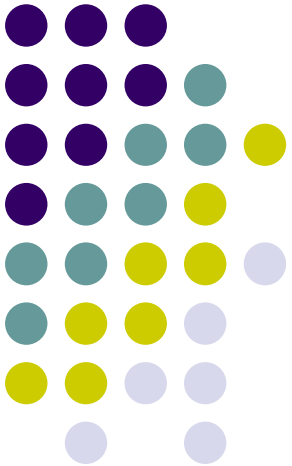
Implementing MD Simulations and Principal Component Analysis to Calculate Protein Motions

Myrna I. Merced Serrano

Computational and Systems Biology Summer Institute

Iowa State University

Ames, IA 50010



MD

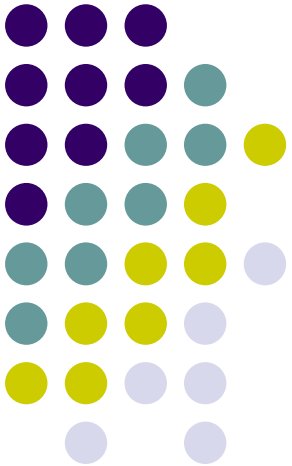
- Molecular Dynamics is a way to study atoms and molecules with computer simulations.
- This study of atoms and molecules is based on the laws of physics.

PCA

- Principal Component Analysis is a technique to reduce multidimensional data sets to lower dimensional data.

Goal

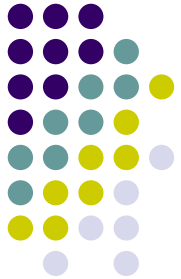
To perform MD Simulations to study large domain protein motions.



Data

- Protein pdb's files:
  - 1AL3→ CysB
  - 1LST→ LAO (closed form)
  - 2LAO→ LAO (open form)
  - 1PDA→ PBGD
  - 2DRI→ RBP
  - 1TFA→ OVOT
  - 2LIV→ LIVBP

Software



- Cygwin
  - Linux environment for Windows.
- NAMD
  - MD code to perform simulations of large biomolecular systems.
- MatLab
  - Programming language and numerical computing environment for algorithm development, data visualization, and data analysis.

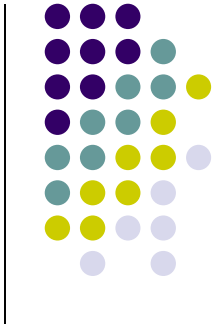
Methods

- Download the pdb files of the proteins from http://www.pdb.org.
- Development of a bash script to generate the files to run the minimizations in NAMD.
- Minimize each protein for 2,500 steps.
- Development of a bash script to generate the configuration file to run the MD.
- Run each protein for 3ns (1,500,000 steps).
- Development of a bash script to generate the input matrix for the PCA.
- Analyze each trajectory with the PCA in MatLab.
- Plot the results of the PCA (MatLab and Excel).

Script to generate the files

```
#Script to generate the files for the minimization in NAMD

#!/bin/bash

COMMON_DIR=/home/mymese/inv_summer_2007/data/common
DISTANCE_PROT_BOX=5

#write de pdb protein file
echo -e "mol load pdb $1.pdb\\n set prot [atomselect top protein]\\n \$prot writepdb $1_p.pdb\\n quit\\n" |
     /usr/vmd/vmd -dispdev text

#create the pgn file
/bin/sed -n -e "s/MOL/$1/" -e "w $1.pgn" $COMMON_DIR/template.pgn

#run the pgn file
echo -e "source $1.pgn\\n quit\\n" | /usr/vmd/vmd -dispdev text

#put the protein in a box of water
echo -e "package require solvate\\n solvate $1_psfgen.psf $1_psfgen.pdb -t $DISTANCE_PROT_BOX -o $1_wb \n quit\\n" |
     /usr/vmd/vmd -dispdev text

#load psf and pdb into VMD and calculate the center of the box
box_center=`echo -e "mol load psf $1_wb.psf\\n mol load pdb $1_wb.pdb\\n set prot [atomselect top \"all\"]\\n
     measure center \\$prot\\n quit\\n" | /usr/vmd/vmd -dispdev text | tail -3 | head -1 | cut -d ' ' -f3-5`

#load psf and pdb into VMD and calculate the minmax of the box
box_minmax=`echo -e "mol load psf $1_wb.psf\\n mol load pdb $1_wb.pdb\\n set prot [atomselect top \"all\"]\\n
     measure minmax \\$prot\\n quit\\n" | /usr/vmd/vmd -dispdev text | tail -3 | head -1 | cut -d ' ' -f3-8`

#values of box center and measures of minmax
echo "Box center: $box_center"
echo "Box minmax: $box_minmax"

#copy the psf and pdb to common directory
cp $1_psfgen* $COMMON_DIR/.
cp $1_w* $COMMON_DIR/.

#create the configuration file
/bin/sed -n -e "s/MOL/$1/" -e "s/BOX_CENTER/$box_center/" -e "w $1_conf_wb.conf" $COMMON_DIR/template_wb.conf
```

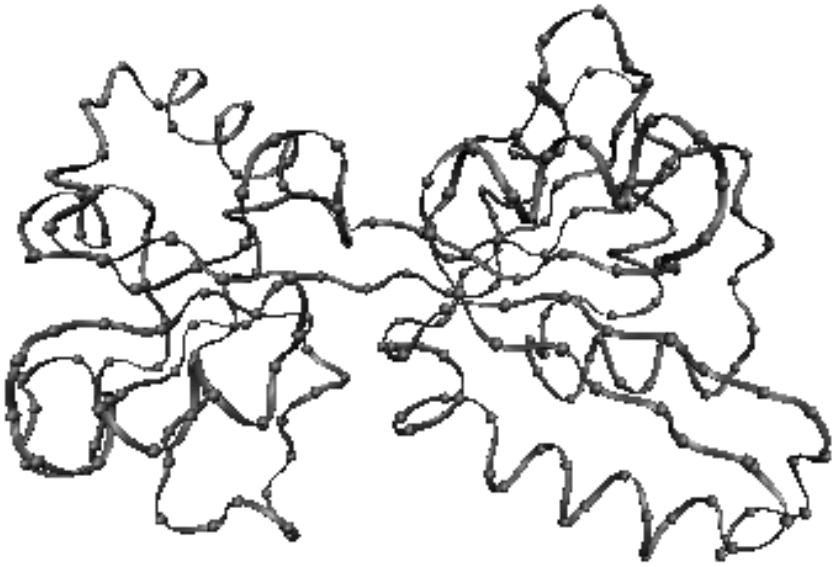
Script to generate the matrix



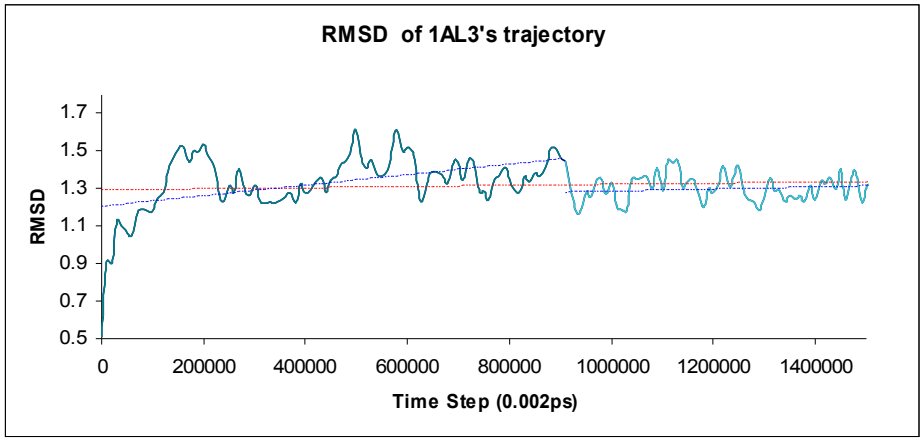
```
#Script to generate the matrix for the PCA

#!/bin/bash

#You need to change the name of the input file's names
PSF_FILE=$1_wb.psf
DCD_FILE=$1

#load the dcd trajectory file and saved in a pdb format
echo -e "mol load psf $PSF_FILE dcd $DCD_FILE.dcd \\n animate write pdb DCD_$DCD_FILE.pdb \\n quit\\n" | /usr/vmd/vmd -dispdev text

#remove the residues in the trajectory
grep "CA" DCD_$DCD_FILE.pdb > DCD_CA_$DCD_FILE.coords

#first residue
CA=`head -1 DCD_CA_$DCD_FILE.coords | cut -d ' ' -f1-14`

#count the number of residues
grep -c "CA" $1_wb.pdb > res.temp

#count the number of frames
grep -c "${CA}"  DCD_CA_$DCD_FILE.coords > fram.temp

#set the number of residues
total_residues=`cat res.temp`
echo total: $total_residues

#set the number of frames
total_frames=`cat fram.temp`
echo total: $total_frames

#remove the coordinates of the CA
cut -c32-54  DCD_CA_$DCD_FILE.coords > CA_$DCD_FILE.coods

#create the matrix with the coordinates of the CA
cat CA_$DCD_FILE.coods | for i in `seq 1 $total_frames`; do
      for i in `seq 1 $total_residues`; do
                read coords;
                echo -n "$coords  " >> matrix_$DCD_FILE.txt;
      done
      echo " " >> matrix_$DCD_FILE.txt;
done

#remove the not necessary files
rm CA_* DCD_CA* *.temp
```

Analysis

- In this study we worked with 7 different proteins, here are the results of 1AL3.

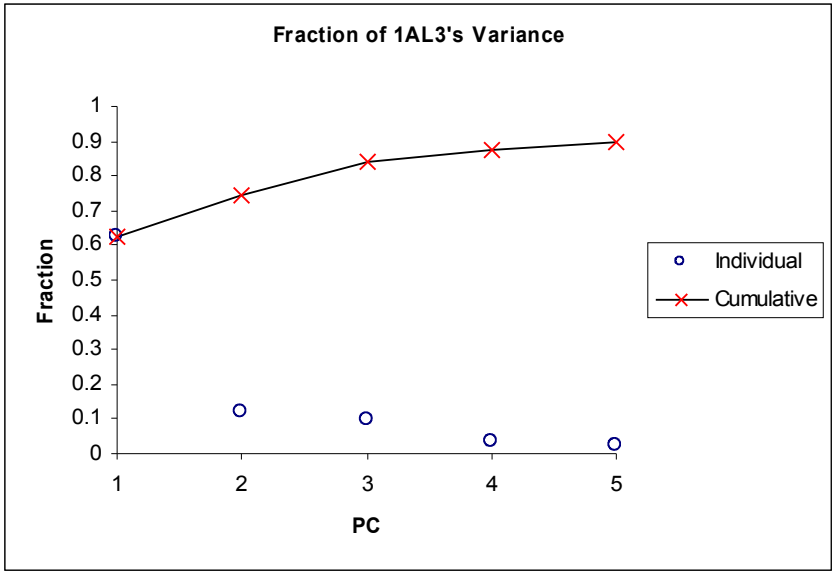


RMSD of the trajectory



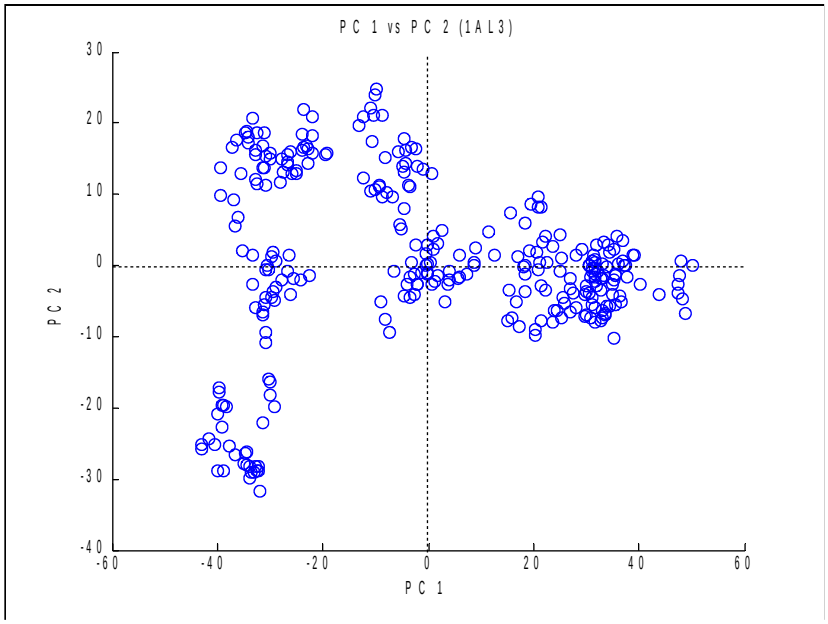
- This figure shows the RMSD for the 3ns of 1AL3's trajectory.
- The protein has relatively small fluctuations over time.

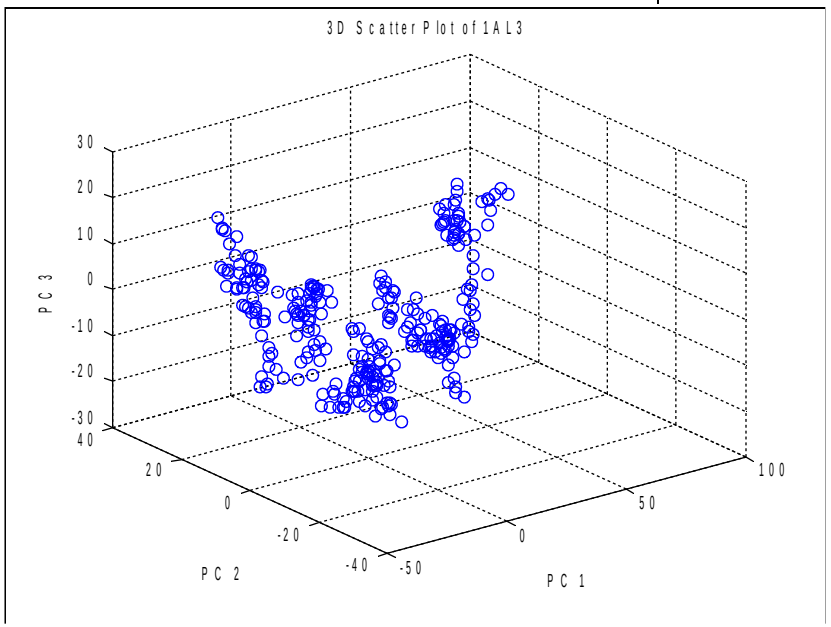
Fraction of Variance



- MatLab code:
  - PCA
    - `[pc, score, latent, tsquare] = princomp(matrix_1AL3);`
  - Individual Fraction
    - `latent` = the eigenvalues of the `matrix_1AL3` covariance matrix
    - `latent(1:5)/sum(latent)`
  - Total Cumulative Fraction
    - `sum(latent(1:5))/sum(latent)`

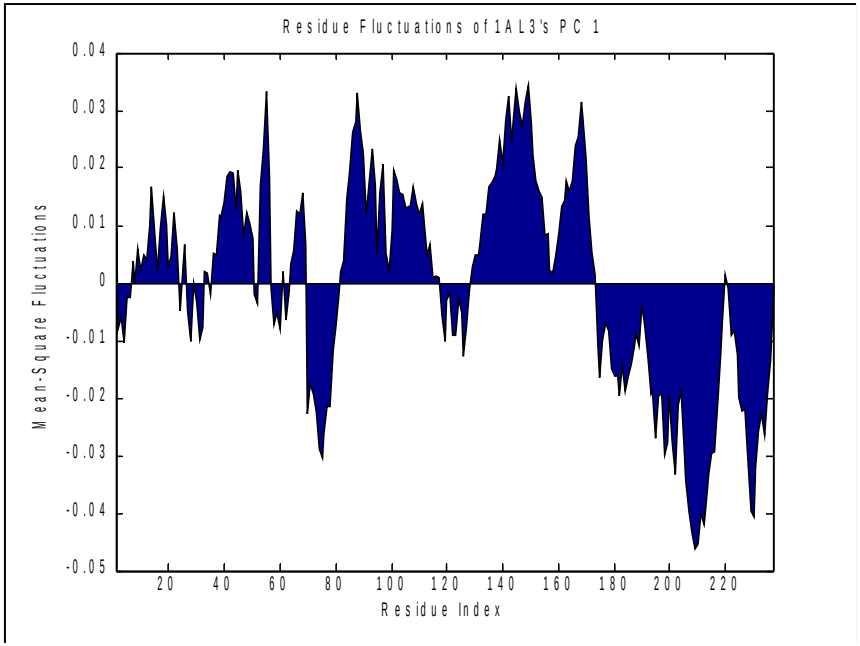
Scatter Plots of the Data

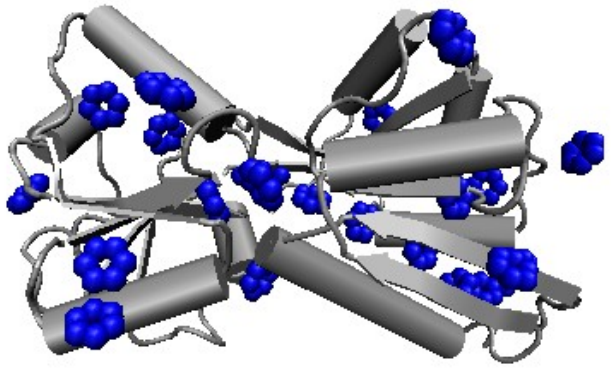




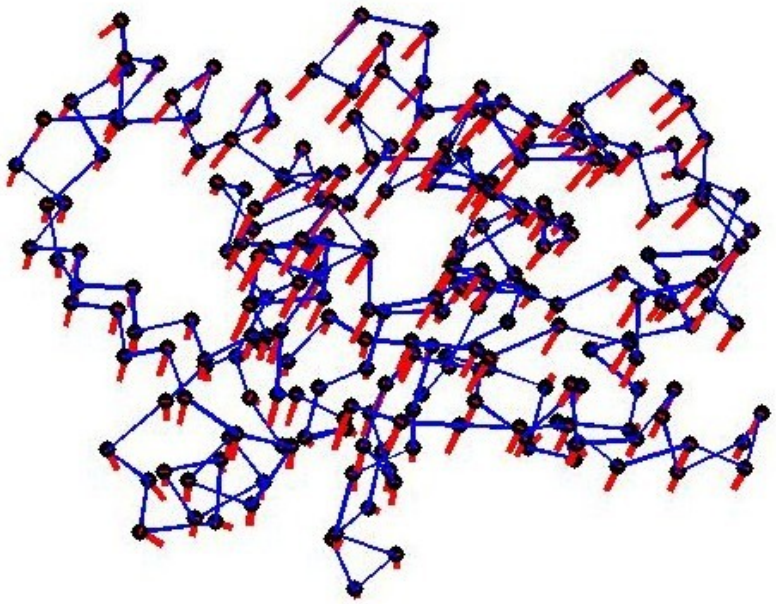
- MatLab code:
  - `score` = Z-scores
  - `scatter(score(:,1), score(:,2))`
  - `scatter3(score(:,1), score(:,2), score(:, 3))`

PC1 Residue Fluctuations



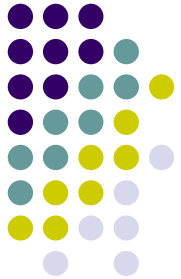


Visualization of the PC 1 Motion



Picture and video by Lei Yang.

Preliminary Conclusions



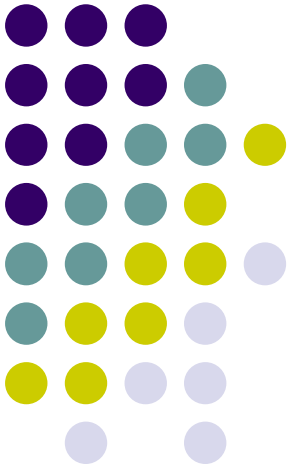
- These results suggest that minimized protein structures are stable.
- Principal components can be used to extract the important motions of a protein; however, long simulation times are required to obtain larger motions.

Acknowledgements

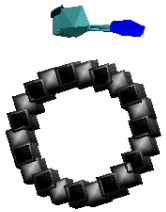
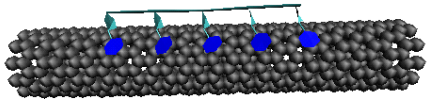
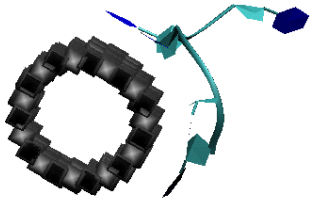
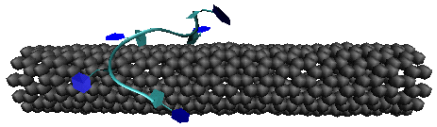
- Lei Yang
- Dr. Xuefeng Zhao
- Dr. Robert Jernigan
- CyBlue Staff
- BBSI 2007

And now what?

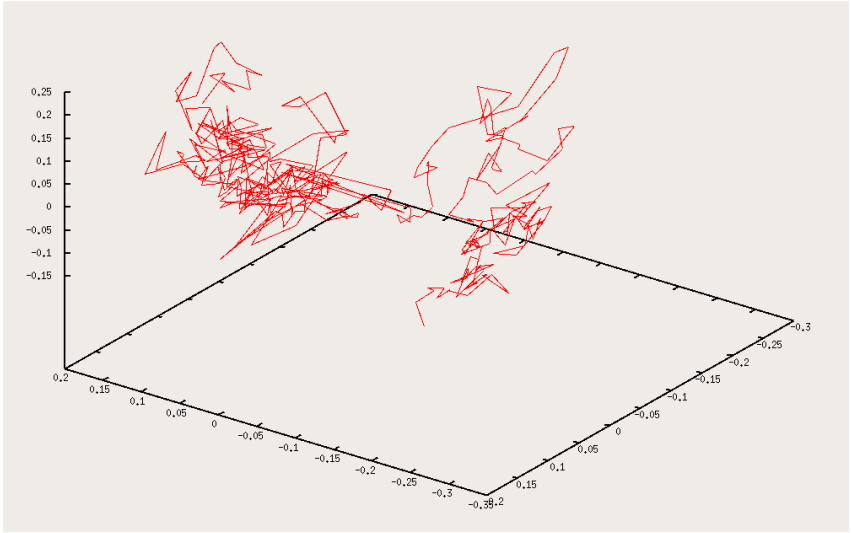
Principal Components Analysis in the trajectories of DNA-CNT Hybrids.



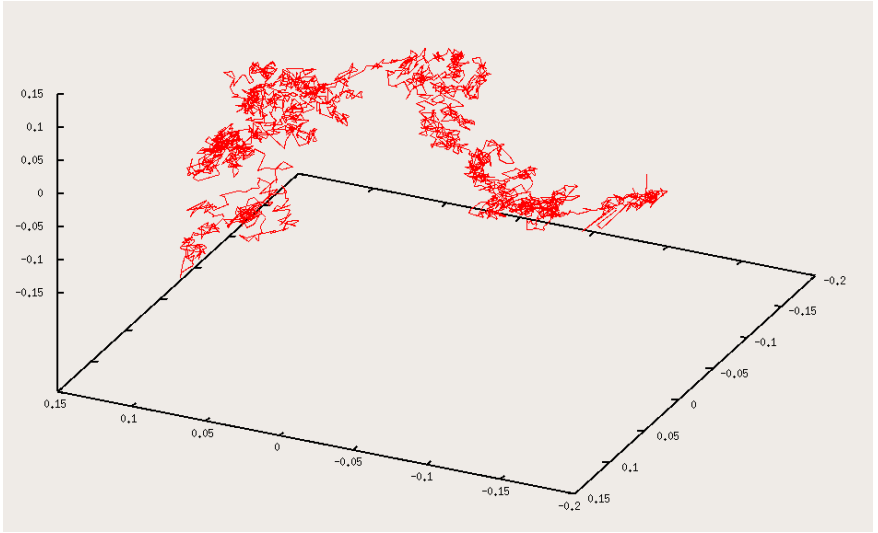
DNA-CNT Hybrid (5 monomers)

	Orthographic View	Perspective View
Initial Structure		
Final Structure		

Some preliminary results



Poly-C of 5 monomers



Poly-C of 15 monomers

- There is a transition to the final structure.
- The process is irreversible.

Graphics by José Sotero