ExTASY: A Python-based Extensible Toolkit for Advanced Sampling and Analysis in Biomolecular Simulation

Ardita Shkurti1, Charles Laughton*1, Ramon Goñi2,3, Iain Bethune4, Elena Breitmoser5, Shantenu Jha5, Vivek Balasubramaniam6, Cecilia Clementi6,7, Ben Leimkühler8, Panos Parpas9, Mauro Maggioni10

1 School of Pharmacy, Centre for Biomolecular Sciences, The University of Nottingham, University Park, Nottingham, NG7 2RD, UK
2 Department of Life Sciences, Barcelona Supercomputing Center, J. Girona 29, Barcelona, 08034, Spain
3 Joint BSC-CRG-IRB Program in Computational Biology, Barcelona, Spain
4 Edinburgh Parallel Computing Centre (EPCC), The University of Edinburgh, Edinburgh, EH9 3FD, UK
5 Electrical and Computer Engineering, Rutgers University, Piscataway, New Jersey, NJ 08854, USA
6 Department of Computer Science, University of Oxford, Oxford, OX1 3QY, UK
7 Department of Chemistry, Rice University, Houston, TX 77005, USA
8 Department of Theoretical Biological Physics, Rice University, Houston, TX 77005, USA
9 School of Mathematics, The University of Edinburgh, Edinburgh, EH9 3FD, UK
10 Department of Mathematics, Computer Science, and Electrical and Computer Engineering, Duke University, Durham, NC 27708-0320, USA

*Contact: charles.laughton@nottingham.ac.uk

ExTASY - Extensible Tools for Advanced Sampling and analysis — is a project to provide the biomolecular simulation community with a flexible and extensible software toolkit of advanced sampling methods for molecular simulation, targeted primarily at High Performance Computing (HPC) infrastructures (www.extasy-project.org).

PCA
Principal Component Analysis (PCA) applied to molecular simulation data:
- Reduces sampling data dimensionality in order to capture the dominant modes of motion of the molecular system;
- Gives insight into structural and dynamical behaviour of molecules;
- Enables highly compressed data storage of simulation trajectory files.

Pyhton wise
- The Python-implemented CoCo-MD workflow maps molecular data structures into PCA space, identifies new data structures in the same space and converts the newly discovered data structures back into the original dimensionality space of the MD data;
- MDAnalysis – a Python package – is used from CoCo-MD to convert MD data into numpy arrays that are subsequently processed in the code;
- scipy.linalg (linear algebra) is used to determine the PCA space of MD data – dimensionality-reduced space of the biomolecular system's shapes dynamics;
- scipy.ndimage (multi-dimensional image processing) is used to determine regions most distant from any sampled so far in the PCA space.

Results
- Results from the use of PCA to investigate the space of molecular shapes and transitions among them in the course of the MD simulations are shown for the Major Urinary Protein extending our previous work [2];
- Average of 100 independent simulations for each scenario;
- Dotted lines are the clusters boundaries of the MD projections. Each of the clusters is assumed to belong to a specific molecular shape.
- The red circles A to J are the 10 clusters identified in the PCA projections and the simulation data of Penta Alanine;
- The labels of the green arrows show the number of steps that it takes before the cluster pointed by the arrow is visited from cluster A. Each step is associated with a print of a snapshot of atomistic coordinates in the course of the MD simulation.
- The CoCo method shows an excellent performance (yellow highlighted steps labels) on visiting the distant clusters much earlier than traditional MD.

References