Identification of conformational states and transitions in between from molecular dynamics trajectories and vibrational signatures

Sandro Wrzalek, Irtaza Hassan, Petra Imhof

Institute of Theoretical Physics, Freie Universität Berlin, Arnimallee 14, 14195 Berlin, Germany

We used the Markov state modeling (MSM) technique to identify meta-stable conformations of biomolecules such as small peptides and the time scales associated with the slowest processes sampled in trajectories from classical molecular dynamics (MD) simulations. The vibrational spectra of representative conformations are calculated from the first-principle MD simulations in solution at finite temperature in the frequency range that is most sensitive to the peptide conformation, so-called amide bands (1300–1800 cm-1), see the example of the small floppy peptide Alanine-Leucine (AL)[1]. This combined approach to analyse the vibrational signature is extended to Alanine-Leucine-Alanine and Alanine-Leucine-Alanine-Leucine (ALAL). Generally, all computed spectra show two prominent bands which are assigned to the stretch vibrations of the carbonyl and carboxyl group, respectively. Variations in bandwidths and exact maxima are likely due to small fluctuations in the backbone torsion angles. The detailed vibrational spectroscopic analysis of the most probable, beta sheet-like conformations of the selected peptides (i.e., AL, ALA, ALAL) is performed with the time-frequency analysis based on the wavelet transform using the trajectories from the first principle MD simulations. Particularly, we analysed how the instantaneous frequencies of carbonyl groups (C=O) present in the peptides are affected by the local solvent environment. It is clear from the wavelet analysis of carbonyl bonds that due to change in the hydrogen bonding state or simply due to change in the local solvent environment, the state of the carbonyl bond changes which leads to change in its instantaneous frequency.

Although, MSMs work great to identify the meta-stable conformations for small molecules like peptides, they can become unreliable for molecules with more degrees of freedom, like DNA and RNA, since Markov models are, in conjunction with the demand for large data sets, highly sensitive to the preprocessing steps. Especially, commonly used coordinate projection methods like Principal Component Analysis (PCA) and Time-lagged independent component analysis (TICA) are risky: First, the state space may not be reducible to a set of only a few basis vectors leading to important information being lost (or still too many dimensions). The latter becomes difficult for clustering the data set. PCA is furthermore problematic since it rotates and reduces the space based on the variance, so it likely removes dimensions containing the slowest (least sampled=lowest variance) processes. Our approach tackles this problem by first reducing dimensions of different type (Cartesian, curvlinear, binary) within their space and combine them later with the Multi-View clustering method proposed in [2]. Our approach for polar coordinates defines two time windows with length l, separated by a constant lag time τ. Those windows move through a trajectory, while comparing their distributions, via Kullback–Leibler divergence, with each other. If the distance of two distributions exceeds a defined threshold, κ, a transition in the time frame $2l+\tau$ will be considered. A possible new state (pns) is defined as the part of the trajectory, after and before a new transition. A pns has to be distinguished from a repeated state to qualify for a new state. That distinction of states can be considered in a clustering manner: To validate if the pns is a truly new state, its distribution will be compared to that of the already known states - if the distribution is different the pns will become a new state, else the part of the trajectory will be assigned to the closest new state. Data-sensitive clustering methods like (H)DBScan can be helpful. Binary coordinates can be reduced by using auto encoders (Neural Networks) to encode those binary vectors in fewer dimensions.

[1] I. Hassan, L. Donati, T. Stensitzki, B. Keller, K. Heyne, and P. Imhof, CPL, 2018 698:227–233.
[2] S. Kanaan-Izquierdo, PhD thesis - Universitat Politecnica de Catalunya, 2017, Multiview pattern recognition methods for data visualization, embedding and clustering