Time-series segmentation and latent representation of musical instruments

Gregory Szep

King's College London

August 14, 2018

Abstract

Music information retrieval tasks serve as faithful benchmarks for time-series analysis pipelines due to the availability of strongly labelled training data such as MusicNet. Clustering algorithms in spectral sub-spaces, hidden Markov models and causal convolutional neural networks are compared in their ability to transform time-series to a continuous latent space that clusters eleven orchestral instruments. The latent space is evaluated quantitatively with precision-recall metrics obtained by comparing the instrument prediction from a segment of audio to the ground truth obtained from musical scores, and qualitatively by generating samples of audio for given regions in the latent space.

1 scientific publication

Select a scientific publication about non-equilibrium molecular dynamics of biomolecules and discuss its relevance respect to the course you attended and this research assignment.

2 Analysis of Umbrella Sampling simulations

Every student has access to the raw data of umbrella sampling simulations for two systems. These umbrella sampling simulations have been ran having the aim to calculate the Potential of Mean Force (PMF) of the mechanical unfolding of two peptides: one being in the initial conformational state of an a-helix; the other being in the initial conformational state of a b-hairpin. Every student is supposed to calculate the PMF for both the systems

2.1 Describe the simulated systems and the type of performed simulations you analyze

2.2 Calculate and display (plot) the PMF for the two systems. Explore the relevance of the bootstrap parameter respect to evaluation of the PMF's error using the script $\mathbf{g}_w ham_s cript_n ew.x$

Bootstrapping is a resampling method that quantifies statistical uncertainty by dividing the data into N subsets, hence without additional measruements uncertainty is obtained by *pulling the data up by its own bootstraps*. By default the `g_wham` code forms subsets with replacement over the complete histograms along the reaction coordinate [1], leaving us with N as a hyperparameter.

This hyperparameter is subject to the bias-variance trade-off. Chooseing a large N results in relatively unbiased and uncertain estimates, while a small number of subsets give rise to more certain yet biased estimates. The former is preferred but comes with a computational cost. It appears that the uncertainty bounds converge to a fixed value after around $N \geq 20$ so there is little sense in setting N at any other value than $N = 20$.

2.3 Quantify and discuss the histograms' overlap as obtained from the analysis of the umbrella sampling simulations

3 Conductance of outer membrane protein F

Outer membrane protein F (ompF) is present in E. Coli

Every student has access to the raw data of non-equilibrium molecular dynamics simulations with an external electric field switched on. This set up corresponds in having a specific applied voltage across the simulation box. These non-equilibrium molecular dynamics simulations have been ran having the aim to test/check if this type of simulations are able to reproduce the feature of the protein OMPF of being weakly cationic selective in salts of monovalent cations (K+,Cl-).

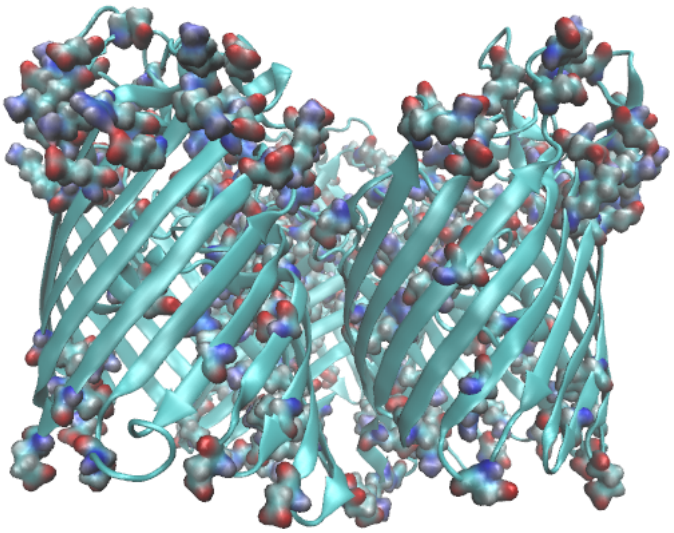

Figure 1: Side-view of structure: three β-barrels transmembrane channel. Highlighted are charged residues, which dominanate the outer membrane side.

Every student will find raw data for the following applied voltages: $\pm$100mV, $\pm$200mV, $\pm$1000mV. For $\pm$100mV: every student will have only 1 trajectory to analyse (the file dcd). For $\pm$200mV: every student will have 6 trajectories to analyse (the file dcd). For $\pm$1000mV: every student will have 3 trajectories to analyse (the file dcd). Every student is supposed to follow the pipeline of analysis proposed/provided during the tutorial that permits to estimate the relevant properties of interest of the system. Read carefully the file README before to start the analysis! Calculate and discuss ion flux through the channel as well as its conductance, putting the results in relation with the current literature.

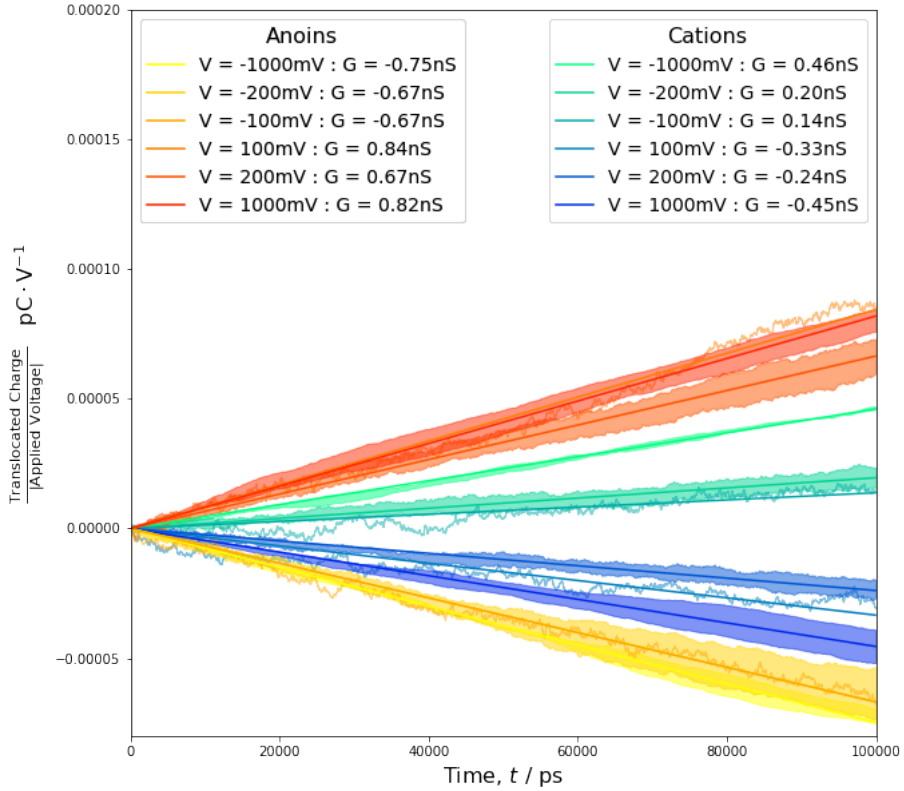

Figure 2: Conductance estimates G for anions and cations across outer membrane protein F (ompF) under various applied voltages V

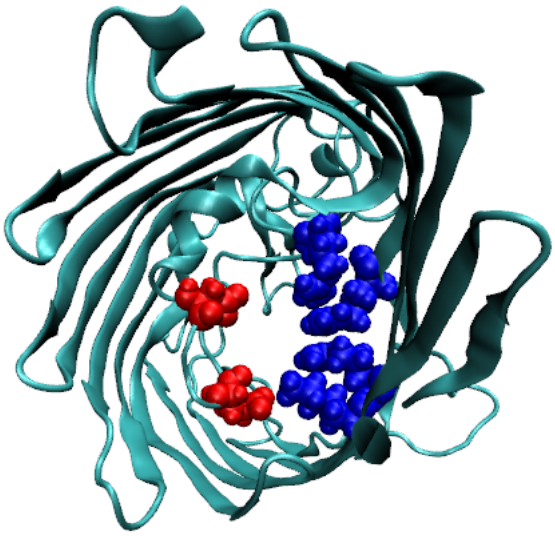

Figure 3: Single ompF β-barrel from inner membrane side, highlighting positive COOH (red) and negative NH_2 (blue) partial charged groups that facilitate cationic selectivity

The student is supposed to address all the points of the assignment described above, but has the freedom to focus in discussing details of methodology and selected approaches either to the "Analysis of Umbrella Sampling Simulations" or to the "Analysis of the non-equilibrium molecular dynamics simulations with external electric field".

References

[1] J. S. Hub, B. L. de Groot, and D. van der Spoel, "g_wham—A Free Weighted Histogram Analysis Implementation Including Robust Error and Autocorrelation Estimates," *Journal of Chemical Theory and Computation*, vol. 6, pp. 3713–3720, dec 2010.