8-2022

State-Based Biological Communication

Nathan Clement
ncleme2@clemson.edu

STATE-BASED BIOLOGICAL COMMUNICATION

A Thesis
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science
Chemistry

---

by
Nathan John Clement
August 2022

Accepted by:
Dr. Brian Dominy, Committee Chair
Dr. Dvora Perahia
Dr. Steve Stuart
Dr. Emil Alexov
Dr. Hugo Sanabria

ABSTRACT

Allostery (*1*) is the process through which proteins self-regulate in response to various stimuli. Allosteric interactions occur between nonadjacent spatially distant residues (*1*), and they are exhibited through the correlated motions (*2*) and momenta of participating residues. The location of allosteric sites in proteins can be determined experimentally but computational methods to predict the location of allosteric sites are being developed as well (*2-4, 10*). Experimental and computational methodologies for locating allosteric sites can be used to design specific targeted drug delivery (*5-6, 19*), but these methods have not yet fully explained a mechanism for allosteric communications.

An allosteric pathway is a chain of residues that "communicate" by frequently colliding into one another. The frequency of collisions causes members of the chain to transfer kinetic energy amongst each other preferentially (*3-4*). Allosteric pathways begin and end at protein binding sites. An allosteric event occurs when an external molecule interacts with a binding site. An allosteric process is triggered by an allosteric event (*7-9*), and it is a consequence of a protein's free energy landscape changing in response to the stimuli (*12*). The protein begins to assume a new conformation due to the changes in its free energy landscape, and as its structure changes its functionality also changes as the system approaches a new equilibria. At equilibrium, a protein's conformational ensemble remains stable, and the residues participating in an allosteric pathway remain fairly constant (*3*).

The frequent collisions along allosteric pathways lead to quantifiable mathematical patterns in the physical states (position, momentum, internal energy) of

allosteric pathway residues over time. Non-allosteric pathway residues also collide with other residues but will not display a discernable pattern in physical state with other residues over time. Current computational methods for quantifying patterns in physical state to identify allosteric pathways utilize Percolation Theory (*3*), Isotropic Heat Diffusion (*4*), Direct Cross Correlation (*2*), and Information Theory (*11*, *13*). This work strives to enhance the Information Theoretic approach for locating allosteric sites and use this new perspective to develop a model to describe protein communication. The Information Theoretic approach has been chosen due to its ability to capture dynamic, nonlinear relationships, at relevant biological temperatures. Mutual Information (MI) quantifies the information that two variables share (*5*), and it will be used in this work to examine signaling relationships between a protein's residues at equilibrium. There is evidence to suggest that allosteric signals travel along energy pathways through transfers of kinetic energy between colliding residues (*3*-*4*). This work hypothesizes that a pattern of collisions forms during equilibria via repetitive kinetic energy transfers between residues along an allosteric pathway. If the energy transferred during this process functions as a repetitive biological 'signal' then there will be quantifiable patterns in physical state data that Mutual Information can be used to characterize analytically.

DEDICATION

This work is dedicated to my fiancé, Calista, and my soon to be born daughter, Karolina. Writing this thesis is the most difficult thing that I have ever done, and without either of you I would never have been able to finish it. It was your love and support that kept me going when I struggled. The both of you inspired me to push my limits and explore the boundaries of what is theoretically possible. I love you dearly.

PS: I can’t wait to meet you peanut!

ACKNOWLEDGMENTS

When I started college, my goal was to obtain an “easy” business degree that I could then leverage into a good job. I took my first chemistry class and Dr. Jason Robinson, my professor, encourage me to participate in STEM outreach at a local elementary school. The experience changed my perspective on science, and after it I decided to change my major to engineering.

I transferred to a different campus and took chemistry, math, and physics classes. I also began participating in undergraduate mathematics research to strengthen my resume. Part of my project required collaboration from a computational chemist and in looking for one I met Dr. Patricia Todebush. Dr. Todebush took me under her wing and eventually convinced me to switch my major to chemistry. Dr. Todebush was my biggest advocate during my undergraduate degree, and her mentorship led me to many great opportunities… most notably Clemson University!

I first attended Clemson as an NSF REU student the summer before I graduated with my bachelor's degree. My advisor was Dr. Brian Dominy, and while I worked in his lab, I realized that I wanted to pursue research professionally. After graduating, I returned to Clemson and designed the experiments that became this thesis. Dr. Dominy guided me through some of the most difficult years of my life. Without his mentorship I would not be the scientist that I am today, and this project would never have been finished.

Lastly, I want to acknowledge Clemson University for a generous allotment of compute time on the Palmetto cluster.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

CHAPTER ONE

AN INFORMATION-THEORETIC EXPLORATION OF PROTEIN SIGNALING

Introduction:

Allostery is the process induced in a protein when a chemical signaling event (ex. binding a small molecule) occurs at one location which causes information to flow to another part of the protein and a change the function, dynamics, or conformation of a distal location (*1,2 16*). There is a great deal of interest in determining a physical mechanism for allostery as well as developing methodologies for locating allosteric residues in proteins (*2-4,10,24*). Allostery has a broad range of possible applications from drug delivery (*3*) to self-assembly (*6*). Allostery is often coined "The Second Secret of Life" (*1*) and developing a more sophisticated model to predict and manipulate allosteric effects would significantly advance the field of biochemistry. For example, having the ability to predict allosteric effects combined with machine learning ligand synthesis would enable the creation of ligands that are designed to alter a particular function of a particular protein with high specificity and efficacy. In that case, the only parameter needed to design such a ligand would be the protein's conformation necessary for a specific function and a library of allosteric effects to build from. Many different methods for locating allosteric residues have been explored in the literature: Anisotropic Heat Diffusion (*4)*, Percolation Theory (OHM webserver) (*3*), Direct Cross-Correlation (*2,4,17*), and Information Theory (*2*).

Anisotropic Heat Diffusion:

Protein sequences are heterogeneous in nature which causes them to display a broad range of conformations and functions. Signal propagation through proteins appears to arise from closely packed secondary structures that collide frequently with one another and transfer kinetic energy (*4*). Kinetic energy transferred in this way is defined in this work as a signal.

As a protein changes conformation, each of its residues assume a new position relative to each other. In their new positions, the residues have a new likelihood of colliding into one another due to their altered proximity from other residues. If a conformational change alters the residue's positions drastically the new likelihood of collision can lead to changes in the signaling pathway. Signaling pathways have been shown to evolve over time (*3*) and upon ligand binding (*4,10,18*). New signaling pathways have also been documented to form after an old pathway was disrupted by a mutation (*8,15*).

Anisotropic heat diffusion is a simple and computationally inexpensive method for locating energetically anisotropic signaling pathways via the following steps (*4*).

1. The protein is minimized and equilibrated to 10 K to minimize atom movements.
2. A localized area of it is heated in a 300 K bath.
3. A short gas phase molecular dynamics simulation is performed to measure the heat diffusion.

By following the above steps, only the heated residue can move which causes it to collide with other residues and transfer heat to them in the form of kinetic energy. Several different residues were tested to evaluate the directionality and reproducibility of the energy pathway found (*4*).

1. The active site residue was heated, and energy flowed from active site to distal site.
2. The distal site was heated, and energy flowed from distal site to active site.
3. A surface residue that was not a part of the energy pathway was heated, and energy did not flow.

Direct cross correlation (explained below) was used to verify that the energetic pathway located by anisotropic heat diffusion was observable in a biologically relevant setting. Additionally, direct cross correlation was used to calculate the speed a signal propagates through a protein for comparison to the value obtained from anisotropic heat diffusion.

Percolation Theory:

Percolation Theory is another method that utilizes residue collisions to quantify signal propagation. In this method, the active site residues are perturbed many times and the frequency of kinetic energy transfers caused by collisions between residues that propagate through the system (*3*).

Groups of residues that collide frequently with one another are expected to be allosteric and exhibit correlated motions. Percolation Theory is strictly a static analytical

method which suggests that it could miss out on dynamic causes for allosteric interactions. For example, several snapshots were taken from a molecular dynamics simulation and then analyzed with OHM. It was reported that the allosteric pathways change "moderately" over the course of the simulation but not quantify by how much (*3*).

Direct Cross-Correlation:

Direct Cross-Correlation (DCC) is the first truly dynamic method of locating allosteric residues introduced thus far. DCC quantifies the strength of a linear relationship between two variables for each pairwise combination of residue time series data (*2,14,17*). The data for each pairwise combination of residues can be described in one of three ways: positively correlated, negatively correlated, or uncorrelated. DCC's output is a symmetric matrix which is visually interpreted with a heatmap whose axes each correspond to residue number. Each colored square on the heatmap describes the correlation coefficient for 1 pairwise combination of residue time series data.

DCC can be used to either to quantify signal propagation at equilibrium or to evaluate the accuracy of a new method being developed (*2,3,14,17*). In the former case, DCC with the addition of a time delay can be used to search for signals that are not instantaneously transmitted (*3*).

Information Theory:

Correlated motions and by extension signals propagating through a protein may be exhibited through more types of mathematical relationships than strictly linear (*11,13*). Mutual Information, an Information Theoretic Method, quantifies of the information that two variables share (*11,13*). In this case, information describes the predictability of one

variable's identity based off the identity of another (*11,13*). For example, if two different residue's trajectories are completely dependent upon one another, then one residues trajectory can be used to confidently compute the identity of the other and vice vera.

MD simulations paired with Information Theoretic analysis can be used to provide a holistic and dynamic analysis of protein signaling at equilibrium. Outside of the difference in quantification scope, the MI algorithm works in much the same way as DCC, but with one other major difference. MI can only be used on discrete data while DCC can be used on continuous data which makes DCC easier to implement with confidence.

MI is quantified for each pairwise combination of residue time series data over the course of a MD simulation. The MI output is a symmetric matrix which is visually interpreted with a heatmap whose axes each correspond to residue number. Each colored square on the heatmap describes the information for one pairwise combination of residue time series data. Theoretically, MI should provide more information about the system than DCC, but the comparison is easier to understand if the DCC is squared so that both heatmaps are on the same scale.

Methods:

Equilibrium Molecular Dynamics Simulations:

Equilibrium molecular dynamics simulations were performed with NAMD 2.14 on allosterically activated Ras grown in $Ca(C_2H_3O_2)_2$ (pdb 3K8Y) (*20*), allosterically activated Ras which was then deactivated by soaking it in $Mg(C_2H_3O_2)^2$ (pdb 3LBN) (*20*), inactive Ras grown in $CaCl_2$ (pdb 2RGE) (*20*), and Y32F mutant Ras (pdb 3K9N)

under the CHARMM36 (*26*) forcefield. Modified input files from the CHARMM-GUI Solution Builder (*23-24*) were used to prepare each system and are documented below. Each Ras system had structurally important Ca^{2+} and Mg^{2+} cations coordinated with it, and notably 3K8Y also coordinated an acetate anion $(C_2H_3O_2)^-$. Each X-Ray crystallographic structure was also coordinated with GNP, but a stable ligand could not be generated for simulation. 2RGE and 3LBN had residues (61-68 and 62-63 respectively) with missing atomic information that was modeled with GalaxyFill (*27*).

NAMD's conjugate gradient and line search algorithm (*28-29*) was used to minimize each system in a periodic rectangular water box with explicit TIP3 H_2O for 10,000 steps. CHARMM-GUI calculated the number of K^+ and Cl^- ions necessary to neutralize each system with a KCl concentration of 0.15 M, and it also selected each water box's dimensions to ensure that their edges were 10 Å away from the solvent (*23-24,30*). The dimensions for each water box are shown in the table below.

3K8Y (Å)	2RGE (Å)	3LBN (Å)	3K9N (Å)
X = Y = Z = {min = 0, max = 67}	X = Y = Z = {min = 0, max = 67}	X = Y = Z = {min = 0, max = 67}	X = Y = Z = {min = 0, max = 66}

Table 1.1: Water Box Dimensions

Each Ras structure was then slowly heated to 303.15° K over 610 ps. The systems were equilibrated in the NVT ensemble using Langevin dynamics to maintain constant temperature. CHARMM-GUI (*23-24*) selected heavy atoms to be placed under harmonic restraints that slowly relaxed over the course of a 154 ns equilibration. Initially, the harmonic restraints relaxed over the course of 50 ns (*31*), but the resulting systems were

unstable which led to clashes that made the systems blow up. The systems that equilibrated longer did not experience these difficulties.

The Van der Waals interactions for equilibration and production had a cutoff of 12 Å and CHARMM force switching began implementation at 10 Å. The cutoff of 12 Å was chosen because it is the cutoff distance that was used to develop the CHARMM36 forcefield (*30*). CHARMM force switching is NAMD's switching algorithm for the CHARMM36 forcefield. Full electrostatic interactions were implemented with Particle Mesh Ewald that utilized an interpolation order of 6 and grid spacing of 1.

During production, a constant pressure of 1.01325 bar (1 atm) was maintained by NAMD's Nosé-Hoover Langevin piston pressure control (*28*-*29*) and a constant temperature of 303.15° K was maintained by Langevin dynamics.

Signals have been documented to propagate through a protein at ~14 Å/ps (*4*) and have been weakly detected with time-delayed direct cross-correlation from a 500ns molecular dynamics simulation (*4*). If signal strength grows stronger with the length of the simulation it is quantified from, then during a 2.5 μs simulation previously observed signals become easier to detect and signals that were too weak to detect in a shorter simulation could also be observed.

Each Ras structure was simulated at equilibrium for ~ 2.5 μs and snapshots were recorded every 10 ps. A 10 ps snapshot time was selected signal propagation will still be observable and the number of frames recorded is minimized. Reducing the number of frames recorded also reduces the computational cost of analysis performed on the resulting trajectory.

Each protein's alpha carbon trajectory coordinates were superimposed on their PDB coordinates using Bio3D's implementation of the Kabsch algorithm prior to performing an RMSD to evaluate equilibration time (*32*). Alpha carbons were chosen to significantly reduce the computational cost of analysis and attempt to conserve information about each residue's position. Portions of each trajectory were also viewed to ensure that the RMSD values were representative of the movie.

Locating Allosteric Pathways with Information Theoretic Analysis:

Mutual Information (MI), an Information Theoretic method (*13*), was selected to detect and quantify signal propagation in a protein at equilibrium. MI is defined below as:

$$H(X) = -\sum_{i=1}^{n} P(x_i) log P(x_i)$$

Equation 1.1: Shannon Entropy

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Equation 1.2: Mutual Information

Where I(X;Y) is the information shared between discrete variables X and Y, H(X) and H(Y) are the Shannon entropies of each variable, H(X,Y) is the joint entropy of the two variables, and $P(x_i)$ is the probability of state n occurring (*13*). I(X;Y) can be a broad range of numbers which makes it difficult to compare with other methods. The MI of each pairwise combination of residue trajectory data was computed (using the R package InfoTheo (*33*) and resulted in a unique set of MI values for each residue that describe its informatic relationship with every other residue in the protein. Each residue's MI set contains a different range of information values which makes it difficult to intuit the

meaning of a given MI because there is no point of reference. These difficulties were overcome by changing each residue's set of MI values to a scale of zero to one using symmetric uncertainty (SU) (*34*) which is defined below:

$$U(X,Y) = 2 * \frac{I(X;Y)}{H(X) + H(Y)}$$

Equation 1.3: Symmetric Uncertainty

A SU of zero indicates that two variables are statistically independent, and a SU of one indicates that two variables are identical. Biochemically speaking, an SU of one between two different residues (X and Y) indicates that a statistical relationship exists where X's trajectory could be used to compute Y's perfectly and vice versa. In this case, signals would perfectly propagate between those two residues. On the other hand, a SU of zero between two different residues indicates that no information propagates between them.

If signals propagate via kinetic energy transfers caused by collisions between residues (*34*), then the SU of spatially close residue pairs should increase with their frequency of collision. Chains of spatially close residues pairs whose members frequently collide with one another would propagate signals from one point to another along the chain. By extension, the members of a signaling chain would show increased SU with one another and could be more formally described as allosteric pathways. These relationships can be located visually by plotting the pairwise SU values in a heatmap.

A SU heatmap is originally a triangular matrix which is made symmetric to help the viewer observe patterns in the data. Both axes of the SU heatmap are residue number, and the diagonal line through the center of it represents a residue's information with itself

which is maximal. Every other pixel on the heatmap denotes the SU between the two residues who intersect at that point. The SU of a pixel can be determined by comparing the pixel's color with the color key.

Heatmaps aide in visualization and analysis of matrix data but become increasingly difficult to read as the number of indices grows larger. R's native agglomerative hierarchical clustering algorithm (hclust) was used to identify clusters of residues with similar pairwise SU values (*35-36*). The hierarchical clustering method used the complete linkage method to identify clusters (*35-36*). The SU heatmaps could then be reordered based off the clustering results, and groups of allosteric residues that make up an allosteric pathway can easily be visually identified.

Discretization Methods:

Unfortunately, MI can only be used to quantify the information between discrete variables, and the output of a MD simulation is continuous. To use MI to analyze a MD trajectory it must be "binned" or in other words transformed from continuous to discrete. If the binning process is not done rigorously, information will be lost, and the MI will not represent the data accurately (*11*). A non-representative MI heatmap would yield incorrect allosteric residues after clustering and give no new insight into the system.

Three types of trajectory data for each residue were analyzed to compute each protein's pairwise information values: position, total potential energy, and RMSD. Position was separated from the equilibrium simulation as Kabsch algorithm superimposed coordinate data for each alpha carbon. Theoretically, binning results can be improved by minimizing the number of variables being binned which results in a

decrease in the space between datapoints (*37-39*). Total potential energy and RMSD were selected for this reason in an attempt to raise cluster quality. Total potential energy was also selected because signal transmission is hypothesized to result from kinetic energy transfer (*34*). If signal transmission occurs due to kinetic energy transfers between residues, then MI could be used to track signaling over a residue's potential energy trajectory as those transfers occur. Energy is also a state variable, and thus would resist translation, rotation, and solvent bombardment noise unlike positional data. The total potential energy for each residue was calculated with the VMD and NAMD extension NAMD Energy using the CHARMM36 forcefield. The solvent, ions, and heteroatoms were excluded from each of these calculations. RMSD for each residue was computed from Kabsch algorithm superimposed coordinate data for each alpha carbon.

Three discretization methods were chosen to evaluate which method and data type combination conserves information the most effectively during binning. The selected methods are Clustering Large Applications based on RANdomized Search (optimized k-Medoids) (*40-42*), Histogram binning, and Equal Width/Frequency Binning. These methods are unsupervised machine learning classification methods that cover a range of computational complexity. Each of these methods requires k number of microstates as an input. Each method and its heuristic for selecting k will be explained in more detail below.

Equal Width/Frequency Binning:

Equal width/frequency binning are simple, computationally inexpensive methods. The heuristic for both of these methods approximates the Freedman-Diaconis rule and is defined below (*33*).

$$k_{univariate} \approx (n)^{\frac{1}{3}}$$

Equation 1.4: Approximated Freedman-Diaconis Rule

Where n is the total number of observations. Unfortunately, this heuristic is only designed with univariate data in mind, and if $k_{univariate}$ is used to bin each of the three variables in position data the computed SU is artificially high. This is caused by how Shannon entropy creates sample spaces (S). If the univariate rule is applied the 3d data, there will be three equal k values for that data.

$$k_{v1} \approx (n)^{\frac{1}{3}}$$

$$k_{v2} \approx (n)^{\frac{1}{3}}$$

$$k_{v3} \approx (n)^{\frac{1}{3}}$$

Equation 1.5: Approximated Freedman-Diaconis Rule for 3D Data Set

$$S_{3d} = k_{v1}k_{v2}k_{v3} = n$$

Equation 1.6: Sample Space Generated from Equation 1.5

With the original rule, the number of possible states in the data equals the number of observations in the data. In this case, S_{3d} needs to equal $k_{univariate}$ to properly approximate the Freedman-Diaconis rule.

$$k_{v1} \approx (n)^{\frac{1}{9}}$$

$$k_{v2} \approx (n)^{\frac{1}{9}}$$

$$k_{v3} \approx (n)^{\frac{1}{9}}$$

Equation 1.7: Approximated Freedman-Diaconis Rule Modified for 3D Data Set

$$S_{3d} = k_{v1}k_{v2}k_{v3} = n^{\frac{1}{3}}$$

Equation 1.8: Sample Space Generated from Equation 1.7

Equal width binning separates a variable into k equal width bins over its range of data. In equal width binning each bin is populated based on the distribution of the data. Equal width binning is extremely susceptible to outliers that make the bin width larger and lowers binning accuracy (*33*,*38-39*).

Equal frequency binning creates k bins that are each populated by the same number of data points. In this method bin width has no constraints placed upon it. Equal frequency binning is also susceptible to outliers which cause the bin boundaries to be placed incorrectly (*33*,*38-39*).

Equal frequency/width binning both serve as the first step for more sophisticated supervised machine learning classification methods. Supervised machine learning methods can be useful for discretizing large datasets, but if these methods do not create accurate bins for trajectories, then there is no reason to explore that subset of supervised classification methods.

Histogram Binning:

Histogram binning is another computationally inexpensive binning method that has been used previously to discretize trajectory data for MI analysis (*2*). Histogram binning is simple to implement but can lead to the overestimation of MI (*11*). The R hist() function was used to bin trajectories and it utilized the built-in Freedman-Diaconis rule, "FD", to heuristically select k (*35-36,43*). The Freedman-Diaconis rule can scale to large numbers of observations, and it is defined below.

$$Bin\ width = 2 * \frac{IQR(x)}{\sqrt[3]{n}}$$

Equation 1.9: Freedman-Diaconis Rule

Where IQR(x) is the interquartile range of the variable and n is the number of observations. Histogram binning also applies the same heuristic adjustment to 3d position data as described above.

Clustering Large Applications based on RANdomized Search (CLARANS):

The CLARANS method is a modified version of k-medoids clustering that is designed to create unique clusters for large datasets. Unique clusters are comprised of datapoints that are a minimum distance away from their cluster's medoid and are not close to another cluster's boundary or medoid (*40*). CLARANS is resistant to outliers and less computationally expensive than other k-medoids methods but compared to the prior two discretization methods it is significantly more computationally expensive. One other benefit that CLARANS is that, unlike the previous two methods, each residue can have its own k value heuristically determined which allows for more mobile residues to display more states and less active residues to display fewer.

K values were heuristically determined with two methods: the Calinski-Harabasz Index from the R package (*44*) "fpc" and the Silhouette Coefficient from the R "cluster" package. The CLARANS method was also implemented from the R "cluster" package with the pamonce = 6 method (*42,45*). The k with the largest variance ratio criterion is the most appropriate when using the Calinski-Harabasz Index. The average silhouette value for all points in a dataset was also used to select k. Silhouette coefficients for each datapoint can range from negative one to one, with one indicating that the point is closer to the points in its cluster than the points in another cluster and negative one indicating that a data point is closer to the members of another cluster than the members of its own. In this method, the K value with the highest average silhouette coefficient is the optimal value.

For both heuristics, an optimal k value can be selected by evaluating a range of k's and selecting the one that best fits the data as described above. Ten k values were tested for the CLARANS method: 4, 5, 6, 7, 8, 20, 30, 40, 50, and 100. Originally, k = 2 and k =3 were also part of the tested range but could not be performed due to algorithmic memory restrictions. The range was expanded several time during development until the average silhouette values approached a steady minima, and the CH values peaked.

Validation Methods:

The accuracy of the information theoretic approach to allosteric residue identification was compared with two previously established methods: percolation theory via dokhlab's Ohm webserver (*3*) and Direct Cross Correlation (DCC) (*2, 17*).

Percolation theory can be used to locate allosteric residues and quantify the expected strength of their correlations in a heatmap (*3*). In this theory, a protein's active site is perturbed many times and the resulting potential collisions are propagated and counted to determine which residues will frequently collide with one another and thus communicate. The downside of percolation theory is that it is an entirely static method and any dynamic changes to an allosteric pathway are excluded (*3*), but it does not require a simulation to locate allosteric sites which drastically cuts down on analysis time. Percolation theory's accuracy can be very high or low depending on the system (*3*), which suggests that dynamics may play a variable role in allosteric pathway formation.

DCC quantifies the strength of the linear relationship (Pearson correlation coefficient) between two variables and can be used in a pairwise fashion for residue trajectories in the same way as MI (*2, 17*). The downside of DCC is that it misses any non-linear relationships between two variables, but it is less computationally expensive than MI and can natively analyze continuous data, so it does not lose any information to discretization.

DCC values range from negative one to one. A DCC of one indicates that two variables are perfectly correlated, a DCC of negative one indicates that two variables are perfectly anticorrelated, and a DCC of zero indicates the two variables share no correlation. The DCC matrices were squared so that overall linear correlation could be compared with SU on the same scale. This also made it possible to visualize if the information theoretic approach found any non-linear signals that DCC missed.

R’s native agglomerative hierarchical clustering algorithm and complete linkage method were also used to identify groups of allosteric residues for both percolation theory and DCC (*35*-*36*).

Results and Discussion:

Ras Alpha Carbon RMSD:

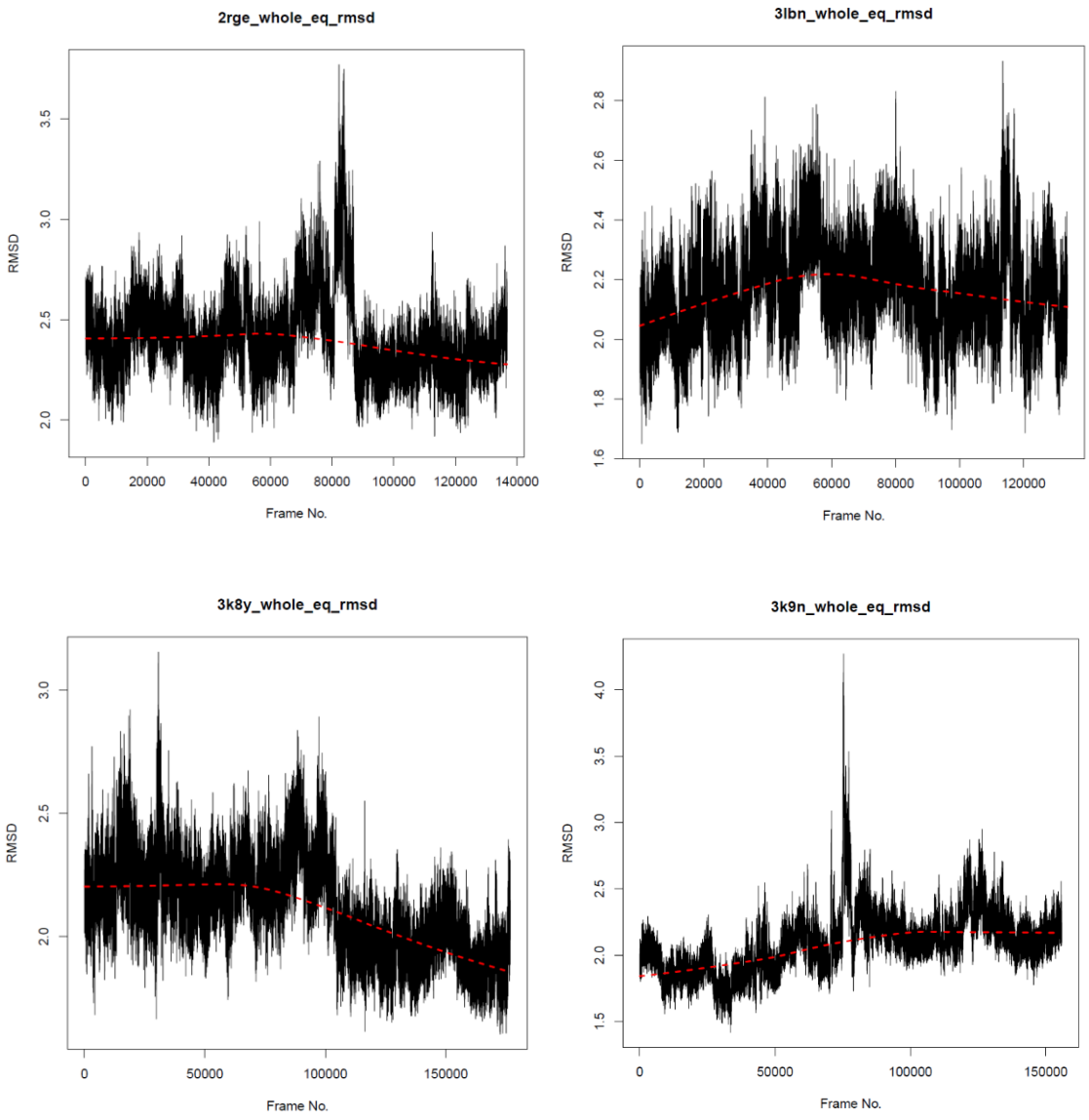

Figure 1.1: Alpha Carbon RMSD of Equilibrium Trajectory

3K8Y's production run lasted for a total of 2.47 us and it established equilibrium after 0.7 us. ~0.95 us later its RMSD dropped from an average of 2.3 to 2 and then stabilized again. The drop in RMSD indicates that the system underwent a conformational change at this point. The RMSD relaxed further, and the protein did not move erratically in the trajectory. It appears that the protein entered a new equilibrium which is supported by further analysis below.

2RGE's production run lasted for a total of 2.07 us and it established equilibrium after 0.7 us. ~0.8 us later its RMSD rose from an average of 2.5 to 3. Then it quickly dropped to 2.5 again. After briefly touching 2.5 the RMSD shot to 3.8. After quickly peaking at 3.8 the RMSD dropped to 2.2 and remained there. The rapid fluctuations in RMSD indicate that the system underwent a conformational change at this point

3LBN's production run lasted for a total of 2.44 us and it established equilibrium after 1.1 us. 3LBN took the longest to reach equilibrium but remained at a stable RMSD of 2.2 for the rest of the simulation. 3LBN likely maintained one conformation for the whole simulation after reaching equilibrium due to the minimal changes in its average RMSD.

3k9n's production run lasted for a total of 2.06 us and it established equilibrium after 0.5 us. ~0.75 us later its RMSD rose sharply from an average of 2 to a brief peak around 4.2. After the peak the RMSD quickly dropped back down to 2.3 and remained there for the rest of the simulation. The rapid spike in RMSD indicates that the system underwent a conformational change at this point. The protein appears to have experienced a brief clash then the system relaxed again.

Protein	EQ1 - before spike (μs)	EQ2 - after spike (μs)	Net Equilibrium (μs)
2RGE	0.700	0.469	1.369
3K8Y	1.000	0.768	1.768
3K9N	0.600	0.760	1.560
3LBN	N/A	N/A	1.338

Table 1.2: Equilibration Trajectory Lengths

Microstate k comparison:

The position, total potential energy, and RMSD equilibrium trajectories for each residue of each Ras structure were discretized with histogram binning, equal width/frequency binning, and CLARANS to explore which data type and discretization combination produced the most accurate identification of allosteric residues. All the equilibrium trajectories in Table 2. were discretized with equal width binning, equal frequency binning, and histogram binning because the low computational cost of these methods allowed for them to be applied to the whole equilibrium trajectory. The heuristics (see Methodology) for selecting a k value for these methods is also much less computationally expensive than those for CLARANS.

Protein PDB	Production	Equal Width/Freq	Histogram
2RGE	EQ 1	39 ± 0	136 ± 28
	EQ 2	38 ± 0	134 ± 27
	Net EQ	53 ± 0	177 ± 43
3K8Y	EQ 1	44 ± 0	152 ± 30
	EQ 2	40 ± 0	134 ± 28
	Net EQ	56 ± 0	186 ± 49
3K9N	EQ 1	39 ± 0	132 ± 30
	EQ 2	38 ± 0	129 ± 30
	Net EQ	53 ± 0	176 ± 43
3LBN	Net EQ	51 ± 0	164 ± 38

Table 1.3: K Value Average and Standard Deviation

Equal width/frequency binning is the simplest discretization method utilized and also applies the simplest heuristic to evaluate k, and every residue's k value being the same is a byproduct of that. Each k generated by the Freedman Diaconis Rule for histogram binning has the same number of observations, but the interquartile range for each residue is different which leads to a large deviation in possible k values.

Both heuristics for selecting k for CLARANS required that a range of possible k values be tested and evaluated. The tested range was, (4, 5, 6, 7, 8, 20, 30, 40, 50, 100). The range of k values was expanded several times during development until the average silhouette values approached a steady minima, the Calinski-Harabasz values peaked. For a vast majority of the residues, the testing range listed above contained a Calinski-

Harabasz value peak and selecting k was trivial. Although some of the residues had two peaks very close together (10^{-3}-10^{-4}) and the algorithm selected the k with the slightly higher value. The Calinski-Harabasz plot of these residues contained a valley instead of a peak which indicates that more k values should have been tested.

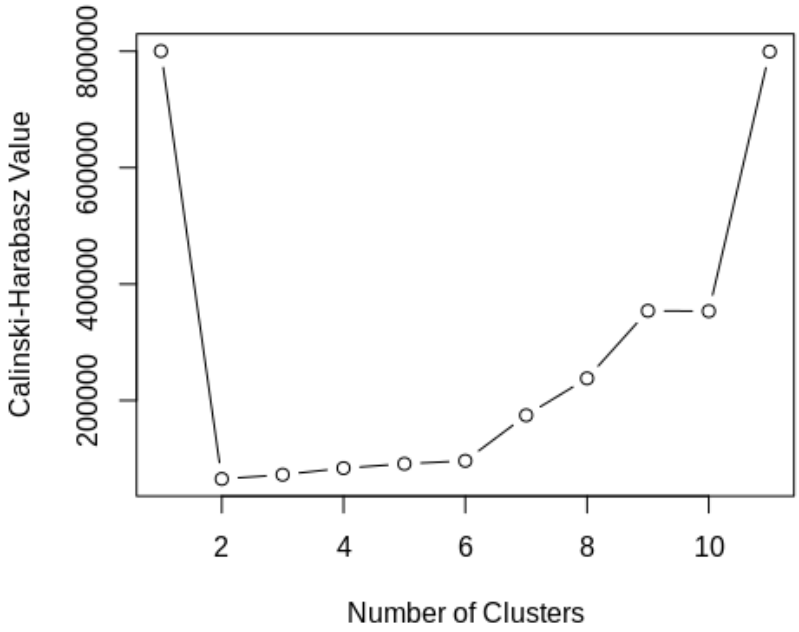

Figure 1.2: Calinski-Harabasz Plot

Silhouettes scores were also used to select a k for each residue. Silhouette scores can range from negative one to one. A silhouette score of one means an individual data point is closer to the data in its cluster than the data in another and therefore belongs perfectly inside the cluster. A silhouette score of zero represents that a data point is closer to the members of another cluster than its own and therefore does not belong in its cluster. A k with an average silhouette score greater than 0.8 is considered to represent a data set well, and because each residue's trajectory was clustered individually the average of the average k for each residue describes the quality for the clustering of the protein as

a whole. Unfortunately, none of the tested k values for any residue had an average silhouette score greater than 0.55 which indicates none of the selected k values represent the trajectories well. The average and standard deviation of k for each production and data type is listed in the figure below.

Protein PDB	2RGE		3K8Y		3K9N	
Production	EQ 1	EQ 2	EQ 1	EQ 2	EQ 1	EQ 2
CLARANS Position	4 ± 1	4 ± 1	5 ± 1	4 ± 1	4 ± 1	4 ± 1
CLARANS Energy	89 ± 24	88 ± 24	89 ± 24	92 ± 22	88 ± 24	89 ± 23
CLARANS RMSD	58 ± 41	62 ± 42	67 ± 41	70 ± 40	55 ± 45	54 ± 43

Table 1.4: K Value average and Standard Deviation for CLARANS

Overall, based off of the low average silhouette values and similar CH indices, none of the discretization methods led to highly representative bins for each dataset. Although, the discretized data sets created by the CLARANS method did yield SU heatmaps with similar features to previously established allosteric residue detection methods and they will be described below.

Data Type and Mutual Information:

The potential energy data yielded minimal information values for each protein, equilibrium trajectory, and discretization method. No signals were observed in any energy derived mutual information heatmap because all of the SU values ~0. It is possible that there was not enough significant energy transfer occurring over the duration of the simulation to lead to any observable signaling patterns via the net potential energy of each residue.

The RMSD data yielded low information values when compared to the positional data. Each Ras structure displayed minor characteristic areas of information, but only when clustered with CLARANS. Larger and more detailed versions of these information areas are observed in perturbation propagation correlation heatmaps, direct cross correlation heatmaps, and the CLARANS mutual information analysis for the positional trajectory data, but no groups of allosteric residues could be extracted from any of the data. When RMSD is clustered with histogram binning, equal width binning, or equal frequency binning no relationships were observed.

The positional data captured the most pairwise relationships and also yielded areas of correlation that matched percolation theory and DCC. Overall, the CLARANS method paired with positional data quantified the most information out of the tested discretization methods and data type combinations. Unfortunately, the CLARANS method does not scale well to large datasets which inhibited the length of simulations that it was used to analyze in this project. Interestingly, the positional data was the only data type where some signals were visible between discretization methods which suggests that the relationships it stores are a robust characteristic of the data set. The CLARANS method also appears to have quantified more areas of low correlation than the other discretization methods. More analysis is required to determine if the areas of higher correlation in the other methods are noise or statistically significant.

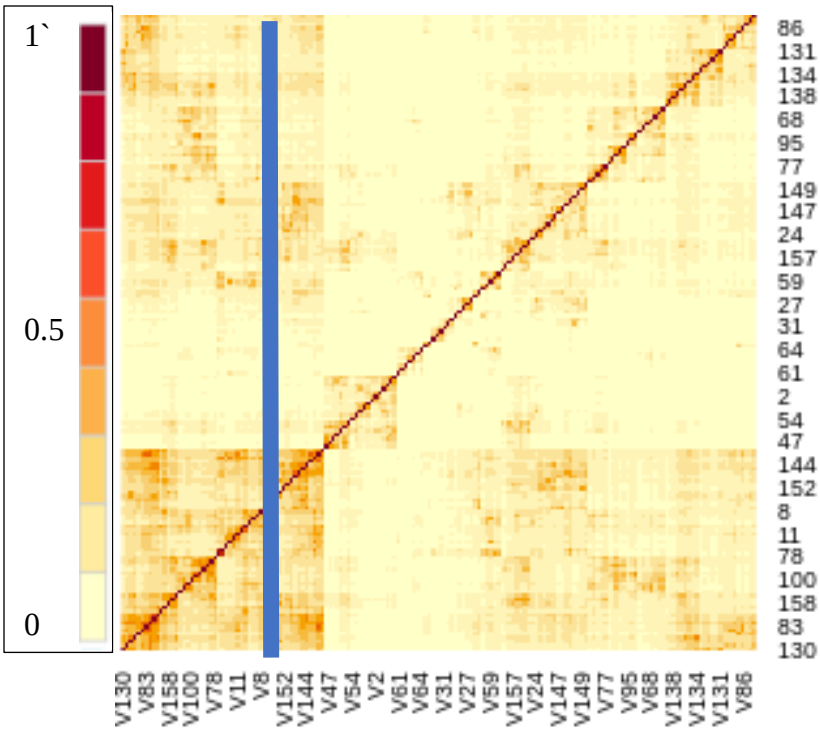

Figure 1.3: Clustered Percolation Theory

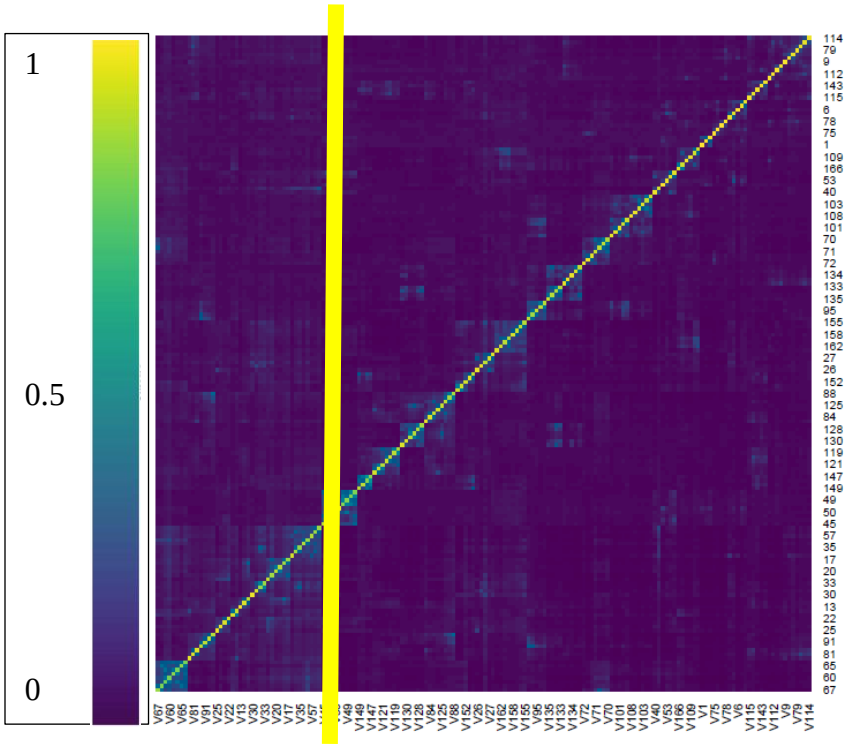

Figure 1.4: Clustered Symmetric Uncertainty

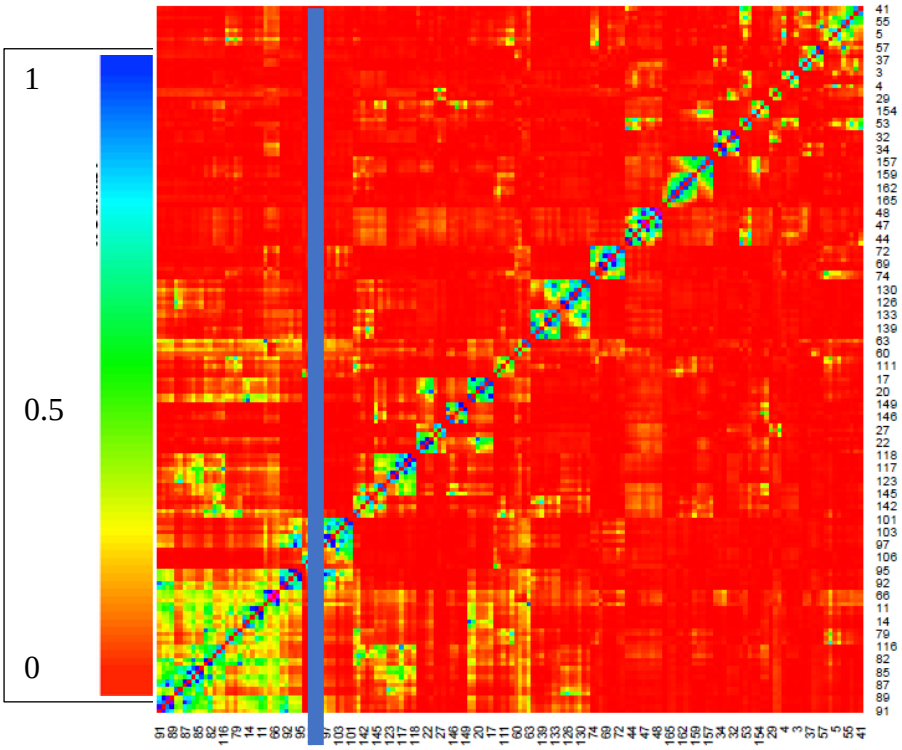

Figure 1.5: Clustered Direct Cross-Correlation Squared

Allosteric Site Identification:

Each clustered heatmap is read in the same way as an unclustered heatmap. The axes are still both residue number, but the indices are no longer in numerical order. The residues in the clustered heatmaps have been sorted by into clusters by the similarity of their pairwise analytical method values. A vertical line has been drawn on each heatmap to show where the most resolved clusters lie. The left side of the line is one cluster and the right side is the other. The SU heatmap, squared DCC heatmap, and OHM allosteric correlation prediction heatmap each had an optimal number of two agglomerative clusters present.

Each method with two optimal clusters had a smaller cluster that contained anywhere between 30-50 highly related residues and another larger cluster containing the rest of the residues which were less related on average. Each analysis type also had several small subsets of clusters that can be seen as small boxes along the diagonal running through the heatmaps that are present in both clusters. The small subset clusters were all very close together, and their dendogram heights were too small to warrant creating another cluster. These small subsets were generally comprised of groups of sequence adjacent residues that are members of the same secondary structures. α helices and β sheets are held together by strong hydrogen bonds which makes them rigid and causes the residues that make them up do display these kinds of correlated motions.

Each 3K8Y CLARANS position trajectory snapshot also had two well resolved agglomerative clusters with one cluster being smaller but higher in information and the other being larger and lower in information comparatively. Many smaller subclusters were also present in these heatmaps whose dendogram heights were too small to warrant creating another cluster to place them in.

Unfortunately, none of the clustered heatmap methods returned very accurate results. 3K8Y has nine experimentally located allosteric residues (residue numbers 97, 100, 101, 106, 107, 108, 109, 111, and 137) (*3*). The squared DCCM method identified the most residues that were experimentally verified as allosteric and identified the least number of falsely positive allosteric residues. Interestingly, when OHM accounted for the active site it identified all of the allosteric residues with perfect accuracy, but when it did not account for the active site, its accuracy dropped significantly with only 5

experimentally validated residues successfully grouped together and 48 falsely positive identifications.

	Actual	Squared DCCM	OHM Correlations	OHM Percolation	EQ1 XYZ	EQ2 XYZ
True Positive	9	6	5	9	2	0
False Positive	N/A	11	48	0	34	25
False Negative	N/A	3	4	0	7	9
Net Positive	N/A	17	53	9	36	25

Table 1.5: Summary of Allosteric Residue Identifications

The information theoretic approach was the least accurate method tested. Its most accurate results (EQ1) only successfully identified 2 experimentally validated allosteric residues and it grouped them with 34 falsely positive identifications. 46 residues switch agglomerative clusters between EQ1 and EQ2 which suggests that 3K8Y was either in the process of entering a new equilibrium or in the process of restoring its old equilibrium after the conformational event that occurred 1 μs into the production trajectory.

Conclusions:

It is possible that rare and/or uncommon signals become easier to observe as a simulation lengthens. There were no agglomerative hierarchical clusters present in the equal width, equal frequency, or histogram binned positional trajectory heatmaps for 3K8Y EQ1 and EQ2. Although, when the entire production was analyzed all three methods developed two agglomerative clusters in their SU heatmaps. Much like in the other clustered heatmap results there was a smaller cluster of residues that shared more

information and a larger cluster of residues that shared less information. These heatmaps are less accurate for identifying allosteric residues than the clustered CLARANS method, but the increase in signal strength supports exploring ways to expand the scale of the data discretized by that method. The CLARANS algorithm will not be able to cluster the entire production because it has a hard cap at ~65,000 datapoints, but cap currently on this project is ~35,000 datapoints. Approximately doubling the number of observations may increase signal strength significantly like in the other discretization methods.

The average silhouette value for the k's selected in the CLARANS method were ~0.5 for each residue. The optimal k that was selected for CLARANS fell between the equal width/frequency Freedman-Diaconis bin value and the histogram Freedman-Diaconis bin value. It is possible that too small of a k range was tested which is supported by the low average silhouette values and the strange spike that occurred at the end of the Calinski-Harabasz index plots. A significantly larger range of clusters needs to be tested before moving away from CLARANS as a discretization method to evaluate if there is a more optimal k that was missed. A more representative k would yield higher quality clusters which could increase allosteric residue identification accuracy by preserving more information from the trajectory during clustering.

It is also possible that a more specialized unsupervised discretization method may yield more representative clusters and by extension SU. This information theoretic approach misidentified most of the experimentally validated allosteric residues which could be due to information that was lost during discretization. In this situation, the SU

gained from a discretization method that is optimized for time series data may be worth the increased computational cost.

Improving the quality and length of MD simulations is likely the most important improvement that can be implemented in this project. CHARMM-GUI was a very useful tool to learn how to perform advanced MD simulations, but many of the input files did not work as intended and learning how to repair them cost a significant amount of time. There are likely still inconsistencies and bugs in the input files that were used to run these simulations which could be why three out of the four systems experienced a conformational event approximately half-way through their production runs. The next iteration of this project will have its input files built entirely by hand so that the quality of the input files is not in question. Analyzing simulations that maintain a steady equilibrium for several microseconds would also add a great deal of flexibility to the analysis and provide a means to evaluate the impact of simulation length on signal strength.

It is also possible that discrete MI is not useful for identifying a protein's allosteric sites. The information lost by binning may be too detrimental even with the most advanced algorithms. A possible avenue to work around binning information loss is to instead estimate continuous MI to quantify the information of pairwise trajectory data more directly. Estimating continuous MI is very computationally expensive, but if its results are more representative of the system then it is likely a more fruitful direction for future study than discrete MI and its increasingly complex discretization methods. The power and accessibility of GPU resources on clusters is growing rapidly, and in the last

two years many developers have built GPU parallelization backends that can be utilized to increase the speed of algorithms.

APPENDICES

Appendix A

Supplementary Plots

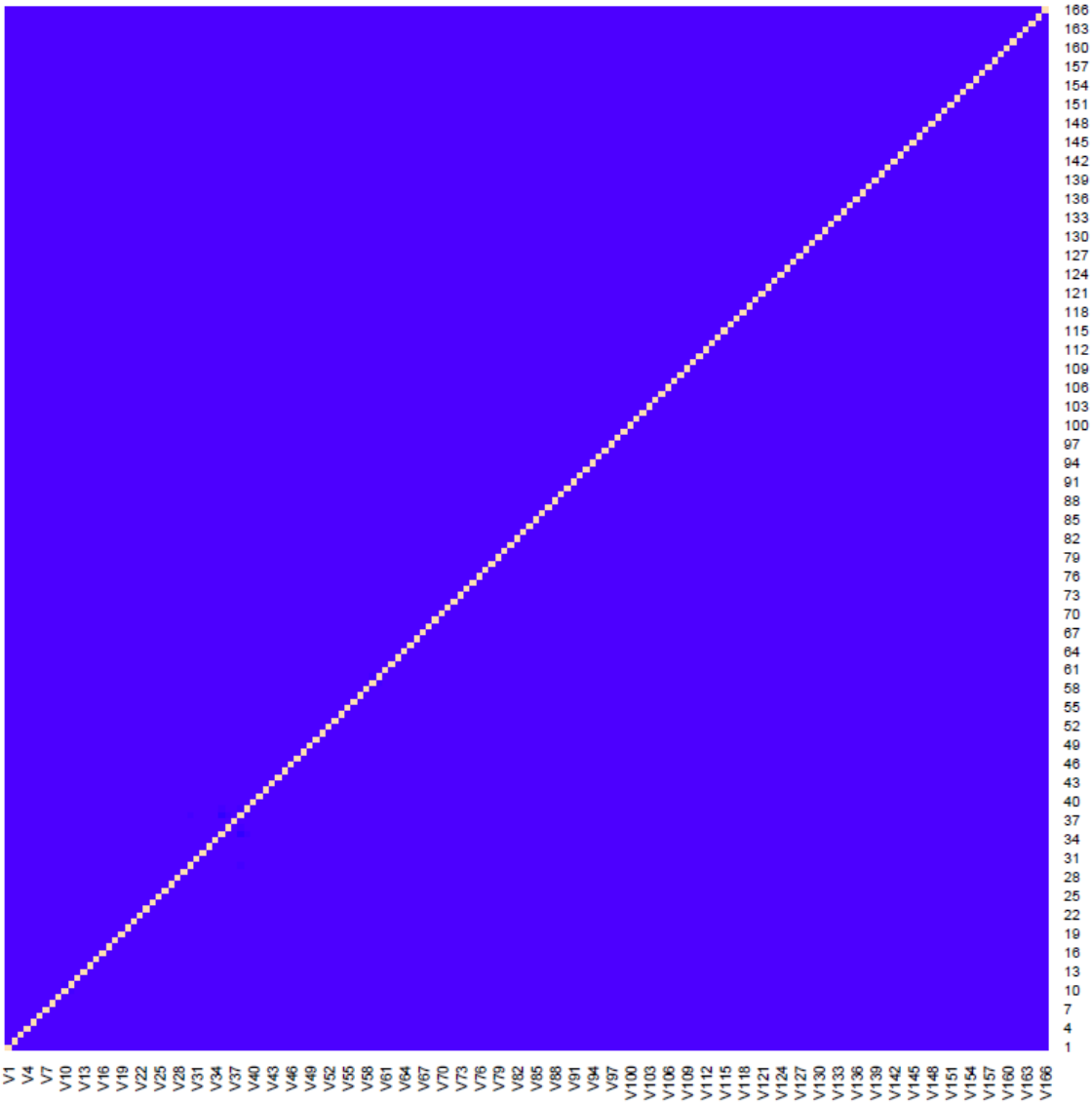

Figure A-1: 3K8Y SU Generated from Potential Energy

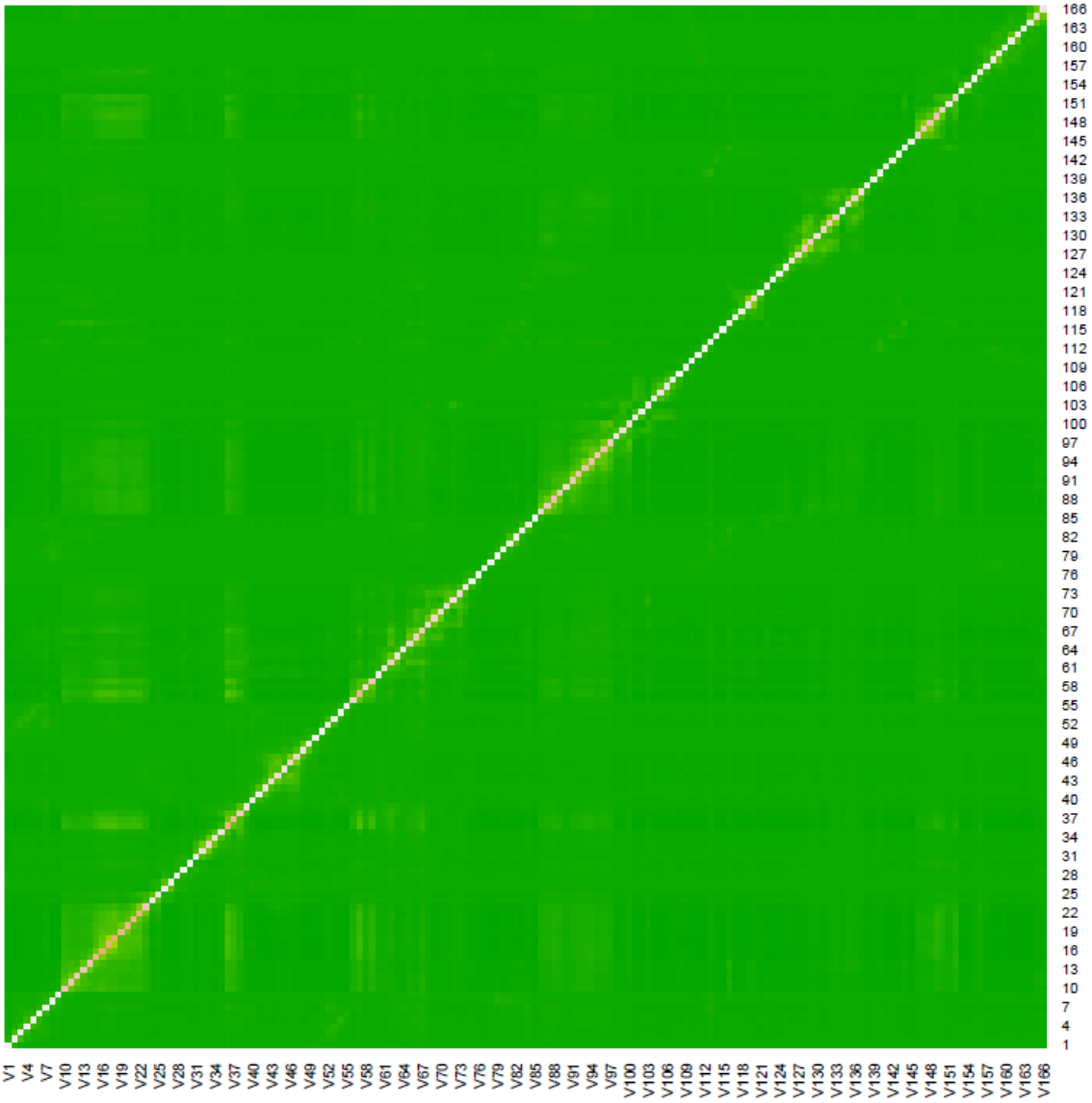

Figure A-2: 3K8Y SU Generated from RMSD

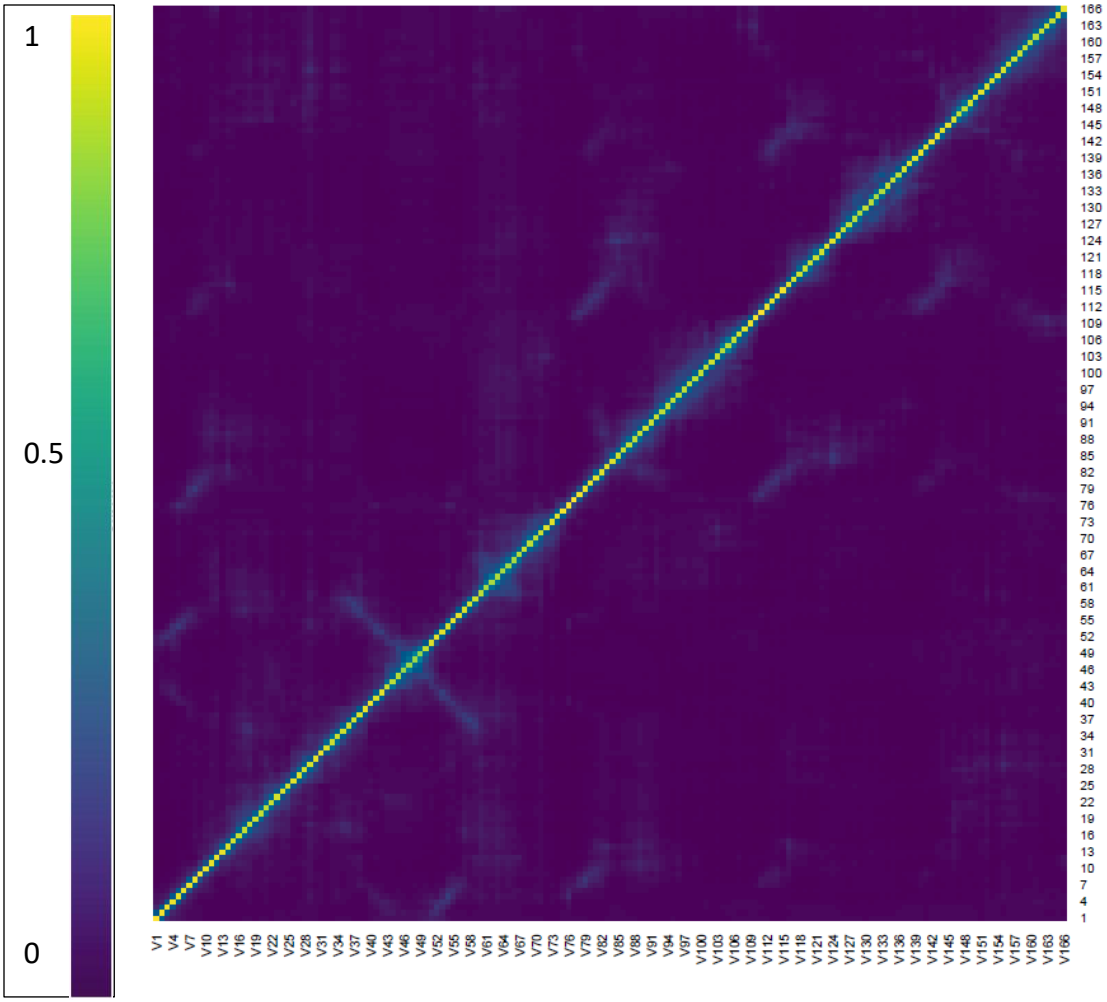

Figure A-3: 3K8Y SU Generated from Position

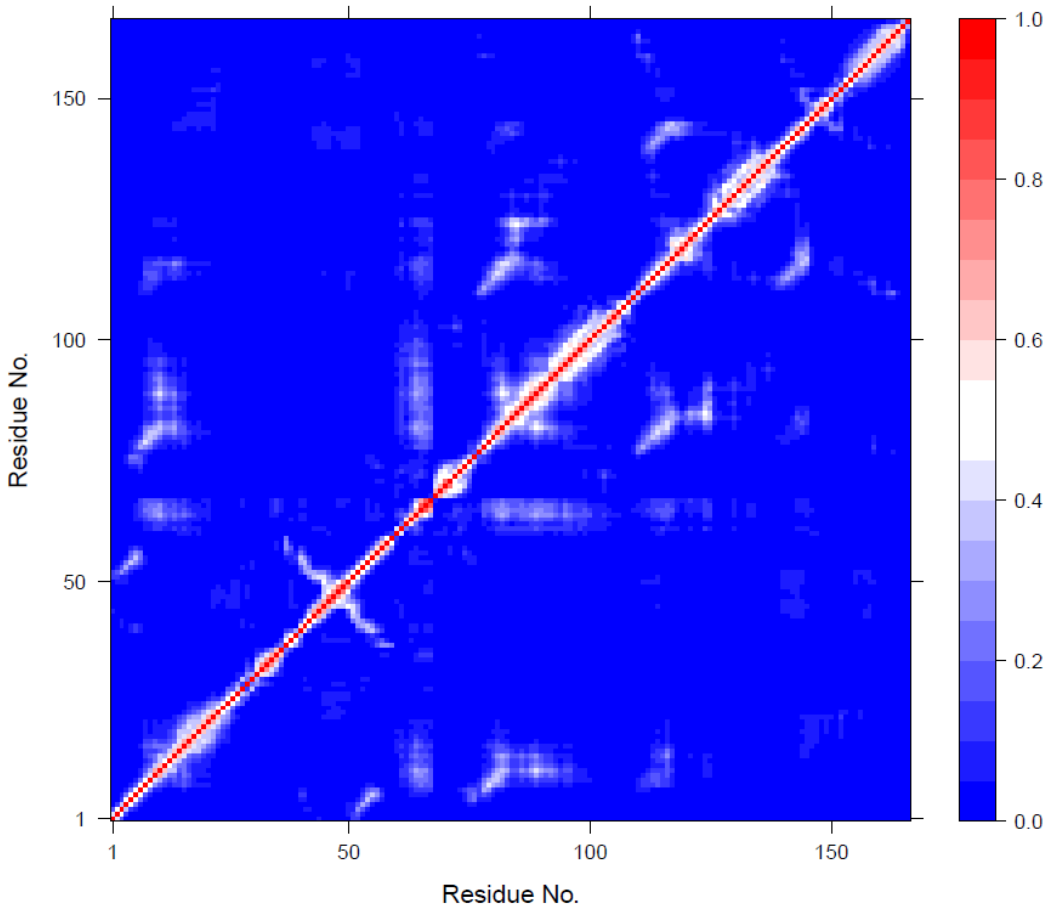

Figure A-4: 3K8Y Direct Cross-Correlation Squared

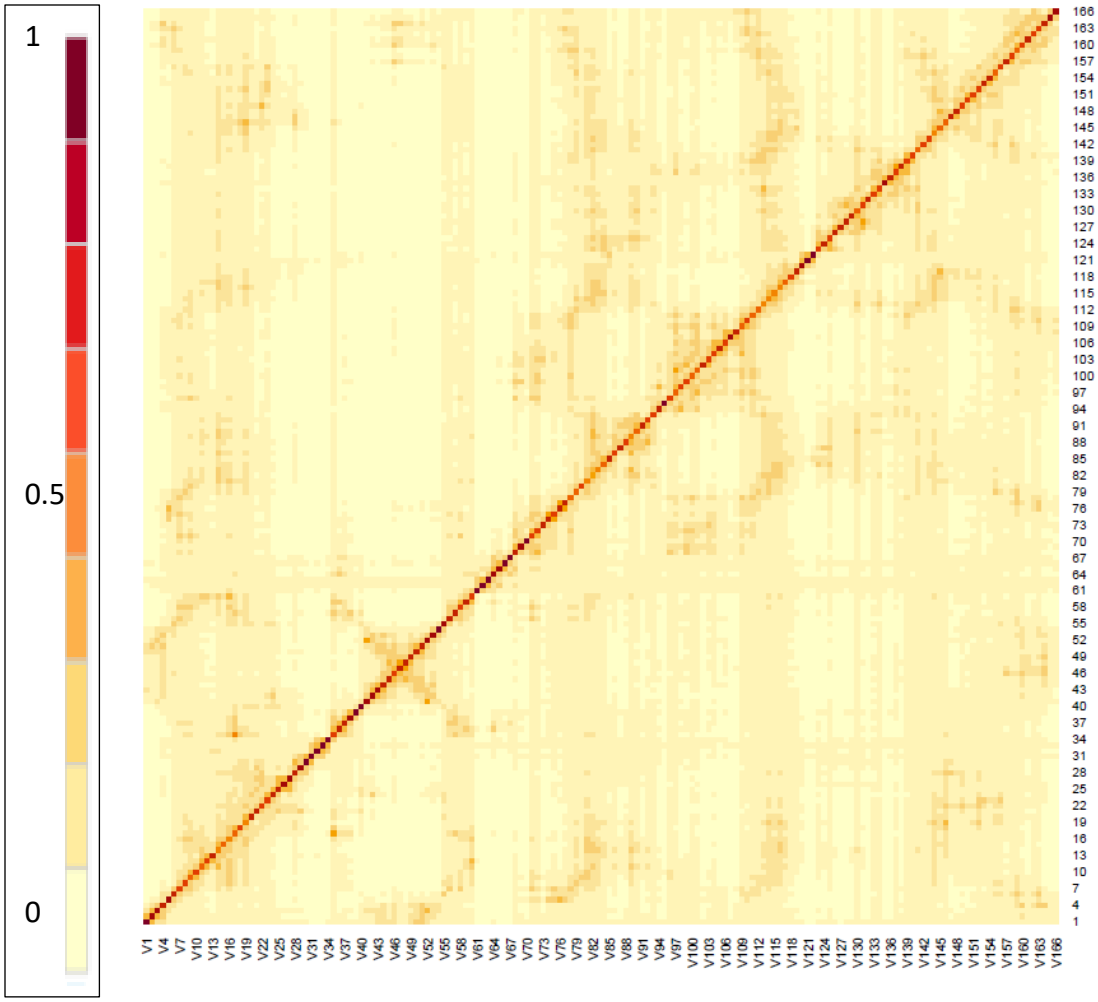

Figure A-5: 3K8Y Correlation Predictions Generated with OHM

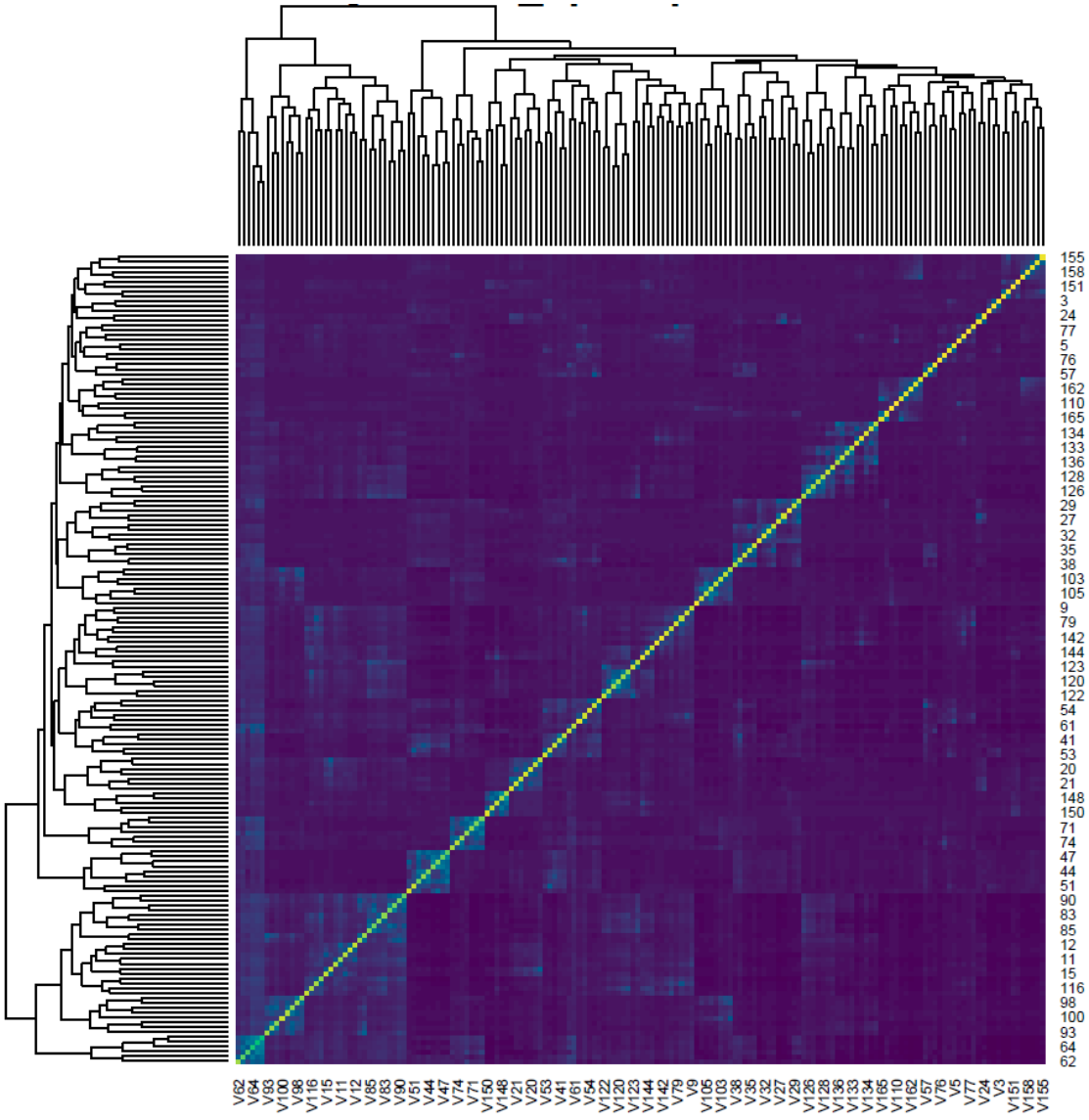

Figure A-6: 3K8Y EQ1 SU Generated with CLARANS Microstates

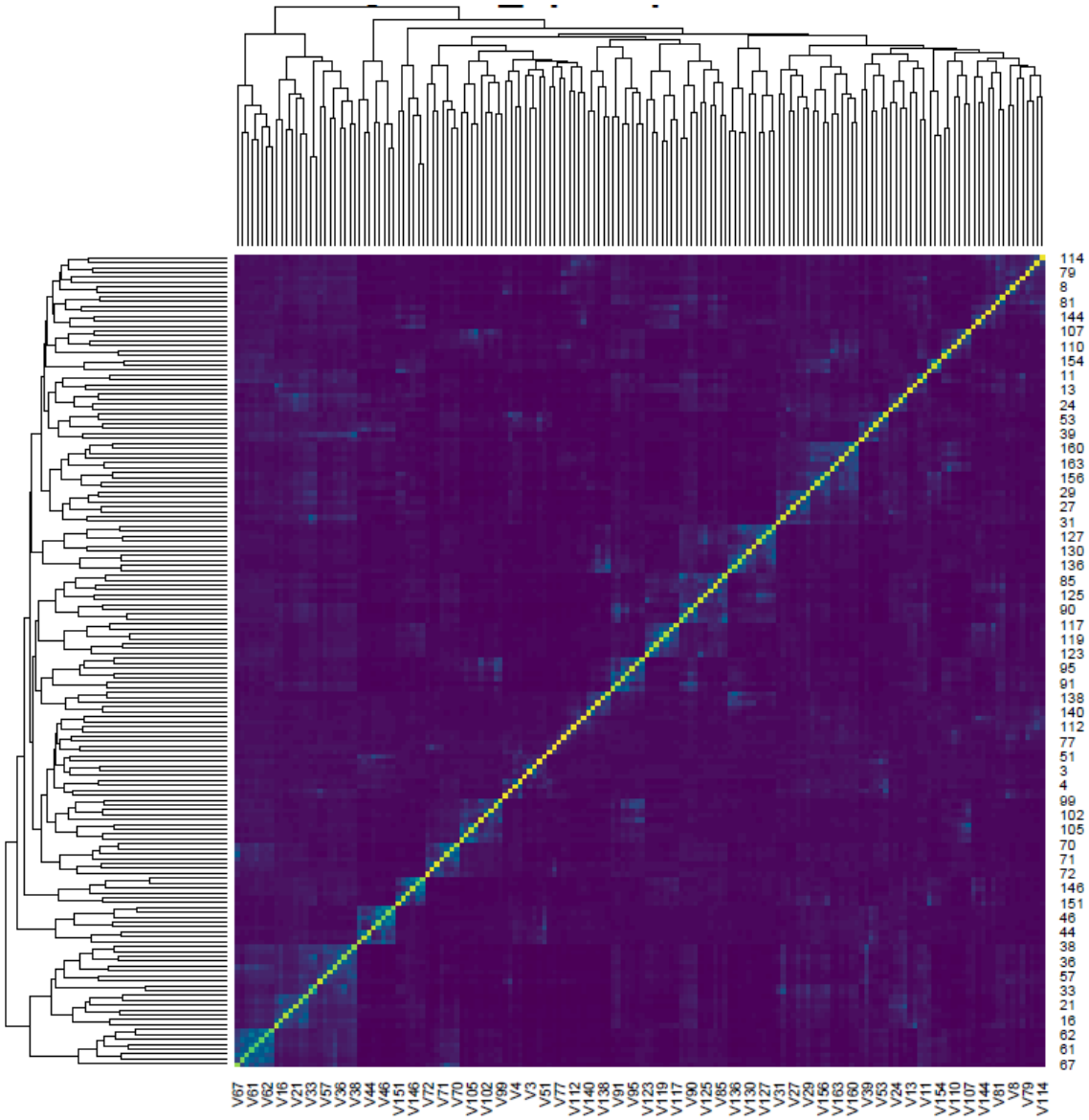

Figure A-7: 3K8Y EQ2 SU Generated with CLARANS Microstates

REFERENCES:

(1) Liu, J.; Nussinov, R. Allostery: An Overview of Its History, Concepts, Methods, and Applications. *PLoS Comput. Biol.* **2016**, *12* (6), 3–7. https://doi.org/10.1371/journal.pcbi.1004966.

(2) McClendon, C. L.; Friedland, G.; Mobley, D. L.; Amirkhani, H.; Jacobson, M. P. Quantifying Correlations between Allosteric Sites in Thermodynamic Ensembles. *J. Chem. Theory Comput.* **2009**, *5* (9), 2486–2502. https://doi.org/10.1021/ct9001812.

(3) Wang, J.; Jain, A.; McDonald, L. R.; Gambogi, C.; Lee, A. L.; Dokholyan, N. V. Mapping Allosteric Communications within Individual Proteins. *Nat. Commun.* **2020**, *11* (1), 1–13. https://doi.org/10.1038/s41467-020-17618-2.

(4) Ota, N.; Agard, D. A. Intramolecular Signaling Pathways Revealed by Modeling Anisotropic Thermal Diffusion. *Journal of Molecular Biology*. 2005, pp 345–354. https://doi.org/10.1016/j.jmb.2005.05.043.

(5) Lindsley, C. W. 2013 Philip S. Portoghese Medicinal Chemistry Lectureship: Drug Discovery Targeting Allosteric Sites. *J. Med. Chem.* **2014**, *57* (18), 7485–7498. https://doi.org/10.1021/jm5011786.

(6) Chen, J.; Vishweshwaraiah, Y. L.; Dokholyan, N. V. Design and Engineering of Allosteric Communications in Proteins. *Curr. Opin. Struct. Biol.* **2022**, *73*, 102334. https://doi.org/10.1016/j.sbi.2022.102334.

(7) Leander, M.; Yuan, Y.; Meger, A.; Cui, Q.; Raman, S. Functional Plasticity and Evolutionary Adaptation of Allosteric Regulation. *Proc. Natl. Acad. Sci. U. S. A.* **2020**, *117* (41), 25445–25454. https://doi.org/10.1073/pnas.2002613117.

(8) Köhler, C.; Carlström, G.; Gunnarsson, A.; Weininger, U.; Tångefjord, S.; Ullah, V.; Lepistö, M.; Karlsson, U.; Papavoine, T.; Edman, K.; Akke, M. Dynamic Allosteric Communication Pathway Directing Differential Activation of the Glucocorticoid Receptor. *Sci. Adv.* **2020**, *6* (29). https://doi.org/10.1126/sciadv.abb5277.

(9) Buhrman, G.; Holzapfel, G.; Fetics, S.; Mattos, C. Allosteric Modulation of Ras Positions Q61 for a Direct Role in Catalysis. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (11), 4931–4936. https://doi.org/10.1073/pnas.0912226107.

(10) Amor, B. R. C.; Schaub, M. T.; Yaliraki, S. N.; Barahona, M. Prediction of Allosteric Sites and Mediating Interactions through Bond-to-Bond Propensities. *Nat. Commun.* **2016**, *7*. https://doi.org/10.1038/ncomms12477.

(11) Steuer, R.; Kurths, J.; Daub, C. O.; Weise, J.; Selbig, J. The Mutual Information: Detecting and Evaluating Dependencies between Variables. Bioinformatics 2002, 18 (SUPPL. 2), 231–240. https://doi.org/10.1093/bioinformatics/18.suppl_2.S231.

(12) Jain, A.; Hegger, R.; Stock, G. Hidden Complexity of Protein Free-Energy Landscapes Revealed by Principal Component Analysis by Parts. *J. Phys. Chem. Lett.* **2010**, *1* (19), 2769–2773. https://doi.org/10.1021/jz101069e.

(13) Shannon, C. E. A Mathematical Theory of Communication. *Math. Theory Commun.* **1949**.

(14) Harding, C. J.; Cadby, I. T.; Moynihan, P. J.; Lovering, A. L. A Rotary Mechanism for Allostery in Bacterial Hybrid Malic Enzymes. *Nat. Commun.* **2021**, *12* (1). https://doi.org/10.1038/s41467-021-21528-2.

(15) Daura, X. Advances in the Computational Identification of Allosteric Sites and Pathways in Proteins. *Adv. Exp. Med. Biol. USER* **2019**, *1163*, 141–169.

(16) Tsai, C. J.; Nussinov, R. A Unified View of "How Allostery Works." *PLoS Comput. Biol.* **2014**, *10* (2). https://doi.org/10.1371/journal.pcbi.1003394.

(17) Fenwick, R. B.; Orellana, L.; Esteban-Martín, S.; Orozco, M.; Salvatella, X. Correlated Motions Are a Fundamental Property of β-Sheets. *Nat. Commun.* **2014**, *5* (May), 1–9. https://doi.org/10.1038/ncomms5070.

(18) Huang, Q.; Song, P.; Chen, Y.; Liu, Z.; Lai, L. Allosteric Type and Pathways Are Governed by the Forces of Protein-Ligand Binding. *J. Phys. Chem. Lett.* **2021**, *12* (22), 5404–5412. https://doi.org/10.1021/acs.jpclett.1c01253.

(19) Hub, J. S.; De Groot, B. L. Detection of Functional Modes in Protein Dynamics. *PLoS Comput. Biol.* **2009**, *5* (8). https://doi.org/10.1371/journal.pcbi.1000480.

(20) Altunkaya, A.; Bi, C.; Bradley, A. R.; Rose, P. W.; Prli, A.; Christie, H.; Costanzo, L. Di; Duarte, J. M.; Dutta, S.; Feng, Z.; Green, R. K.; Goodsell, D. S.; Hudson, B.; Kalro, T.; Lowe, R.; Peisach, E.; Randle, C.; Rose, A. S.; Shao, C.; Tao, Y.; Valasatava, Y.; Voigt, M.; Westbrook, J. D.; Woo, J.; Yang, H.; Young, J. Y.; Zardecki, C.; Berman, H. M.; Burley, S. K. The RCSB Protein Data Bank: Integrative View of Protein, Gene, and 3D Structural Information. *Nucleic Acids Res.* **2017**, *45* (271–281), 271–281. https://doi.org/10.1093/nar/gkw1000.

(21) Buhrman, G.; Wink, G.; Mattos, C. Transformation Efficiency of RasQ61 Mutants Linked to Structural Features of the Switch Regions in the Presence of Raf. *Structure* **2007**, *15* (12), 1618–1629. https://doi.org/10.1016/j.str.2007.10.011.

(22) Chang, F.; Steelman, L. S.; Lee, J. T.; Shelton, J. G.; Navolanic, P. M.; Blalock, W. L.; Franklin, R. A.; McCubrey, J. A. Signal Transduction Mediated by the Ras/Raf/MEK/ERK Pathway from Cytokine Receptors to Transcription Factors: Potential Targeting for Therapeutic Intervention. *Leukemia* **2003**, *17* (7), 1263–1293. https://doi.org/10.1038/sj.leu.2402945.

(23) Jo, S.; Kim, T.; Iyer, V. G.; Im, W. CHARMM-GUI: A Web-Based Graphical User Interface for CHARMM. *J. Comput. Chem.* **2008**, 174–182. https://doi.org/10.1002/jcc.

(24) Lee, J.; Cheng, X.; Swails, J. M.; Yeom, M. S.; Eastman, P. K.; Lemkul, J. A.; Wei, S.; Buckner, J.; Jeong, J. C.; Qi, Y.; Jo, S.; Pande, V. S.; Case, D. A.; Brooks, C. L.; MacKerell, A. D.; Klauda, J. B.; Im, W. CHARMM-GUI Input Generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM Simulations Using the CHARMM36 Additive Force Field. *J. Chem. Theory Comput.* **2016**, *12* (1), 405–413. https://doi.org/10.1021/acs.jctc.5b00935.

(26) Brooks, B. R.; Brooks, C. L.; Mackerell, A. D.; Nilsson, L.; Petrella, R. J.; Roux, B.; Won, Y.; Archontis, G.; Bartels, C.; Boresch, S.; Caflisch, A.; Caves, L.; Cui, Q.; Dinner, A. R.; Feig, M.; Fischer, S.; Gao, J.; Hodoscek, M.; Im, W.; Kuczera, K.; Lazaridis, T.; Ma, J.; Ovchinnikov, V.; Paci, E.; Pastor, R. W.; Post, C. B.; Pu, J. Z.; Schaefer, M.; Tidor, B.; Venable, R. M.; Woodcock, H. L.; Wu, X.; Yang, W.; York, D. M.; Karplus, M. CHARMM: The Biomolecular Simulation Program. *J. Comput. Chem.* **2009**, *30* (10), 1545–1614. https://doi.org/10.1002/jcc.21287.

(27) Coutsias EA, Seok C, Jacobson MP, Dill KA. A kinematic view of loop closure. J Comput Chem. 2004 Mar;25(4):510-28. doi: 10.1002/jcc.10416. PMID: 14735570.

(28) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W.; McGreevy, R.; Melo, M. C. R.; Radak, B. K.; Skeel, R. D.; Singharoy, A.; Wang, Y.; Roux, B.; Aksimentiev, A.; Luthey-Schulten, Z.; Kalé, L. V.; Schulten, K.; Chipot, C.; Tajkhorshid, E. Scalable Molecular Dynamics on CPU and GPU Architectures with NAMD. *J. Chem. Phys.* **2020**, *153* (4), 1–33. https://doi.org/10.1063/5.0014475.

(29) Humphrey, W.; Dalke, A.; Schulten, K. VMD: Visual Molecular Dynamics. *J. Mol. Graph.* **1996**, *14* (October 1996), 33–38.

(30) Practical considerations for Molecular Dynamics. (n.d.). Retrieved July 12, 2022, from https://computecanada.github.io/molmodsim-md-theory-lesson-novice/LICENSE.html

(31) Menzer, W. M.; Xie, B.; Minh, D. D. L. On Restraints in End-Point Protein-Ligand Binding Free Energy Calculations. *J. Comput. Chem.* **2020**, 573–586. https://doi.org/10.1002/jcc.26119.On.

(32) Bio3D: An R package for the comparative analysis of protein structures. Grant, Rodrigues, ElSawy, McCammon, Caves, (2006) Bioinformatics 22, 2695-2696

(33) Patrick E. Meyer (2022). infotheo: Information-Theoretic Measures. R package version 1.2.0.1. https://CRAN.R-project.org/package=infotheo

(34) Han, J.; Kamber, M.; Pei, J. *Data Mining: Data Mining Concepts and Techniques*; 2014. https://doi.org/10.1109/ICMIRA.2013.45.

(35) RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL http://www.rstudio.com/.

(36) R Core Team (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

(37) Koch, I. Analysis of Multivariate and High-Dimensional Data. *Anal. Multivar. High-Dimensional Data* **2012**, 1–504. https://doi.org/10.1017/CBO9781139025805.

(38) Miller, H. J.; Han, J. *Geographic Data Mining and Knowledge Discovery*; Miller, H. J., Han, J., Eds.; Taylor and Francis, 2001.

(39) Ng, Raymond T, H. J. Efficient and Effective Clustering Data Mining Methods for Spatial. *Proc. 20th Int. Conf. on Very Large Data Bases,* **2002**, 144–155.

(40) Freedman, D.; Diaconis, P. On the Histogram as a Density Estimator:L2 Theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **1981**, *57* (4), 453–476. https://doi.org/10.1007/BF01025868.

(42) Learning, M. Encyclopedia of Machine Learning. *Encycl. Mach. Learn.* **2010**. https://doi.org/10.1007/978-0-387-30164-8.

(43) Schubert, E.; Rousseeuw, P. J. Faster K-Medoids Clustering: Improving the PAM, CLARA, and CLARANS Algorithms. *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)* **2019**, *11807 LNCS*, 171–187. https://doi.org/10.1007/978-3-030-32047-8_16.

(44) Christian Hennig (2020). fpc: Flexible Procedures for Clustering. R package version 2.2-9. https://CRAN.R-project.org/package=fpc

(45) Rousseeuw, P. J. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* **1987**, *20* (C), 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

(46) Microsoft Corporation and Steve Weston (2022). doParallel: Foreach Parallel Adaptor for the 'parallel' Package. R package version 1.0.17. https://CRAN.R-project.org/package=doParallel

(47) Microsoft and Steve Weston (2022). foreach: Provides Foreach Looping Construct. R package version 1.5.2. https://CRAN.R-project.org/package=foreach

(48) Douglas Bates, Martin Maechler and Mikael Jagan (2022). Matrix: Sparse and Dense Matrix Classes and Methods. R package version 1.4-1. https://CRAN.R-project.org/package=Matrix