CHARMM Analysis Tools

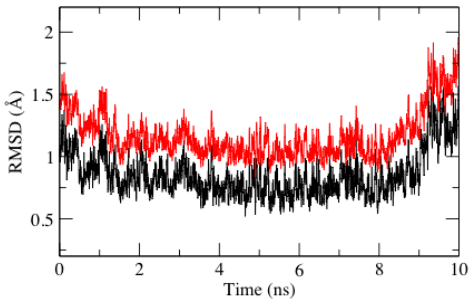
Jianhan Chen

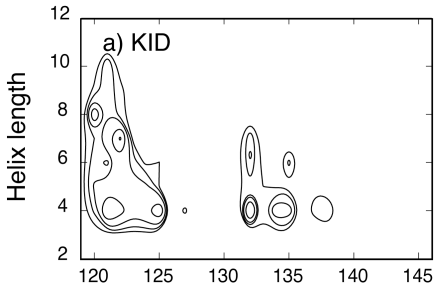
Kansas State University

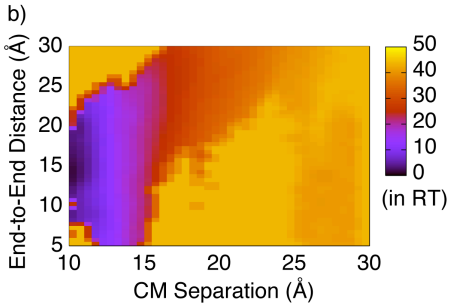
Credits: This lecture and associated tutorials were built upon materials originally created by **Lennart Nilsson (**Karolinska Institutet) for the MMTSB/CTBP Workshop of 2006.

Overview and Purposes

- Main CHARMM analysis modules
- Examples of "typical" analysis
  - Structural properties, solvent, clustering …
  - Provide a better understanding of the philosophy and basic flow of CHARMM analysis
  - Improve ability to sort through/understand the CHARMM documents to locate the tools for the specific analysis that is needed
- The number of analysis tools in CHARMM still dazzle me
  - Find help: tutorials, test cases, CHARMM forum, "experts" …
  - Treasure hunting?





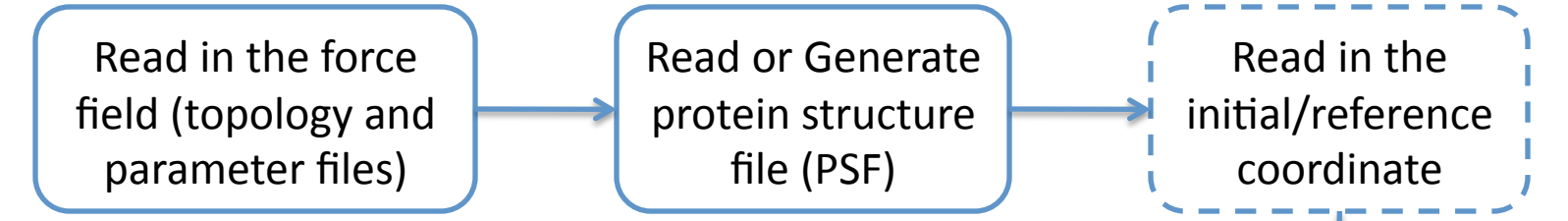


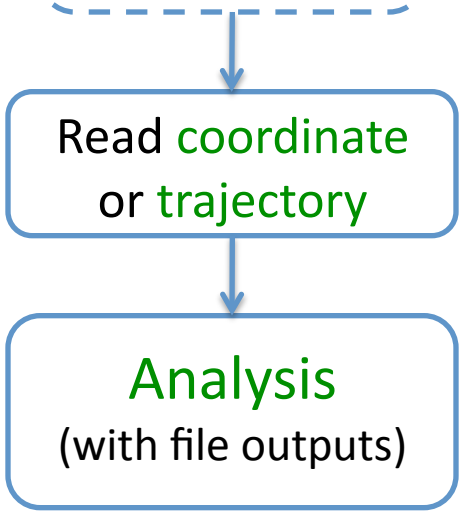
What is “analysis”?

- The process of extracting certain properties from simulation
  - Often for the purpose of assessment, validation or prediction
  - The important focus of connecting to experiment
- Some 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, Rg, …) – convergence?
  - Properties of a single structure – representative?
  - Properties of an ensemble – how many?
  - Time-dependence – time scale?
- This lecture (and associated tutorials) only focus on the technical aspects.

Basic Flow of Operations in CHARMM

- Remember that CHARMM only provides functional blocks.
- The user needs to write programs, or, CHARMM "input scripts", to specify the composition and flow of operations





Almost all CHARMM analysis is done by post-processing of a coordinate or trajectory (a collection of coordinates) file.

A trajectory may be processed recursively frame by frame or as a whole within the input script depending on the analysis tool.

```
* Compute CA rmsd between two coordinates
*

!Read topology and parameter (aka force field)
read  rtf card name toppar/top_all22_prot.inp
read para card name toppar/par_all22_prot.inp

!Read PSF and initial/reference coordinates
read psf  card name 3gb1_solv.psf
read coor pdb resid name 3gb1_solv.pdb

!Save a copy in the COMParison coordinate set
coor copy comp

!Read another set of coordinates
read coor pdb resid name nptprod.pdb

!Compute the heavy atom and CA RMSD values
coor orient rms select type CA end
set carms = ?rms

write title name rmsd.dat
* CA rms = @carms
*

stop
```

title

(vs comments)

force field

PSF
reference coor.

coor. to analyze

RMSD analysis

Output

Extract from “rmsd-singlepdb.inp”

Internal and User-Defined Variables

- Internal variables: listed in subst.doc
  - Values assigned by related charmm operations (e.g., analysis).
  - Accessed by `?VARIABLE_NAME`
  - Not directly modifiable in input scripts
- User defined variables (see miscom.doc)
  - Created and manipulated directly within an input script
  - To create or modify:
    ```
    set  VARIABLE = VALUE
    calc VARIABLE = …
    ```
  - Accessed by `@VARIABLE_NAME`
- Special user accessible variables `@1 – @9`
  - Can be used in system command calls
- SCALar: access to internal data (charges, mass, …) and allow simple array arithmetic within the input scripts (scalar.doc)

```
read  rtf card name toppar/top_all22_prot.inp
read para card name toppar/par_all22_prot.inp
read psf  card name 3gb1_solv.psf
read coor pdb resid name 3gb1_solv.pdb
coor copy comp

!Open file unit of trajectory input
open read unit 13 file name nptprod.dcd

!Open the output file and write header
open write unit 11 card name rmsd-rgyr-correl.dat

!Invoke CORREL mode
correl maxtime 1000
   !request RMS with orient and radius of gyration
    enter v1 rms orient
    enter v2 gyra

   !specify the trajectory to process
    traj firstu 13 nunit 1

   !write the time series to a file
    edit v1 veccod 2
    write v1 dumb time unit 11
end  ! Exit CORREL

stop
```

force field

PSF
reference coor.

traj to analyze

Output file

Request RMSD
and Rg analysis

Analyze!

Write output

Extract from “rmsd-rgyr-traj-correl.inp”

Main CHARMM Analysis Modules

- CORMAN: single coordinate sets, averages of trajectories (corman.doc)
  - Basic format: `COOR XXXX`
- QUICK: quick analysis of single coordinate sets (miscom.doc)
- Energies: ENERgy and INTEaction energy (energy.doc)
- CORREL: time-series and correlation functions from trajectories (analyze trajectories) (correl.doc)
  - A set of sub-commands within the CORREL block.
- Solvent analysis: COOR ANAL, RDFSOL (rdfsol.doc)
- NMR analysis: NMR (nmr.doc)
- Quasi-harmonic modes: VIBRAN (vibran.doc)

Trajectories

I: Single Structure Analysis

- Structural geometry data (`COOR` or `QUICk`)
  - Distance, angle, dihedral, radius of gyration, least-squares-plane through set of atoms, sugar conformations (puckering), helix orientation, geometry center or center of mass, …
- Energy data (`ENERgy` or `GETE` or `INTEraction`)
  - Energy: all terms accessible by internal variables (e.g., ?ener)
  - Forces, contributions from each atom (or subsets of atoms), interaction energy between specified sets of atoms
- Composite data
  - Secondary structure analysis: `COOR SECS` (DSSP-based)
  - Hydrogen bonds (`COOR HBONd`), RMSD, solvent accessible surface area (`COOR SURF`), volume, cavities, contacts, inertia tensor, ….
- Single structure analysis can be put in a loop to analyze trajectories

Geometric Examples

`quick 34 41` prints distance between atoms 34 and 41

`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 that only dependent on atoms in one set are not computed:

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

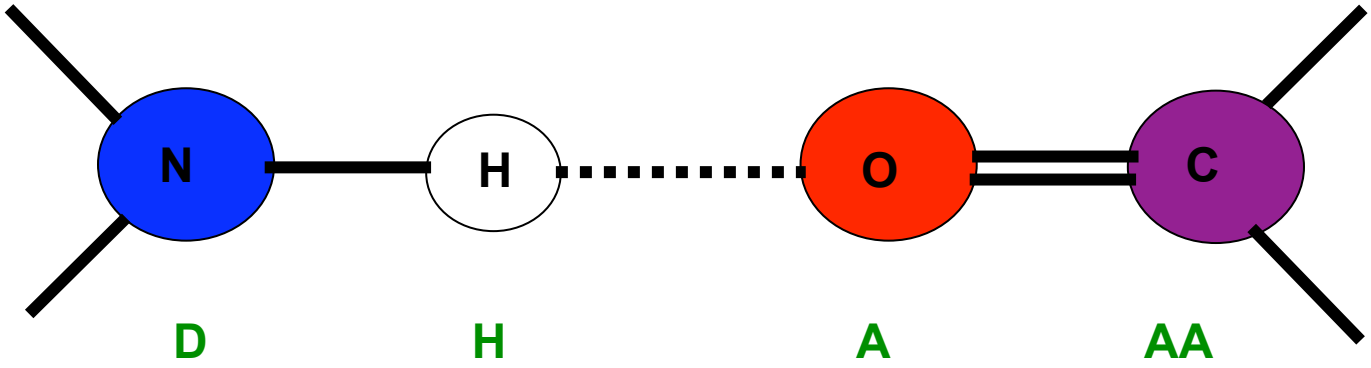
after an energy evaluation the force (kcal/mol/Å) on each atom is available:

```
coor force comp  ! copy force into comparision set
scalar xcomp show sele type ca end  ! print Fx on all CA atoms
```

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

```
analysis on                  ! turn on the parition analysis
skipe bond
energy
scalar econt show            ! print atomic contributions
echo ?ener ?elec ?grms       ! energy.doc
```

Hydrogen Bonding Analysis



- Several parameters can be used to characterize the h-bond:
  - r(D-A), r(H-A), ∧(D-H-A), ∧(H-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)
  - These specifications are used by `COOR HBONd` analysis facility

Hydrogen Bond Criteria

Table II. Comparison between

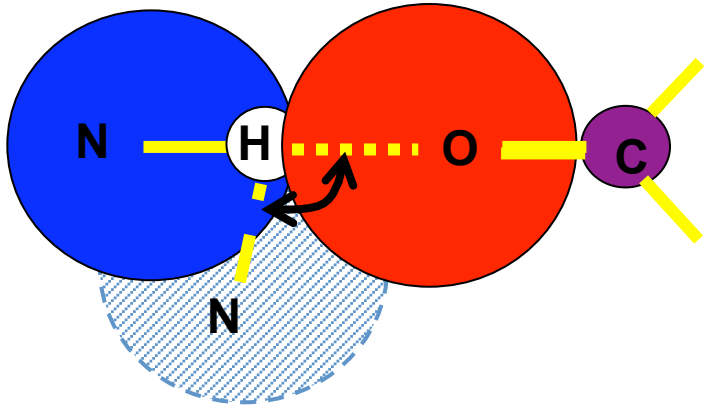
criterion	*i*–*i*+4 (%)
O–H < 2.4 Å	84.1
O–H < 2.6 Å	78.5
O–N–H < 35°	
O–H < 2.5 Å	85.3
N–H–O > 120°	
O–H < 2.4 Å	68.9
N–H–O > 135°	
O–N < 3.4 Å	91.8
N–H–O > 90°	
O–N < 3.4 Å	91.7

Ref. De Loof, H., Nilsson, L., and Rigler, R. (1992) JACS *114*, 4028-35.

- 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.

H-Bond Analysis for a Single Structure

- 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 segi prot end sele segi 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 segi prot end sele segi 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.

Some PBC can be handled. The tool `COOR CONTact` does not use acceptor/donor lists in the PSF (purely distance based).

Protein Secondary Structure Analysis

- Proteins usually have a high content of secondary structure elements, α-helices and β-sheets
- 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 default hydrogen bond criterion here is r(H-A)<2.6Å.
- Example:

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

Finds secondary structures in the first selection, within the context of the second selection, e.g., 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 (`?nalpha ?nbeta`), and returns flags in `WMAIN` array (0: coil, 1: alpha, 2: beta).

II: Analysis of Trajectories

- To obtain averages/distributions (`CORMAN` etc), time series (`CORREL` etc)
- Several modules/commands can directly access trajectories (coordinates or velocities), eg `CORREL` and `NMR`
- Trajectory specification: used in multiple major commands

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

Access *k* binary 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.

Example: Compute average structure and RMS fluctuations (in `WMAIN`):

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

H-Bond Analysis of Trajectories

```
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.

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:

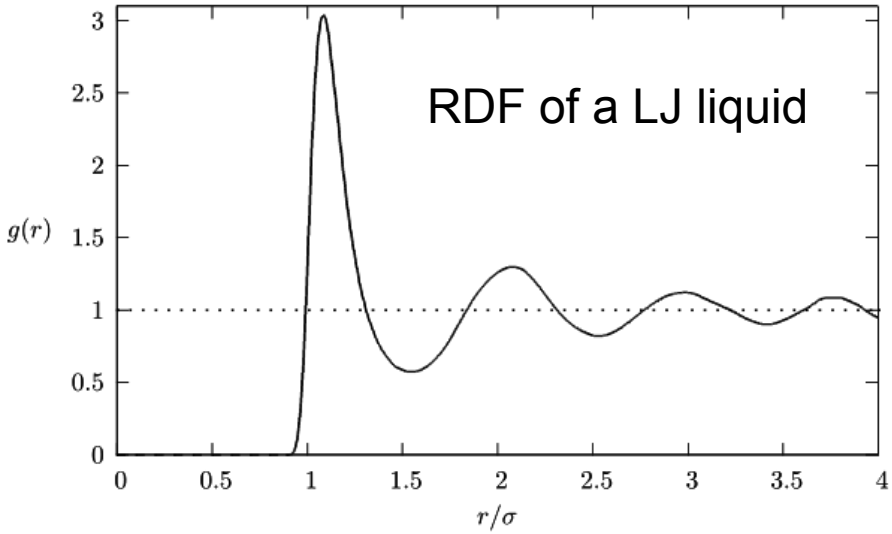
```
coor hbond first 51 nunit 3 skip 2500 –
sele segid prot end sele segid wat irhi 21 ithi 22
```

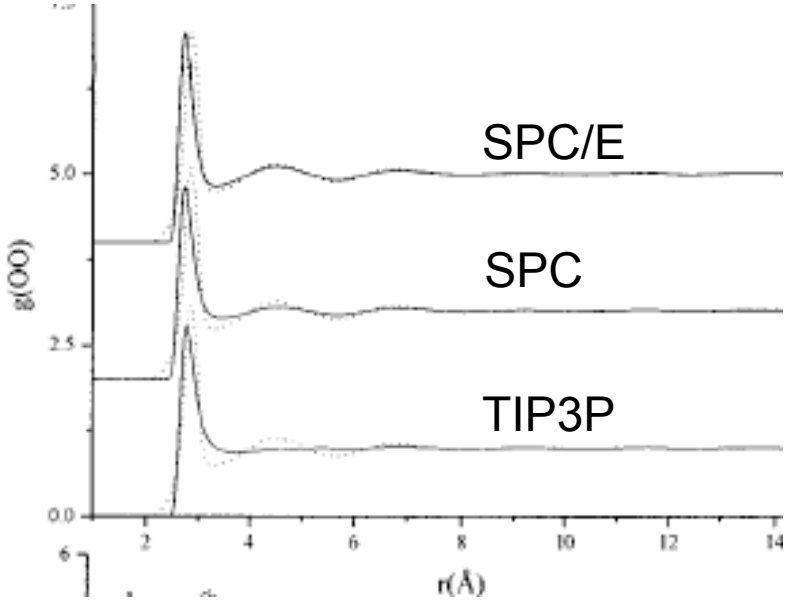
Solvent Structural Analysis

- Radial distribution function (**RDF**)

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

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.





RDF Analysis in CHARMM

CHARMM tools: `RDFSOL` and `COOR ANALysis`

Example 1: g(r) for waters; the program defaults are used to calculate the density using selected atoms within 10 A (`RDSP` keyword) of the reference point (0,0,0) (`REF` keyword)

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

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

RDF Analysis in CHARMM

CHARMM tools: `RDFSOL` and `COOR ANALysis`

Example 2: g(r) for water oxygens wrt backbone amide hydrogens

```
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
```

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

Several types of analysis (with some exceptions) may be combined in a single "`coor analysis`" command (see corman.doc).

COOR ANAL: Hydration Analysis

Calculate number of solvent molecules within a specified distance of a site (invoked by specification of `rhyd` and/or `ihydn`):

`?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 (if `IHYDN>0`).

```
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
```

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

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

COOR ANAL: 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 $\langle(\mathbf{r}(t)-\mathbf{r}(0))^2\rangle$
- CHARMM calculation of MSD (invoked by `imsd`)

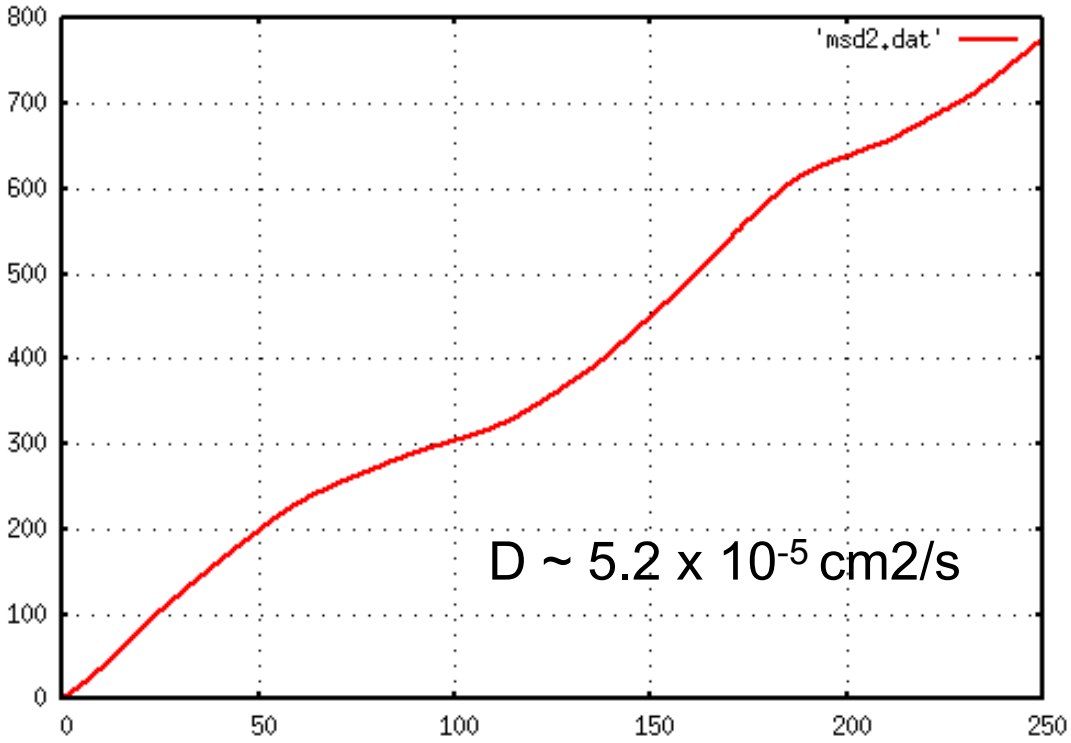
```
open unit 21 read unform name pept500.cor
open unit 31 write form name  pept500.msd
coor anal select type oh2 end  -
  firstu 21 nunit 1 skip 10 –
  imsd 31 rspin 0.0 rspout 999.9 ncors 100 -
  xbox @6 ybox @7 zbox @8
```

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

A Sample MSD Function

"msd2.dat"

time (ps)	MSD (Å^2)
0.000000	0.000000
0.200000	0.666211
0.400000	1.471468
0.600000	2.197925
0.800000	2.877727
1.000000	3.546650
1.200000	4.227092
1.400000	4.904824
1.600000	5.579034
….	



Calculated from a 1 ns simulation of a TIP3P water box.

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
```

```
read  rtf card name toppar/top_all22_prot.inp
read para card name toppar/par_all22_prot.inp
read psf  card name 3gb1_solv.psf
read coor pdb resid name 3gb1_solv.pdb
coor copy comp

!Open file unit of trajectory input
open read unit 13 file name nptprod.dcd

!Open the output file and write header
open write unit 11 card name rmsd-rgyr-correl.dat

!Invoke CORREL mode
correl maxtime 1000
   !request RMS with orient and radius of gyration
    enter v1 rms orient
    enter v2 gyra

   !specify the trajectory to process
    traj firstu 13 nunit 1

   !write the time series to a file
    edit v1 veccod 2
    write v1 dumb time unit 11
end  ! Exit CORREL

stop
```

force field

PSF
reference coor.

traj to analyze

Output file

Request RMSD and Rg analysis

Analyze!

Write output

Extract from "rmsd-rgyr-traj-correl.inp"

```
read  rtf card name toppar/top_all22_prot.inp
read para card name toppar/par_all22_prot.inp
read psf  card name 3gb1_solv.psf
read coor pdb resid name 3gb1_solv.pdb
coor copy comp

!Open file unit of trajectory input
open read unit 13 file name nptprod.dcd

!Open the output file and write header
open write unit 11 card name rmsd-rgyr-corman.dat

!Process a trajectory using a loop
traj firstu 13 nunit 1
set inx = 1
label nextframe
    traj read
    coor orient rms select .not. resn tip3 end
    coor rgyr select .not. resn tip3 end
    write title unit 11
    *@inx ?rms ?rgyr
    *
incr inx by 1
if inx le 100 goto nextframe

stop
```

force field

PSF
reference coor.

traj to analyze

Output file

Traj specification

Start of loop

RMSD and Rg analysis

Write output

End of loop

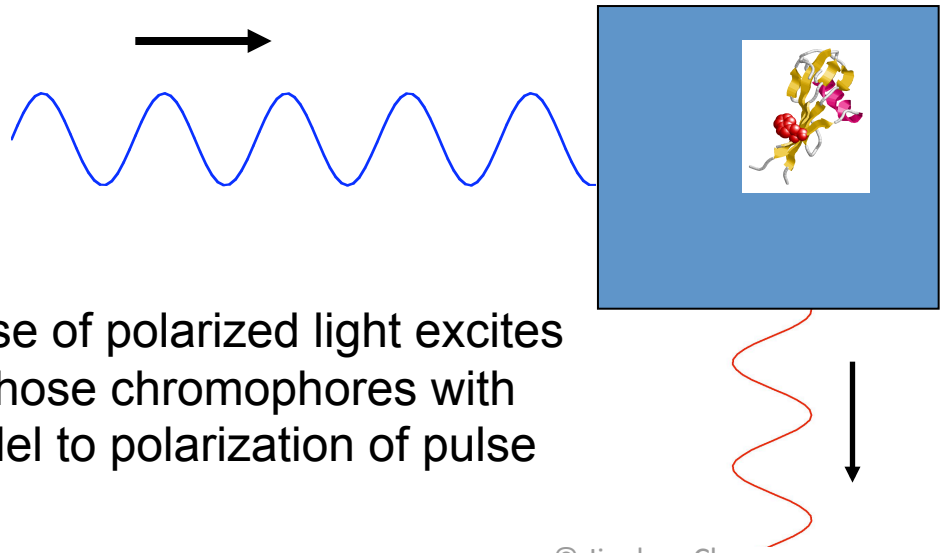
Extract from "rmsd-rgyr-traj-corman.inp"

Correlation Function Analysis

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

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

  - Provide direct links to experimental observables (time scale, order parameter, spectral density, etc) (MSD is a correlation function!)
- **Example**: time-dependent fluorescence anisotropy



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

Time-dependent fluorescence anisotropy, r(t)

- r(t) is related to rotational diffusion of the chromophore.

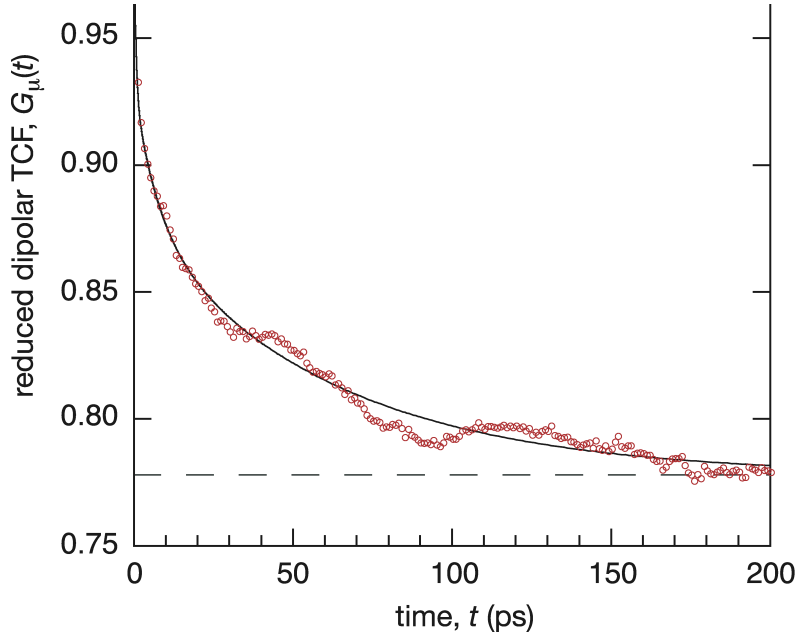
$$r(t) \propto \frac{I_{II}(t) - I_{\perp}(t)}{I_{II}(t) + 2I_{\perp}(t)} = \left\langle P_2(\hat{\mu}_{abs}(0) \bullet \hat{\mu}_{em}(t)) \right\rangle$$

  - $P_2(x) = (3x^2 - 1)/2$ is 2nd order Legendre polynomial.
  - For a rigid spherical molecule, r(t) ~ exp (-t/τ).

- Effective correlation time

$$\tau_e = \frac{\int_0^{\infty} (r(t) - r(\infty))dt}{r(0) - r(\infty)}$$

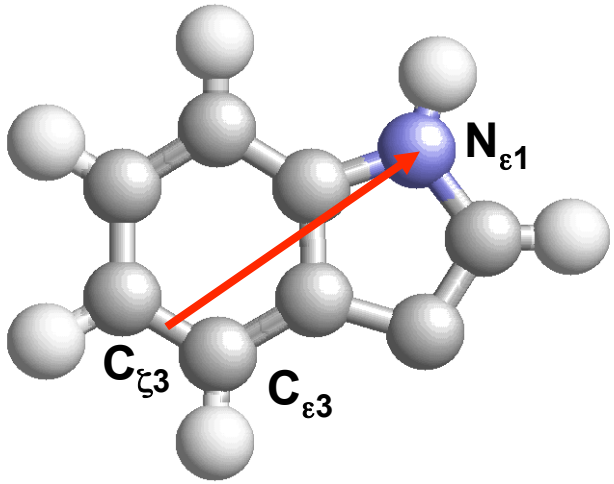
  - $\tau_e = \tau$ for single exponentials



Computing r(t) in CHARMM CORREL

```
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                    ! Normalize vector
corfun la la p2                      ! Compute p2
write corr dumb time unit 31
END
```

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



NMR Relaxation Analysis

- Implemented with NMR module (nmr.doc)
- Compute relaxation constants (R_1, R_2, NOE) at a specified field strength, given an estimate of molecular tumbling time.
- All relaxation constants are related to the orientational correlation function of the bond vector (e.g., N-H). The theory is quite involved.
- Besides the relaxation constants, also print out order parameters, effective correlation times. One can request output of the bond vector time series r(t) and/or the correlation functions c(t).
- A few references:
  - Bruschweiler et al., *JACS* 114, 2289 (1992)
  - Chen et al., *J. Biomol. NMR,* 29, 243 (2004)

Example of NMR Relaxation Analysis

```
NMR
  RTIMES select type N  end select type HN end

  DYNA firstu 51 nunit 1 -
     rtumbl 9000.0 hfield 11.74 cut 1.5 tmax 5000 -
     ilist 21 iwrite 11 C(t) R(t)  ! Output options
END
```

This example was to process a 10-ns trajectory of DHFR in TIP3P water.

`RTIMES` requests relaxation analysis for all vector between atom types N and HN within `cut 1.5` to be computed.

Assuming that the molecular tumbling time is 9 ns (`rtumbl 9000.0`) and the magnetic field strength is 11.74 T (`hfield 11.74`).

`tmax 5000` is the length of correlation functions. It is an important parameter and can have a large influence on the results. One needs to examine the correlation function `C(t)` to check if the choice is appropriate. `tmax` typically should not be greater than half of the total simulation length.

Clustering Structures

- Multiple methods are available in CHARMM for clustering
- `CORREL` contains a cluster algorithm (see correl.doc)
  - Mainly intended for clustering based on RMSDs of a set of backbone dihedrals, but could in principle for any time series (caution needs to be taken in mixing various properties with very different variants).
- `RMSDyn` can be used for clustering based pair-wise RMSDs betweens structures in a trajectory (see dynamc.doc)
  - Generate a 2D p-q projection for visualization
- The MMTSB Tool Set also contains a tool (enscluster.pl) for structure clustering.
  - Need to convert trajectory to MMTSB structure ensemble
  - Multiple clustering algorithms available
  - more convenient post-processing: many "routine" analysis can be quite conveniently done with an array of tools!

Example of Clustering with CORREL

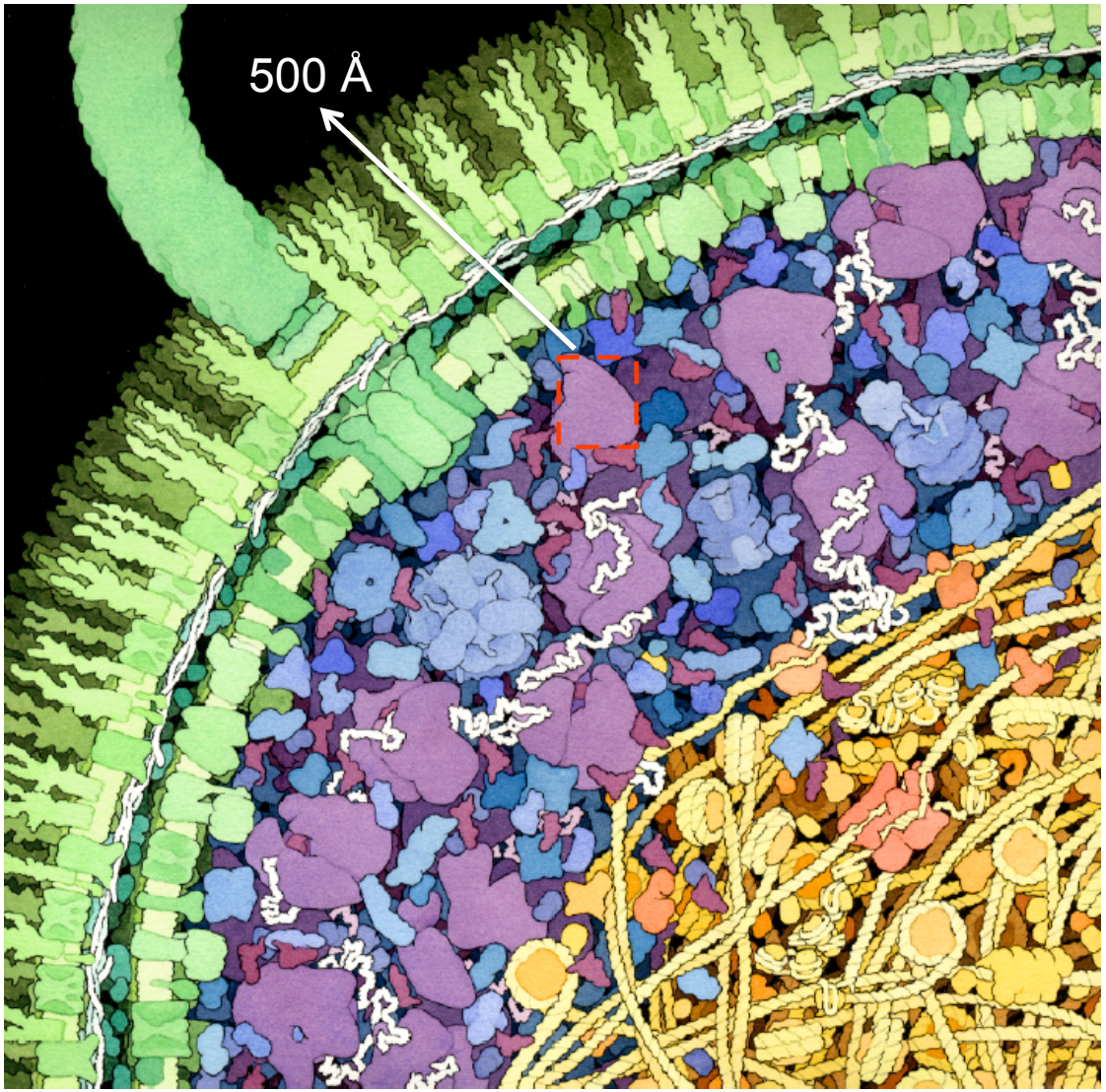
```
correl maxtimesteps 5000 maxseries 150 maxatoms 750

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
edit s30 veccod 3  ! Combine time series for clustering
cluster s30 angle radius 30.0

end
```

The CLUSter command clusters these data into groups with similar time series values, with each cluster being defined by a "cluster center". The cluster centers are output to `UNICluster`, and a list of time points and assigned clusters is given in the cluster membership file (`UNIMember`).

Credit: Lennart Nilsson

Centre for Structural Biochemistry
Department of Biosciences and Nutrition
Karolinska Institutet
http://www.csb.ki.se/md/md.html

Inconclusive Experiments



see also, Bionanotechnology D.S. Goodsell 2004 Wiley

Extreme simplification

Limited force field accuracy

Large gaps in timescales