Bioinformatics, 33(17), 2017, 2768–2771
doi: 10.1093/bioinformatics/btx349
Advance Access Publication Date: 31 May 2017
Applications Note

OXFORD

Structural bioinformatics

MD-TASK: a software suite for analyzing molecular dynamics trajectories

David K. Brown[1], David L. Penkler[1], Olivier Sheik Amamuddy[1], Caroline Ross[1], Ali Rana Atilgan[2], Canan Atilgan[2] and Özlem Tastan Bishop[1,*]

[1]Research Unit in Bioinformatics (RUBi), Department of Biochemistry and Microbiology, Rhodes University, Grahamstown 6140, South Africa and [2]Faculty of Engineering and Natural Sciences, Sabanci University, Tuzla 34956, Istanbul, Turkey

*To whom correspondence should be addressed.
Associate Editor: Alfonso Valencia

Received on December 16, 2016; revised on May 18, 2017; editorial decision on May 29, 2017; accepted on May 30, 2017

Abstract

Summary: Molecular dynamics (MD) determines the physical motions of atoms of a biological macromolecule in a cell-like environment and is an important method in structural bioinformatics. Traditionally, measurements such as root mean square deviation, root mean square fluctuation, radius of gyration, and various energy measures have been used to analyze MD simulations. Here, we present MD-TASK, a novel software suite that employs graph theory techniques, perturbation response scanning, and dynamic cross-correlation to provide unique ways for analyzing MD trajectories.
Availability and implementation: MD-TASK has been open-sourced and is available for download from https://github.com/RUBi-ZA/MD-TASK, implemented in Python and supported on Linux/Unix.
Contact: o.tastanbishop@ru.ac.za

1 Introduction

Molecular dynamics (MD) is used to understand the movement of atoms of macromolecules in a simulated cell environment. MD simulations produce trajectories depicting the motions of atoms by presenting the atomic coordinates at specified time intervals, allowing for investigation into changes over time. Traditionally, these trajectories are used to analyze macromolecule dynamics by calculating well-established measures such as root mean square deviation, root mean square fluctuation (RMSF), radius of gyration and energy-based approaches including the molecular mechanics/Poisson Boltzmann surface area (MM/PBSA) and the molecular mechanics/generalized Born surface area (MM/GBSA) (Kollman *et al.*, 2000).

Applications of MD simulations are broad, ranging from determining the stability of macromolecular complexes to understanding allosteric behavior of proteins. Although the measures mentioned above are informative, sometimes additional approaches are required. For example, changes in residue interaction networks (RINs) have been investigated in the context of mutation analysis (Bhakat *et al.*, 2014; Doshi *et al.*, 2016; Brown *et al.*, 2017). Another method, perturbation response scanning (PRS), following MD simulations, is used to identify residues important for controlling conformational changes (Atilgan and Atilgan, 2009).

Although traditional measures are incorporated into MD programs, they can be limited depending on the research questions. Hence, individual research labs often write custom scripts to further analyze trajectories. This might be challenging to some, especially if it requires mathematical knowledge and complex scripting. To serve this need, we present MD-TASK, an easy-to-use tool suite with detailed documentation, for analyzing MD trajectories. MD-TASK includes what we call dynamic residue network (DRN) (which combines the RINs of the frames in an MD trajectory) analysis, PRS, and dynamic cross-correlation (DCC), none of which are found in commonly used MD packages.

2 Materials and methods

2.1 Implementation

MD-TASK was developed using Python for Linux/Unix-based systems. Various non-standard Python libraries, including NumPy,

© The Author 2017. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

SciPy, Matplotlib (Hunter, 2007), MDTraj (McGibbon *et al.*, 2015) and NetworkX, were used in the suite. Thus, MD-TASK supports any formats the underlying Python libraries support including .binpos (AMBER), LH5 (MSMBuilder2), PDB, XML (OpenMM, HOOMD-Blue), .arc (TINKER), .dcd (NAMD), .dtr (DESMOND), hdf5, NetCDF (AMBER), .trr (Gromacs), .xtc (Gromacs), .xyz (VMD) and LAMMPS. Further, the igraph package for R was used to generate residue contact maps.

2.2 Network analysis

RINs can be analyzed using graph theory. In a RIN, each residue in the protein is a node in the network. An edge between two nodes exists if the C_β atoms (C_α for Glycine) of the residues are within a user-defined cut-off distance (usually 6.5–7.5 Å) of each other. MD-TASK constructs a DRN and uses this to calculate the changes in betweenness centrality (BC) and average shortest path (L) to residues over the trajectory.

2.2.1 Betweenness centrality

BC is a measure of how important a residue is for communication within a protein. The BC of a node is equal to the number of shortest paths from all nodes to all others that pass through that node. MD-TASK uses an implementation of Ulrik Brandes algorithm in the NetworkX library for calculating BC (Brandes, 2001).

2.2.2 Average shortest path

For a given residue, L is the sum of the shortest paths to that residue, divided by the total number of residues less one. MD-TASK uses a custom algorithm in the NetworkX library for quickly calculating the shortest path between all residues (NetworkX Developers, 2017). The average shortest path describes how accessible a residue is within the protein.

2.2.3 Residue contact map

Residue contact maps are generated by monitoring the interactions of a residue throughout a simulation, yielding a network diagram with the residue of interest [e.g. single nucleotide polymorphism (SNP)] at the center, and residues that it interacts with arranged around it. Edges between the residue of interest and the other residues are weighted based on how often the interaction exists.

2.3 Perturbation response scanning

Given the atomic coordinates for initial and target states, together with an equilibrated MD trajectory of the initial structure, the algorithm performs a residue-by-residue scan of the initial conformation, exerting external forces on each residue, and records the subsequent displacement of other residues using linear response theory and a variance-covariance matrix obtained from the MD trajectory (Atilgan *et al.*, 2012). The quality of the predicted displacement is assessed by correlating the predicted and experimental difference between the initial and target states. This results in a correlation coefficient for each residue, where a value close to one implies good agreement with the known experimental change. Residues whose perturbation invokes a conformational displacement closest to the target structure are reported as hot residues.

Table 1. Results of MD-TASK performance tests

Script	Average time (s)
calc_network.py (–calc-L)	37 298
calc_network.py (–calc-BC)	62 109
calc_delta_BC.py	16 852
calc_delta_L.py	1864
avg_network.py	1713
compare_networks.py	4230
delta_networks.py	2289
contact_map.py	19 806
calc_correlation.py	39 703
prs.py	95 480

2.4 Dynamic cross-correlation

DCC is calculated using the following formula:

$$C_{ij} = \frac{\langle \Delta r_i \cdot \Delta r_j \rangle}{\sqrt{\langle \Delta r_i^2 \rangle} \cdot \sqrt{\langle \Delta r_j^2 \rangle}}$$

with Δr_i the displacement from the average position of atom i, and $\langle\rangle$ the time average over the whole trajectory (Di Marino *et al.*, 2015). MD-TASK generates a heat-map depicting the DCC between the C_α atoms of selected frames in a trajectory to identify relative residue movements.

3 Performance

The results of a performance test are presented in Table 1. Tests were conducted on the 'example_small.dcd' trajectory, a 599 residue, 2000 frame trajectory provided in the 'example' sub-directory of the MD-TASK Github repository. Tools were set to iterate through the trajectory at 100 frame intervals and then executed 10 times. These executions were timed using the Linux 'time' utility, which includes the time to start up the Python interpreter, providing an accurate measure of what users will experience in practice. An average time was then calculated for each tool. The PRS script used 50 random forces per residue.

Tests were conducted on a PC running Ubuntu 16.04 with an Intel Core i5-6300U CPU with 4 logical cores running at 2.4 GHz and 8 GB RAM.

4 Applications

The network measures, BC and L, have been used in minimized protein structures (Ozbaykal *et al.*, 2015) for Alanine scanning. One suggested use, here, is SNP analysis over MD simulations. The network analysis scripts of MD-TASK were applied to analyze renin-angiotensinogen system (Brown *et al.*, 2017). The data indicated that combination of RMSF values with network analysis can be informative. Further, a SNP analysis protocol combining traditional MD analysis tools with DRN is proposed in a recent review article (Brown and Tastan Bishop, 2017).

PRS identifies single residues playing an active role in protein conformational changes between an initial and target state (Atilgan and Atilgan, 2009). Perturbative methods are instrumental in uncovering different protein functions such as the effect of pH on the distribution of conformations of calmodulin (Atilgan *et al.*, 2011) and ferric binding protein (Atilgan and Atilgan, 2009). By disclosing nonevident relationships, PRS was used to suggest new experiments to explore allosteric communication, e.g. in HSP70 (Penkler *et al.*, 2017).

As an example of MD-TASK outputs, we used HIV protease. Sequence of closed conformation of crystal structure (4ZIP) was used as wild type (WT) target sequence to model its structure in open conformation by means of homology modeling. Two major

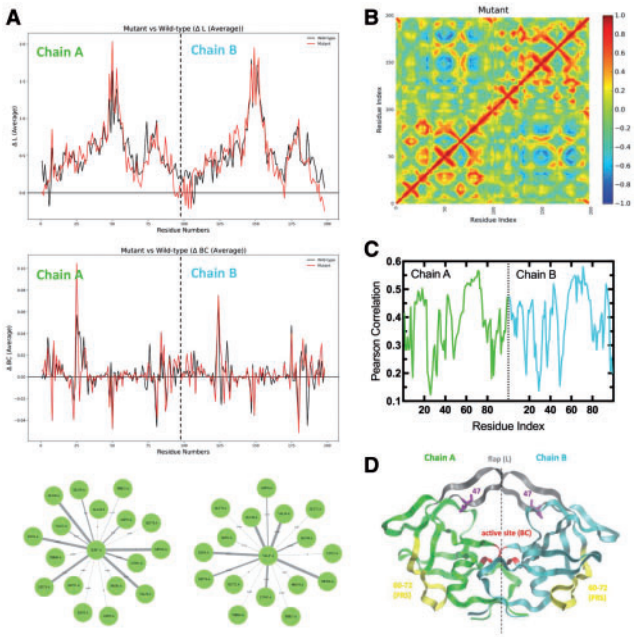

Fig. 1. Outputs for **(A)** Network analysis: average ΔL; average ΔBC; residue contact maps (from top to bottom); **(B)** DCC; **(C)** PRS (plot not generated by MD-TASK); **(D)** HIV-protease with significant regions highlighted

(V32I, I47V) and one minor (V82I) mutations were then introduced, also using homology modeling, to create the open conformation mutant structure. In both cases 1TW7 was used as the template. Network analysis was performed for 40 ns MD runs for both WT and mutant proteins in open conformation (Fig. 1A). DCC was calculated for the mutant protein (Fig. 1B). For PRS (Fig. 1C), an equilibrated 20 ns section of the mutant trajectory was used. The PDB structure, 3S54, which represents the closed conformation with the same mutations, was used as the end state during PRS calculations. Residues having the highest change in reachability (highest ΔL) during the course of the trajectory are 46–56, comprising the flap. This property gives HIV-protease the flexibility to expand the active site cavity and diminish the effect of inhibitors while staying functional (Martin *et al.*, 2017). The peaks in ΔBC correspond to active site residues 25–27, supporting the hypothesis that high BC positions are responsible for interdomain communication (Ozbaykal *et al.*, 2015). While these properties are similar in both WT and mutant, the neighborhood structure of position 47 slightly shifts, as shown in the residue contact maps. DCC (Fig. 1B) demonstrates that the intrachain motions are highly correlated, while the motions of the chains with respect to each other are anticorrelated (residues 1–99 versus 100–198). The PRS results (Fig. 1C) display that there are no single residues whose perturbation directly leads to the conformational change between the equilibrated WT structure at the 20 ns snapshot and the triple mutant, as the maximum correlations are ~0.6. Nevertheless, highest correlations span residues 60–72 implying an allosteric communication between this region and the flap motions. The specific residues are mapped onto the protein structure in Figure 1D. Figure 1 summarizes how MD-TASK provides a means to analyze the trajectories, and gives a bird's eye view of various factors that may be effective in the dynamics of a protein.

5 Conclusion

MD simulations have become an important tool in structural bioinformatics. Here, we present a new tool suite for analyzing MD trajectories using DRN analysis, PRS, and DCC. To the best of our knowledge, MD-TASK is the first downloadable tool suite for analysis of different properties along MD simulations not commonly found in other MD packages.

Acknowledgements

We thank Tandac Furkan Guclu and Gizem Ozbaykal for helpful discussions.

Funding

This work is supported by the National Institutes of Health Common Fund under grant number U41HG006941 to H3ABioNet, and by the National Research Foundation (NRF), South Africa, [grant number 93690]. The content of this publication is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

Conflict of Interest: none declared.

References

Atilgan,A.R. *et al.* (2011) Subtle pH differences trigger single residue motions for moderating conformations of calmodulin. *J. Chem. Phys.*, **135**,

Atilgan,C. *et al.* (2012) Network-based models as tools hinting at nonevident protein functionality. *Annu. Rev. Biophys.*, **41**, 205–225.

Atilgan,C. and Atilgan,A.R. (2009) Perturbation-response scanning reveals ligand entry-exit mechanisms of ferric binding protein. *PLoS Comput. Biol.*, **5**, 10005–10044.

Bhakat,S. *et al.* (2014) An integrated molecular dynamics, principal component analysis and residue interaction network approach reveals the impact of M184V mutation on HIV reverse transcriptase resistance to lamivudine. *Mol. Biosyst.*, **10**, 2215–2228.

Brandes,U. (2001) A faster algorithm for betweenness centrality*. *J. Math. Sociol.*, **25**, 163–177.

Brown,D.K. *et al.* (2017) Structure-based analysis of single nucleotide variants in the renin-angiotensinogen complex. *Glob. Heart*, pii: S2211–8160(17)30006-6.

Brown,D.K. and Tastan Bishop,Ö. (2017) Role of Structural Bioinformatics in Drug Discovery by Computational SNP Analysis. *Glob. Heart*, pii: S2211-8160(17)30009-1.

Doshi,U. *et al.* (2016) Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. *Proc. Natl. Acad. Sci. USA*, **113**, 4735–4740.

Hunter,J.D. (2007) Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.*, **9**, 90–95.

Kollman,P.A. *et al.* (2000) Calculating structures and free energies of complex molecules: combining molecular mechanics and continuum models. *Acc. Chem. Res.*, **33**, 889–897.

Di Marino,D. *et al.* (2015) Characterization of the differences in the cyclopiazonic acid binding mode to mammalian and P. Falciparum Ca2+ pumps: a computational study. *Proteins*, **83**, 564–574.

Martin,P. *et al.* (2017) 'Wide-Open' structure of a multidrug-resistant HIV-1 protease as a drug target. *Structure*, **13**, 1887–1895.

McGibbon,R.T. *et al.* (2015) MDTraj: a modern open library for the analysis of molecular dynamics trajectories. *Biophys. J.*, **109**, 1528–1532.

NetworkX Developers. (2017) NetworkX all_pairs_shortest_path_length function.

Ozbaykal,G. *et al.* (2015) In silico mutational studies of Hsp70 disclose sites with distinct functional attributes. *Proteins*, **83**, 2077–2090.

Penkler,D. *et al.* (2017) Perturbation response scanning reveals key residues for allosteric control in Hsp70. *J. Chem. Inf. Model*, doi: 10.1021/acs.jcim.6b00775.