A Streamlined Method to Compute and Visualise Markov Communities in Biomolecules

Robert Miles Ryan Reese
Supervised by: Dr Sophia Yaliraki† and Dr Jun Yang‡

†Department of Chemistry, Imperial College London, White City Campus, W12 0BZ, United Kingdom
‡Department of Chemistry, The University of Hong Kong, Pok Fu Lam, Hong Kong

Overseas Research Fellowship (ORF)
2023 for Science Students

Poster No.: C2
Name: Robert Miles Ryan Reese
University No.: 3035743584
Student's Major: Chemistry
Student's Minor: Mathematics, Computer Science

Imperial College London

香 港 大 學
THE UNIVERSITY OF HONG KONG

Abstract & Introduction

Over the past few years, a significant amount of effort has been placed on developing and applying computational approaches that can reveal fundamental properties of biomolecules. Recently published examples include software capable of predicting allosteric sites[1] and signalling pathways[2] within proteins. Here, we present a Python package known as *VMC* (**Visualise Markov Communities**), which intends to streamline the process of colouring multiscale communities of atoms within proteins.

These communities of atoms are found by performing computational analysis using Markov Stability (**MS**) on the energy-weighted atomistic graph representation of the protein. MS is an established and computationally-efficient graph-partitioning method that can be used to identify robust atom communities encoded in the atomic protein structure.[3] These robust partitions correspond to biochemically relevant substructures of the protein. As the quality of a partition in MS is linked to a stochastic Markov process taking place on the graph, all levels of spatial resolution can be explored.[4] This allows critical functional and structural subunits such as amino acids, secondary structures, and larger subdomains of the protein to be uncovered in an unsupervised manner.[5] This application of MS is well-established and has been applied to study multi-scale organisation in biomolecular structures such as myosin,[6] caspase-1,[7] and the neuronal network of *C. elegans*.[8]

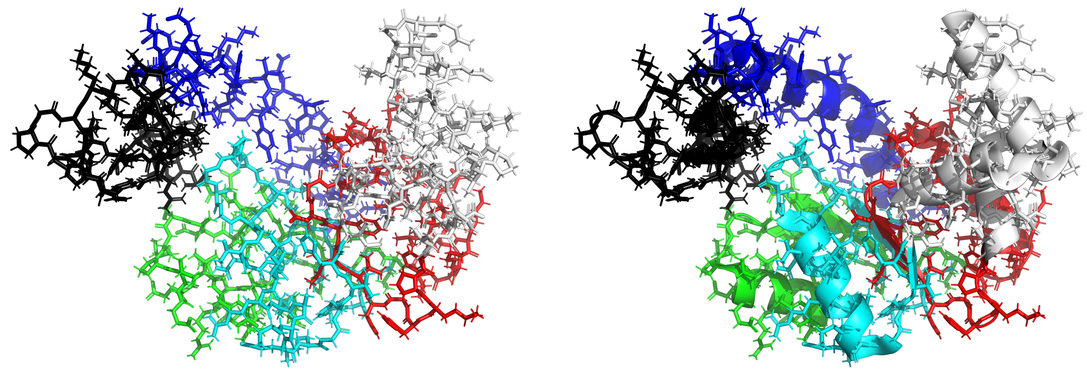


Figure 1. Communities of atoms in the structure of Apo Adenylate Kinase from *Aquifex Aeolicus* (PDB ID: 2rh5) found using Markov Stability and visualised using *VMC*. Without **(left)** and with **(right)** protein ribbons and helices.

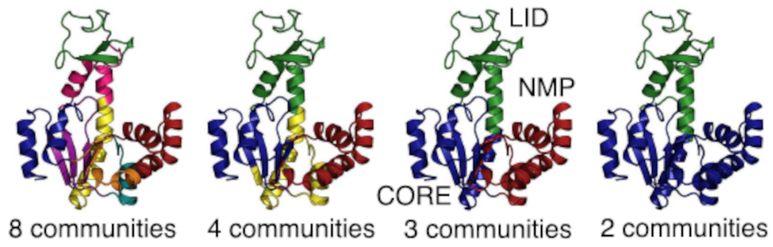


Figure 2. The relevant partitions of Adenylate Kinase found at large Markov times by Delmotte et al.[6] The 8-community partition of the protein structure agrees with previous hinge analyses, while the 3-community partitions corresponds to experimentally verified functional domains of AdK

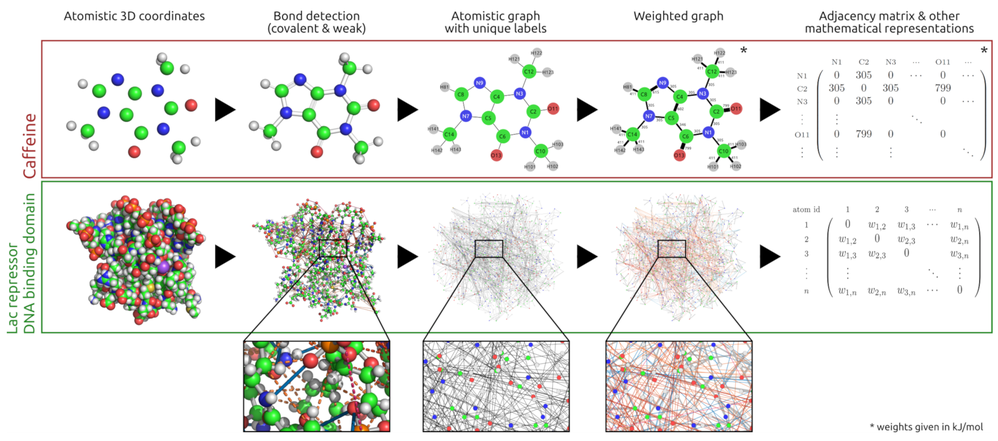


Figure 3. Method overview for *BagPype*: a Python package for the construction of atomistic, energy-weighted graphs from biomolecular structures.[10]

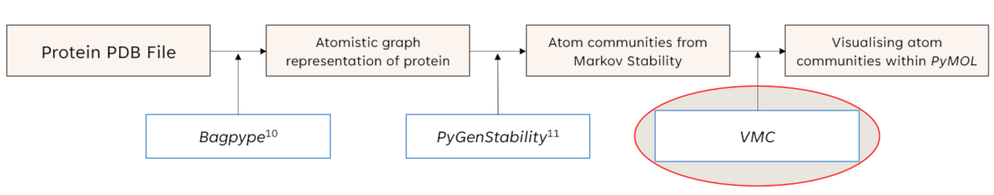


Figure 4. The intended scheme that streamlines the process of uncovering the multiscale structure of proteins using Markov Stability. Here, *VMC* completes the final visualisation step of the scheme.

Scheme & Methods

Thus far, no software has been published that allows the obtained graph partitions to be highlighted directly on the protein structure. The results of MS analysis are only relevant if the location and distribution of atom communities can be visualised in their proper context. Assisted by the open-source molecular visualisation system *PyMOL*[9], *VMC* solves this issue by providing a convenient interface to visualise atoms on the protein according to their respective communities.

Within our scheme, we have utilised the Python package *BagPype*[10], which automates the construction of the protein's graph representation using the atomistic cartesian coordinates contained in its PDB file. The atomistic graph is parameterised with energies representing weak and covalent bonds, and then encoded into a weighted adjacency matrix. We have combined this with a recently published Python package implementation of MS under the name *PyGenStability*[11], which provides a convenient interface for C++ implementations of the Leiden graph-partitioning algorithm. *VMC* is therefore intended to act as the final visualisation step of a 3-part scheme.

Compared to other methods, the absence of course-graining means that robust atom communities encoded in the protein sturucture emerge directly from the detailed physico-chemical information at the smallest atomic scale. Moreover, as MS follows the time evolution of a Markov diffusion process taking place on the graph, no community scale has to be imposed *a priori*. The entire multiscale structure of the protein can therefore be disclosed.

Implementation & Examples

VMC is a multi-functional Python software package that provides a suite of visualisation tools. The primary goal is usability and accessibility for fellow researchers. For instance, given a set of MS results at a particular Markov timescale, atoms can be coloured according to their communities in *PyMOL* using a single built-in function. **(Figure 5)**

The software also allows multiple Markov timescales to be visualised at once, in which case, snapshots of atom communities at each timescale are automatically named and saved in PNG format. These snapshots can then be compiled into an animation using another built-in function, allowing users to visualise Markov time acting as a resolution parameter that finds increasingly coarser partitions corresponding to larger communities. **(Figure 6)**

Finally, the package also includes a function that allows the results of MS analysis to be plotted and viewed from within the same *PyMOL*[9] window. The function also automatically labels the most robust partitions as detected by *PyGenStability*[10] and displays the number of atom communities at these Markov timescales. **(Figure 7)**

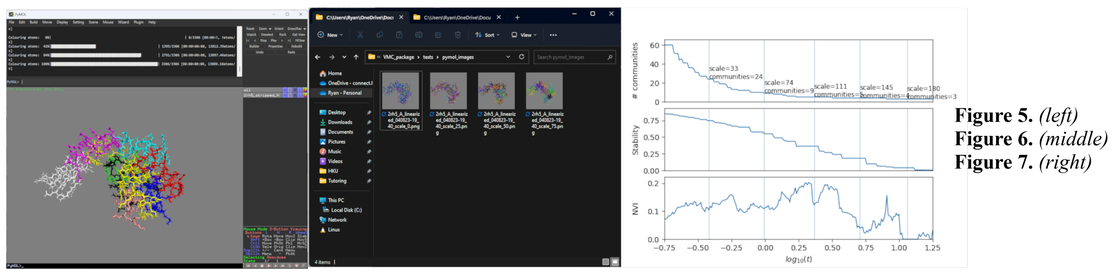
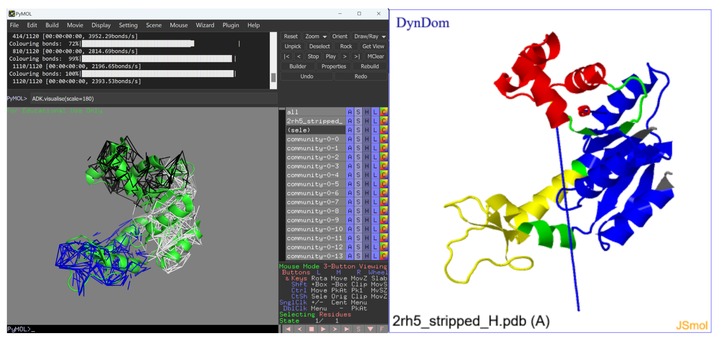


Figure 5. *(left)*
Figure 6. *(middle)*
Figure 7. *(right)*

Figure 8. One of the most robust graph partitions as selected by *PyGenStability* divides the atoms of AdK (PDB ID: 2rh5) into three subgraphs that exactly correspond to the three widely accepted subdomains of AdK.

(Left) Visualisation using *VMC*.
(Right) The three subdomains of AdK as predicted by the molecular dynamics-based program *DynDom*[12].



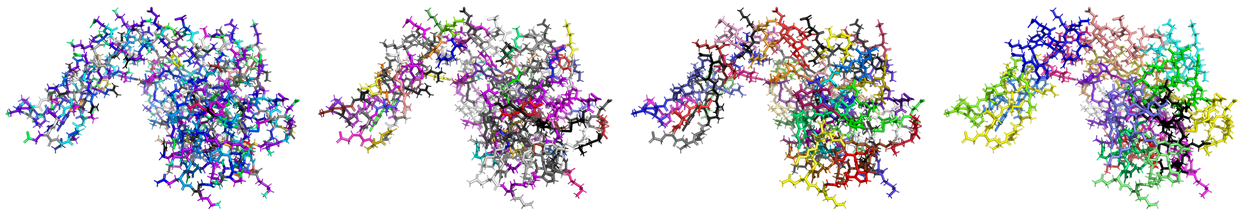


Figure 9. Communities of atoms in the structure of AdK (PDB ID: 2rh5) found using Markov Stability and visualised using *VMC* at Markov timescales 0/100 *(left)*, 25/100 *(middle-left)*, 50/100 *(middle-right)*, 75/100 *(right)*. Note that the number of communities decreases as the timescale of the Markov process increases.

Outlook

A recently published paper attempted to quantify edge-to-edge relations within graph networks. The paper investigated the effect of instantaneous bond fluctuations propagating throughout proteins and introduced a quantity known as bond-to-bond propensity.[13] To continue our project, we have decided to combine this idea with community detection using MS, thereby finding robust communities of bonds within the protein structure. Hence, analogous visualisation functions for communities of bonds, have been created. However, given that this is a novel direction whose applications have not been properly investigated, it has been excluded from the open-source Python package. Hopefully, after further research into these bond communities, the entirety of the code can be released as part of a separate paper in the future.

Currently, *VMC* has been published as an official Python package on the Python Package Index so that the software can be freely and conveniently installed by other users. Similarly, online documentation for *VMC* has been released, which includes guides, examples, and explains supplementary tools that users may find helpful.

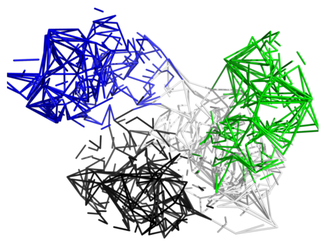


Figure 10. An example of bond communities in the structure of AdK (PDB ID: 2rh5) being visualised using *VMC*.

References & Source Code

1. Mersmann, Sophia F. et al. ProteinLens: a web-based application for the analysis of allosteric signalling on atomistic graphs of biomolecules. Nucleic Acids Research 49, W551-W558 (2021). https://doi.org:10.1093/nar/gkab350
2. Wu, N., Yaliraki, S. N. & Barahona, M. Prediction of Protein Allosteric Signalling Pathways and Functional Residues Through Paths of Optimised Propensity. Journal of Molecular Biology 434, 167749 (2022). https://doi.org:https://doi.org/10.1016/j.jmb.2022.167749
3. Peach, R. L. et al. Unsupervised Graph-Based Learning Predicts Mutations That Alter Protein Dynamics. bioRxiv (2019). https://doi.org:10.1101/847426
4. Delvenne, J.-C., Yaliraki, S. N. & Barahona, M. Stability of graph communities across time scales. Proceedings of the National Academy of Sciences 107, 12755-12760 (2010). https://doi.org:10.1073/pnas.0903215107
5. Schaub, M. T., Delvenne, J.-C., Yaliraki, S. N. & Barahona, M. Markov Dynamics as a Zooming Lens for Multiscale Community Detection: Non Clique-Like Communities and the Field-of-View Limit. PLOS ONE 7, e32210 (2012). https://doi.org:10.1371/journal.pone.0032210
6. Delmotte, A., Tate, E. W., Yaliraki, S. N. & Barahona, M. Protein multi-scale organization through graph partitioning and robustness analysis: application to the myosin-myosin light chain interaction. Phys Biol 8, 055010 (2011). https://doi.org:10.1088/1478-3975/8/5/055010
7. Amor, B., Yaliraki, S. N., Woscholski, R. & Barahona, M. Uncovering allosteric pathways in caspase-1 using Markov transient analysis and multiscale community detection. Mol Biosyst 10, 2247-2258 (2014). https://doi.org:10.1039/c4mb00088a
8. Bacik, K. A., Schaub, M. T., Beguerisse-Díaz, M., Billeh, Y. N. & Barahona, M. Flow-Based Network Analysis of the Caenorhabditis elegans Connectome. PLOS Computational Biology 12, e1005055 (2016). https://doi.org:10.1371/journal.pcbi.1005055
9. Schrodinger, LLC. The PyMOL Molecular Graphics System, Version 2.0 (2017).
10. Song, F., Barahona, M. & Yaliraki, S. N. BagPype: A Python package for the construction of atomistic, energy-weighted graphs from biomolecular structures. (2022). https://doi.org:10.5281/zenodo.6326081
11. Arnaudon, A. et al. barahona-research-group/PyGenStability. (2023). https://doi.org:10.5281/zenodo.7898442
12. Lee, R. A., Razaz, M. & Hayward, S. The DynDom database of protein domain motions. Bioinformatics 19, 1290-1291 (2003). https://doi.org:10.1093/bioinformatics/btg137
13. Amor, B. R. C., Schaub, M. T., Yaliraki, S. N. & Barahona, M. Prediction of allosteric sites and mediating interactions through bond-to-bond propensities. Nature Communications 7, 12477 (2016). https://doi.org:10.1038/ncomms12477

The open-source code and documentation are hosted on GitHub under a GNU General Public License at <**https://github.com/Ryan-Reese/VMC**>.