin 2500 skip 50 stop 5000
```

Snapshots can be anlyzed in a loop

```
open read unform unit 51 name myfile_4.trj
open read unform unit 52 name myfile_5.trj
open unit 21 write form name inte.dat
echu 21
echo time/ps etot evdw ecoul (kcal/mol) xaver/A #hbonds
set i 1
traj first 51 nunit 2 begin 2500 skip 50 stop 5000
label loop
  traj read !read next snapshot as specified above
  update [nbond-spec]
  interaction sele segid prot end sele segid dna end
  coor stat sele segid dna2 end
  coor hbond sele segid sub1 .and. –
   .not. (type HN .or. type O) end sele segid dna2 end
  echo ?time ?ener ?vdw ?elec ?xave ?nhbond
  incr i by 1
  if i .lt. 50 goto loop
```

CORREL

The CORREL module (correl.doc) allows **extraction** of various time-series from a trajectory, **manipulation** of these time-series and calculation of **correlation** functions.Time-series in CHARMM are structured data sets with a number of properties.

```
correl [maxtime n] [maxseries m] [maxatoms k]
 enter name1 type [optional type-dependent info]
 enter name2   …
  .
  .
 traj nfirst int nunit int begin int skip int end int
 mantime name action
 edit name …
 read name ! Can be simple data file, eg results from
             !  interaction energy calculations

 corfun name1 name2
 write name unit int

end
```

Hydrogen bonds from a trajectory

```
coor hbond first 51 nunit 3 skip 2500 –
sele segid prot end sele segid wat verbose
```

Computes all hydrogen bonds between the two selections for each specified coordinate frame.

For each acceptor/donor in first selection: print average number of hydrogen bonds and the average lifetime over the trajectory. Note that SKIP can influence the lifetime estimate. 5ps resolution means that intermittent breaks < 5ps are less likely to count as a real disruption.

The verbose keyword has two effects:

i) a more detailed summary with atom identifications from the second set is printed at the end.

ii) each time an instance of a hydrogen bond is broken information about this event, including the duration of this particular hydrogen bond, is printed.

For hydrogen bonds to solvent, use a recentered trajectory, or the COOR HBOND support for some PBC types.

Distance (unit irhi) and time (unit ithi) distributions of hydrogen bonds:

```
coor hbond first 51 nunit 3 skip 2500 –
sele segid prot end sele segid wat irhi 21 ithi 22
```

r(A-H) *vs* time

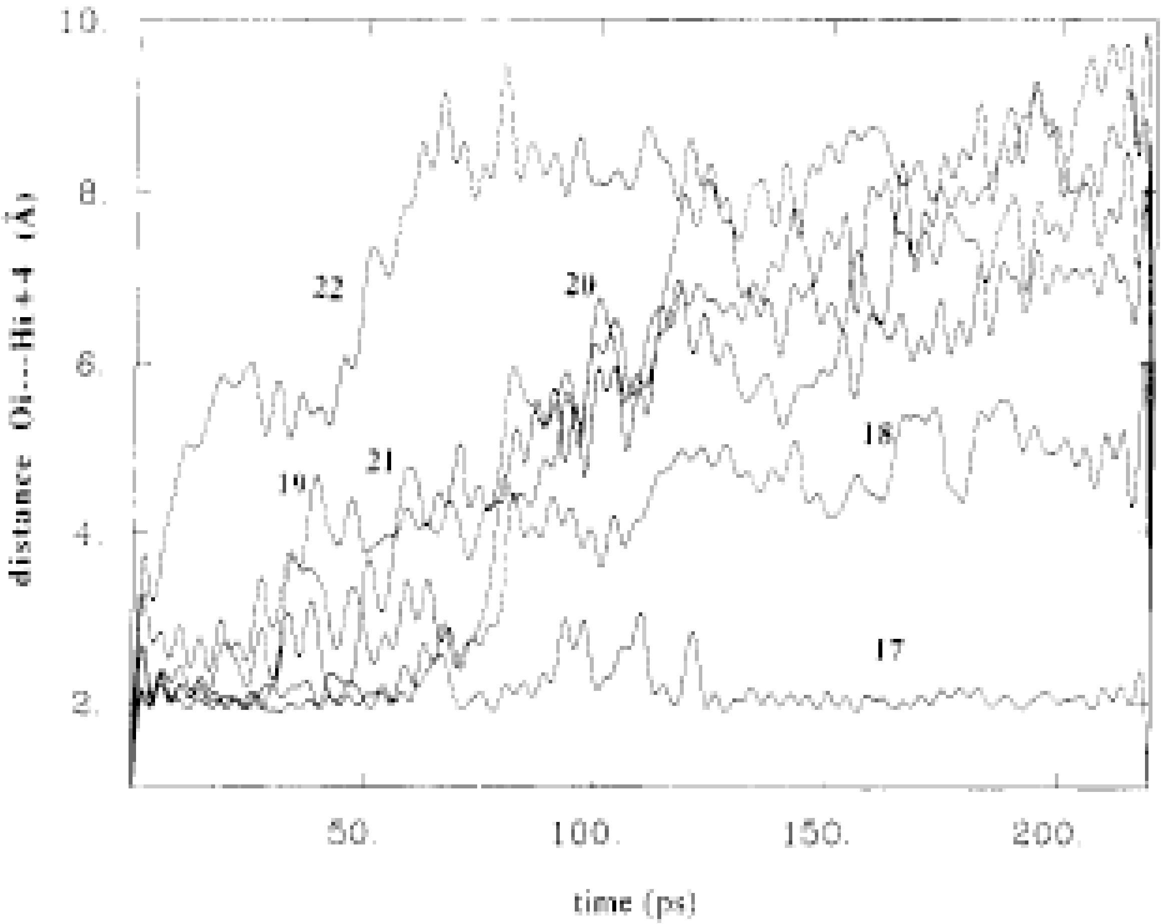


Figure 9. Oxygen(*i*)–hydrogen(*i*+4) distances for six neighboring backbone oxygen atoms starting at residue 17 of the H_2O simulation. Data points, collected every 100 fs, were filtered as in Figure 6.

De Loof et al (1992) JACS

Solvent structure

Radial distribution functions:

$$g_{AB}(r) = \frac{N_{AB}(r,\Delta r)}{\rho_B V_S(r,\Delta r)}$$

where $N_{AB}(r,\Delta r)$ is the average number of *B* sites found in a shell, Δr thick, at distance r from the *A* sites, V_S the volume of this shell, and ρ_B the average number density of B sites in the system.

CHARMM can compute ρ_B, or it can be given by user.

Radial distribution functions in CHARMM

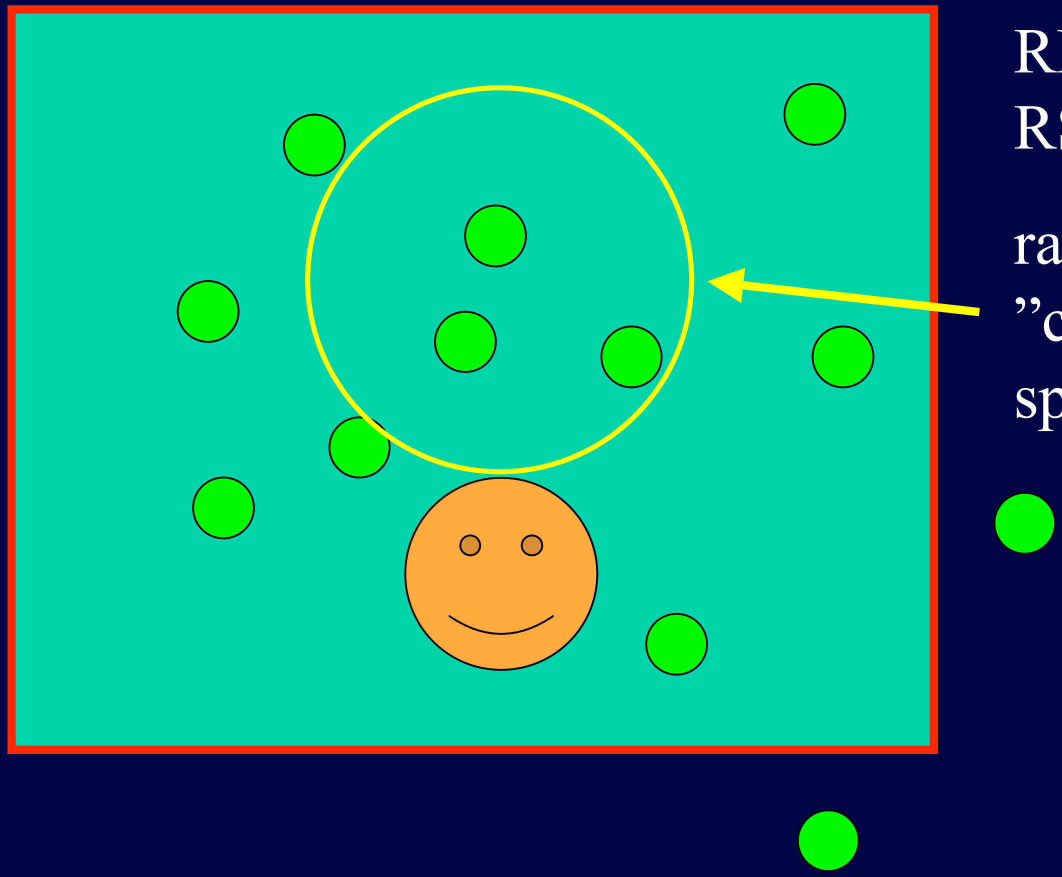
CHARMM modules: RDFSOL and COOR ANALysis

g(r) for waters; the program defaults are used to calculate the density using selected atoms within 10A (RDSP keyword) of the reference point (0,0,0) (REF keyword)

```
open unit 21 read unform name pept500.cor
open unit 31 write form name  pept500.goo
open unit 32 write form name  pept500.goh
open unit 33 write form name  pept500.ghh
! WATEr gets all three g(r) functions computed
coor anal water select type OH2 end –
  firstu 21 nunit 1 skip 500 -
  igdist 31 ioh 32 ihh 33 –
  mgn 100 dr 0.1 rsph 999.9  -
  xbox 30.0 ybox 30.0 zbox 30.0
```

Three columns of numbers are written to each *.g** file:
r (Å), g(r), total # of configurations within r.
Excluded volume and PBC effects can be corrected for.

Geometry of setup



Keywords

RDSP &
RSPH

radii of
”control”
spheres

COOR ANALysis, cont´d

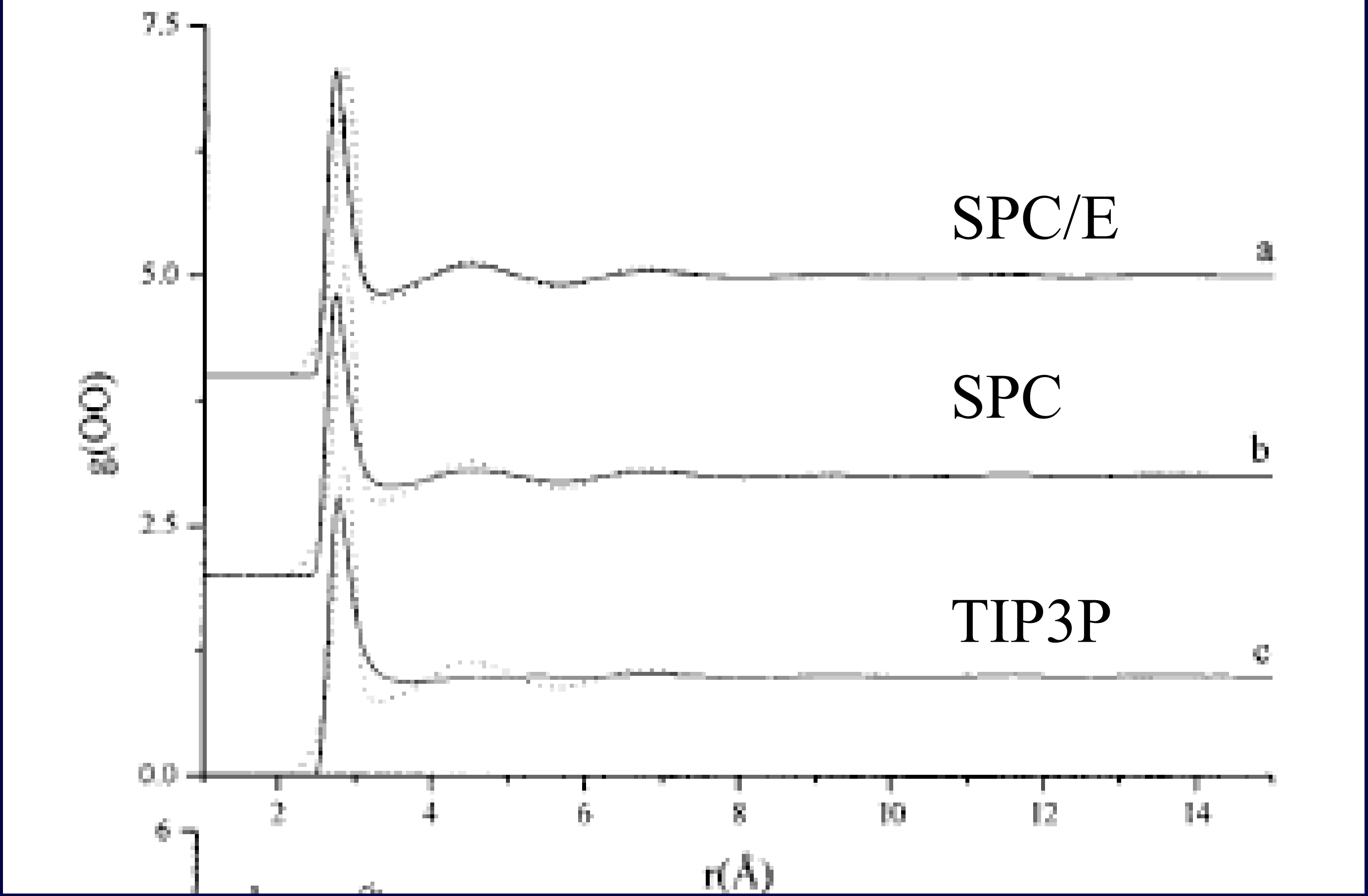
g(r) for water oxygens wrt backbone amide hydrogens

When several solute atoms are specified as the site their average position will be used as the reference position if MULTi is not present

```
open unit 21 read unform name pept500.cor
open unit 31 write form name  pept500.gonh
coor anal select type oh2 end  -
 site select type HN end multi –
 firstu 21 nunit 1 skip 500 -
 isdist 31 mgn 100 dr 0.1 rsph 999.9
```

Several types of analysis (with some exceptions) may be combined in a single ”coor analysis” command.

Water $g_{OO}(r)$



Hydration number

Calculate number of solvent molecules within a specified distance of a site:

?NHYDRR - number of solvent molecules (residues)

?NHYDAR - number of solvent atoms

?NHYDAA - number of solvent atoms within RHYD of solute atoms
(3 water molecules within RHYD of a 7-atom solute → NHYDAA=63)

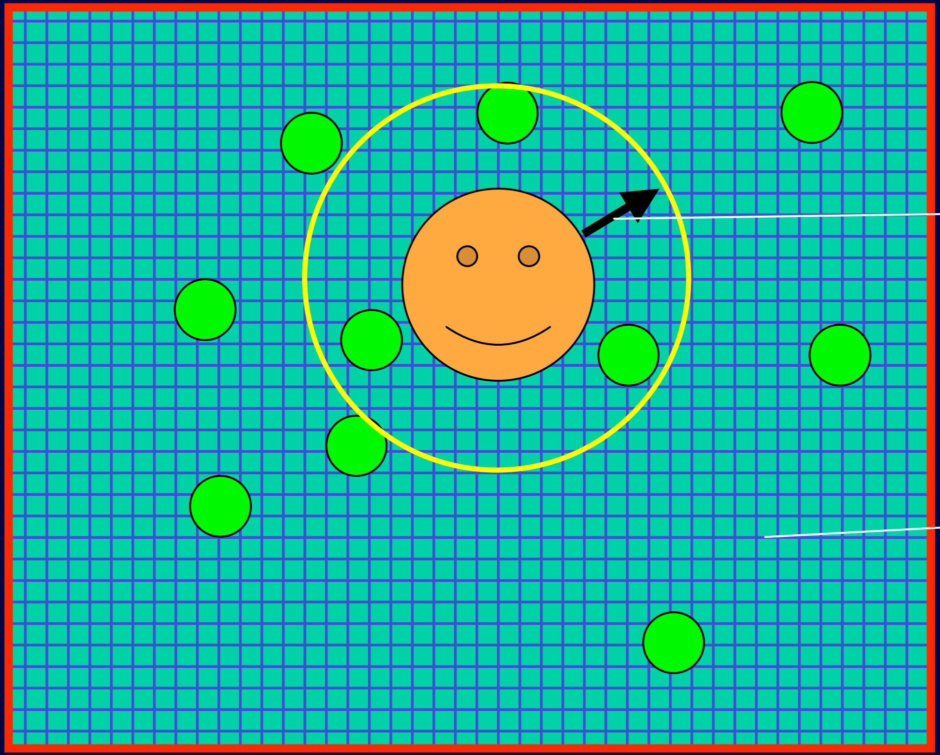
Sets the CHARMM variables to the averages over the trajectory, and prints the values to the logfile; optionally also to a file every timestep.

```
coor anal sele resn tip3 .and. type oh2 end -
    site sele resn asp .and. type od1 show end multi -
    firstu 21 nunit 1 skip 500   rhyd 3.0
```

NB! You need keyword MULTi if the solute (the SITE) has more than one atom.

Use COOR ANAL IHIST to get a 3D distribution (histogram)

Hydration number / 3D histogram



RHYD

Grid for histogram NB! Have to use oriented and recentered trajectory

MMTSB/CTBP Summer Workshop

Solvent dynamics

water self diffusion

The diiffusion coefficient D can be computed using the Einstein relation:

$$\lim_{t \to \infty} MSD(t) = 6Dt$$

MSD is the mean square displacement <(**r**(t)-**r**(0))^2> of a molecule in time *t*.

D is obtained from the slope of MSD vs *t* at ”long” time.

MSD calculation with CHARMM

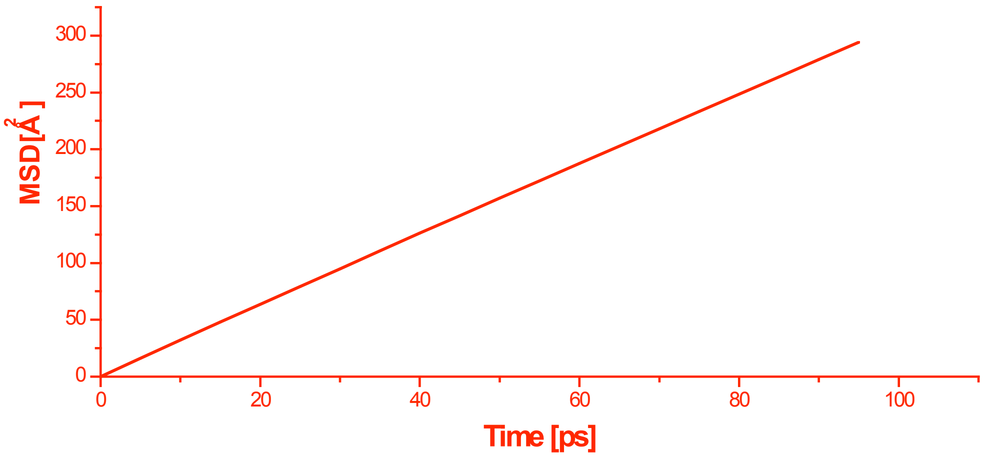
```
open unit 21 read unform name pept500.cor
open unit 31 write form name  pept500.msd
coor anal select type oh2 end  -
  firstu 21 nunit 1 skip 10 –
  imsd 31 rspin 0.0 rspout 999.9
  ncors 20 xbox @6 ybox @7 zbox @8
```

CHARMM will print an estimate of D, but you should plot MSD(t) and compute D from the slope of the linear part.

Result of MSD calculation

”pept500.msd”

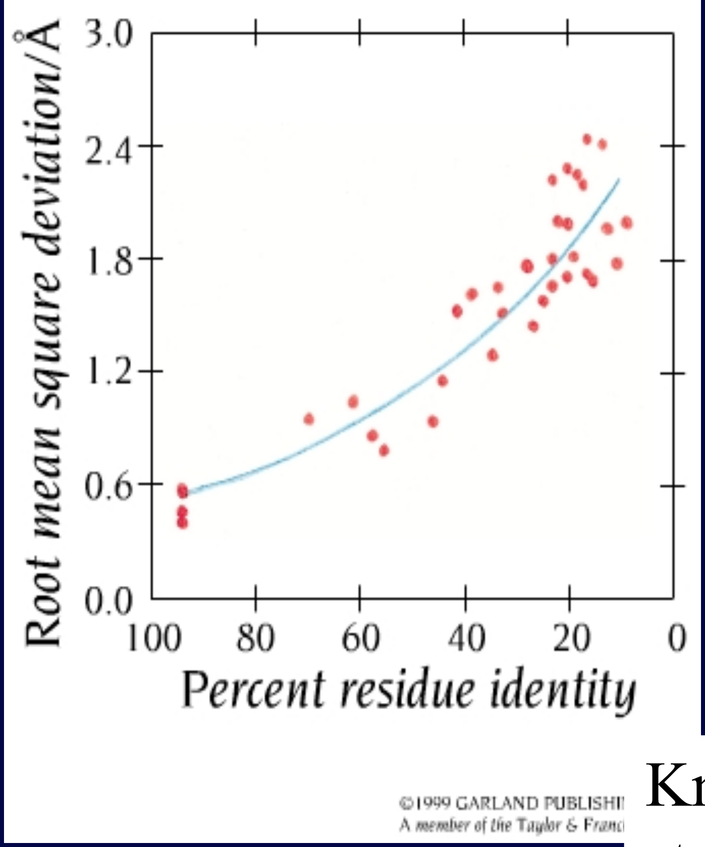
time (ps)	**MSD (Å^2)**
0.00000E+00	0.00000E+00
0.50000E+01	**0.16341E+02**
0.10000E+02	0.32182E+02
0.15000E+02	0.47965E+02
0.20000E+02	0.63677E+02
0.25000E+02	0.79330E+02
0.30000E+02	0.94947E+02
0.35000E+02	0.11055E+03
0.40000E+02	0.12612E+03
0.45000E+02	**0.14158E+03**



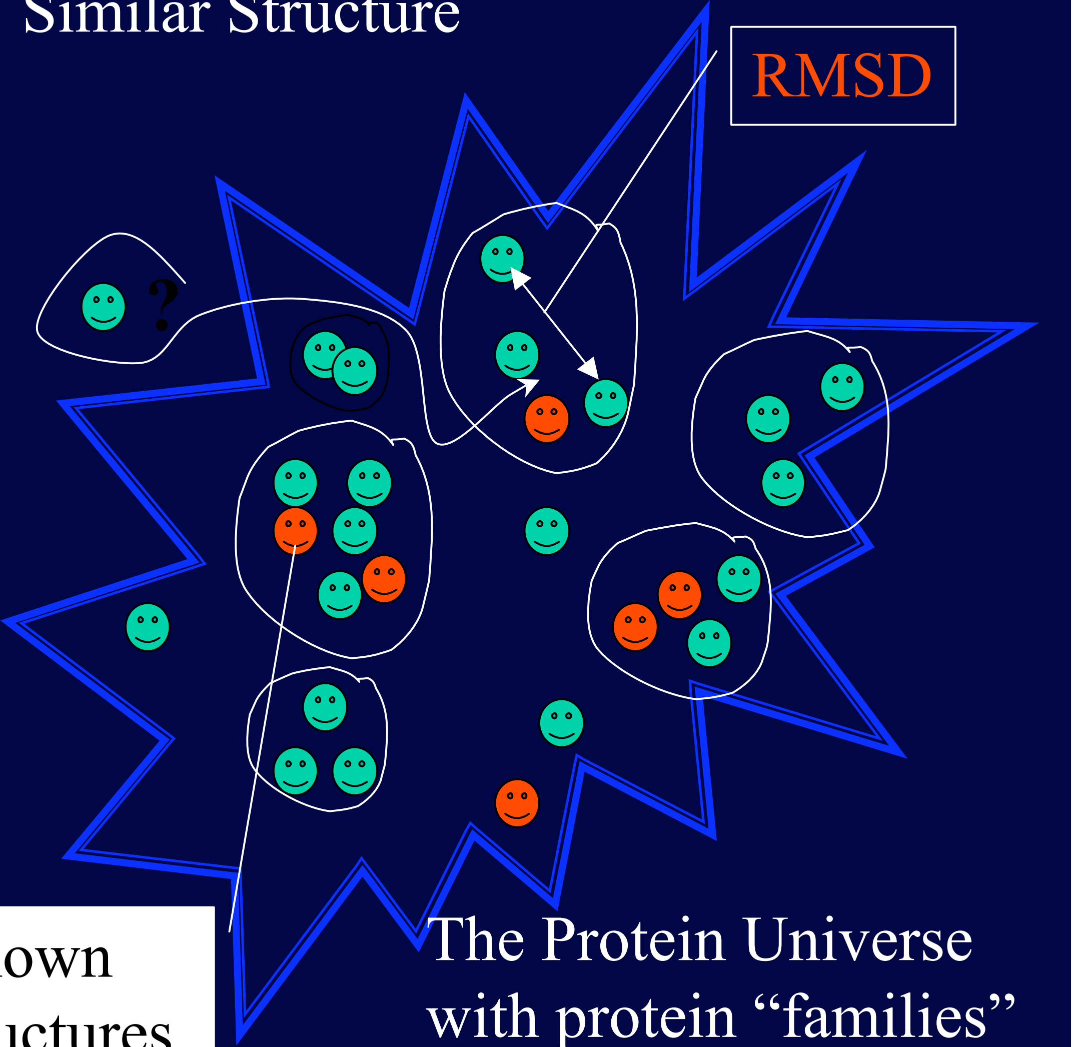
D=(MSD(45)-MSD(5))/(6* (45-5))=
(140-16)/240=0.53 [Å^2/ps]=5.3 10^{-9} [m^2/s]

Clustering structures

Similar Sequence → Similar Structure



©1999 GARLAND PUBLISHI
A member of the Taylor & Franci



RMSD

Known structures

The Protein Universe with protein "families"

Clustering with CHARMM

Cluster structures in a trajectory based on RMSDs in a space spanned by a set of backbone dihedrals; conformers with similar values for the specified dihedrals will be grouped together. The algorithm is briefly described in correl.doc (Karpen, M. E., Tobias, D. T., & Brooks III, C. L. (1993). Biochemistry 32:412-420.)

Any time series can be used, not only dihedrals. Mixing various types is permitted, but raises the issue of how to compare apples and oranges (=”apelsin” in Swedish, lit. ”apple from China”) (*eg*, normalize by variance).

Clustering with CHARMM, example

```
correl maxtimesteps 5000 maxseries 150 maxatoms 750

enter s30 dihe pept 30 n  pept 30 ca  pept 30 c pept 31 n
enter s41 dihe pept 41 n  pept 41 ca pept 41 c  pept 42 n
enter f42 dihe pept 41 c  pept 42 n pept 42 ca  pept 42 c

traj  firstu 62 nunit 3
! COMBINE TIME SERIES PRIOR TO CLUSTERING
edit s30 veccod 3
cluster s30 angle radius 30.0

end
```

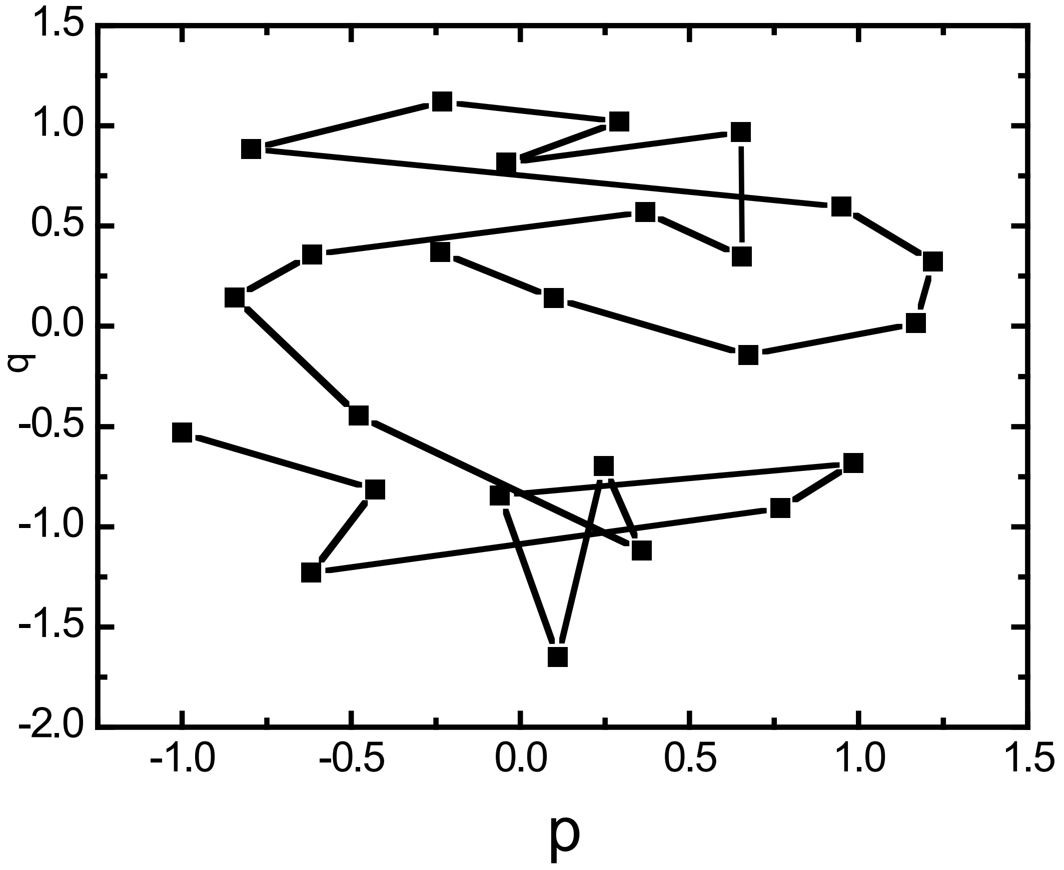
Cluster using projection of RMSDs

To cluster based on pair-wise RMSDs between structures in a trajectory the RMSDyn command can be used (dynamc.doc) (Levitt, M. J.Mol.Biol. (1983) 168, 621-657):

```
rmsdyn orient rms firstu 62 nunit 2 –
   begin 1000 SKIP 500 STOP 30000 –
   PQunit 21 PQseed 123
```

Each structure *i* in the trajectory is assigned a ”coordinate pair” (p_i,q_i) such that the distances between all the points thus obtained in the *p*-*q* plane approximates the true set of pairwise RMSD values.

p-q plot



Correlation function analysis

Time series data, f(t), and many relaxation phenomena, are often easy to characterize by correlation functions:

$$C(\tau) = \langle f(t) \cdot f(t+\tau) \rangle$$

The angular brackets denote averaging over all possible time-origins,*t*, in the trajectory; for a stationary process *t* is often omitted (replaced by ”0”); τ is the lag-time.

Example

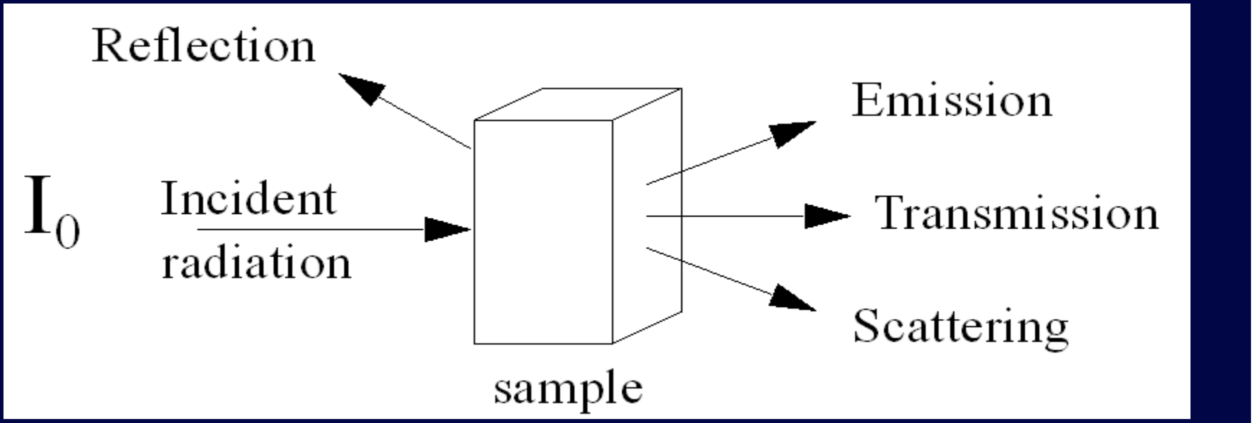
time-dependent fluorescence anisotropy, *r(t)*

A chromophore (Trp) is excited with a short pulse of vertically polarized light. Emission of vertically and horizontally polarized light is monitored as a function of time after the excitation. Initially most of the the emission is parallel to the excitation, but with time a random distribution is established.

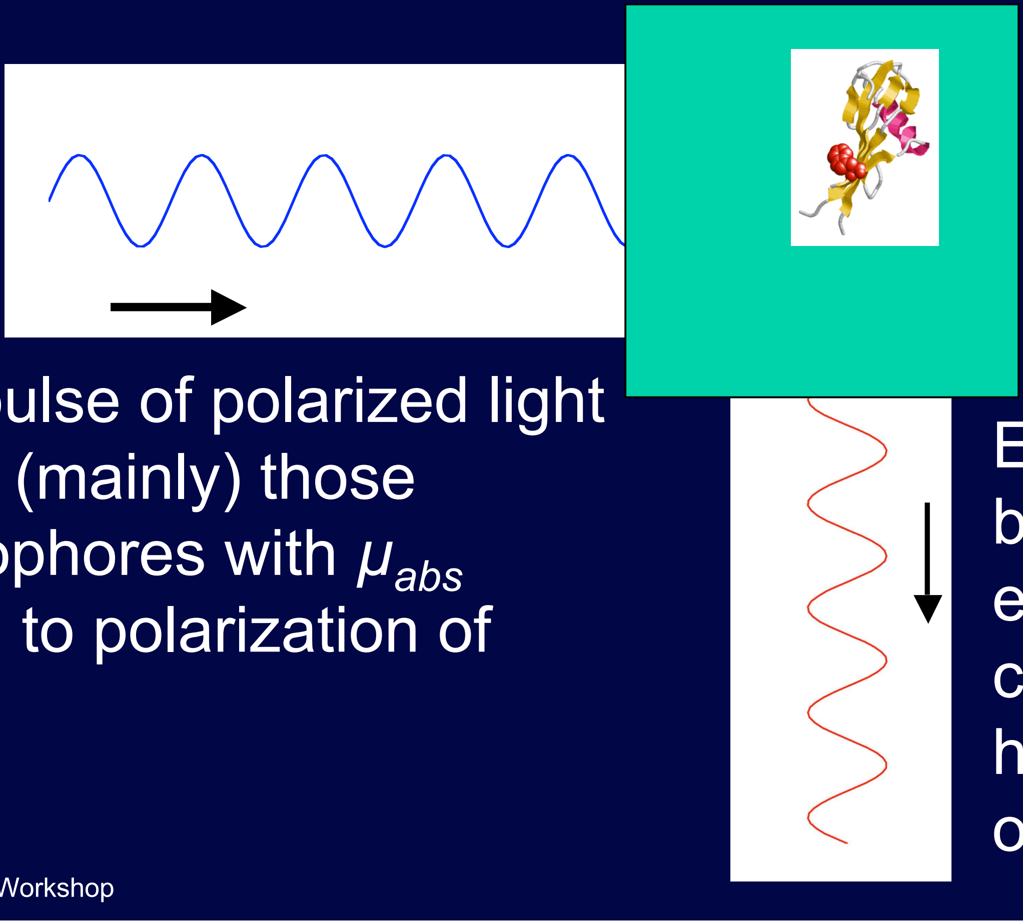
$$r(t) \propto \frac{I_{II}(t) - I_{\perp}(t)}{I_{II}(t) + 2I_{\perp}(t)} = \left\langle P_2(\widehat{\mu}_{abs}(0) \bullet \widehat{\mu}_{em}(t)) \right\rangle$$

r(t) tells us about rotational diffusion of the chromophore (possibly also about the host macromolecule). For a rigid sphere r(t) is a single exponential, *r(t)*=0.4exp{-t/τ}, characterized by the decay time τ; $P_2(x)$ is the second order Legendre polynomial $P_2(x)=(3x^2-1)/2$

Optical spectroscopy



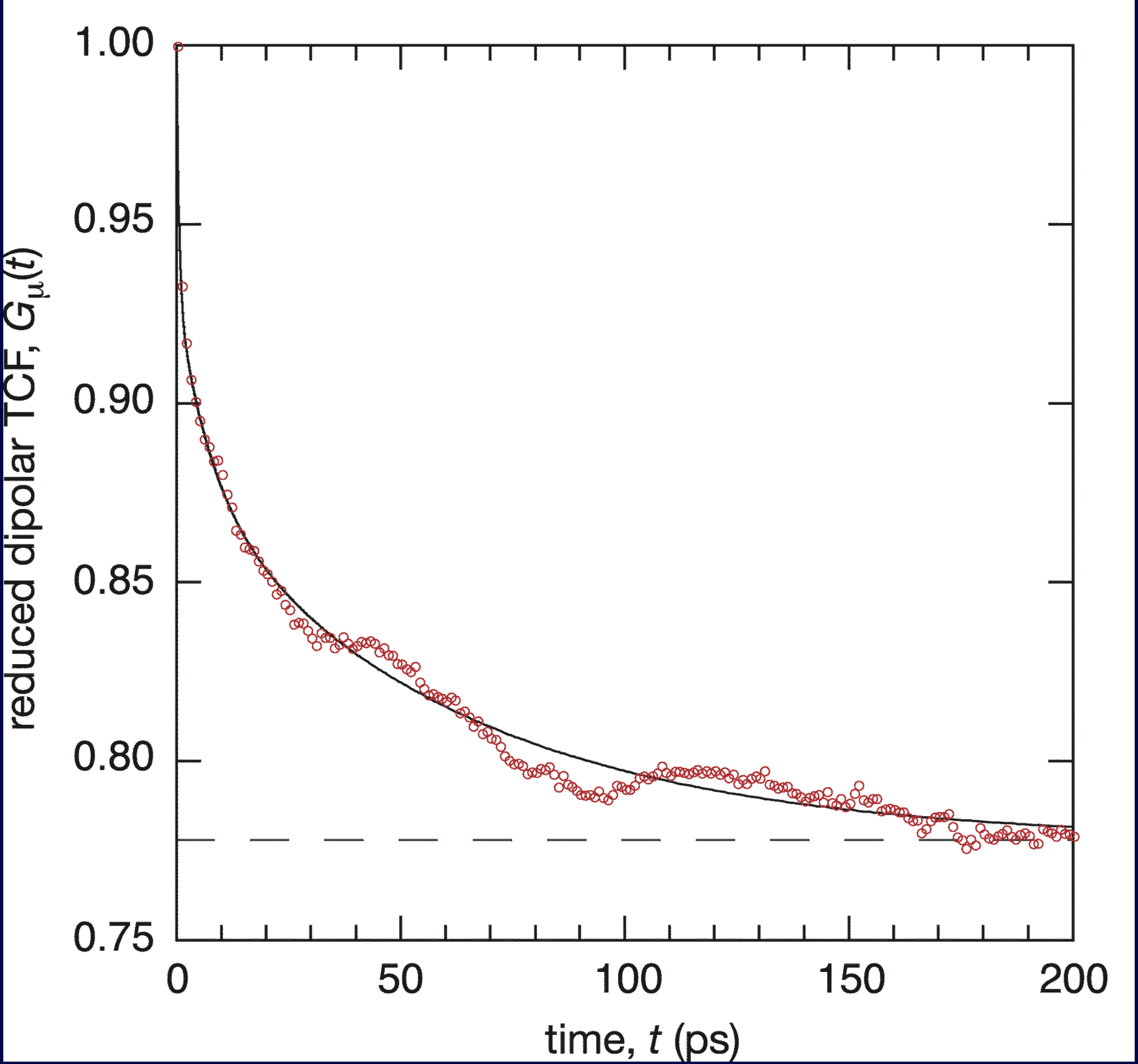
Time-resolved fluoresence



Short pulse of polarized light excites (mainly) those chromophores with μ_{abs} parallel to polarization of pulse

Emission occurs briefly (ps-ns) after excitation, when chromophore may have changed orientation

Example of calculated *r(t)*

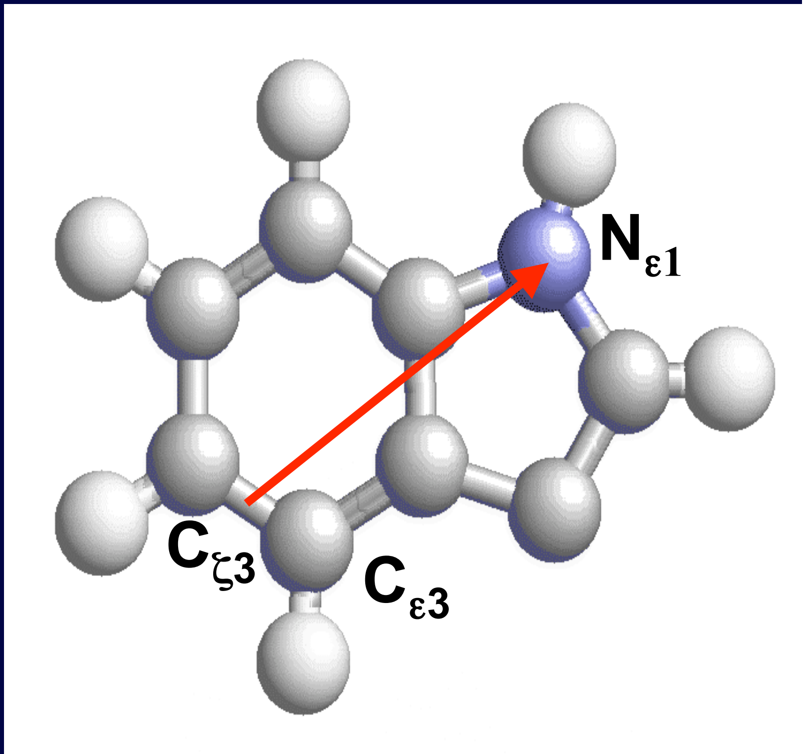


Computing *r(t)* in CHARMM

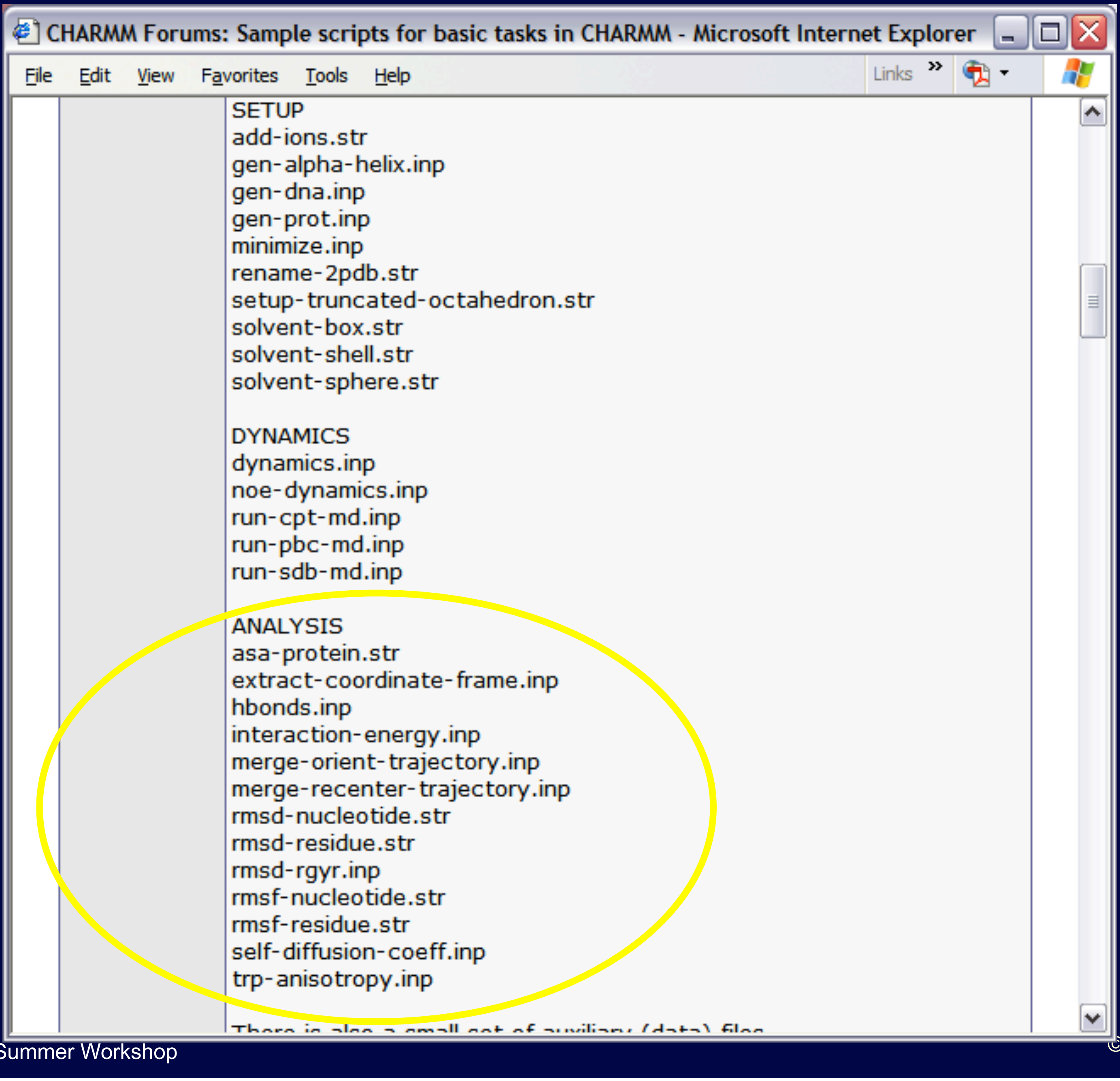
```
correl maxtime 2000 maxseries 10 maxatom 20
ENTER LA  VECTOR XYZ prt 8 NE1  prt 8 CZ3 -
                 prt 8 NE1  prt 8 CE3

traj firstu 62  nunit 2
mantime la normal
corfun la la p2
write corr dumb time unit 31
*hi
*
```

Computes rank 2 correlation function for a unit vector along the average of vectors NE1-CZ3 and NE1-CE3.
This approximates the L_a transition dipole in Trp; in reality a second dipole is also involved (corfun la lb p2).



www.charmm.org -> Forums -> Script Archive


CHARMM Forums: Sample scripts for basic tasks in CHARMM - Microsoft Internet Explorer
File
Edit
View
Favorites
Tools
Help
Links

SETUP
add-ions.str
gen-alpha-helix.inp
gen-dna.inp
gen-prot.inp
minimize.inp
rename-2pdb.str
setup-truncated-octahedron.str
solvent-box.str
solvent-shell.str
solvent-sphere.str

DYNAMICS
dynamics.inp
noe-dynamics.inp
run-cpt-md.inp
run-pbc-md.inp
run-sdb-md.inp

ANALYSIS
asa-protein.str
extract-coordinate-frame.inp
hbonds.inp
interaction-energy.inp
merge-orient-trajectory.inp
merge-recenter-trajectory.inp
rmsd-nucleotide.str
rmsd-residue.str
rmsd-rgyr.inp
rmsf-nucleotide.str
rmsf-residue.str
self-diffusion-coeff.inp
trp-anisotropy.inp

There is also a small set of auxiliary (data) files

INCONCLUSIVE EXPERIMENT:
PAVLOV'S CAT