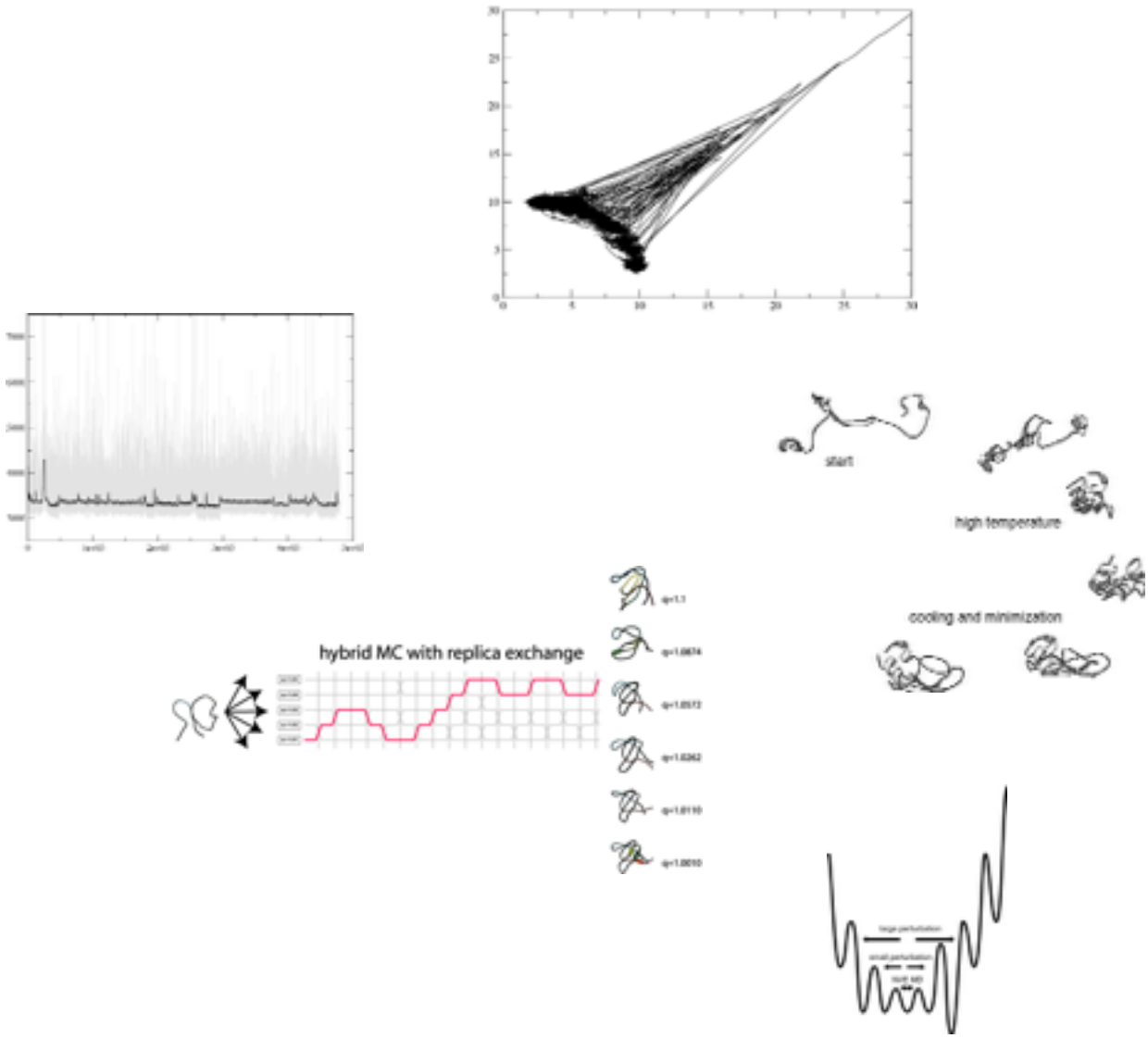
MICHAEL NILGES, INSTITUT PASTEUR



Practical

AlgoSB'13

Domaine d'Ariane, Impasse de Carpette, 31700 Mondonville

November 25-29, 2013

OVERVIEW

Objective

In the first part of the practical you will analyse trajectories that were pre-calculated. You should gain some basic familiarity with the program CNS (Brunger et a., 1998) and with the notion of sampling.

Goals

The analysis will allow you to check the sampling of conformational space and the similarity to the reference structure (in this case, the x-ray crystal structure of the same molecule).

Solution

You will use the program CNS to prepare the first basic input file that is the basis for all calculations of a molecular structure with CNS. You will then use a prepared input file that allows you to analyse the energetics and the distribution of the structure around a reference structure.

Practical outline

The trajectories analysed in this practical were calculated with a variant of the Bayesian method presented in the lecture. The trajectories are rather short and will not explore as much conformational space as the published examples obtained with ISD. The calculation method used is a combination of Monte Carlo with Molecular Dynamics simulations and replica exchange. At each Monte Carlo (MC) step, a short Molecular Dynamics (MD) trajectory generates new coordinates, which are then accepted or rejected by the Metropolis criterion. ISD samples the coordinates by Hybrid Monte Carlo (HMD), that is, a new coordinate proposal state is generated by a short MD trajectory, with random assignment of impulses at the beginning of the trajectory, where the Metropolis criterion compares total energies (kinetic + potential) at the beginning and the end of the trajectory. ISD samples the weight on the data (the width of the log-normal distribution postulated to describe the data in the forward model) independently from the coordinates. This is combined with Replica Exchange (RE), where different replicae differ by a modification of the energy (sampling a Tsallis rather than a Boltzmann distribution). This is quite inefficient and convergence is slow.

In the current calculation, we follow a concept described recently (Nielmeyer 2012), called Non-Equilibrium Candidate Monte Carlo (NCMC). This concept alternates perturbation and relaxation steps. In our case, the perturbation is the random rotation about a part of the torsion angles followed by the selection of the weight on the data. Rather than sampling the weight of the data independently, we choose an optimal weight for the obtained coordinates (as described in Habeck 2006), and add an appropriate term to the total energy that compensates for small weights. The relaxation step consists then in a short NVE MD trajectory, which does not change the total energy (apart from due to numerical errors). The Metropolis criterion compares total energies between subsequent endpoints of trajectories. The scheme is combined with replica exchange (but we use only two replicae), where the difference between them is an overall scaling factor of the data.

If you are interested, you can look at the implementation of this sampling scheme

1. Read the file *generate.inp.* Try to understand what the different sections are for. Look at the files that are read in by this input file.
2. Run this file with the program CNS by typing
   `CNS < generate.inp > generate.out.`
   Look at the output file - are there any errors or warnings? The result of this operation is a file called ***GB1.psf***. This file is the basis for all sub sequent calculations with CNS and can also be read into vmd to define the molecular topology of the molecule. You can look at this file (rather ancient format, compatible with CHARMM and NAMD, but you should not modify it in any way.
3. The pre-calculated trajectories were calculated with a script similar to **`sample.inp.`** This would take several hours / days to execute… but you can edit the file and try to start a calculation. You will need to edit file to specify the location of the data file etc. .
4. There are a number of trajectories pre-calculated in the subdirectory ***./trajectories***. These trajectories were calculated with different completeness of the data, ranging from 10 to 100 %. All the data are in the subdirectory ./data. The data completeness is indicated by a number in the filename (10, 20, 30, 40, 50, 60, 80, 100). Choose one of the trajectories (or several) and visualise them with VMD: either by typing
   `vmd 2pgb_hydro.pdb yourtrajectory.dcd` or
   `vmd GB1.psf yourtrajectory.dcd.`
   The first option has the advantage that it will display the reference structure, which can also be used to define the native secondary structure.
5. In this case (i.e., you have the reference structure), once the trajectory is read in, put the slider to the first image and chose “new cartoon” in the graphics submenu to display the structure. The protein has the ubiquitin fold (a parallel/ antiparallel beta sheet with one helix across the sheet).
6. Superimpose the trajectory on the reference structure by using the menu item
   `Extensions—Analysis—RMSD Trajectory Tool .`
7. You can follow the sampling and the development of the structure by choosing different visualisation options (e.g., HBonds).
8. Analyse the trajectory with the CNS input script ***energy.inp***. Edit the script (which contains more information to enter filename of the trajectory of you choice and run CNS:
   `CNS < generate.inp > generate.out`
   The script generates a file with the same name and extension .***energy***. The file contains a header row and then one row for each frame in the trajectory:

```
# count ener totenergy noeenergy totalnoe esigma dataweight vdw rmsxray rmsmirror
1 755.13 2700.47 503.754 2449.1 1945.34 0.22488 3.03271 44.109 44.2177
2 1645.22 3396.97 503.753 2255.51 1751.76 0.484558 87.8094 14.6344 13.5484
3 1506.98 3253.99 503.752 2250.76 1747.01 0.493767 78.2779 14.586 13.5705
4 1430.55 3175.41 503.753 2248.61 1744.86 0.497997 31.3972 14.5767 13.5786
5 1405.58 3147.05 503.753 2245.22 1741.47 0.504737 34.9573 14.4341 13.5557
6 1406.46 3148.33 503.753 2245.62 1741.87 0.503932 27.3155 14.387 13.4588
7 1406.46 3148.33 503.753 2245.62 1741.87 0.503932 27.3155 14.387 13.4588
8 1406.46 3148.33 503.753 2245.62 1741.87 0.503932 27.3155 14.387 13.4588
```

```
9 1427.03 3166.69 503.753 2243.42 1739.66 0.508362 32.1057 14.3113 13.4273
```

9. Plot different columns to analyse the progression of the trajectory, sampling of different regions of conformational space, and correlations between them, e.g. (note the - sign)
   `grep -v '#' | awk '{print $3, $9}' | xmgrace -`
10. The last column (10) is the RMS difference to the mirror image of the structure (why would this be a good minimum for the data given ?).
11. In general we will not have a reference structure when we determine a new structure. Then, we need to analyse the sampling and energetics of the trajectory with different means - this is the interest of Self Organising Maps.