Application of Machine Learning to Study Characteristics of Yeast Heat Shock Protein Hsp104

Anvar Normurodov, Tharushi Rajaguru, and George Stan

Department of Chemistry, University of Cincinnati, OH 45221

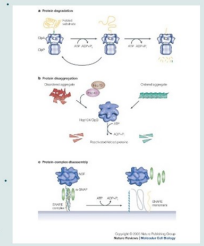
Newly produced proteins or already folded proteins may get damaged and lose their 3D structure that is crucial for their biological function. This can happen because of heat shock or stress, leading to neurodegeneration and various other diseases. Proteins belonging to the AAA+ family (ATPases associated with diverse cellular activities) play a crucial role in preserving protein stability. The Hsp104 from Saccharomyces cerevisiae are members of the AAA+ family that support protein quality control by unfolding aberrant or toxic proteins. Though we have identified the purpose of this enzyme, how this protein performs disaggregation is unclear. In this study we implement a clustering method on molecular dynamics simulations. After having simulated the action of the Hsp104, we identified and picked out most representative states. After identifying biochemically relevant descriptors based on the literature, we apply machine learning algorithms to identify the structural features that best describe the identified conformations.

INTRODUCTION

AAA+ proteins

- AAA+ - ATPases associated with diverse cellular activities use energy from ATP to perform various cellular tasks
- Contain conserved nucleotide binding motifs in the nucleotide binding domains:

  Walker A: GxxxxGKT – bind nucleotide
  Walker B: hhhhD(D/E) – hydrolyze nucleotide
- Functions in protein quality control, DNA replication, vesicular transport



Hanson, P., Whisstaer, S. AAA+ proteins: have engine, will rework. *Nat Rev Mol Cell Biol* **4**, 119–129 (2003). https://doi.org/10.1038/nrm1024

HSP104 disaggregation nanomachine

- Hexameric double ring structure
- Found in Eukaryotes
- Involved in thermotolerance and produced in response to various stress factors such as hydrogen peroxide, ethanol, and cold shock
- Substrate threading and unfolding mediated by ATP driven power strokes
- Each monomer has two conserved nucleotide binding domains and pore loop

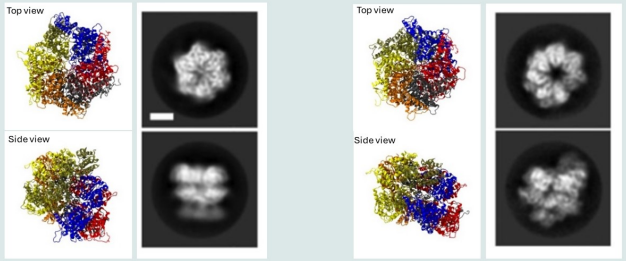
Shorter J, Southworth DR. Spiraling in Control: Structures and Mechanisms of the Hsp104 Disaggregase. Cold Spring Harb Perspect Biol. 2019 Aug 1;11(8):a034033. doi: 10.1101/cshperspect.a034033. PMID: 30745294; PMCID: PMC6671941.

Cryo-EM structures

Cryo-EM at resolution of reveals following conformational states:

Extended (PDB ID: 5VYA)
4.Å

Closed (PDB ID: 5VJH)
4Å

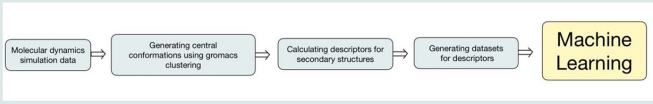


Gates SN, Yokom AL, Lin J, Jackrel ME, Rizo AN, Kendsersky NM, Buell CE, Sweeny EA, Mack KL, Chuang E, Torrente MP, Su YJ, Shorter J, Southworth DR. Ratchet-like polypeptide translocation mechanism of the AAA+ disaggregase Hsp104. Science. 2017 Jul 21;357(6348):273-279. doi: 10.1126/science.aan1052. Epub 2017 Jun 15. PMID: 28619716; PMCID: PMC5771629.

GOAL OF THIS STUDY

The main aim of this study is identifying conformational states of Hsp104 using Gromacs clustering and characterizing the identified states using machine learning techniques

METHODS



GROMACS RMSD Clustering

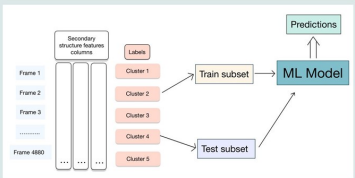
Simulating the behavior of the HSP104 complex with a substrate outputs a trajectory with many frames (4880 in our case).

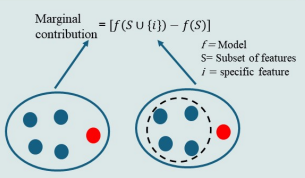
- Trajectory contains xyz coordinates of every single atom over a specified timescale
- To further identify most representative structures, we performed GROMOS clustering which is based on root-mean-square deviation values.

$$RMSD = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(r_i - r_i^{ref})^2}$$

	Frame 1	Frame 2	Frame 3	Frame 4	.....	Frame 4880
Frame 1	0	0.127	0.8	0.155	.....	0.6
Frame 2	0.127	0	0.12	0.15	.....	0.61
Frame 3	0.8	0.12	0	0.13	.....	0.63
Frame 4	0.155	0.15	0.13	0	.....	0.61
.....	.....	.....	.....	.....	.....	.....
Frame 4880	0.6	0.61	0.63	0.61	.....	0

Machine Learning Approaches





Algorithms

1. Linear Discriminant Analysis (LDA)
2. Decision Tree
3. Gaussian Naïve Bayes (NB)
4. Support Vector Machines (SVM
5. Random Forest (RF)
6. Gradient Boosting (XGB)

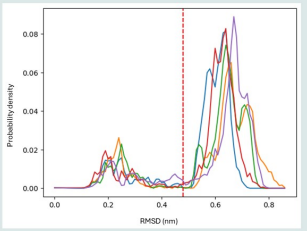
Features	Description
RSA	Relative solvent accessibility
Coulomb	Electrostatic interactions between each secondary structure and all the other secondary structures
VDW	van der Waal interactions between each secondary structure and all the other secondary structures
COM	Center of mass of each secondary structure to the center of mass of the protein
Phi and psi	Backbone torsion angles of each secondary structure

Hayatshahi HS, Ahuactzin E, Tao P, Wang S, Liu J. Probing Protein Allostery as a Residue-Specific Concept via Residue Response Maps. J Chem Inf Model. 2019 Nov 25;59(11):4691-4705. doi: 10.1021/acs.jcim.9b00447. Epub 2019 Oct 21. PMID: 31589429.

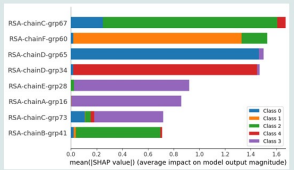
RESULTS

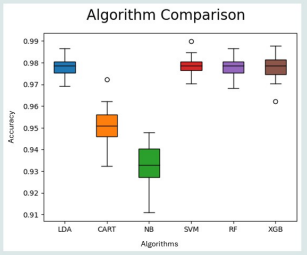
Clusters

- A range of cutoff values was tested for clustering calculations:
- Central structures obtained via clustering enabled us to identify central conformations around which most of the frames tend to cluster.
- These central frame structures are further used as the most contributing structures needed for the supervised learning

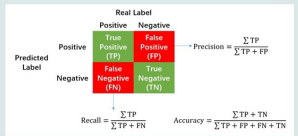


Machine learning





Metric	Value for XGB on RSA
Accuracy	97.8%
Precision	98.54%
Recall	97.2%
F1 Score	97.8%
Kappa	97.1%
MCC	97.2%
AUC	99.9%



J. Ma, Y. Ding, J. C. P. Cheng, Y. Tan, V. J. L. Gan and J. Zhang, "Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective," in IEEE Access, vol. 7, pp. 148059-148072, 2019, doi: 10.1109/ACCESS.2019.2946401.

Conclusion

- Machine learning analysis on secondary structures corresponding to different clusters showed that Gradient Boosting was highest performing algorithm in terms of accuracy.
- SHAP analysis can be used to identify the most significant features for each descriptor.

Future work

Subsequent steps following this study would be to identify key residues participating in key conformational states, as well as studying the effect of the number of clusters, trying different clustering methods and different cutoff values for clustering for machine learning analysis.

Acknowledgements