Plotting Tools for a Molecular Dynamics Trajectory Database

Jeremiah Boswell, Anik Karan, Anthony Agee and Dr. Thomas C. Bishop
Department of Biochemical Engineering, Louisiana Tech University

LA-SiGMA
Louisiana Alliance for Simulation-Guided Materials Applications

Abstract

A typical problem in molecular dynamics simulations is the analysis of the trajectories. Numerous tools exists for extracting specific observables from trajectories but few tools exist for managing the statistical analysis and plotting of data as part of an overall simulation and analysis workflow. In order to achieve such an integrated workflow we utilized a programming language known as R that would potentially be able to process many files at once and produce various graphical interpretations from the observables extracted from the raw trajectories. Eventually, this summary of the data will be available to the world through a web interface, ideally using the iBiomes software.

Introduction

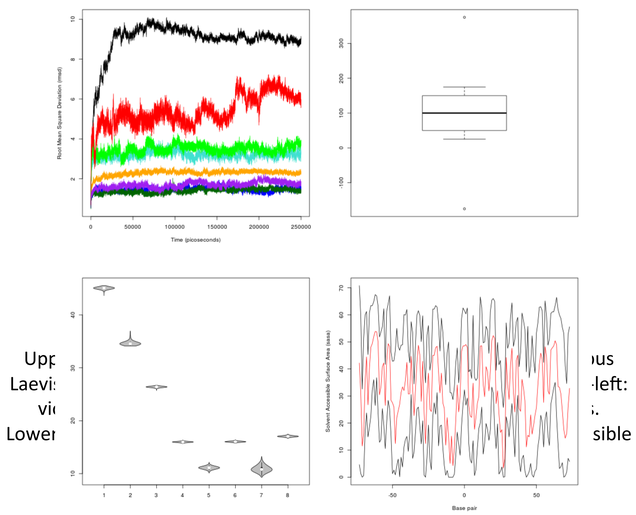
The Bishop Lab at Louisiana Tech is developing workflows for the simulation and analysis of small ensembles of mononucleosomes. A mononucleosome is a strand of DNA, 147 base long pairs wrapped around eight histones (proteins). Nucleosomes compact DNA into chromatin in all higher organisms. In order to study the behavior of nucleosomes, genomic data is given as input for a simulation program known as NAMD. Running on remote supercomputers, NAMD simulates the motion of the molecules and records their positions as a function of time on the scale of nanoseconds. The time evolution or trajectory is stored as a large dynamics coordinate dump (.dcd) file. In order to process these files, various analysis scripts were written using Visual Molecular Dynamics (VMD). These analysis scripts are then run on the .dcd files to generate arrays of data (.dat) files. These .dat files contain various observables from our trajectories, but share the similarity of space-separated columns of numerical data. Our goal is to develop a set of tools in R for the plotting and analysis of any file containing an arbitrary NxM array of data.

R is a statistical package and interpreted programming language that is available for the Windows, Macintosh, and Linux platforms. It is also an full-fledged environment for data analysis. From the official website, R is actually based on the S programming language which was developed by Bell Laboratories, and falls under the C family of programming languages. Unlike most programming languages that have complex structures but are designed to be obvious in function, R takes a “maximum processing per statement” approach. In practice, a robust programming language that is written to complete a task spanning multiple lines with straight-forward statements could have an R equivalent consisting of a single line, utilizing nested specialized methods.

Methods

The different syntactical style of the R programming language, called for initially spending resources on reading documentation on various functions, from control statements to plotting methods. This information was then used to create simple programs that would produce a certain type of plot for a static data file. The functionality of these “proof of concept” programs were later combined into a larger source code file, which would take a static number of arguments from the command line using Rscript, a utility that allows for the use of R non-interactively and more like a scripting language. This larger code file was modified further, including the ability to handle an arbitrary number of input files and also the dynamic nature of the data within these files. Due to the large quantity of data to be processed, this program would undergo further development, with the eventual goal of being able to produce image files containing arrays of multiple plots.

Results



Upp
Laevi
vi
Lower
ous
left:
sible

Discussion

The resulting plots showcase the intended functionality of the program working with a few of the different types of observables data files. For the top left graph, the top line represents the root mean square values (rmsd) for the DNA of 1KX5 Xenopus Laevis, and the lower lines represent the rmsd values alternating between the protein and the protein core for the different histones. The upper-right plot shows, from the furthest extent towards the center: outliers, upper and lower quartiles, and the median representations on a boxplot. The outliers are greater than/less than 1.5 times the value of the upper/lower quartiles. The upper and lower quartiles have 25% of the data being greater than/less than their values, respectively. Finally, the median has 50% of the data being greater than its value, and 50% of the data being less than its value. The bottom-left plot shows radius of gyration data in the form of violin plots. Violin plots are boxplots that have kernel density plots on their sides and connect outliers to the rest of the data representation, giving them the shape that resulted in the name of this plot type. For the bottom-right graph the black lines represent the minimum and maximum values of the solvent accessible surface area (sasa) data for 1KX5 Xenopus Laevis, and the red line represents the mean (or average) of this data.

Conclusion

As the program to create these plots is still a work in progress, more time will need to be invested to reach the goal of automatically generating these arrays of plots automatically as opposed to working with R interactively on a per-case basis. Other ways to view trends in the data may need to be accounted for, which could result in the program being broken up into a number of separate tools. It is worth mentioning that viewing the plot of the range for each timestamp of a helix parameter observable one after the other showed a distinct pattern in the peaks of the graph. In order to look into this for making a conclusion on the biological data being studied, this will most likely end up being a functionality to be added, where a movie made up of frames of data versus time can be generated with a simple command.

Acknowledgements

This material is based upon work supported by the National Science Foundation under the NSF EPSCoR Cooperative Agreement No. EPS-1003897 with additional support from the Louisiana Board of Regents.

LBRN
Louisiana Biomedical Research Network