Mucin: 4. Data curation

1. Page 1

The data produced in these runs needed to be curated in the following ways before it could be easily usable.

1. Some runs had to be stopped and restarted. The trajectories for these runs were concatenated back together.

2. In the mannosylated and unglycosylated runs, a few extra atomic positions (part of the solvent) were accidentally saved. These were removed.

3. A hardware failure caused a few trajectory frames to become corrupted. These were identified and removed.

The attached script, make_cpptraj_files_to_fix_trajectories.bash, carries out these steps with the help of a C program, fix_coords, that was written specifically for this situation.

Created: 19 Apr 2013 15:29:41 GMT , Updated: 22 Apr 2013 21:08:15 GMT

The trajectories were concatenated using cpptraj by loading all restarted trajectories before processing and writing out a new trajectory.

Created: 22 Apr 2013 21:05:56 GMT , Updated: 22 Apr 2013 21:10:31 GMT

To remove the extraneous coordinates, a dummy topology file containing the same, incorrect number of atoms was constructed. All extra atoms were assigned to fake water molecules. Cpptraj was then instructed to remove all water molecules from the trajectory and write out a new trajectory. It was necessary to add 8 dummy water molecules to the mannosylated peptide and 28 dummy waters to the unglycosylated peptide.

Created: 22 Apr 2013 21:09:20 GMT , Updated: 23 Apr 2013 15:22:47 GMT

It was not possible to use cpptraj to remove the corrupted frames -- and it was only possible at all because the original trajectory was saved in netCDF format. So, a C program (attached) was written. Because GLYLIB does not have ready access to a NetCDF reader, cpptraj was instructed to write out a plain-text trajectory-coordinate file. This file was used as input to the program.

The program does the following:

1. Loads the relevant topology file.
2. Scans through the trajectories and measures all bond distances.
3. Deletes any frame containing a bond less than 0.95 of the smallest equilibrium bond length or larger than 1.15 of the largest equilibrium bond length.
4. Saves information about deleted frames.
5. Saves good frames to one file and bad frames to another for later inspection.

Created: 22 Apr 2013 21:11:59 GMT , Updated: 22 Apr 2013 21:17:35 GMT

make_cpptraj_files_to_fix_trajectories.bash

This file runs cpptraj and fix_coords.
Created: 22 Apr 2013 21:18:46 GMT

remove_bad_coords.c

This is the source code for fix_coords.
Created: 22 Apr 2013 21:19:25 GMT

leapin_plus8waters

This leap input generated the dummy topology file used to remove extra coordinates from the mannosylated runs.
Created: 23 Apr 2013 15:24:01 GMT

leapin_plus28waters

This leap input generated the dummy topology file used to remove extra coordinates from the unglycosylated runs.
Created: 23 Apr 2013 15:24:54 GMT