A Machine learning study on the fluctuation of RNA molecule and its relation to Magnesium binding

Erdong Ding, Vaishnavi Mallagundla

October 24, 2023

1 Introduction

Molecular dynamics (MD) simulations has been applied to study the dynamics and the thermodynamics for large biomolecules for decades. The system will evolve under Newton's equation with energy exchange with a heat source to keep the temperature a constant. After the simulation, a trajectory is obtained with encodes the conformation distribution of the system. Among all the biomolecules, RNA is a special one because of it usually has multiple local minima in the free energy surface and a highly negatively charged backbone which requires a lot of cations (especially magnesium) to compensate the charge. Basically, the magnesium tends to bind to specific sites on the surface of RNA, which can be measured by the drastic change in the diffusion manner of magnesium. The typical time scale of binding is 10 ns before the cation becomes diffusive again and then bind to another association site. It is shown that the fluctuation of the RNA molecule near its local minimum is likely to be coupled with the binding of cations and might imply the potential conformation change pathway that can not be captured by unbiased MD simulations(*1*, *2*).

2 Problem statement

Previously, the fluctuation of RNA and its correlation with cations has been studied by so called "joint PCA", simultaneously for the Cartesian coordinates of the RNA phosphorus atoms and for the occupation counts for the ions in the association sites (positions where ions binds RNA) (*2*). I have recently conducted similar simulations on a different RNA system. However, because the intrinsic complexity, the largest eigenvector from PCA only covers around 20% of the variance. And non-linear effect seems to necessary to study the complicated energy landscape. Another potential issue of this "joint PCA" is that the occupation of cations is a Boolean value, which means the correlation between RNA fluctuation and Mg binding can not be well-expressed by definition.

Here, we will separate this problem into two sub-problems. First, we will address the PCA issue by using the Variational Autoencoders (VAE) to conduct a dimension reduction for the RNA system, and obtain a automatic clustering in the new low-dimensional space. Then, the Magnesium density in each clusters will be measured and a density peak clustering will be conducted to find the Magnesium association cites for different RNA clusters.

3 Technical approach

The VAE and its variant method(*3–5*) have been widely used in the field of molecular dynamics for dimension reduction and reaction coordinates determination. Here, we are going to implement a recently developed VAE method, GMVAE(*3*), in pytorch (the original package was written with TensorFlow) and applied the method for the dimension reduction. The GMVAE method ha introduced another Gaussian distributed variable y to acknowledge the multi-basin nature of the physical system to encourage the separation of metastable states.

It is important to choose the input variables for the VAE training. Previous choices include pairwise distances between heavy atoms as well as the dihedral angles (*3*), scaled heavy backbone atom Cartesian coordinates (*5*), contact matrix of a coarse-grained biomolecule model (*4*). It inspired two possible inputs for the model training. The first is to align the trajectory to the crystal structure and then use the Cartesian coordinates of all the Phosphorous atom (67*3 variables).The second is to use a contact map method developed by our group. Here, we do not measure the entire pairwise distance matrix but select representative contacts based on the crystal structure. We can measure the distance between these contacts for each frame of the simulation as a measurement of how the conformation deviates from the crystal structure and use these distance as inputs (531 variables for this RNA).

After the dimension reduction is finished, we anticipate to find a collective variable that captures the fluctuation of the RNA and constructs a clustering of RNA for different metastable states. The Magnesium distribution of each cluster will be collected, and the density peak clustering (*6*) will be conducted to determine the Magnesium association sites for each RNA cluster and see the difference of Magnesium binding.

4 Results so far

All the input data has been prepared.

References

1. R. L. Hayes *et al.*, en, *J. Am. Chem. Soc.* **134**, 12043–12053, ISSN: 0002-7863, 1520-5126, (2022; https://pubs.acs.org/doi/10.1021/ja301454u) (July 2012).
2. H. Lammert, A. Wang, U. Mohanty, J. N. Onuchic, en, *J. Phys. Chem. B* **122**, 11218–11227, ISSN: 1520-6106, 1520-5207, (2023; https://pubs.acs.org/doi/10.1021/acs.jpcb.8b06874) (Dec. 2018).
3. Y. B. Varolgüneş, T. Bereau, J. F. Rudzinski, en, *Mach. Learn.: Sci. Technol.* **1**, Publisher: IOP Publishing, 015012, ISSN: 2632-2153, (2023; https://dx.doi.org/10.1088/2632-2153/ab80b7) (Apr. 2020).
4. D. Bhowmik, S. Gao, M. T. Young, A. Ramanathan, *BMC Bioinformatics* **19**, 484, ISSN: 1471-2105, (2023; https://doi.org/10.1186/s12859-018-2507-5) (Dec. 2018).
5. H. Tian *et al.*, *Frontiers in Molecular Biosciences* **8**, ISSN: 2296-889X, (2023; https://www.frontiersin.org/articles/10.3389/fmolb.2021.781635) (2021).

6. A. Rodriguez, A. Laio, en, *Science* **344**, 1492–1496, ISSN: 0036-8075, 1095-9203, (2023; `https://www.science.org/doi/10.1126/science.1242072`) (June 2014).