User documentation for DROIDS 2.0 – a GUI-based pipeline for comparative protein dynamics

Gregory A. Babbitt[1*], Jamie S. Mortensen[2], Erin E. Coppola[2], Lily E. Adams[1], Justin K. Liao[2]

1. T.H. Gosnell School of Life Sciences, Rochester Institute of Technology, Rochester NY
2. Biomedical Engineering, Rochester Institute of Technology, Rochester NY

*corresponding author email address: gabsbi@rit.edu

Overview

DROIDS 2.0 is an open source software project aiming to visualize and quantify the impact of one of the longest time scale processes in the universe (i.e. molecular evolution) on one of the shortest time scale processes in the universe (i.e. molecular motion). Specifically, we want to know how molecular evolution over 100s of millions of years impacts the functional molecular motions that play out over a few femtoseconds in real time. A primary motivation of this project is to combine GPU accelerated biophysical simulations and GPU graphics to design a gaming PC into a 'computational microscope' that is capable seeing how mutations and other molecular events like binding, bending and bonding affect the functioning of proteins and nucleic acids. DROIDS-1.20 (**D**etecting **R**elative **O**utlier **I**mpacts in molecular **D**ynamic **S**imulation) is a GUI-based pipeline that works with AMBER16/18 (Assisted Model Building with Energy Refinement), Chimera 1.11 and CPPTRAJ to analyze and visualize comparative protein dynamics on GPU accelerated Linux graphics workstations. DROIDS employs a robust and nonparametric statistical method (multiple test corrected KS tests on all backbone atoms of each amino acid) to detect significant changes in molecular dynamics simulated on two homologous PDB structures. Quantitative divergence in atom fluctuation (i.e. calculated from vector trajectories) are displayed graphically and mapped onto movie images of the protein dynamics at the level of individual residues. P values indicating significant changes are also able to be similarly mapped. DROIDS is useful for examining how mutations, epigenetic changes, or binding interactions affect protein dynamics. DROIDS was produced by student effort at the Rochester Institute of Technology under the direction of Dr. Gregory A. Babbitt as a collaborative project between the Gosnell School of Life Sciences and the Biomedical Engineering Dept. Visit our lab website (https://people.rit.edu/gabsbi/) and download DROIDS from Github at https://github.com/gbabbitt/DROIDS-2.0---free-software-for-comparative-protein-dynamics

We will be posting video results periodically on our YouTube channel https://www.youtube.com/channel/UCJTBqGq01pBCMDQikn566Kw

A single page Quick Start Guide (pdf) and full Installation Guide and User Manual are available with the download. They outlines the processes encountered in each of the three main GUI interfaces. It is strongly advised that users be comfortable with how to prepare PDB files for molecular dynamic (MD) simulation using GPU accelerated AMBER 16 (pmemd.cuda). DROIDS assists with modifying .pdb files

named in the GUI for AMBER simulation, however the user should become very familiar with the programs running at these steps (i.e. antechamber, pdb4amber, and teLeap) and read through all output at the DROIDS terminal to ensure that the structures are properly prepared for MD simulation. You must consult the AMBER documentation for this knowledge. The DROIDS GUI provides automation of teLeap, a program for pdb file setup, but care must be taken to read output on the Linux terminal for any errors. The programs 'antechamber' and 'pbd4amber' are used by DROIDS in modifying files for MD and are generally prior to starting teLeap in DROIDS. Please consult the Amber16 user manual for more details. Typically preparation includes (A) removing mirrored images and other chemical artifacts (done manually in Chimera prior to DROIDS), (B) performing a structural alignment (using Chimera MatchMaker and Match->Align when prompted by DROIDS) followed by subsequent saving of a Clustal format file (.aln), (C) adding H atoms and removing crystallographic waters (use pdb4amber button in DROIDS to dry and reduce), (D) estimating and loading force field parameterization regarding important ligands if a protein-ligand interaction is modeled (use antechamber button). Then finally (E) run teLeap button in DROIDS to setup topology and coordinate files for simulation. For v2.0 we have added script to check the file sizes of teLeap output files and recommend whether the process likely failed of succeeded at this step. teLeap is nicely verbose, so warnings on terminal when running teLeap button is very helpful for any indications of problems specific to your structural models. For many at this stage of model prep, it is not unusual to go back to modify the original .pdb file and run through the prep stages again. Be sure to view your models in Chimera using the 'all atom' preset so that you do not miss small molecules that might trip up the MD setup. Amber is designed not to run unless all atoms in your system can be properly parametrized by the force field you have chosen. Many force fields are available to try in the amber16/dat/leap/cmd folder. Many are appropriate only for certain macromolecules, and analysis of binding interaction will require several are loaded. ALSO NOTE: AMBER 16 software must be licensed from the University of California. More details about purchasing and installation can be found at http://ambermd.org/. DROIDS is tested on Linux Mint 18.1 and Ubuntu 16.04 and is offered freely under the GPL 3.0 license and is available on GitHub https://github.com/gbabbitt/DROIDS-1.0

Primary software dependencies are Amber16, Ambertools16 or 17, CUDA 8.0, UCSF Chimera 1.11 (or higher), perl-tk (perl 5.10), python-tk, python-gi, R-base, R-dev, ggplot2, dplyr, gridExtra, FNN, e1071 (R packages), evince (Linux pdf viewer), GStreamer (Linux movie viewer) and descriptive.pm (a perl statistical module provided with our download). Hardware dependencies are only to have a high end Nvidia graphics card with proper drivers. All testing of DROIDS 2.0 was done on the Nvidia Titan Xp and GTX 1080 GPUs. Linux versions of Chimera are available at (https://www.cgl.ucsf.edu/chimera/). All other dependencies are able to be addressed via the usual Debian package downloads. CUDA drivers for Linux are available from Nvidia. NOTE: do not use CUDA 7.5 with Amber16. CUDA 8.0 is currently supported. DROIDS now supports setups with dual GPU's (with no SLI connection) which allows the two sets of simulations intended to be compared (i.e. query and reference proteins) to be run simultaneously. As of 2018, we recommend using a dual Titan Xp build, which allows most comparative protein dynamics

to be run overnight (e.g. assuming <500 residue polypeptide chains, 10ns equilibration time and 50-100 production runs at 1/3 to ½ ns each.

DROIDS is activated by entering 'perl DROIDS.pl' at the Linux terminal opened from within the DROIDS folder. DROIDS v2.0 initially starts with a small GUI requesting user to add paths to Chimera and Amber's force field data files (e.g. amber16/dat/leap/cmd). As Amber16 is typically installed to the Desktop, this path will be different on different machines. Make sure you edit the path appropriately before attempting to run DROIDS. The GUI will create a paths.ctl file. Once this file is created for your individual machine, it can be saved and dropped into DROIDS folders prior to each run. The typical bashrc file can be used similarly, but this GUI was added to make this initial setup simpler for less experience Linux users. Once the paths GUI is closed, the main DROIDS v2.0 GUI will appear. Here the user is directed to choose one of the various types of comparative analysis that can be done, choose MD sim software, and indicate whether the machine is running a single or dual GPU. Upon clicking 'run DROIDS' the user is taken to the first main GUI for setup, running MD, and parsing of MD simulation output. The second main GUI controls the DROIDS statistical analyses and the last main GUI controls the image color-mapping and movie rendering and viewing options. These three steps are described in more detail in the sections below.

IMPORTANT NOTE: When running DROIDS on many protein comparisons, we find that explicitly solvated systems (i.e. PME method) tend to yield more conservative results regarding the significance of the KS test when compared to implicitly solvated comparisons (i.e. GB method). This is likely expected due to the many more degrees of freedom under the PME option. We recommend that users explore both methods of solvation when using DROIDS. Implicitly solvated protein comparisons run relatively fast and are useful for an initial investigation, however comparison of explicitly solvated systems may yield more interesting local variation in mutational impacts.

Specific analyses now offered in DROIDS v2.0

DROIDS v2.0 now offers 10 different pipelines intended for specific types of comparative analysis. Examples with .pdb files are provided. First time users should run the examples provided in the exampleFiles folder first, to get a sense of what setup required and what output is delivered. The 10 analysis and example files are listed here.

1. ***Analysis of self-stability of dynamics on a single protein*** – this compares MD of a protein to itself and is useful for finding regions of protein that are less stable. Try it on 1ubq.pdb and notice the lack of stable dynamics near the c-terminal tail, where ubiquitin is ligated to 'tag' proteins for degradation.
2. ***Analysis of mutational impacts on a protein*** – here the user can create mutant versions of a given protein by replacing one or several AAs using automatically optimized selections from the

Dunbrack rotamer library and comparatively quantify the local mutational impacts on MD using KL divergence in atom fluctuation. This option is great for simulating studies of site-directed mutagenesis.

3. ***Analysis of evolutionary divergence in MD on a protein*** – here the user can analyze divergence in MD using PDB files for an ortholog pair. This option is interesting when applied to questions of thermostability. For example, compare thermostable Taq DNA polymerase (4n56.pdb) to its less stable cousin in E. coli (1kfd.pdb).
4. ***Analysis of relative impacts of mutation in a disease system*** – This is one of the most interesting and complex pipelines DROIDS offers. Users need an ortholog pair as in #3 and a known disease mutation as in #2. The MD impacts of the ortholog and disease systems are compared with a single KS test applied to local regions around mutations in each system to determine if the disease has a significant molecular dynamic component to it. If it does, then the GUIs further allow the simulation and classification of novel variants as being neutral or deleterious in accordance with a support vector machine learning algorithm trained on local atom fluctuations around mutation sites in the ortholog and disease systems. This is particularly useful for examining mutational tolerance and whether specific genetic backgrounds can influence clinical manifestation of a given disease, or even a drug's ability to combat a disease under a specific genetic background. Try the CFTR (ATP binding domains of cystic fibrosis gene) human/zebrafish ortholog (5uak.pdb and 5uar.pdb) and insertion of ARG at position 70 to simulate the effect of the deltaF608 mutation found in 90% of clinical cases.
5. ***Analysis of impact of DNA protein interaction upon binding*** – This option allows users to identify and visualize where DNA binding in the system occurs by comparing the dynamics of protein in the bound and unbound states. Binding is identified via dampened atom fluctuation in the bound model. Try the TATA binding protein example using 1ytb_bound and 1ytb_unbound (where the DNA chains were removed).
6. ***Analysis of the impact of mutation(s) on DNA-protein interaction*** – Here site-directed mutagenesis in both cis and/or trans can be simulated on the DNA bound protein system and mutational impacts on binding observed. This is particularly useful for questions around gene regulatory evolution on a given transcription factor. Try mutating 1ytb_bound in regions where strong binding is indicated in analysis #5.
7. ***Analysis of the comparison of two DNA-protein interactions*** - Like analysis #3, this option allows comparison of DNA-binding homologs directly from two PDB files. This is useful for analyzing more distant evolutionary divergences in transcription factors.
8. ***Analysis of the impact of protein-ligand interaction with a drug, toxin or activator*** – Like analysis #5, this option allows user to examine how a given protein binds a particular ligand by comparing bound and unbound protein dynamic states. Binding mechanisms are identified by reduced atom fluctuation in the results. Try examples provided to demonstrate binding of HIV

drug sustiva (efavirenz) to the drug target of viral reverse transcriptase (1fk9.pdb). Note: requires three files (1fk9_bound, 1fk9_unbound and 1fk9_ligand).

9. ***Analysis of the impact of mutation(s) on protein-ligand interaction*** – Like #6 this option allows user to put mutations onto the protein-ligand system and analyze the effects on MD. This is potentially useful for examining how genetic backgrounds can influence the working of a drug or toxin.
10. ***Analysis of impact of an epigenetic modification on a protein*** - Because DROIDS comparative analysis can be applied to any structural change to a protein, and not only genetically based ones, the effect of epigenetics on MD can also be compared. This option allows user to compare effects of protein methylation/acetylation, phosphorylation, protonation or disulfide bridging on molecular dynamics. Try it out on the example of a disulfide bridge engineered onto lysozyme (1l35 compared to 1lyd) or several bridges in proinsulin (1kqp and 1kqp_SSbonds)

NOTE: If you use DROIDS for published work please use the following citation

Babbitt et al., DROIDS 1.20: A GUI-Based Pipeline for GPU-Accelerated Comparative Protein Dynamics, Biophysical Journal (2018), https://doi.org/10.1016/j.bpj.2018.01.020

The DROIDS pipeline

The DROIDS pipeline is run as a series of linked Perl-Tk scripts that are controlled at the command line. The Quick Start Guide lists the steps shown schematically in Figure 1. The user starts the pipeline by placing the two PDB files to be compared in the DROIDS main folder, opening a terminal, and typing 'perl DROIDS.pl. After the paths.ctl file is created, the main GUI opens allowing choice of analysis, and specification of hardware and software. After this, the user is guided through three main GUI's each for (1) Amber MD simulation, vector trajectory analysis and file preparation and parsing for DROIDS, (2) DROIDS comparative statistical analysis of protein dynamics and graphical plotting in R, and finally (3) PDB structure color-mapping and movie rendering in Chimera and subsequent movie viewing in the DROIDS movie viewer. We now offer alternative GUI for computer builds with dual GPU cards. This will run MD on both homologous protein structures at the same time. This GUI interface is designed to control and run all stages of the MD simulations of both the query and reference PDB structures that will be needed for later DROIDS analysis. This includes typical teLeap setup of the PDB file, structural alignment of the query and reference proteins, and an energy minimization, heating and equilibration run on each PDB. These runs are followed by N number of sampling runs with N specified by the user. Random spacer runs precede each sampling run so as to minimize the impact of initial conditions on the MD sampling (i.e. minimize differences merely due to chaos in the MD runs). Afterwards, MD is run, users will collect atom info and flux data using buttons that run typical cpptraj commands that loop through each sampling run. The last step includes the parsing of the vector trajectory output to the structurally-based sequence alignment in performed earlier in Chimera. Some analyses in DROIDS call

for choice of ‘strict’ vs ‘loose’ homology (which determines upon which amino acids the DROIDS statistics will be applied). Loose homology should be chosen when evolutionary distances between the PDB files are large. Strict homology should be chosen when sequences are nearly identical (e.g. examination of one or several specific mutations). After parsing, a second GUI will pop up and lead users through DROIDS statistical analysis and graphical output. Here users run the statistical comparisons and choose method of multiple test correction. At this point a third GUI will pop up and allow color-mapping and graphics options to be applied to the static and moving images of the reference PDB. The statistical test employed by DROIDS is a KS test applied specifically to the collective backbone MD of each amino acid residue (i.e. atoms N, CA, C and O masked during cpptraj).

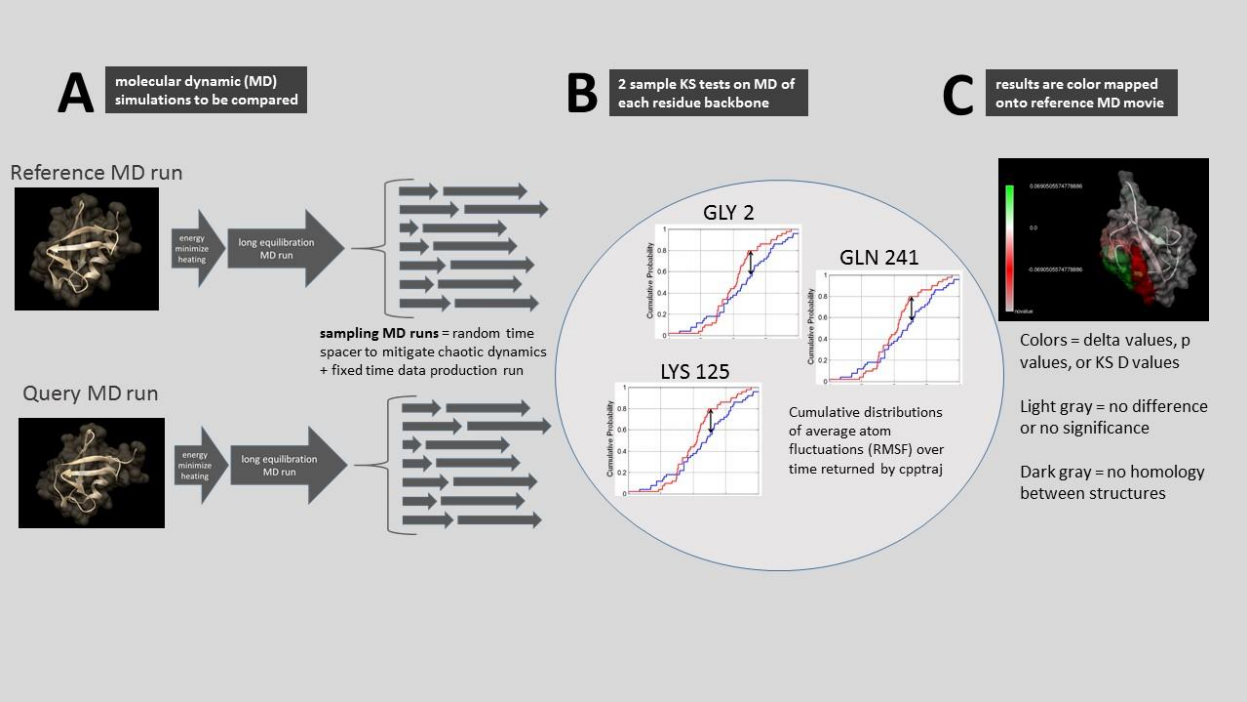

Figure 1. A schematic representation of DROIDS comparative molecular dynamic analysis software. DROIDS 1.20 is a software tool for multiple test corrected amino acid-level pairwise comparison of molecular dynamics of two comparable PDB structures. The three main phases of analysis include (A) MD sampling runs and vector trajectory analysis, (B) statistical comparison via multiple test corrected KS tests, and (C) visualization results on static and moving images.

Running MD with Amber via DROIDS

This MD GUI interfaces allows the user to set the most important parameters for the MD (e.g. name the force field, set run times of each phase, choose a solvation method, add salt conc) as well as determine how many sampling MD runs on each protein will be analyzed in later analysis. For most proteins, I often take 50-100 sampling runs at 0.5ns each, after a single equilibration phase of 10-50ns…depending upon

how stable the structure behaves. Users are guided through creation of a structurally-based sequence alignment using Chimera MatchMaker and Match->Align, followed by setup of topology and coordinate files using teLeap. Then the script automates the energy minimization, heating, equilibration and MD production sampling runs on the two homologous structures and reports the progress to the Linux terminal. This part of the analysis takes the longest (e.g. the two comparative runs on two typical implicitly solvated systems may take 24-48 hours to run on the GTX 1080 card). Explicit solvated systems may run 2-3X longer. Details about the MD are hard coded into the portion of the script that writes the control file (i.e. the control subroutine). These settings can be easily changed by users with some experience with Amber commands and perl scripting. The default assumes constant temperature (300K) and pressure during production. Note that MD output is produced in the form of binary files (.nc file type extension) rather than text (i.e. .mdcrd file type). This is to allow the saving of hard drive space and proper file type for cpptraj analysis that follows. These files are not 'readable' in any sort of text editor. Jobs are scheduled to the GPU by means of a while loop that periodically pgreps the process ID's produced by pmemd.cuda. The GPU will not automatically control job scheduling the way a CPU will. So we have added a GPU surveillance button that opens terminals that monitor the load on the GPU as well as current running processes. If the user interrupts a script and starts another job, this will not terminate the previous run. If the user sees that two pmemd.cuda processes are running at once, then the data is likely corrupt as the GPU is attempting to run both jobs at the same time. We include a 'kill' button which will pkill all pmemd.cuda jobs. This is handy when restarting DROIDS after previous interruption. It is recommended that user keep surveillance open at all times alongside the main terminal when running then MD wrapping script (GUI_START_DROIDS.pl). See Figure 2 for how this should look on your desktop. Before each sampling MD run, a random time length spacer is generated uniformly distributed between 0 and 0.5 x length of the sampling run. The purpose of this step is to average out the effect of chaotic dynamics that may be observed if the initial starting conditions were always exactly taken after the equilibration step has finished. A typical DROIDS analysis might consist of 0.5ns heating, 10-50ns of equilibration and 50 x 0.5ns of sampling runs on each protein. With this setting, most comparisons of protein dynamics can be achieved in 12-48 hours of run time using a dual GPU machine with GTX 1080.

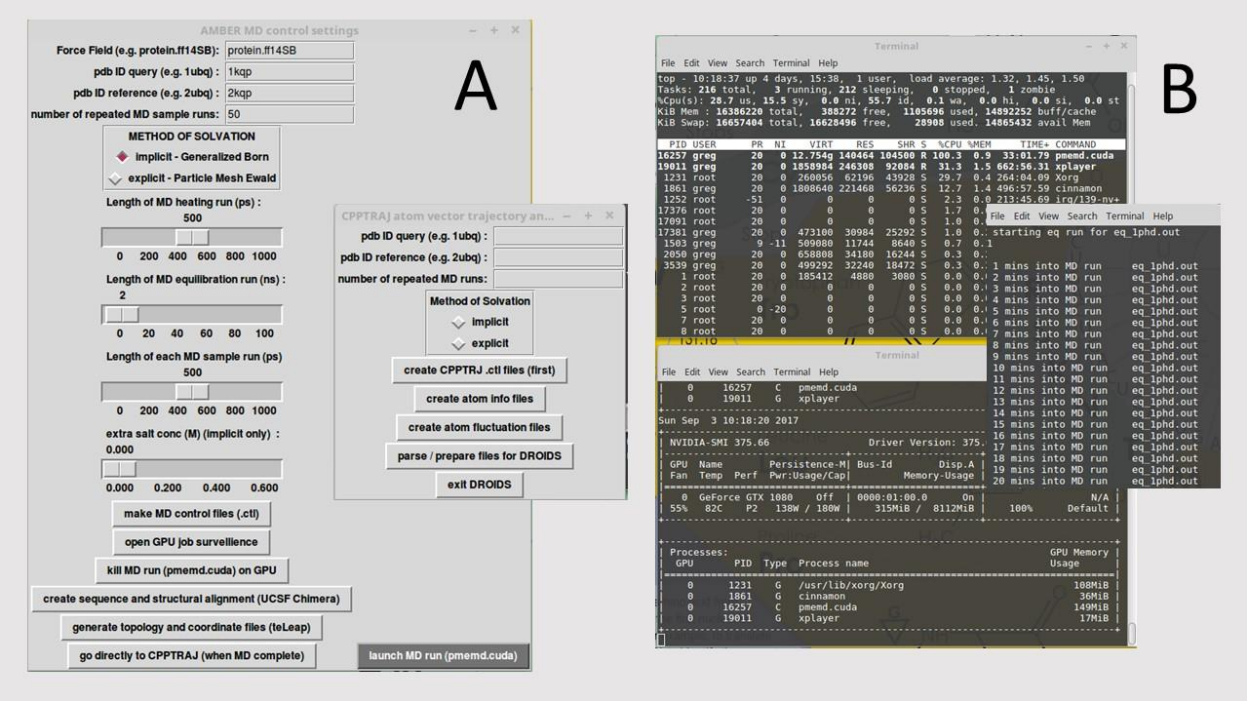

Figure 2. (A) The DROIDS 1.20 GUI interfaces for controlling molecular dynamic simulations and sampling conditions in Amber16 and subsequent cpptraj analysis. (B) Linux terminal windows showing the progression of the MD simulations as well as general surveillance of GPU loads and process IDs.

Calculation and collection of atom fluctuations

After the end of the MD simulations the user is guided through vector trajectory analysis using cpptraj (Ambertools16/17). See Figure 2 (inset). The buttons are run from top to bottom and include making control files, collecting atom information, calculating atom fluctuations, and lastly, preparing and parsing the cpptraj output for subsequent DROIDS analysis. The

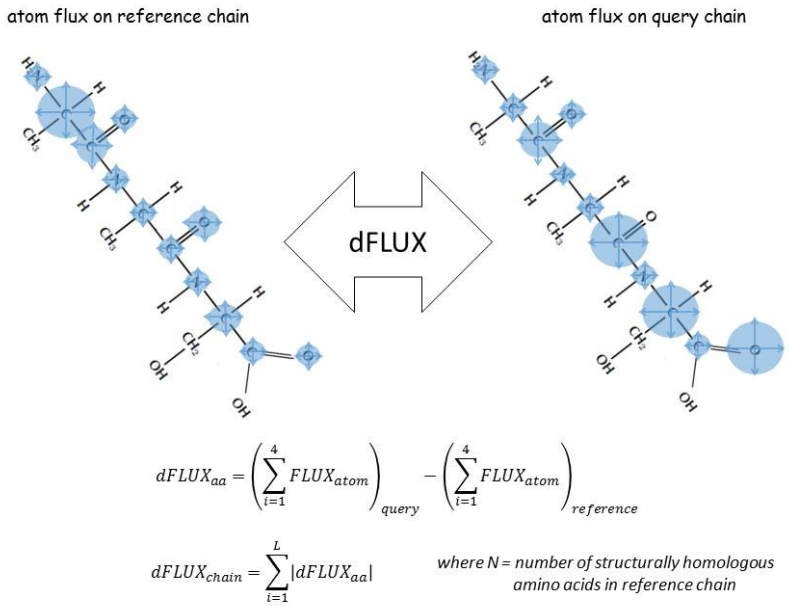

Figure 3. A schematic representation hypothetical differences atom fluctuation (dFLUX). Functional analysis of destabilization due to mutation and or evolution of functional thermostability can be addressed using dFLUX. In the DROIDS color mapping, dFLUX is averaged over the 4 backbone atoms of each amino acid. Global dFLUX for the whole chain is simply the sum of absolute dFLUX over the length of the polypeptide chain. Version 2.0 also allows dFLUX to be defined using symmetric Kullback-Leibler divergence between the distributions of atom fluctuation. This option provides a richer view of differences when color-mapping dFLUX.

setup we use under the hood is designed to return amino acid averaged motions collected only over the backbone of the polypeptide chain (i.e. N, CA, C, O). Fluctuation is very rapid (10-20 femtoseconds on most bonds) and largely harmonic and thus is relevant to comparative studies of protein stability (i.e. evolution of thermostability, functional epigenetic modifications, or disease-related genetic mutations that globally destabilize function. During initial setup (start GUI), the user is also guided from the terminal through the creation of a structural alignment of both protein structures using Chimera's MatchMaker and Match -> Align tools. The user is directed to save the resulting sequence alignment as a Clustal format file (.aln) using the name of the reference PDB ID in the title as follows Nxxx_align.aln (e.g. ubiquitin would be 1ubq_align.aln). Not that it is very important that the user trims the chains to the same length after alignment so that data is collected correctly from homologous amino acids. In GUI 2, the user is now also asked to specify whether the DROIDS statistics and mapping are to be conducted using 'loose' or 'strict' homology. Strict homology will only conduct MD comparisons on the backbone atoms of the protein when the aligned amino acid residues are identical. Loose homology will compare backbone MD even when residues are different as long as the structural alignment file identifies then as homologous. Note: atoms in sidechains are always excluded from all analyses via a mask used in cpptraj. When pipelines use strict homology on a protein comparison without a large evolutionary distance a brighter

color selection (i.e. red or yellow) is used for nonhomologous regions as a way to label interesting mutations in the resulting images and movies of the dynamics. Under structural comparisons of greater evolutionary distances, where the underlying protein sequences are likely to be quite different, loose homology will be used along with a less conspicuous color (i.e. usually gray) to mark regions in the protein comparison that lack true homology (i.e. are poorly aligned) MatchMaker provides user ability to choose appropriate substitution matrices and gap penalties to reduce the problem of poor alignment. DROIDS automatically excludes these regions from analysis. NOTE: at the end of parsing, a folder named 'atomflux' should appear with individual files for each comparison per residue. The number of files in this folder should correspond to the number of residues in the reference protein that have homologous residues in the query protein. If there are far fewer files in the atomflux folder than expected, this is most likely due to the fact the sequence at PDB does not exactly match the structure. Occasionally, one will need to trim the alignment file to match the structure, and then rerun the parsing again.

Comparative analysis and visualization of mutational impacts on protein dynamics

The statistical analysis is the heart of comparative protein dynamics using DROIDS. The initial steps include making choices about the type of analysis you want, then producing the control files you need. Then you run the KS tests in R on the next button. R graphics will show analyses as a popup in the pdf viewer. After this step the user will generate Chimera 'attribute' files for color mapping. Color mapping generally scales in saturation with the strength of the delta shift in atom motion (fluctuation or correlation) between the two sets of MD runs. Regions lacking homology are darker gray. If you are only changing the mapping options (i.e. data types – delta, p, or D values, color schemes or scaling of plots), you do not need to rerun the statistical tests. If you change statistical test options (i.e. motion type, p value cutoff, or multiple test correction), you will need to rerun the KS tests again. As the number of KS tests equals the number of amino acids on the chain, correction for multiple testing is highly recommended. Multiple test correction methods included as options in DROIDS are the Bonferroni correction or Benjamini-Hochberg estimation of false discovery rate. The dFLUX values of the query runs can be scaled to the absolute dFLUX values of the reference runs if the user is more interested in relative difference rather than absolute difference. Be sure to choose color schemes that correspond to the data type as indicated on the screen. Color gradients can be auto-scaled (highest to lowest value) or fixed at one of several options. When statistical options are changed (excepting p value corrections) a new DROIDS results folder is generated for each set of tests. After the Chimera attributes are stored for mapping, the user can generate color mapped static structures in Chimera and/or render movies with the appropriate color mapping shown from 6 points of view X1, X2, Y1, Y2, Z1, and Z2…or alternatively with 2 points of view incorporating a smooth vertical and horizontal roll during playback. These movies can be viewed

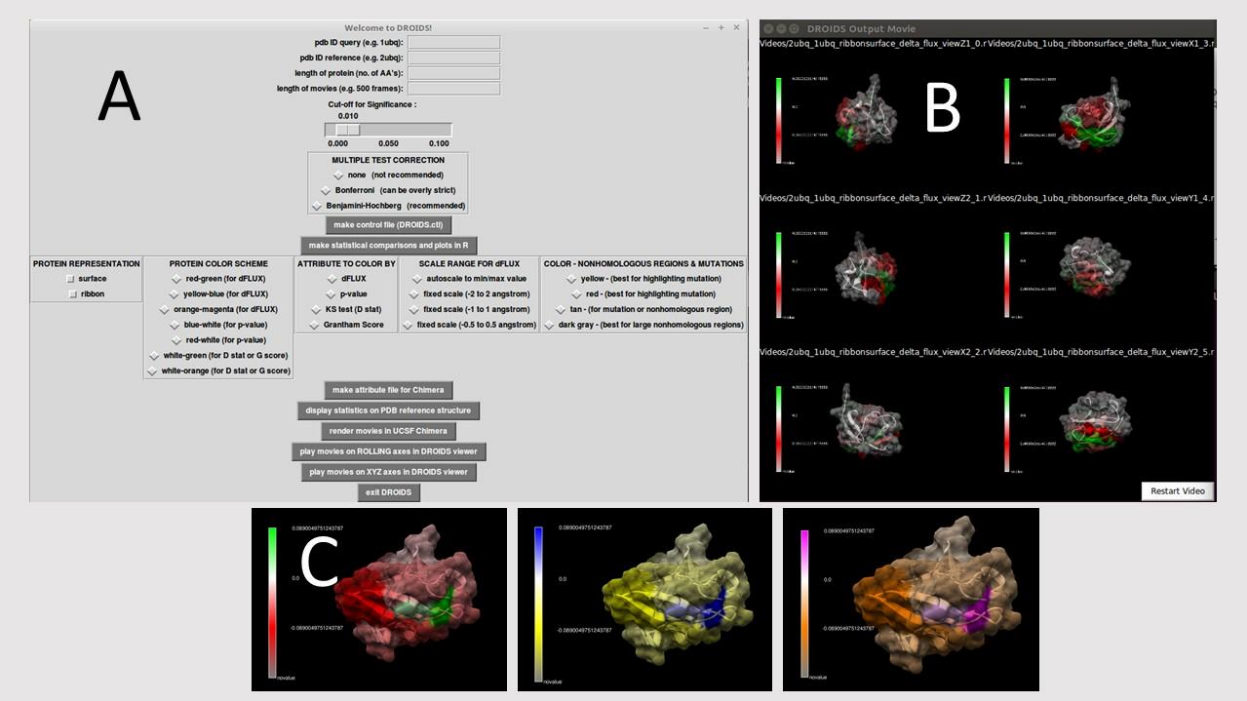

Figure 4. (A) The final DROIDS 1.20 GUI controlling the KS statistics, multiple test correction method and graphics options. (B) A movie viewer showing six points of view (front, back, left, right, top, bottom) is also provided. (C) Color options are quite numerous. Bivariate color options for dFLUX are shown here. Univariate coloring options for p or D values of the KS test are also provided.

simultaneously in concert in the DROIDS movie viewers. While the colors mapped correspond to the overall analysis, the movie dynamics correspond to only the first MD sampling run taken on the reference PDB structure.

Current and future uses for DROIDS 1.20

The potential uses for DROIDS are many. Some ideas we have imagined during its development include the visualization of the functional effects of natural and artificial mutation at the protein sequence level (i.e. amino acid replacement and site-directed mutagenesis). Population variation associated with disease related malfunction (i.e. nsSNV – nonsynonymous single nucleotide variant) might also be analyzed. The functional impacts of post-translational modifications (e.g. disulfide bridging or phosphorylation) and epigenetic modifications (e.g. acetylation and methylation) will also be of considerable interest on computers that can handle larger molecular systems. Functional consequences of natural evolutionary divergences created through the processes of speciation, gene duplication and genetic drift / genomic decay can also be compared. The study of functional binding interactions (protein-ligand, protein-DNA and protein-protein) is now available with DROIDS v2.0. Currently, the code is limited to single chain comparisons. Null comparisons are also useful. These are when the exact duplicate of the same PDB files are run through DROIDS. Because molecular dynamics can diverge

wherever the system does not settle into potential energy wells, a null comparison on a single structure using DROIDS can show users where the MD is potentially failing to replicate reproducible biophysics. V2.0 offers a specific option for this sort of analysis as well. The v2.0 edition of DROIDS offers many options that are more specific to particular areas of molecular evolutionary biology that involve DNA-protein binding (e.g. chromatin dynamics, transcription factor binding, etc.) and that incorporate machine learning based classification and visualization of more disruptive changes to dFLUX that are outside what is typical of functional orthologs.