Summer 6-15-2021

Molecular Dynamics Simulations Provide Insight into Stability of Hyperthermophilic Endoglucanases

Logan E. Sheffield
Stephen F Austin State University, loganesheffield@gmail.com

Molecular Dynamics Simulations Provide Insight into Stability of Hyperthermophilic Endoglucanases

Creative Commons License

MOLECULAR DYNAMICS SIMULATIONS PROVIDE INSIGHT INTO STABILITY OF HYPERTHERMOPHILIC ENDOGLUCANASES

By

LOGAN EVERETT SHEFFIELD, Bachelor of Science

Presented to the Faculty of the Graduate School of

Stephen F. Austin State University

In Partial Fulfillment

Of the Requirements

For the Degree of

Master of Science in Biotechnology

STEPHEN F. AUSTIN STATE UNIVERSITY

August 2021

MOLECULAR DYNAMICS SIMULATIONS PROVIDE INSIGHT INTO STABILITY OF HYPERTHERMOPHILIC ENDOGLUCANASES

By

LOGAN EVERETT SHEFFIELD, Bachelor of Science

APPROVED:

_____
Dr. Alexandra Martynova-Van Kley, Thesis Director

_____
Dr. James Briggs, Committee Member

_____
Dr. Brian Barngrover, Committee Member

_____
Dr. Lindsey Porter, Committee Member

_____
Pauline M. Sampson, Ph.D.
Dean of Research and Graduate Studies

ABSTRACT

Endoglucanases play a key role in the industrial production of bioethanol, but the most efficient method requires the utilization of high temperatures and is currently limited by the thermostability of endoglucanases. For this reason, it would be beneficial to discover more high-efficiency, thermostable enzymes to utilize in the hydrolytic process. In this study molecular dynamics simulations were performed on structurally similar endoglucanases with varying levels of thermostability to gain insight on what factors contribute to thermostability in endoglucanases. RMSD, RMSF, PCA, hydrogen bonding and salt bridges were analyzed. Finally, protein energy networks were constructed from nonbonded interaction potentials and analysis was performed using hub population, cluster population, largest community transition profiles and LCC profiles. It was found that the more thermostable endoglucanases exhibited a greater number of hydrogen bonds along with fewer, more segregated electrostatic interactions and a larger network of low-energy van der Waals interactions – likely responsible for providing adequate rigidity to withstand high-temperature conditions while still allowing the flexibility needed for proper catalytic function.

ACKNOWLEDGEMENTS

First and foremost I would like to extend my sincere thanks to Dr. Martynova-Van Kley and the rest of the Van Kley family for their guidance, continuous support and patience throughout my journey here at Stephen F. Austin State University. During my time in Nacogdoches I began to think of you all as my family away from home, and for that I will forever be thankful.

I would also like to thank Dr. Armen Nalian for his invaluable advice which helped me to guide me along my journey into molecular dynamics simulations and programming. Not only were you a prominent resource throughout the planning of my project, you also took on a strong mentor role. You helped me to sharpen my thinking so that I could find the right solution anytime the project took unexpected turns.

I would also like to thank Dr. James Briggs, from the University of Houston, for agreeing to serve as a member on my committee. Your tremendous understanding of this field helped me to stay on track and feel confident in the direction of my research.

My appreciation also goes out to Dr. Barngrover for participating on my committee and helping to give me direction during my planning process, and also to Dr. Porter for agreeing to join during my final semester after a previous member left the university to take a position elsewhere.

I would also like to give my thanks to Ron Havner and Dr. Beatrice Clack for being both mentors and friends to me during my time in the biology department. You both helped to grow my love for science and shape my decision to pursue graduate school.

Another person who greatly helped me through my graduate school journey is Lawrence Ferrell, the custodial worker for the second floor of the Miller Science Building. Our brief but frequent conversations felt like the only constant in my everchanging workday that helped to keep me grounded and clear-headed during many stressful times in the office, and I will always be thankful for that.

Finally, I would like to thank my parents and my brother Cameron for believing in me every step of the way. Your unwaivering support throughout my entire life helped to shape the person I am today.

TABLE OF CONTENTS

LIST OF FIGURES/GRAPHS

LIST OF ABBREVIATIONS

EGAc - Endoglucanase from *Acidothermus cellulolyticus* (moderate thermophile)

EGPh - Endoglucanase from *Pyrococcus horikoshii* (hyperthermophile)

EGXc - Endoglucanase from *Xanthomonas campestris* (mesophile)

fs - femtoseconds

GH5 - glycoside hydrolase family 5

gRINN - get Residue Interaction eNergies and Networks

LCC - largest connected component

MD - Molecular Dynamics

NAMD - Nanoscale Molecular Dynamics

ns - nanoseconds

PCA - principal component analysis

PDB - Protein Data Bank

PEN - protein energy network

RMSD - root-mean-square deviation

RMSF - root-mean-square fluctuation

VMD - Visual Molecular Dynamics

INTRODUCTION

Pyrococcus horikoshii is a hyperthermophilic microbe that produces a highly thermostable β-(1,4)-endoglucanase (EGPh; Figure 1) which has an optimal pH between 5.4 and 6.0, and is capable of retaining 80% activity after heating for 3 hours at 97°C. Identified by Kawarabayasi (1998) as a member of glycoside hydrolase family 5 (GH5), EGPh was later compared to other GH5 members in the presence of 1-ethyl-3-methylimidazolium acetate at various temperatures to observe deactivation mechanisms. Unlike the other sampled GH5 members, however, EGPh did not show any signs of deactivation (Jaegar et al., 2015).

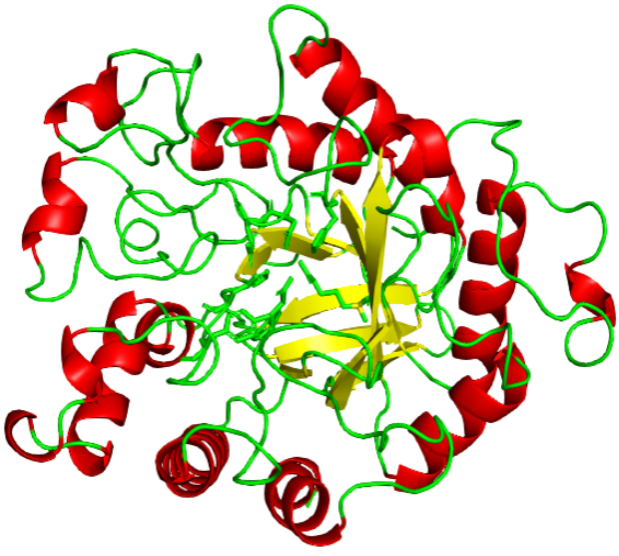

Figure 1: An x-ray structure of EGPh (Kim & Ishikawa, 2011; PDB ID=3AXX). Helices are shown as red, sheets as yellow and loops/turns are shown as green.

As noted by Li et al. (2011), GH5 endoglucanases primarily consist of a catalytic domain, all sharing (ß/α)8 barrel overall topology. There is typically a substrate-binding cleft at the C-terminal end of the barrel into which the cellulosic structures are introduced to the active site (Figure 2). There are seven known conserved residues amongst GH5 members, of which glutamate residues serve as both the proton donor and the nucleophile (Wang et al., 1993).

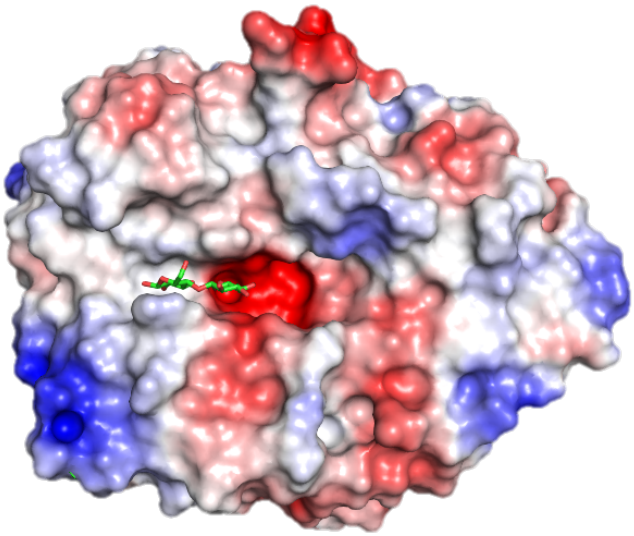

Figure 2: PyMOL-generated model of EGPh modelling electrostatic contact potential. The red color represents negative potential caused by an excess of negative charges near the surface, while the blue color represents positive potential caused by positive charges near the surface. White regions indicate a relatively neutral surface. Cellotetraose (green) is shown within the binding cleft.

Because of the persistent nature of this enzyme, it would be beneficial to observe EGPh alongside known mesophilic endoglucanases of shared structural similarity to gain a better understanding of the factors allowing its operation under higher temperature conditions. This might be accomplished through Molecular Dynamics (MD) simulations,

which would allow the added benefit of visualizing how the enzyme withstands even higher temperatures to predict possible target residues for modification as an effort to further enhance thermostability.

THERMOPHILES: AN OVERVIEW

The ubiquity of microorganisms has been found to persist in a myriad of environments - varying greatly in conditions such as temperature, salt concentration and pH. The term 'extremophile' is used to describe those organisms capable of enduring the harshest of conditions (Rothschild & Mancinelli, 2001). Those organisms capable of withstanding high temperatures are called thermophiles, and can be classified into three groups (Stetter, 2006):

1. Simple Thermophiles: 50-64°C
2. Extreme Thermophiles: 65-79°C
3. Hyperthermophiles: 80°C+

In contrast, mesophiles are those organisms which grow best between 20-45°C (Willey, 2008). Overall, the cellular components of mesophiles and thermophiles are markedly disparate (e.g., differing membrane lipids and guanine/cytosine content; Brock, 1978; Huser et al., 1986). Still, microbes must rely on proteins capable of maintaining stability for the entire range of temperatures experienced within their environment, a characteristic termed thermostability.

While thermostable molecules have been widely utilized for industrial purposes, there is at times a lack of clarity concerning exactly which factors are responsible for differences in molecular thermostability among structurally similar molecules. While it is known that single amino acid mutations may result in decreased thermostability, the act of increasing it is often less simple – in fact, it is uncommon that a single mutation increases the thermostable range by more than 3-5°C (Fontana, 1991). There is a balance between forces preserving the native state of the protein and those disrupting it, in which the former marginally subjugates the latter in the range of 5–20 kcal mol^{-1} (Pace, 1975; Kamerzell & Middaugh, 2008). While these same forces act on protein stabilization amongst psychrophiles (cold-loving organisms) and hyperthermophiles alike, slight variations in the strength or number of interactions can yield a considerable difference in protein stability. This allows for a multitude of possible adjustments that may be used in the stabilization of proteins under various conditions, making it difficult to identify specific changes making great contributions to stability (Goldstein, 2007). Further adding to the complexity, some stabilizing factors are themselves temperature-dependent (e.g., hydrophobic interactions). This balance between forces is responsible for the dynamics of protein systems, including the oscillation of individual atoms as well as movement of entire protein domains.

MOLECULAR DYNAMICS SIMULATIONS

One method of enzyme comparison that has been growing in popularity involves the use of molecular dynamic simulations to provide insight on aspects such as folding pathways, native structure, and atomic interactions contributing to stability (Scheraga et al., 2007). This methodology not only allows for the comparison of temperature-dependent forces, but also temperature-independent differences contributing to the stability of each of the members of the enzyme pairs.

Molecular dynamics (MD) simulations were first used by McCammon et al. (1977) to analyze bovine pancreatic trypsin inhibitor in a vacuum. Their method used an empirical energy function to solve the equations of motion for the atoms (Newton 1687):

$$m_\alpha \vec{r} = -(\frac{\partial}{\partial \overrightarrow{r_\alpha}}) U_{total}(\overrightarrow{r_1}, \overrightarrow{r_2}, \ldots, \overrightarrow{r_N}), \alpha = 1,2, \ldots N$$

Where m_a is the mass of atom α, r_α is its position, and U_{total} is the total potential energy that depends on all atomic positions and, thereby, couples the motion of atoms (Phillips et al., 2005). Improvements to the methodology of McCammon et al. (1977) have included incorporation of counterions, inclusion of explicit solvent molecules surrounding the protein of interest, modifications to the system boundaries, and more realistic modeling of long-range electrostatic forces (Hansson et al., 2002). Implementation of periodic boundary conditions help to minimize problems with boundary effects caused by finite size. When using Ewald summation methods, however, Weber et al. (2000) showed that artifacts can be introduced through the inclusion of periodicity into the calculations for

long-range electrostatic interactions. An effective method to counteract this problem is the reaction field approach, which uses a cutoff radius on polarizable surroundings to correct for pair-wise electrostatic interactions (Zuegg & Gready, 1999).

ANALYSIS OF SIMULATION RESULTS

Numerous analytical tools can be used to assess the output from MD simulations, including root-mean-square deviation and root-mean-square fluctuations—two common methods that compare an atom or a group of atoms to a reference point across a simulation. Solvent accessible surface area is another commonly utilized method and involves taking a measurement of the solvent-accessible surface area of the protein in question over the course of a simulation. Protein dihedral analyses might be used to examine the angles of rotation along the protein structure. This method is useful for determining the arrangement of secondary structure (Benson & Daggett, 2012). Assessment of simulation trajectories using Principal Component Analysis can identify important motions of the proteins (David & Jacobs, 2014). Another method is to calculate the radius of gyration to gain insight on the compactness of the molecule (Lobanov et al., 2008).

Many other methods have been growing in popularity that utilize machine learning to help analyze MD simulations on a deeper level than traditional methods allow. Whereas methods of dimensionality reduction and clustering algorithms are

becoming commonplace in the field (Noe & Nuske, 2013; David & Jacobs, 2014), a method called deep neural networks (LeCun et al., 2015; Schmidhuber, 2015) has greatly expanded the use of machine learning in molecular biochemistry. In human neurons, input signals are received from surrounding neurons through dendrites and, if the signal strength reaches a certain threshold, an action potential is generated along that neuron. In a similar fashion, artificial neurons take input signals their corresponding weights and send an output signal if a threshold is reached. Networks consisting of multiple layers can be analyzed using signals to tune the weights between layers in order to minimize output error. This process allows for a thorough analysis of complex data sets, provided the machine learning algorithm is properly designed. One example of the use of neural networks within molecular dynamics is to reproduce the free-energy surface of molecules (Schneider et al., 2017).

Brinda and Vishveshwara (2005) applied network theory to protein structures to evaluate stability of proteins. They used each amino acid as a node, and the edges of the protein were determined by analyzing the noncovalent interactions between them. Brinda and Vishveshwara (2005) noted that aromatic residues—as well as methionine, histidine and arginine—all act as strong hubs when using high cutoff values, which play a role in the increased stability of thermophilic proteins by helping anneal different secondary structure elements within the protein. The process of generating this type of analysis was simplified when Chakrabarty and Parekh (2016) constructed a server for a network-based analysis of protein structure and folding called Network Analysis of Protein Structures

(NAPS). Additionally, the Bio3D package within the R analytical platform allows for an automated analysis of protein structures and simulation runs, including all of the methods mentioned above (Grant et al., 2006).

Another useful application of network theory to the analysis of MD simulations involves the use of protein energy networks (Vijayabaskar & Vishveshwara, 2010). These energy networks were made by calculating Lennard-Jones and Coulombic interaction sums from 2-nanosecond simulations, which was found to be a sufficient simulation length after comparison with 10-nanosecond runs revealed consistent results. They then compared the results from 12 thermophilic/mesophilic enzyme pairs using weighted graphs that utilize edge weights determined by the interaction energy between amino acids using the following formula:

$$E_{ij} = V_{LJ}(r_{ij}) + V_C(r_{ij})$$

Where $V_{LJ}(r_{ij})$ represents the average potential energy due to Lennard-Jones interactions of residues *i* and *j,* while $V_C(r_{ij})$ represents the potential energy from Coulombic interactions. Vijayabaskar & Vishveshwara, (2010) found that cluster and clique population appeared to be the main factors leading to increased stability of thermophiles, and that thermophiles typically had densely populated hydrophobic cores with local hotspots that help to increase the difference in energy level between folded and unfolded states.

An open-source software called gRINN (get Residue Interaction eNergies and Networks; Sercinoglu & Ozbek, 2018) allows for efficient analysis of residue interaction

energies from simulation runs through an automated interface. The gRINN software also calculates the interaction energy correlations and analyzes the energy networks to help identify functional residues within proteins.

The aim of this study is to conduct MD simulations using a hyperthermophilic endoglucanase and mesophilic relatives and utilize various post-simulation analytical methods to gain insight on the thermostabilizing forces present within the selected molecules.

METHODS

SELECTED ENZYMES FOR THE STUDY

To find enzymes for comparison, protein BLAST was performed on the EGPh crystal structure (PDB ID: 3AXX) using a cutoff value of 100, and results were limited to the top 10 hits. Normal mode analysis—useful for exploring the dynamics of protein families because of the characteristic fluctuations of conserved regions (Grant et al., 2006) – was used to help narrow down results and gain insight on the flexibility of the proteins (Skjaerven et al., 2014). Of the hits provided by BLAST, 1ECE (EGAc) and 4TUF (EGXc) were selected for comparison to EGPh.

To illustrate the structural similarities, the MUSCLE multiple sequence alignment program was used to perform a sequence alignment, followed by a structural alignment of EGPh first to EGAc, and then to EGXc (Edgar, 2004).

MOLECULAR DYNAMICS SIMULATIONS

Visual Molecular Dynamics (VMD) software was used along with the crystal structures of EGPh, EGAc, and EGXc to generate protein structure files. To ensure proper protonation states, proPKA was used to predict the protonation state of each residue at a neutral pH. The enzyme structures were then solvated in water boxes

expanding 10 Å from the protein, followed by ionization to neutrality with sodium chloride using the autoionize plugin in VMD.

Nanoscale Molecular Dynamics (NAMD) software was used to conduct simulation runs, with periodic boundary conditions and parameters from the CHARMM36 All-Atom Additive Protein Force Field (Huang & MacKerell, 2013). The long-range interactions were evaluated using the Ewald Summation method. Minimization was performed first for 1000 steps with the protein fixed, followed by a second 1000-step minimization with all atoms freed. This was followed by a stepwise heating before conducting 100 ns production runs. A two-fs timestep was utilized, allowing for desirable simulation runtimes with minimal loss of information. Temperature control was performed using Langevin dynamics with a coupling coefficient of 1/picosecond.

For each of the selected enzymes, simulations were conducted at 25°C, 50°C, 75°C, 100°C, and 125°C. While water at atmospheric pressure boils at temperatures above 100°C, pressure compensation utilized in the simulations should offset this to allow simulations at and above this temperature.

ANALYSIS OF SIMULATION RESULTS

Upon completion, water was stripped from each of the simulation's output trajectories to allow for manageable file sizes for comparison (~ 10 GB each).

Simulations were visualized using VMD to confirm the integrity of each simulation, then data analysis was performed.

Root-mean-square deviation (RMSD)

RMSD analysis, which measures average overall deviation of a molecule from original starting coordinates, was performed using the output of each simulation using Bio3D in RStudio. RMSD data plots were generated for each simulation by plotting the RMSD against time to confirm proper equilibration of each simulation and to look at how each molecule moved overall throughout each run.

Root-mean-square fluctuations (RMSF)

RMSF analysis, which measures the deviation of each residue from its starting coordinates, was performed using Bio3D in RStudio. RMSF plots were generated to help visualize the contribution of each individual residue to the overall RMSD results and help identify regions of the protein that exhibit significant motions during the simulations (Benson & Daggett, 2012).

Principal component analysis (PCA)

PCA is useful for identifying significant motions in each trajectory and finding changes in motion between trajectories. For each simulation, the two most prominent

principal components were identified using Bio3D in RStudio and plotted in a similar manner to the RMSF plots, with residues plotted on the X axis.

Identification of hydrogen bonds

Protein-protein hydrogen bonding was analyzed for each simulation by calculating the number of hydrogen bonds present per nanosecond at each temperature using the 'HBonds' plugin in VMD. The results were plotted as a function of time using RStudio.

Identification of salt bridges

Salt bridges were identified using the 'Salt Bridges' plugin in VMD using the default cutoff of 3.2 Å. This looked at each simulation for any two oppositely charged residues that ever came closer than 3.2 Å. and designated a salt bridge between them. Next, a data file for each of the identified salt bridges was output into a mother directory for each simulation. The data files contained the distance in angstroms between the two oppositely charged residues for each simulation.

The number of salt bridges present at each timestep was calculated in RStudio using a cutoff of 4.0 Å and plotted as a function of time. This is useful to look for differences in overall salt bridge bonding for each enzyme between simulations.

Next, the 'prevalence' of each salt bridge (i.e., the percent of the time in the simulation run that the salt bridge existed for) was determined by calculating the

percentage of the simulation for which the two involved residues were closer than 4.0 Å to one another. The 30 most prevalent salt bridges for each enzyme were identified by totaling the prevalence of each salt bridge from simulations from each temperature.

Finally, a box-and-whisker plot was generated using RStudio for each salt bridge to compare distances of each bridge at the different temperature runs.

Construction & analysis of protein energy networks

gRINN software was used to generate a protein energy network (PEN) for each simulation run. From gRINN, a data file was generated for each run that contains a list of 'nodes' (residues) along with a weighted edge list calculated via summation of nonbonded interaction potential between the two involved residues.

Hubs, the highly connected nodes in a network (degree >3), were identified using iGraph in RStudio and plotted as a function of 'E', where 'E' is the highest energy that can exist between two residues *i* and *j* to draw and edge between them. While Vijayabaskar & Vishveshwara's paper stated analysis at 25°C was efficient for analysis of thermostability, hubs were analyzed at every simulated temperature for this study to analyze changes in packing efficiency for each of the enzymes.

Clusters, connected components in a network, were identified from each PEN using a depth-first-search (DFS) algorithm, then were plotted as a function of energy in the same manner as the hubs to visualize how segregated the stabilizing units of each enzyme are.

Communities were constructed for each network using k=3 cliques. Cliques are rigid subgraphs in a PEN, while communities are consolidated rigid subgraphs constructed from identified cliques. Once communities were identified, a largest community transition profile was constructed and plotted as a function of ‘E’.

Finally, a largest connected component (LCC) transition profile was obtained for each PEN and plotted as a function of ‘E’ to analyze the overall connectivity of each network.

RESULTS

SELECTED ENZYMES FOR COMPARISON

Of the hits provided by BLAST (Figure 3), EGAc (1ECE) and EGXc (4TUF) were selected for comparison to EGPh based on their different optimal temperatures. 1ECE is a crystal structure for EGAc, an endoglucanase isolated from *Acidothermus cellulolyticus,* a moderate thermophile which has an optimal temperature of 55°C (Ding et al., 2002). The optimal temperature of EGAc has been found to be 81°C (Puhl et al., 2019), while its activity has been seen to drop significantly around 95°C (Sun et al., 2007).

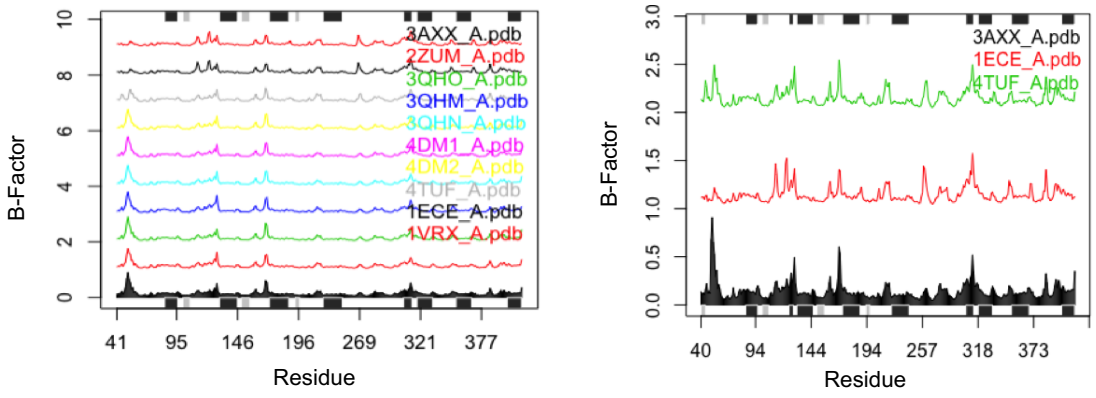

Figure 3: Results from ensemble normal mode analysis of hits from BLAST (left). An extracted secondary structure schematic is shown at the top and bottom of the plot (black representing helices and grey representing sheets). Large fluctuations tend to be predicted for areas containing loops. In the graph to the right, data for all enzymes except those selected for comparison have been omitted. Units are in Angstroms.

4TUF is a crystal structure of EGXc (from mesophilic *Xanthomonas campestris*), which has an optimal temperature of 25-30°C (Puhl et al., n.d.). EGXc has an optimal temperature of 45°C and shows a steady drop in activity as temperature increases above this point (Rosseto, 2016).

To illustrate the structural similarities, the MUSCLE multiple sequence alignment program was used to perform a sequence alignment, followed by a structural alignment of EGPh first to EGAc, and then to EGXc (Figure 4; Edgar, 2004). The stick structures of the seven known conserved residues of GH5 members can also be visualized (Figure 5).

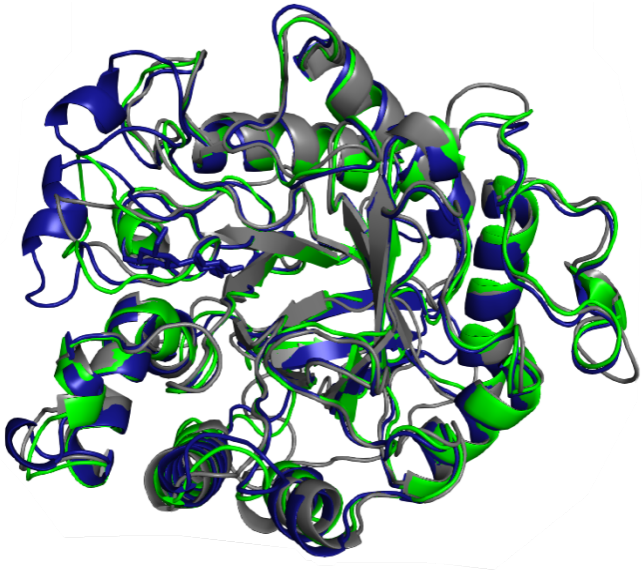

Figure 4: Superimposed structures of EGPh (blue) with EGAc (gray) and EGXc (green). Note, the shared (ß/α)8 barrel topology and also the differences among the turns and loops along the outside of the molecules.

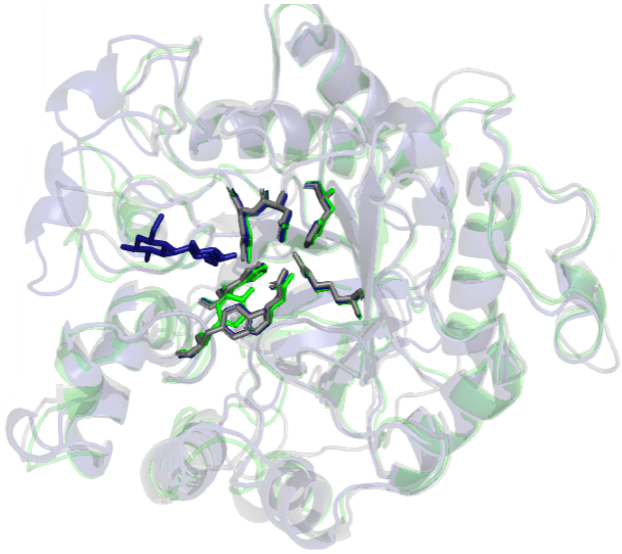

Figure 5: Superimposed structures of EGPh (blue) with EGAc (gray) and EGXc (green). The seven residues conserved amongst GH family 5 members (including the proton donor GLU201 and the nucleophile GLU342) have been displayed in stick form while the other residues were made transparent.

ROOT-MEAN-SQUARE DEVIATION

RMSD analysis, which measures average overall deviation from original starting coordinates, was performed as described in the methods section. Visualization of RMSD plots revealed each enzyme was adequately minimized and equilibrated, as indicated by the levelled off RMSD trajectory. While thermophiles often have lower RMSD at high temperatures than their mesophilic counterparts, EGPh does not seem to follow that pattern (as seen in Figure 6A). EGAc seems to have maintained a degree of rigidity at every temperature, as its RMSD trajectory does not vary much across temperatures when compared to the other two enzymes (Figure 6B). The mesophilic EGXc shows more

deviation than EGAc, but still considerably less than EGPh (Figure 6C). Because EGPh is known to be the more thermostable of the three, it may be inferred that this enzyme employs more flexibility at higher temperatures.

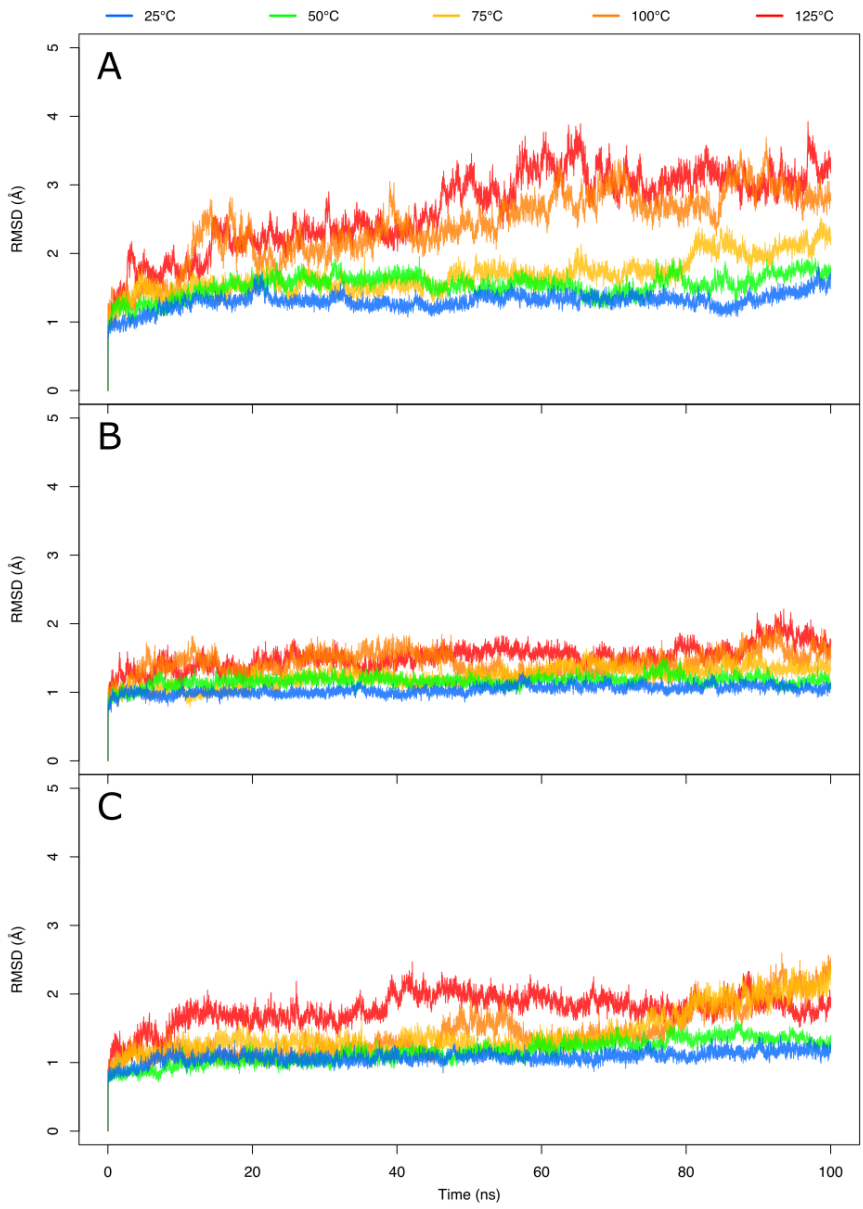

Figure 6: RMSD plot of EGPh (A), EGAc (B), and EGXc (C) at various temperatures.

ROOT-MEAN-SQUARE FLUCTUATIONS

RMSF analysis may be used to measure how much each residue fluctuates from its initial position, and is useful for identifying regions of the protein that exhibit significant motions during the simulations (Benson & Daggett, 2012). RMSF analysis was performed as described in the previous methods section, and the results are shown in Figures 8-10.

Due to the $(ß/α)_8$ barrel topology of GH5 endoglucanases, there are 8 ß-α looped regions throughout the overall structure. Loop 4 forms the left wall of the active site, while loop 6 forms the right wall and helps position the cellulose chain in the active site. The width of this cleft is strongly related to rate of catalysis, with a narrow cleft being correlated to an increased k_{cat}. Loops 1, 3, 5, and 8 all help shape the cleft, with loop 1 specifically responsible for the length of the binding cleft (Glasgow et al., 2020). Loop 5, which lies between and just under the active site walls, is typically the shortest of the loops.

Because the looped regions lack the complex hydrogen bonding patterns present in ß-sheets and α-helices, they exhibit much more movement and are thus visible on RMSF plots as spiked regions. It has been established that motions of these loops are involved in substrate binding and product release, and the flexible motion of loops 6 and 7 specifically has been linked to known to promote proton transfer at the active site (Zheng et al., 2018).

Figure 7 shows that at 100°C (near EGPh's optimal temperature), loops 6 and 7 exhibit a greater degree of motion while still maintaining the original RMSF plot shape. At 125°C, the loop 6 RMSF line exhibits a different shape, indicating some change in the pattern of motion may have occurred. Loop 4, which also works with loop 6 to promote catalysis, exhibits significantly increased movement above 100°C – as does loop 5, which interacts with the substrate as it enters the active sites.

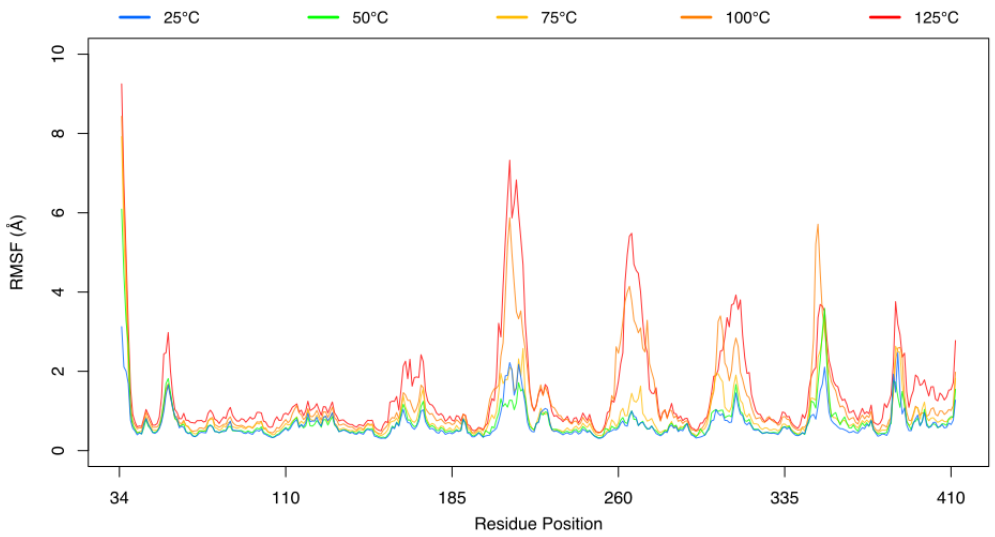

Figure 7: RMSF plot of EGPh at various temperatures.

The moderately thermophilic EGAc displays much less variation in its RMSF plots between different temperatures, but loops 7 and 8 do exhibit a sharp increase in motion above 100°C (Figure 8). At 125°C, loop 7 appears to adopt a different motion, perhaps indicative of a conformational change. Loop 2 also has a moderate spike introduced at 125°C.

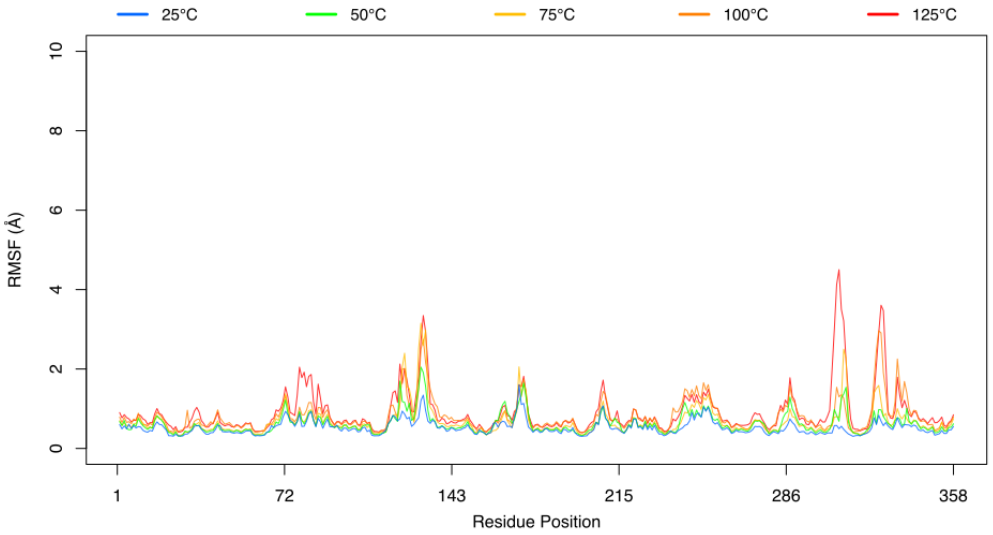

Figure 8: RMSF plot of EGAc at various temperatures.

EGXc also retains much of its motion on the N-terminal half across the temperature changes, but loops 5-8 all show increased RMSF spikes as temperature increases above 50°C (Figure 9).

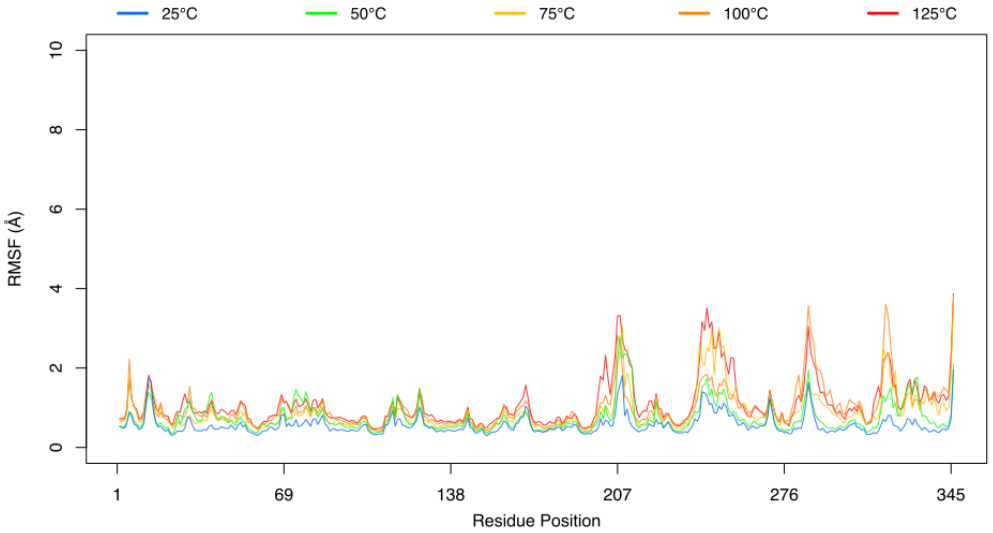

Figure 9: RMSF plot of EGXc at various temperatures.

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is useful for identifying significant motions in each trajectory and may be used for finding changes is motion between trajectories. For each simulation, the most prominent principal component was plotted in a similar manner to the RMSF plots, with residues plotted on the X axis (Figures 10, 11 & 12).

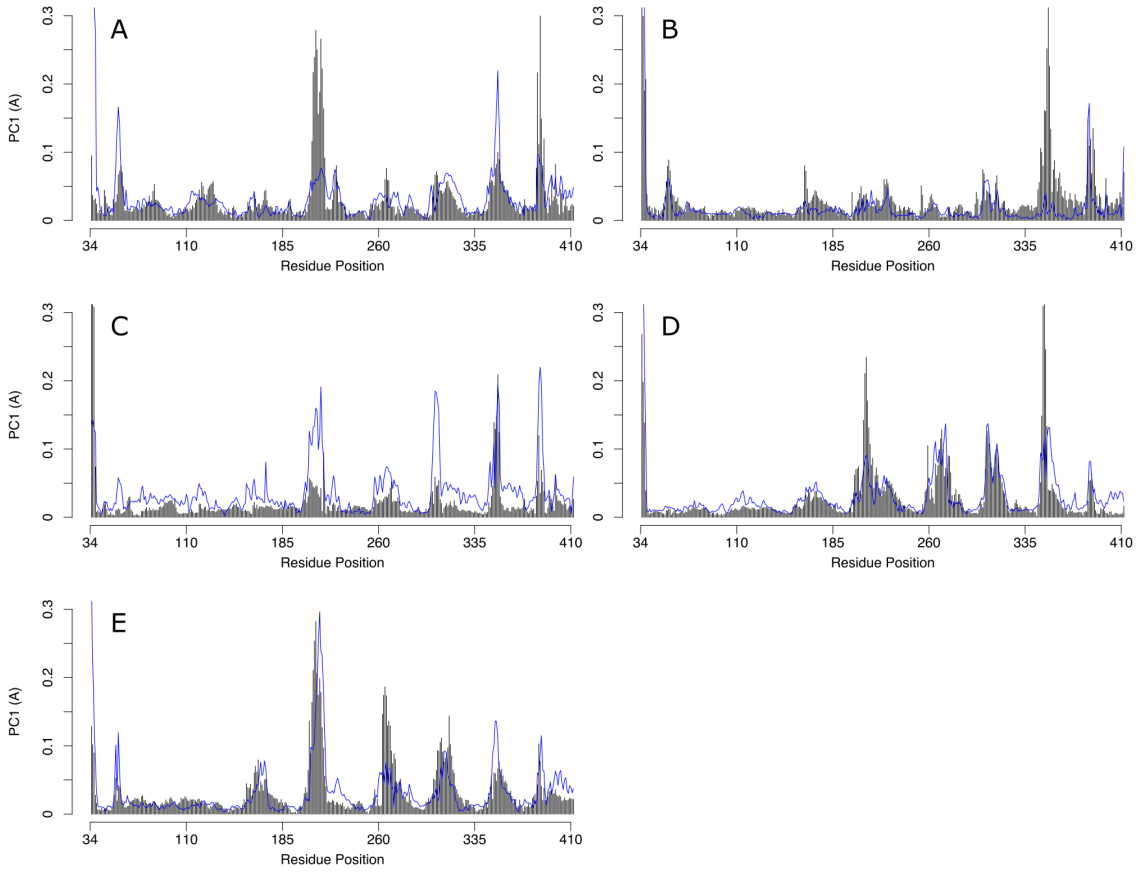

Figure 10: Principal components 1 (black, bar display) and 2 (blue, line display) of EGPh plotted at 25°C (A), 50°C (B), 75°C (C), 100°C (D) and 125°C (E).

For EGPh, loop 5 makes a noticeable change once heated past 100°C. Due to the previously mentioned location and significance of loop 5, it is possible that this change is

disrupting the shape of its binding cleft at the catalytic center and contributing to its loss of function over 100°C.

In EGAc, there is evidence of a change in motion for loops 7 and 8 when heated above 75°C, just as observed with the RMSF analysis (Figure 11). The shape of this plot for loop 8 clearly shows a sharp increase in motion for the C-terminal side of loop 8, which forms part of the right cleft boundary along with loop 6. Because this change is only seen when heated past its optimal temperature, it may be disrupting the proper motion of the enzyme.

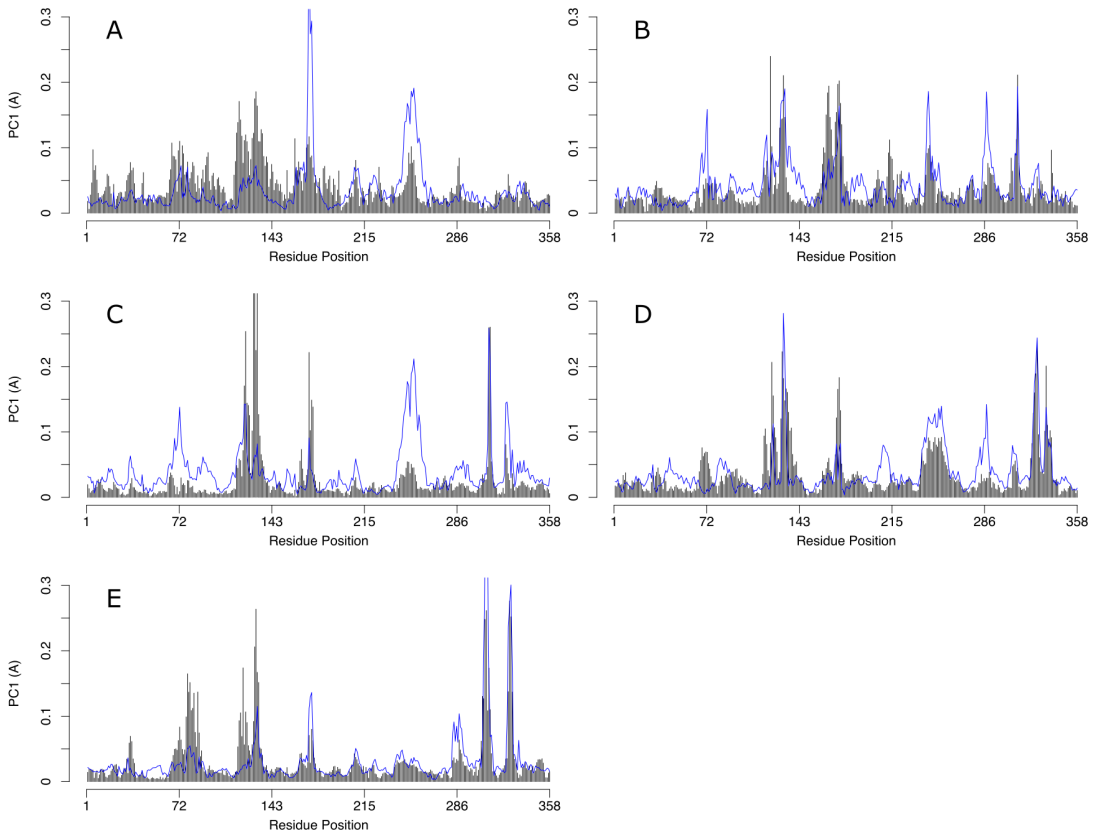

Figure 11: Principal components 1 (black, bar display) and 2 (blue, line display) of EGAc plotted at 25°C (A), 50°C (B), 75°C (C), 100°C (D) and 125°C (E).

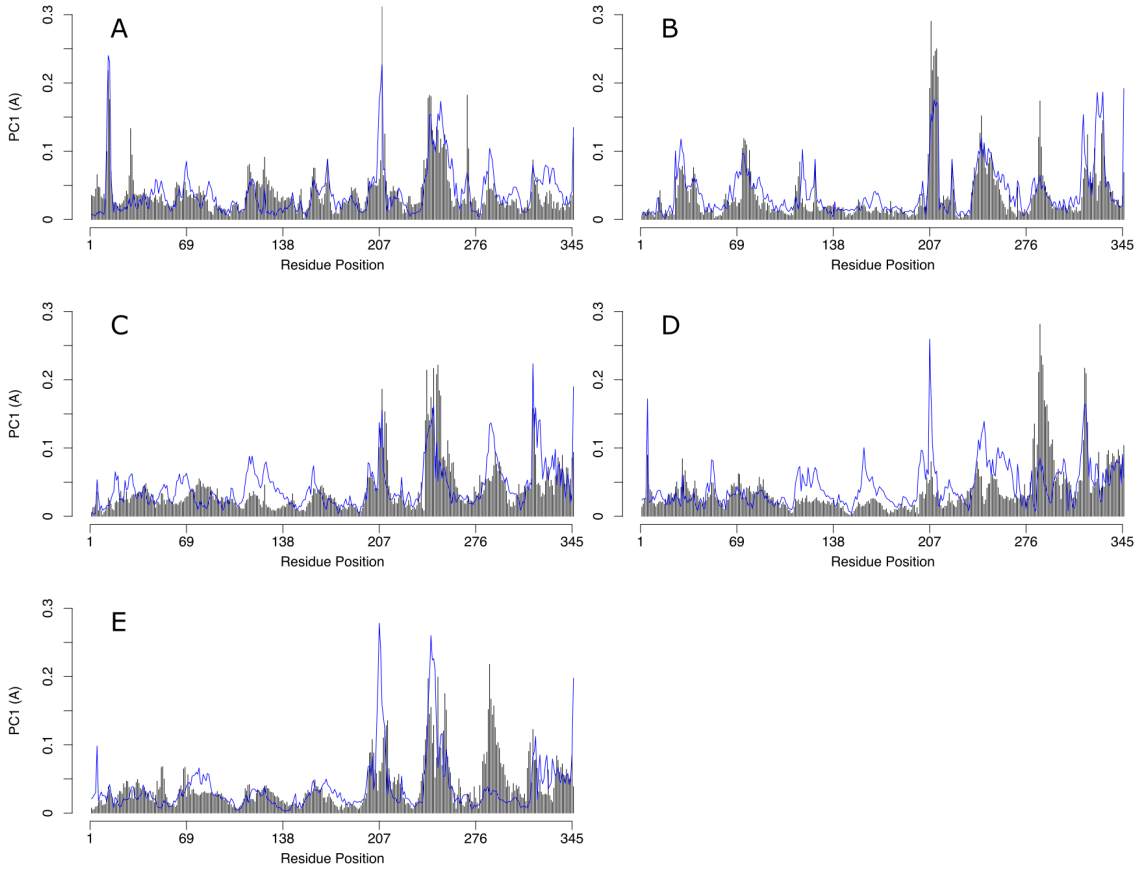

Figure 12: Principal components 1 (black, bar display) and 2 (blue, line display) of EGXc plotted at 25°C (A), 50°C (B), 75°C (C), 100°C (D) and 125°C (E).

In EGXc loops 7 and 8 again show a change in motion when heated above its optimal temperature, with the motions of the first few loops getting overshadowed at temperatures of 75°C and greater in principal component analysis (Figure 12).

HYDROGEN BONDING ANALYSIS

Hydrogen bonding was analyzed for each simulation by calculating the number of hydrogen bonds present per nanosecond at each temperature for each enzyme (Figure 13A-C). Only protein-protein hydrogen bonding was considered for analysis, as it has been reported to be a more significant factor in thermostability (Melchionna et al., 2006).

For each enzyme, the average hydrogen bonding for each simulation was calculated and plotted as a function of temperature (Figure 13D).

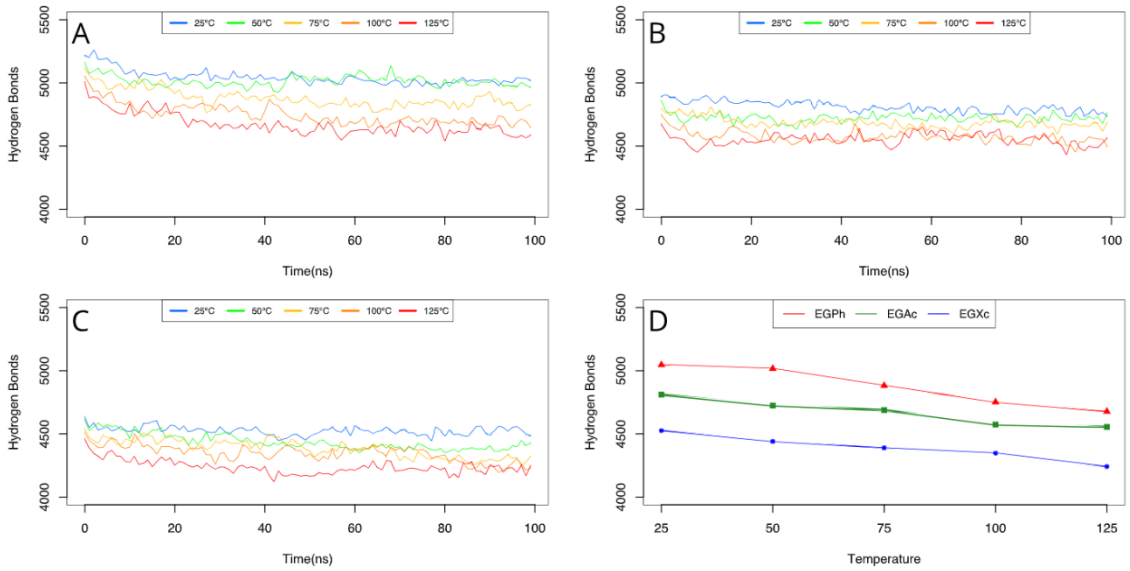

Figure 13: Number of hydrogen bonds at each timestep for EGPh **(A),** EGAc **(B),** EGXc **(C)** and average number of hydrogen bonds at each temperature (in C°) for each endoglucanase **(D)**.

At each temperature there is a clear difference in the number of hydrogen bonds present for each of the three enzymes, with the most thermostable EGPh possessing the most and the mesophilic EGXc possessing the least. The greater amount of hydrogen

bonding present in more thermostable endoglucanases likely helps maintain secondary and tertiary structure needed to sustain proper function.

ANALYSIS OF SALT BRIDGES

Salt bridges were analyzed by counting the number of salt bridges present across the simulation as determined by VMD (with a cutoff value of 3.2 Å). For each salt bridge found, a data file was created containing the distance between the two residues at each timestep. The overall results were first plotted by averaging the number of salt bridges present over each nanosecond and plotting them over time (Figures 14 & 15).

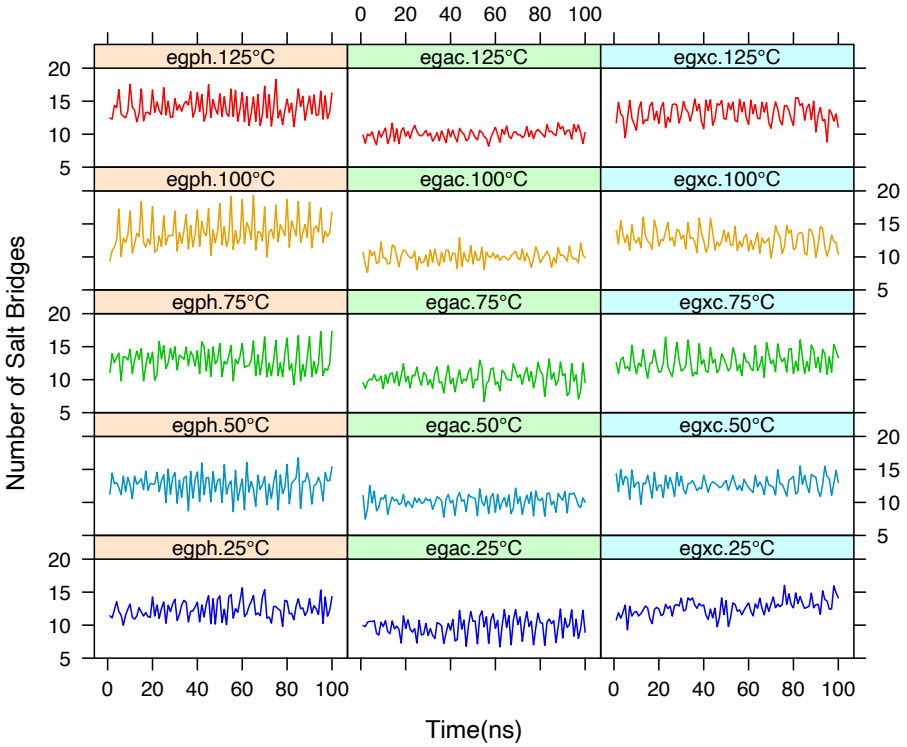

Figure 14: Average number of salt bridges present at each nanosecond, with the results for each enzyme plotted vertically.

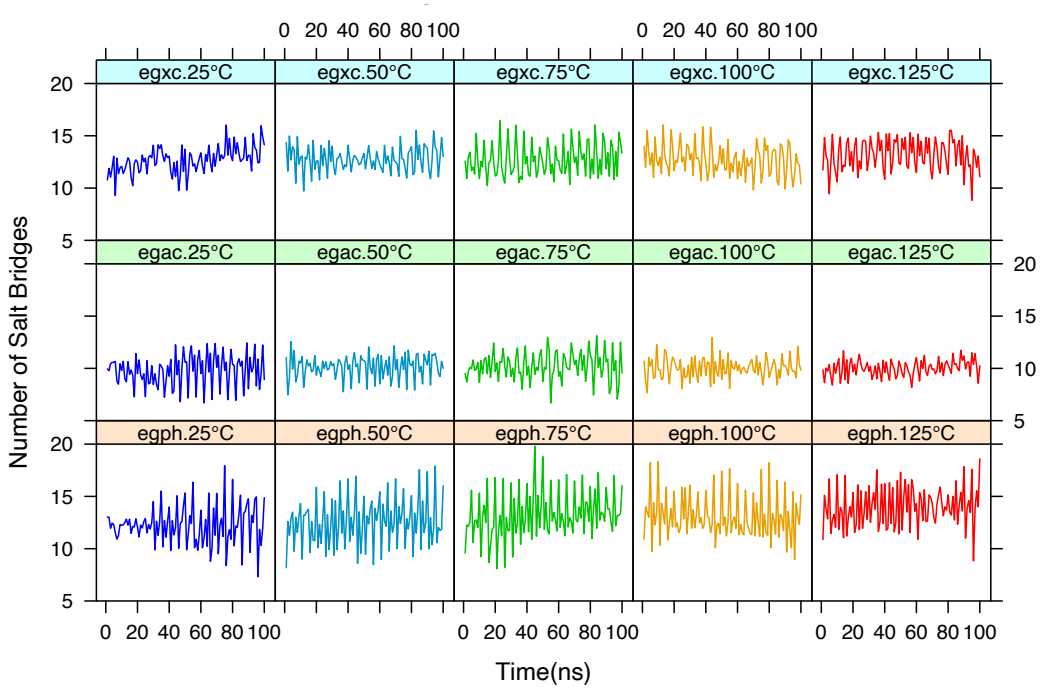

Figure 15: Average number of salt bridges present at each nanosecond, with the results for each enzyme plotted horizontally.

For EGPh, the amount of salt bridges appears to slightly increase as temperature rises through 100°C. The mesophilic EGXc appears to increase its number of salt bridges only from 25°C to 50°C, while the moderate thermophile EGAc retains relatively the same number of salt bridges across each temperature.

Next, for every enzyme/temperature permutation the prevalence of each individual salt bridge was determined by calculating the percentage of the simulation for which it was present. This information was used to identify the top 30 most prevalent salt bridges per enzyme.

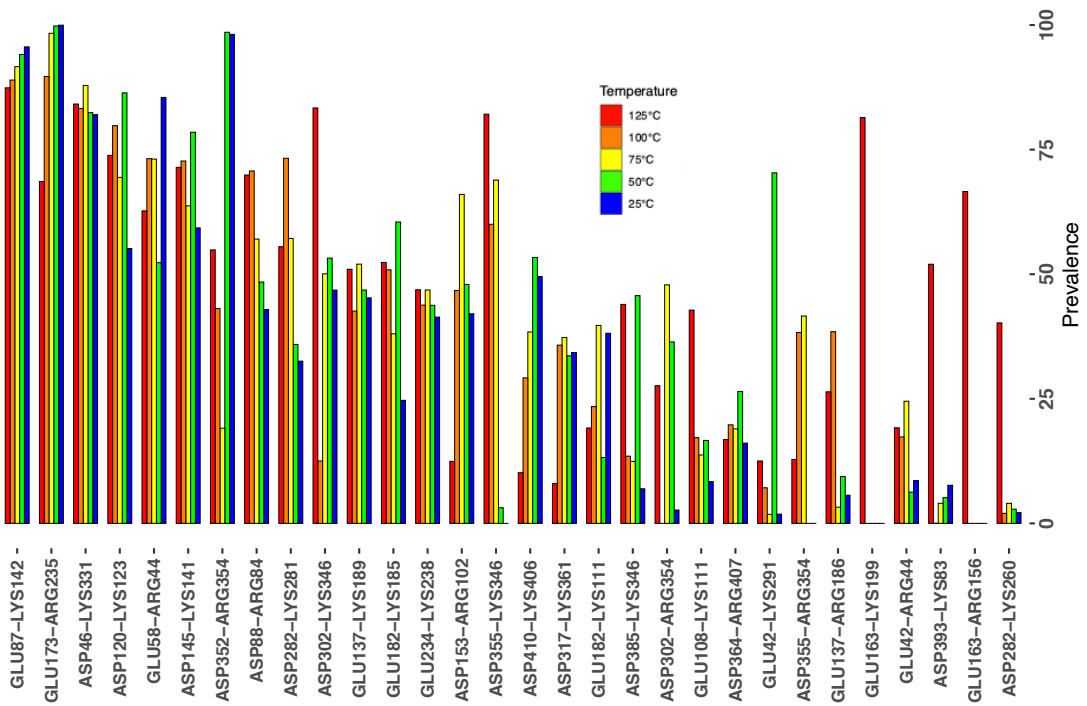

Figure 16: Top 30 most prevalent salt bridges for EGPh.

EGPh Salt Bridges

In the hyperthermophilic EGPh, there are three salt bridges towards the right side of Figure 16 that only significantly appear once heated above 100°C. Because EGPh is known to lack function at this temperature, these salt bridges are not likely to play a part in the proper function of the enzyme. Towards the left side of the figure, Glu173-Arg235 shows a drop in prevalence above 100°C. Further exploration of this salt bridge (Figure 17A) reveals the median distance does not show a large change, but the prevalence decreases gradually from 99.97% at 25°C to 89.68% at 100°C and finally a sharp drop to

68.59% at 125°C. Visual inspection of the location of these residues (Figure 17B) shows they are involved in the binding between loops 3 and 4 in the binding cleft.

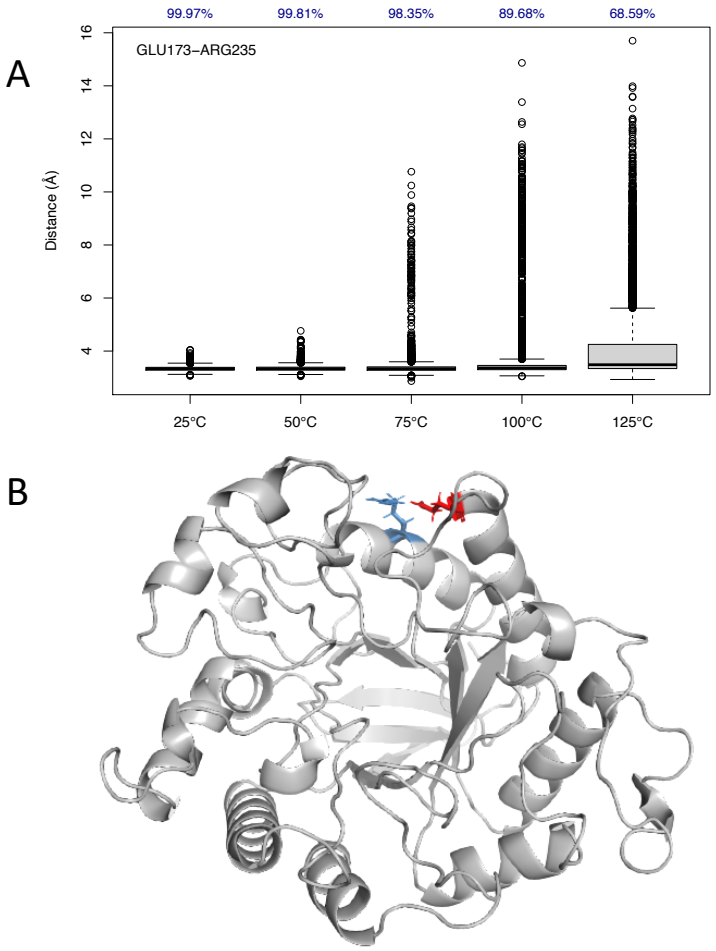

Figure 17: (A): Distance boxplot of the salt bridge between Glu173 and Arg235 at each temperature for EGPh. The prevalence (top, blue) is the percentage of the simulation at which the salt bridge was present. **(B):** Cartoon rendering of EGPh with Glu173 (red) and Arg235 (blue) shown as sticks.

EGAc Salt Bridges

In the moderately thermophilic EGAc, the salt bridge between Asp312-Arg11 appears to greatly drop in prevalence more closely to EGAc's optimal temperature range (Figure 18).

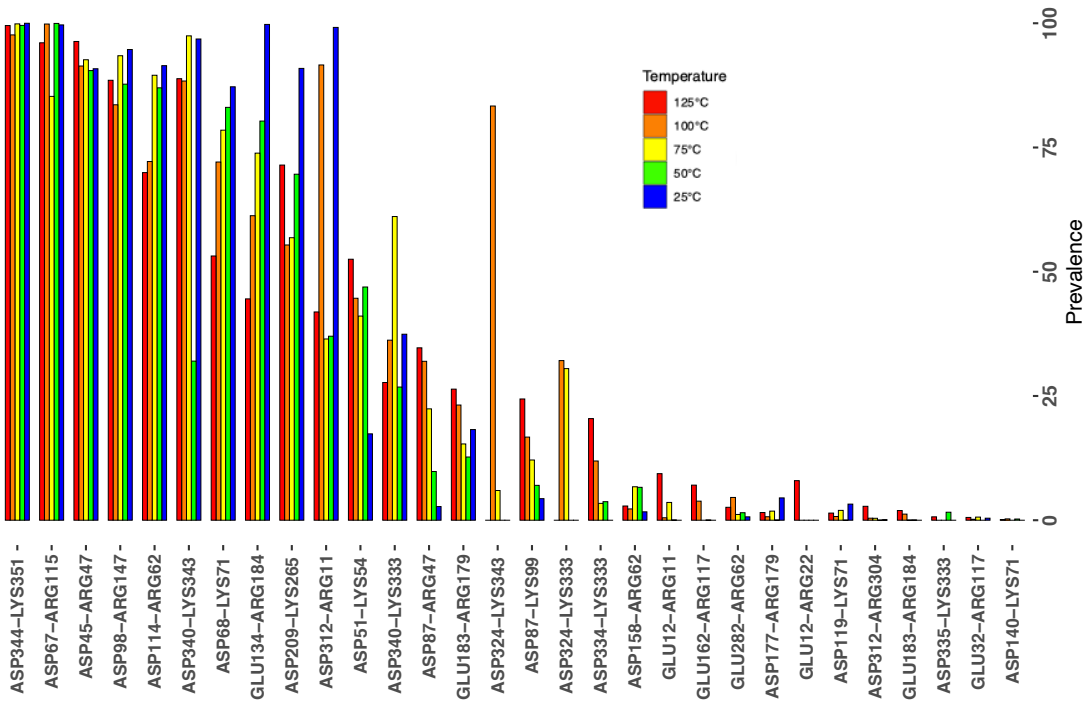

Figure 18: Top 30 most prevalent salt bridges for EGAc.

This is further illustrated in Figure 19, which shows the salt bridge distance for each temperature. This bond is positioned near the base of the enzyme, and its absence likely allows greater flexibility of the C-terminal side of loop 7 (Figure 20), explaining the change in shape of the RMSF plot around this area (refer back to the RMSF section).

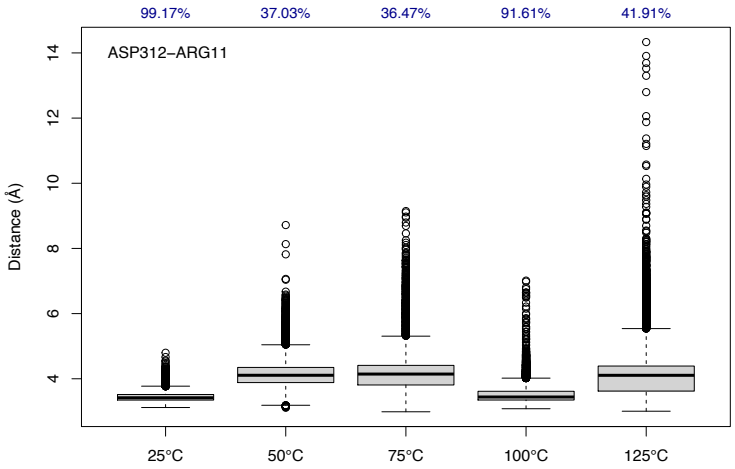

Figure 19: Distance boxplot of the salt bridge between Asp312 and Arg11 at each temperature for EGAc. The prevalence (top, blue) is the percentage of the simulation at which the salt bridge was present.

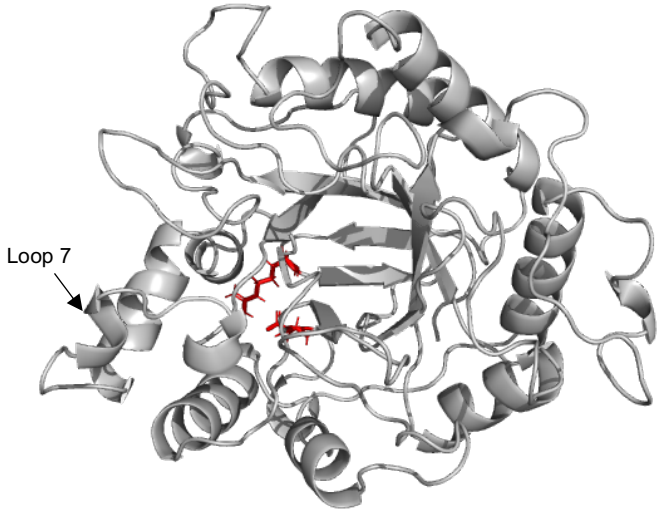

Figure 20: The salt bridge between Asp312-Arg11 **(red)** in EGAc. Loop 7, which is more secured in place with this salt bridge present, is labelled with the black arrow.

Another noteworthy salt bridge in EGAc is Asp324-Lys343, shown in Figure 21. It is only present at 75°C (present 6% of the time) and at 100°C (present 83% of the

time). Visual inspection of this salt bridge reveals that it seems to pull loop 8 outward, thereby opening the catalytic core (Figure 22).

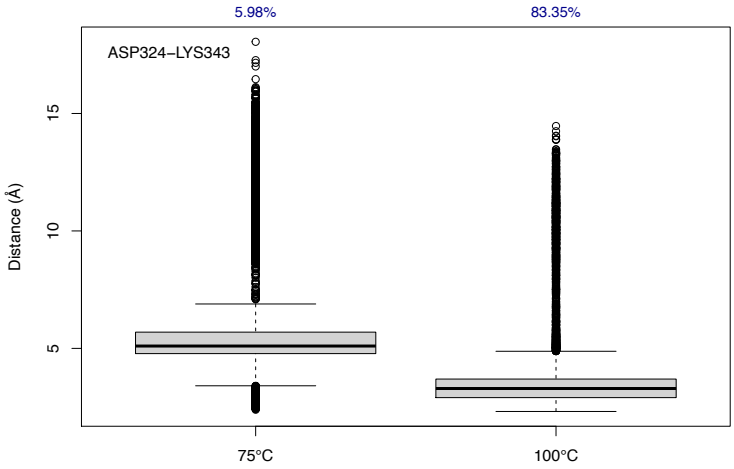

Figure 21: Distance boxplot of the salt bridge between Asp324 and Lys343 at each temperature for EGAc. The prevalence (top, blue) is the percentage of the simulation at which the salt bridge was present.

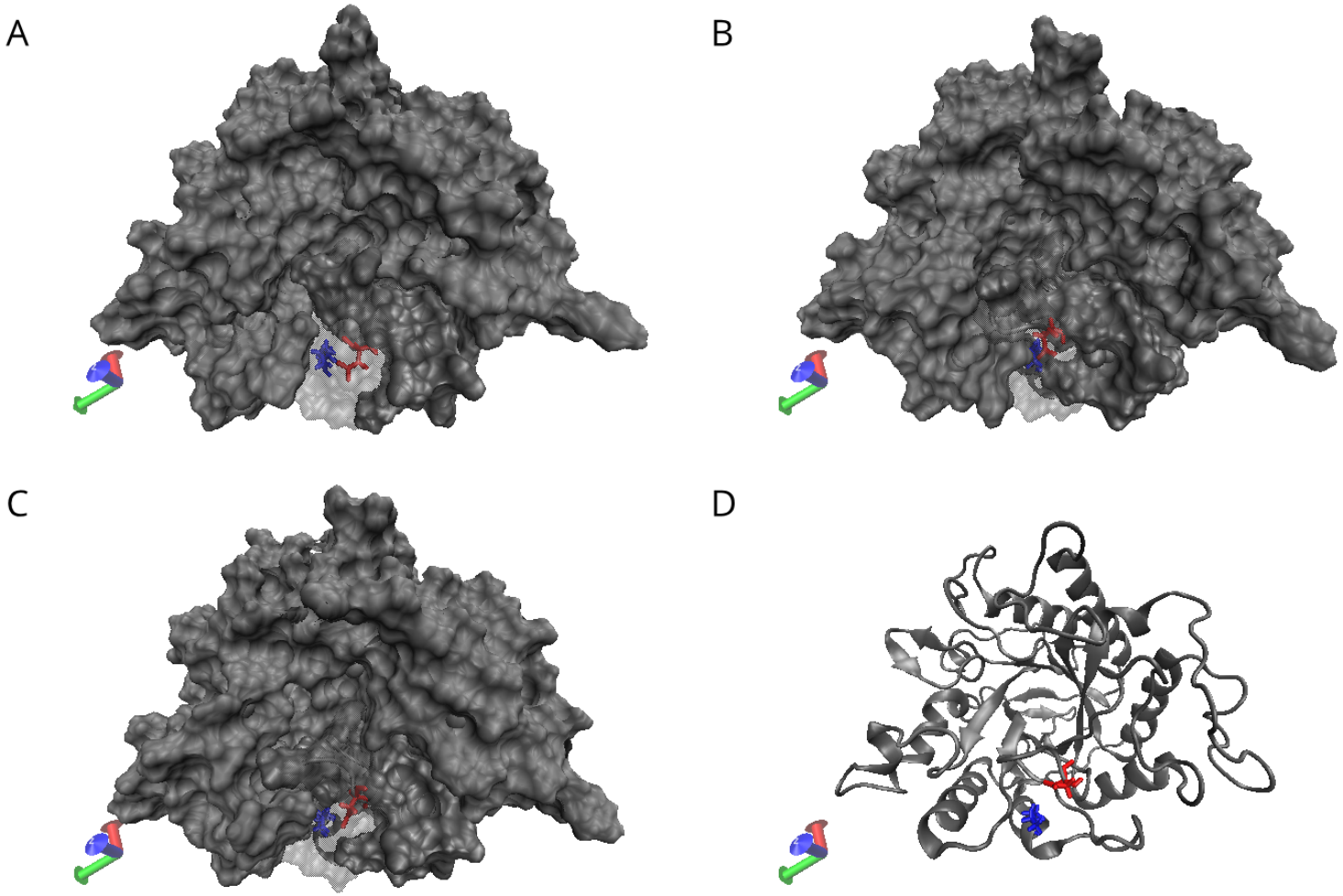

Figure 22: Surface rendering of EGAc at 100°C with Asp324-Lys343 exposed at 0ns (A), 40ns (B), and 75ns (C). A cartoon rendering is also shown at 75 ns (D).

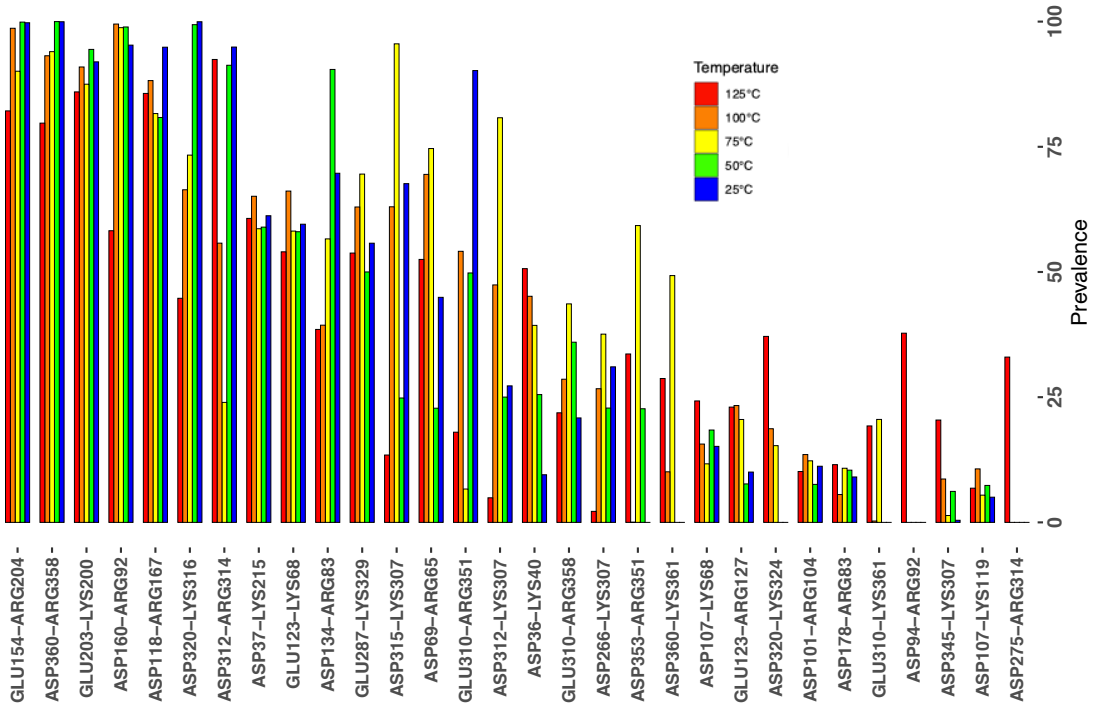

Figure 23: Top 30 most prevalent salt bridges for EGXc.

EGXc Salt Bridges

In the mesophilic EGXc, Asp312 in loop 7 shares a salt bridge with Arg314 for the majority of the simulation at 25°C and 50°C but then shifts to Lys307 at 75°C and 100°C (Figures 23, 24A, 24B). Because Lys307 is closer to the core while Arg314 is towards the outermost part of the loop, this is indicative of a change in loop 7's position and a loss of original conformation (Figure 24C).

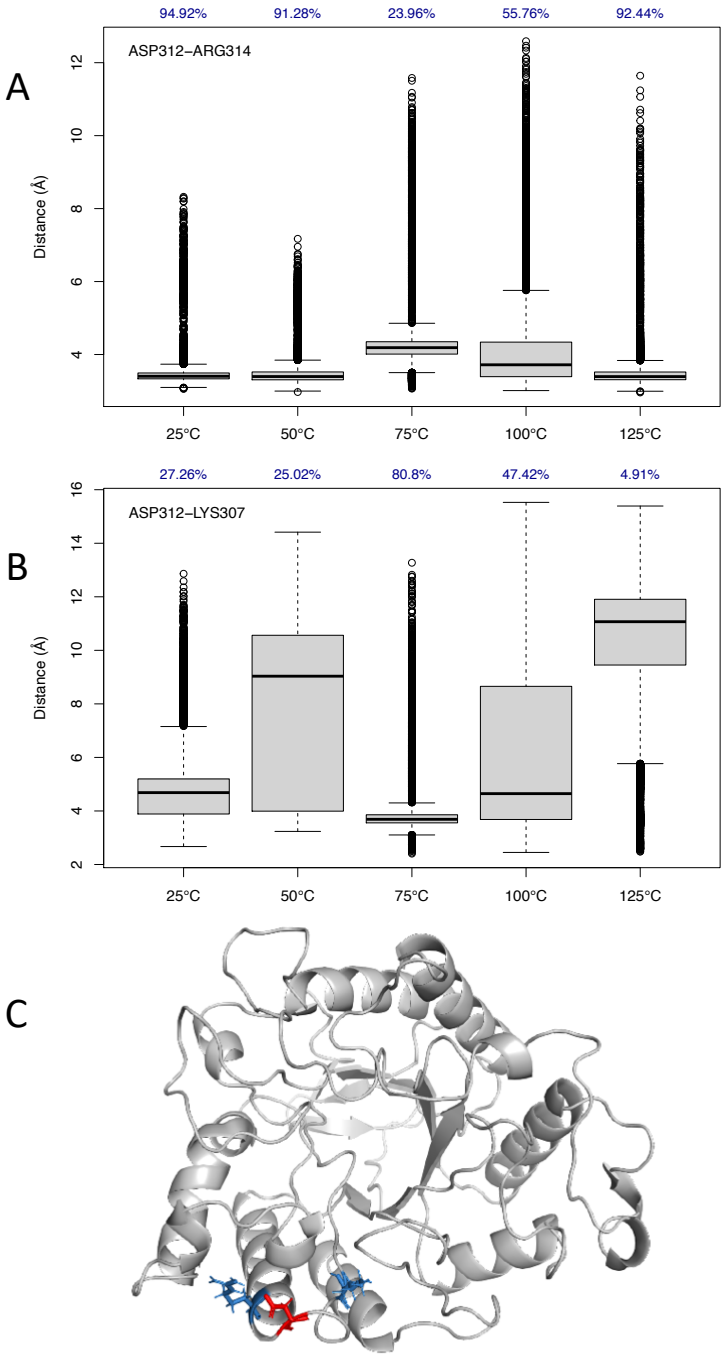

Figure 24: Distance boxplot of the salt bridge between Asp312 and **(A)** Arg314 and **(B)** Lys307 at each temperature for EGXc. The prevalence (top, blue) is the percentage of the simulation at which the salt bridge was present. The location of the three residues is displayed in **C**, with Asp312 colored red, Arg312 blue (left) and Lys307 blue (right).

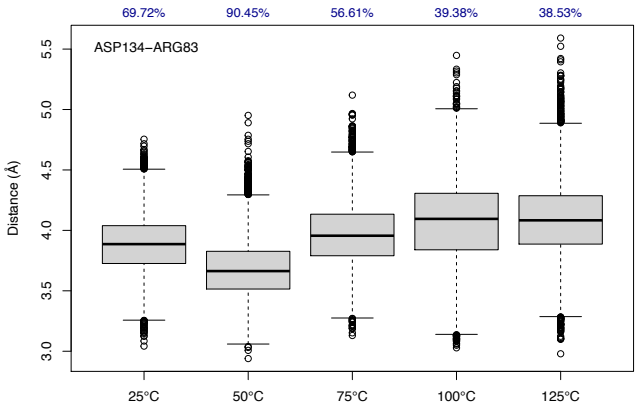

Figure 25: Distance boxplot of the salt bridge between Asp134 and Arg83 at each temperature for EGXc. The prevalence (top, blue) is the percentage of the simulation at which the salt bridge was present.

In the catalytic core, Asp134 maintains a salt bridge with Arg83 for most of the simulation at 50°C, but the salt bridge steadily drops in prevalence as temperature is increased beyond that point (Figure 25). The location of these residues in the catalytic core is shown in Figure 26.

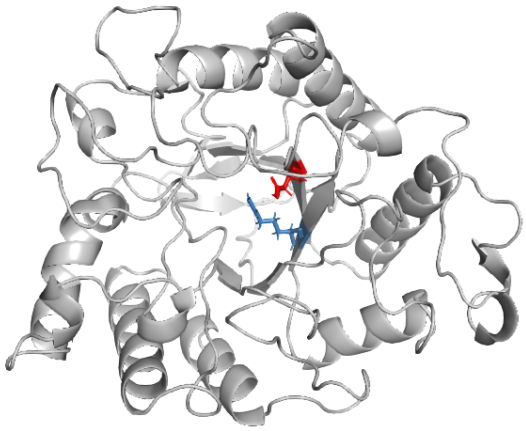

Figure 26: Location of Asp134 (red) and Arg83 (blue) in EGXc.

PROTEIN ENERGY NETWORKS

Protein Energy Networks (PENs) were constructed for the enzymes at each temperature based on simulation runs using residues as nodes, with weighted edges based on average total nonbonded interaction energy. The weaker interaction energies (> -10KJ/mol) are mostly comprised of van der Waals interactions while the stronger interaction energies (< -20KJ/mol) are comprised of electrostatic interactions (Vijayabaskar & Vishveshwara, 2010). Once constructed, the PENS were analyzed to look at hub and cluster population changes, largest community size and largest connected component size. These were then plotted as a function of energy to allow comparison between data sets.

Hub population

Hubs, the highly connected nodes in a network (degree >3), were identified and plotted as a function of energy. This analysis helps visualize an enzyme's "structural resilience… against external perturbations" (Vijayabaskar & Vishveshwara, 2010). While Vijayabaskar & Vishveshwara's paper stated analysis at 25°C was efficient for analysis of thermostability, hubs were analyzed at *every* temperature for this study to analyze changes in packing efficiency for each of the enzymes.

The results show that the hub population of EGPh is greater than its mesophilic and moderately thermophilic counterparts, both at the low energy and at the transition

regions (Figure 27) up until 100°C, at which point the hubs in the transition region drop off. This may suggest a more efficiently packed hydrophobic core in EGPh. As expected, EGXc has less hubs in general when compared to its two more thermostable counterparts.

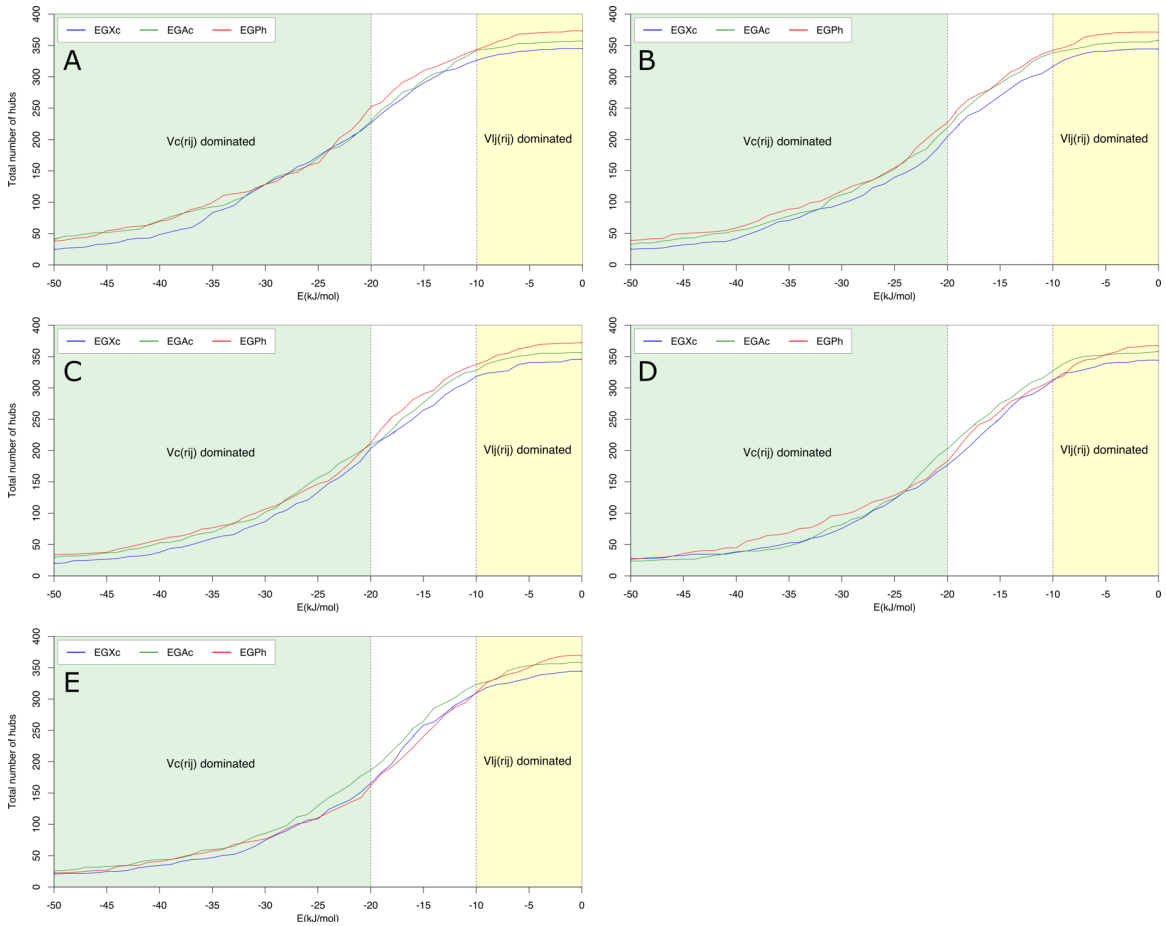

Figure 27: PEN hub population of EGPh (red), EGAc (green) and EGXc (blue) at 25°C (A), 50°C (B), 75°C (C), 100°C (D) and 125°C (E).

.

Cluster population

Clusters, connected components in a network, were identified from each PEN using a depth-first-search (DFS) algorithm, then were plotted as a function of energy in the same way that the hubs were analyzed (Figure 28). Clusters are a good measure of how segregated the stabilizing units of a protein are (Vijayabaskar & Vishveshwara, 2010).

While EGAc has the highest cluster population peak, EGPh has the most high-energy (< ~-40KJ/mol) clusters, showing that EGPh has a better degree of segregation of its high-energy interactions. This higher population of segregated electrostatic clusters at high-energy levels likely provides excellent stabilization of the protein in comparison to its less thermostable counterparts.

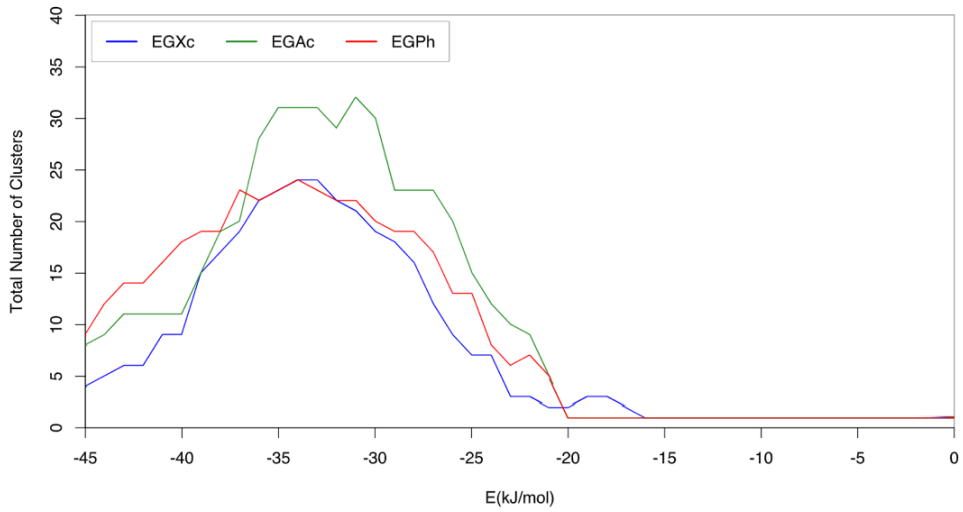

Figure 28: Cluster population at 25°C.

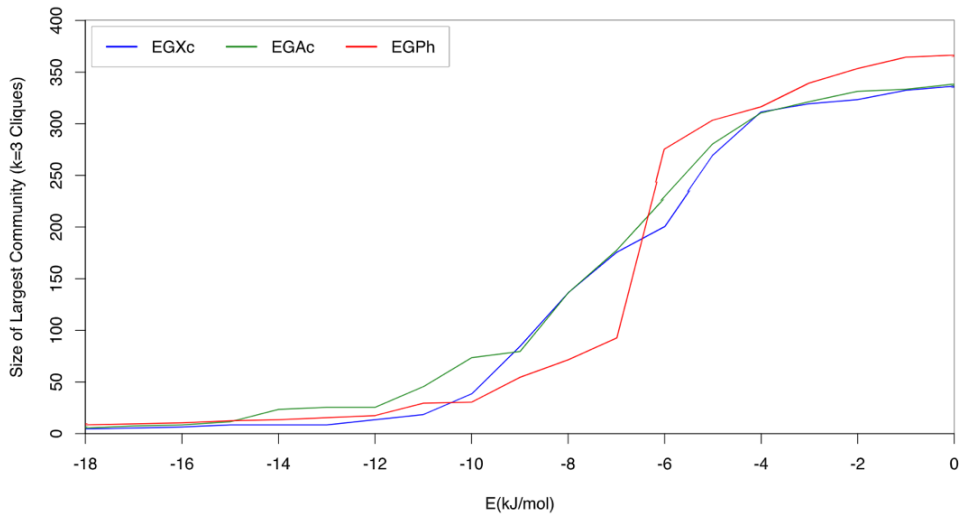

Figure 29: Largest community transition profile at 25°C.

Largest community transition profile

Cliques are rigid subgraphs in a PEN, while communities are consolidated rigid subgraphs constructed from identified cliques. For this study, communities were constructed from k=3 cliques and a largest community transition profile was plotted as a function of energy (Figure 29).

Because of the ubiquity of weak nonbonded interactions in any given molecule, community size is typically very large at low energy cutoffs. As the energy cutoff is increased, the community breaks up into smaller, more numerous communities. According to Vijayabaskar & Vishveshwara (2010), thermophiles typically have larger communities at low energy levels. A large community prescence at only low-energy levels may be interpreted as a lack of presence to electrostatic interactions in stabilization, allowing for a stable but less rigid structure.

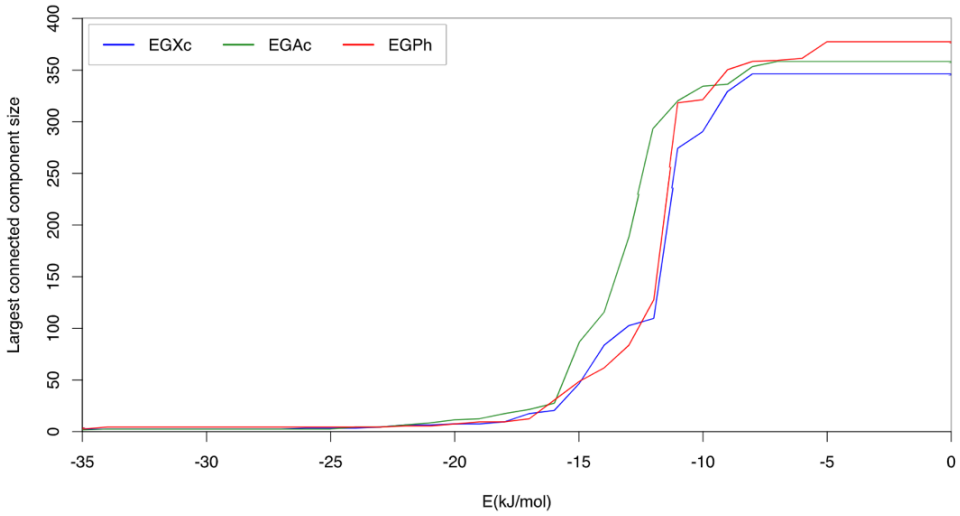

Figure 30: Largest connected component (LCC) transition profile at 25°C.

Largest connected component transition profile

The largest connected component (LCC) of a network is a parameter that may be used to analyze the overall connectivity of a network (Razvi, 2006**)**. An LCC transition profile was obtained for each PEN and plotted as a function of energy (Figure 30). The LCC transition profile for EGPh was larger in the Lennard-Jones-dominated region, consistent with the previous PEN findings. EGAc has a larger LCC across the transition region, while the plot closes in for all three enzymes at the Coulombic-dominated region.

DISCUSSION

Endoglucanases are enzymes that hydrolyze internal β-(1,4)-glycosidic bonds between the glucose monomers of cellulose. This hydrolysis plays a key role in the production of bioethanol, a renewable fuel source with lower greenhouse gas emissions than those of traditional fuels (Acharya & Claudhary, 2012). The most efficient method of accomplishing this bioethanol production is through simultaneous saccharification and fermentation, in which the initial lignocellulosic biomass is exposed to high temperatures in the presence of dilute acid during cellulose hydrolysis (Badieyan et al., 2012). This process is limited by the thermostability of the involved endoglucanase enzymes, however, which currently only allows for a temperature range of 50-55°C (Ando et al., 2002). This low temperature range for endoglucanases requires separation of the saccharification and fermentation processes; thus, it would be beneficial to discover more high-efficiency, thermostable enzymes to utilize in the hydrolytic process. Thus, performing MD simulations on these molecules may help to gain insight on the thermostabilizing forces present within endoglucanases.

It has been shown that thermophiles often have a greater amount of protein-protein hydrogen bonding present than their mesophilic counterparts (Melchionna et al., 2006), which is consistent with the results of these simulations. However, the increased

hydrogen bonding does not appear to cause increased overall rigidity of the molecules, as shown by the increased loop RMSF values in EGPh. While rigidifying proteins is often seen as a method of increasing thermostability, this study has shown that EGPh actually maintains a greater degree of flexibility than its moderately thermophilic and mesophilic GH5 relatives. EGAc and EGXc maintained more overall rigidity than EGPh, although loop 8 of these molecules did exhibit increased RMSF values when heated above their optimal temperatures. Loop 8 works with loop 6 to form the right side of the cleft boundary (Glasgow et al., 2020), so this change may be disrupting the proper shape of the cleft. EGPh and EGAc both showed an increased RMSF range for loop 5, another loop responsible for shaping the catalytic cleft, above their temperature optima. The disruption of the position of these loops when heated seems to greatly change the shape of the binding cleft **(**see salt bridge figures**)** and likely contributes to their loss of function.

Salt bridges, another form of stabilization observed in proteins, do not appear to follow the same pattern as hydrogen bonding in these endoglucanases. At every temperature, the mesophilic EGXc consistently possessed the most salt bridges on average -- followed by the hyperthermophilic EGPh and finally the moderately thermophilic EGAc. Therefore, the number of salt bridges present does not appear to play a vital role in thermostability for these endoglucanases. However, in EGPh the Glu173-Arg235 salt bridge showed a drop in prevalence above its optimal temperature. Due to the location of these residues (Figure 18B), these findings suggest that at temperatures

above 100°C loops 3 and 4 are pulled away from each other, disrupting the structure of the binding cleft. In EGAc, the salt bridge between Asp324 and Lys343 appears at 100°C, which seems to be involved in the distal repositioning of loop 8. This may allow solvent to contact the inner core, which could disrupt proper enzymatic function and thus may be involved in the loss of function at high temperatures. In EGXc, the Asp134-Arg83 salt bridge exhibits a drop in prevalence above 50°C which, due to the location of these residues on loops 1 and 2 in the catalytic core, may indicate a loss of stability and packing efficiency in the core.

It has also been proposed that thermophiles often derive their greater stability not from high-energy bonds, but rather from their weaker non-bonded interactions. Vijayabaskar and Vishveshwara found that using protein energy network analysis on thermophilic and mesophilic protein relatives often revealed an increase in clusters and low-energy cliques (2010), which was observed to hold true for these GH5 endoglucanase enzymes. EGPh possesses more low-energy hubs that fall into the Lennard-Jones region rather than the Coulombic range -- while EGXc has the least, suggesting a greater efficiency of core packing correlates to increased thermostability. While EGAc's PEN has the greatest cluster population at its peak, EGPh has more high-energy clusters. This implies EGPh has more segregation amongst its high-energy interactions. EGPh appears to possess fewer, more segregated electrostatic interactions, along with a larger network of low-energy van der Waals interactions (as seen in its LCC and largest community transition profiles) when compared to the moderately

thermophilic EGAc and the same can be said of EGAc when compared to EGXc. This is likely responsible for providing the adequate rigidity to withstand high-temperature conditions while still allowing the flexibility needed for proper catalytic function.

CONCLUSION

This study has looked at RMSD, RMSF, PCA, hydrogen bonds, salt bridges, and analysis of networks constructed from nonbonded interaction potentials to gain insight on contributing factors to thermostability in endoglucanases. The hyperthermophilic EGPh was seen to have the highest RMSD value, showing an overall greater range of motion than its less thermophilic counterparts which shared a lower, more stable RMSD range relative to EGPh. While RMSF inspection revealed EGAc and EGXc to be more rigid overall than EGPh, loop 8 did show an RMSF increase above their optimal temperatures. In EGPh and EGAc, loop 5 also showed an increase in motion above their optimal temperatures. Because loops 8 and 5 are both directly involved in the shaping of the binding cleft, the disruption of the position of these loops is likely linked to a conformational change in the binding cleft (see salt bridge figures) thus inhibiting proper interaction with the substrate.

Analysis of hydrogen bonding revealed EGPh to have the most hydrogen bonds at each temperature, followed by EGAc and finally EGXc. This suggests there is some positive correlation between thermostability and number of hydrogen bonds in these endoglucanases.

Salt bridges, however, did exhibit this same pattern – the mesophilic EGXc showed a much greater number of salt bridges at every temperature than the moderately thermophilic EGAc. In EGPh and EGXc, there was a steady increase in the number of salt bridges as optimal temperature was approached, while EGAc maintained a relatively constant number of salt bridges at each temperature. For EGPh, the amount of salt bridges appears to slightly increase as temperature rises through 100°C. The mesophilic EGXc appears to increase its number of salt bridges only from 25°C to 50°C, while the moderate thermophile EGAc retains relatively the same number of salt bridges across each temperature. Individual inspection of the prevalence of salt bridges for each enzyme revealed salt bridges that seem to correlate to conformational changes involved in loss of function above optimal temperatures (drop in Glu173-Arg235 prevalence with loops 3 and 4 being pulled apart in EGPh; Asp324-Lys343 forming in EGAc with loop 8 being pulled distally from the core; decline in prevalence of the stabilizing Asp134-Arg83 on loops 1 and 2 in the core of EGXc). However, it is unclear whether the observed changes in these salt bridges are causing conformational changes or are simply a byproduct of it.

Analysis of protein energy networks constructed from nonbonded interaction potentials for each simulation revealed that enhanced core packing efficiency correlates to increased thermostability. Hub analysis showed increased low energy hubs in the more thermostable proteins, while cluster population analysis revealed less overall electrostatic interactions but more high-energy clusters. Inspection of the largest community and LCC

transition profiles revealed less overall electrostatic connectivity in the more thermophilic endoglucanases, with greater low-energy connectivity.

Taking all these findings together, it appears that a greater number of hydrogen bonds along with fewer, more segregated electrostatic interactions and a larger network of low-energy van der Waals interactions is likely responsible for providing the adequate rigidity to withstand high-temperature conditions while still allowing the flexibility needed for proper catalytic function.

FUTURE WORKS

Now that a workflow has been established, the scripts created for analysis of these proteins may be used on repeated simulation runs to ensure reliability of results, and then expanded to other GH5 endoglucanases in later studies to identify structures and sequences that contribute to this pattern. Analyzing more GH5 enzymes will help elucidate whether the patterns observed in this study expand to all similar enzymes or just the selected endoglucanases. Further simulations may also be conducted to model the endoglucanases in the presence of substrates to study the binding and catalysis process at various temperatures. If a method is found that helps to reliably predict GH5 catalytic efficiency through simulation runs, that may be used in conjunction with these analysis scripts to construct a machine-learning-assisted workflow to mass-analyze endoglucanases for efficacy in biofuel production. That knowledge may then be applied to constructing and testing GH5 chimeras for industrial applications.

REFERENCES

1. Acharya, S., & Chaudhary, A. (2012). Bioprospecting Thermophiles For Cellulase Production: A Review. *Brazilian Journal Of Microbiology*, 43(3), 844–856.

2. Ando, S., Ishida, H., Kosugi, Y., & Ishikawa, K. (2002). Hyperthermostable Endoglucanase From *Pyrococcus Horikoshii*. *Applied And Environmental Microbiology*, 68(1), 430–433.

3. Badieyan, S., Bevan, D. R., & Zhang, C. (2012). Study And Design Of Stability In GH5 Cellulases. *Biotechnology And Bioengineering*, 109(1), 31–44.

4. Beadle, B. M., Baase, W. A., Wilson, D. B., Gilkes, N. R., Shoichet, B. K. (1998). Comparing The Thermodynamic Stabilities Of A Related Thermophilic And Mesophilic Enzyme. *Biochemistry*, 38, 2570-2576.

5. Benson, N. C. & Daggett, V. (2012). A Comparison Of Multiscale Methods For The Analysis Of Molecular Dynamics Simulations. *The Journal Of Physical Chemistry B*, 116(29), 8722-8731.

6. Brinda, K. V., & Vishveshwara, S. (2005). A Network Representation Of Protein Structures: Implications For Protein Stability. *Biophysical Journal*, 89(6), 4159–4170.

7. Brock, T. D. (1978). Thermophilic Microorganisms And Life At High Temperatures. *Springer-Verlag*. New York.

8. Chakrabarty, B., & Parekh, N. (2016). NAPS: Network Analysis Of Protein Structures. *Nucleic Acids Research*, 44(W1), W375–W382.

9. David, C. C., Jacobs, D. J. (2014). Principal Component Analysis: A Method For Determining The Essential Dynamics Of Proteins. *Methods Of Molecular Biology*, 1084, 193-226

10. Ding S.Y., Vinzant T. B., Adney W.S., Decker S.R., Baker J.O., Jennings E., Himmel M.E. (2002). New Glycosyl Hydrolases From *Acidothermus*

Cellulolyticus. *National Bioenergy Center, Biotechnology For Fuels And Chemicals Division*. NREL, Golden, CO 80401.

11. Edgar, R. C. (2004). MUSCLE: Multiple Sequence Alignment With High Accuracy And High Throughput. *Nucleic Acids Research*, 32(5), 1792–1797.

12. Fontana, A. (1991). Analysis And Modulation Of Protein Stability. *Current Opinion In Biotechnology*. 2, 551-560.

13. Glasgow, E. M., Kemna, E. I., Bingman, C. A., Ing, N., Deng, K., Bianchetti, C. M., Takasuka, T. E., Northen, T. R., & Fox, B. G. (2020). A Structural And Kinetic Survey Of GH5_4 Endoglucanases Reveals Determinants Of Broad Substrate Specificity And Opportunities For Biomass Hydrolysis. *Journal Of Biological Chemistry*, 295(51), 17752–17769.

14. Goldstein, R. A. (2007). Amino-Acid Interactions In Psychrophiles, Mesophiles, Thermophiles, And Hyperthermophiles: Insights From The Quasi-Chemical Approximation. *Protein Science*, 16(9), 1887–1895.

15. Grant, B. J., Rodrigues, A. P. C., Elsawy, K. M., Mccammon, J. A., & Caves, L. S. D. (2006). Bio3d: An R Package For The Comparative Analysis Of Protein Structures. *Bioinformatics*, 22(21), 2695–2696.

16. Hansson, T., Oostenbrink, C., & Van Gunsteren, W. (2002). Molecular Dynamics Simulations. *Current Opinion In Structural Biology*, 12(2), 190–196.

17. Huang, J., & Mackerell, A. D. (2013). CHARMM36 All-Atom Additive Protein Force Field: Validation Based On Comparison To NMR Data. *Journal Of Computational Chemistry*, 34(25), 2135–2145.

18. Huser, B. A., Datel, B. K. C, Daniel, R. M., And Morgan, H. W. (1986). Isolation And Characterization Of A Novel Extremely Thermophilic Anaerobic *Chemoorganotrophic Eubacterium*. *FEMS Microbiology Letters*. 37, 121-127.

19. Jaeger, V., Burney, P., & Pfaendtner, J. (2015). Comparison Of Three Ionic Liquid-Tolerant Cellulases By Molecular Dynamics. *Biophysical Journal*, 108(4), 880–892.

20. Kamerzell, T. J., & Russell Middaugh, C. (2008). The Complex Inter-Relationships Between Protein Flexibility And Stability. *Journal Of Pharmaceutical Sciences*, 97(9), 3494–3517.

21. Kim, H.-W., & Ishikawa, K. (2011). Functional Analysis Of Hyperthermophilic Endocellulase From *Pyrococcus Horikoshii* By Crystallographic Snapshots. *Biochemical Journal*, 437(2), 223–230.

22. Lecun, Y., Bengio, Y., Hinton, G. (2015). Deep Learning. *Nature*, 521, 436-444.

23. Lobanov, M. Y., Bogatyreva, N. S., Galzitskaya, O. V. (2008). Radius Of Gyration As An Indicator Of Protein Structure Compactness. *Molecular Biology*, 42(4), 623-628.

24. Mccammon JA, Gelin BR, Karplus M (1977). Dynamics Of Folded Proteins. *Nature*, 267:585-590.

25. Melchionna, S., Sinibaldi, R., & Briganti, G. (2006). Explanation Of The Stability Of Thermophilic Proteins Based On Unique Micromorphology. *Biophysical Journal*, 90(11), 4204–4212.

26. Noe, F. & Nuske, F. (2013). A Variational Approach To Modeling Slow Processes In Stochastic Dynamical Systems. *Multiscale Model Simulations*, 11, 635-655.

27. Pace, C. (1975). The Stability Of Globular Proteins. *CRC Critical Reviews In Biochemistry*, 3, 1–43.

28. Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., … Schulten, K. (2005). Scalable Molecular Dynamics With NAMD. *Journal Of Computational Chemistry*, 26(16), 1781–1802.

29. Puhl, A. C., Rosseto, F. R., Stankovic, I., Alvarez, T. M., Squina, F. M., Skaf, M. S., & Polikarpov, I. (N.D.). Substrate Cleavage Pattern, Biophysical Characterization And Crystallographic Structure Of The Major Endoglucanase From *Xanthomonas Campestris Pv. Campestris*.

30. Puhl, A. C., Prates, E. T., Rosseto, F. R., Manzine, L. R., Stankovic, I., De Araújo, S. S., Alvarez, T. M., Squina, F. M., Skaf, M. S., & Polikarpov, I. (2019). Crystallographic Structure And Molecular Dynamics Simulations Of The Major Endoglucanase From *Xanthomonas Campestris Pv. Campestris* Shed Light On Its Oligosaccharide Products Release Pattern. *International Journal Of Biological Macromolecules*, 136, 493–502.

31. Rosseto, F. R. (2016). Biophysical And Biochemical Studies Of A Major Endoglucanase Secreted By *Xanthomonas Campestris Pv. Campestris*. *Enzyme And Microbial Technology*, 7.

32. Rothschild, L.J.; Mancinelli, R.L. (2001). "Life In Extreme Environments". *Nature*. 409 (6823): 1092–1101.

33. Scheraga, H. A., Khalili, M., & Liwo, A. (2007). Protein-Folding Dynamics: Overview Of Molecular Simulation Techniques. *Annual Review Of Physical Chemistry*, 58(1), 57–83.

34. Schmidhuber, J. (2015). Deep Learning In Neural Networks: An Overview. *Neural Networks*, 61, 85-117.

35. Schneider, E., Dai, L., Topper, R. Q., Drechsel-Grau, C., Tuckerman, M. E. (2017). Stochastic Neural Network Approach For Learning High-Dimensional Free Energy Surfaces. *Physical Review Letters 2017*, 119.

36. Serçinoğlu, O., & Ozbek, P. (2018). Grinn: A Tool For Calculation Of Residue Interaction Energies And Protein Energy Network Analysis Of Molecular Dynamics Simulations. *Nucleic Acids Research*, 46(W1), W554–W562.

37. Skjærven, L., Yao, X.-Q., Scarabelli, G., & Grant, B. J. (2014). Integrating Protein Structural Dynamics And Evolutionary Analysis With Bio3D. *BMC Bioinformatics*, 15(1).

38. Sun, Y., Cheng, J. J., Himmel, M. E., Skory, C. D., Adney, W. S., Thomas, S. R., Tisserat, B., Nishimura, Y., & Yamamoto, Y. T. (2007). Expression And Characterization Of *Acidothermus Cellulolyticus* E1 Endoglucanase In Transgenic Duckweed Lemna Minor 8627. *Bioresource Technology*, 7.

39. Vijayabaskar, M., & Vishveshwara, S. (2010). Comparative Analysis Of Thermophilic And Mesophilic Proteins Using Protein Energy Networks. *BMC Bioinformatics*, 11(1), S49.

40. Wang, Q., Tull, D., Meinke, A., Gilkes, N. R., Warren, R. A., Aebersold, R., Withers, S. G. (1993). Glu280 Is The Nucleophile In The Active Site Of *Clostridium Thermocellum* Celc, A Family A Endo-Beta-1,4-Glucanase. *Journal Of Biological Chemistry*, 268, 14096-14102.

41. Weber W, Hünenberger PH, Mccammon JA (2000). Molecular Dynamics Simulations Of A Polyalanine Octapeptide Under Ewald Boundary Conditions:

Influence Of Artificial Periodicity On Peptide Conformation. *Journal Of Physical Chemistry B*, 104:3668-3675.

42. Willey, Joanne M., Linda Sherwood, Christopher J. Woolverton, And Lansing M. Prescott (2008). *Prescott, Harley, And Klein's Microbiology*. New York: Mcgraw-Hill Higher Education, 2008.

43. Zheng, F., Tu, T., Wang, X., Wang, Y., Ma, R., Su, X., Xie, X., Yao, B., & Luo, H. (2018). Enhancing The Catalytic Activity Of A Novel GH5 Cellulase Gtcel5 From *Gloeophyllum Trabeum* CBS 900.73 By Site-Directed Mutagenesis On Loop 6. *Biotechnology For Biofuels*, 11(1).

44. Zuegg J, Gready JE (1999) Molecular Dynamics Simulations Of Human Prion Protein: Importance Of Correct Treatment Of Electrostatic Interactions. *Biochemistry*, 38:13862-13876.

VITA

Logan Sheffield was born in Houston, Texas in 1995. He attended elementary school in Alvin, Texas and graduated from Angleton High School in 2013. In 2018, he received a Bachelor of Science in Biology degree from Stephen F. Austin State University in Nacogdoches, Texas and entered the University's Biotechnology graduate program. During his time in grad school, he received awards for research presented at the ASBMB Experimental Biology Conference in San Diego, California and was employed as a biotechnology lab manager and graduate instructor on campus. In August 2021, he received the degree of Master of Science in Biotechnology from Stephen F. Austin State University.

Permanent Address: 1453 S Bluebonnet Ln
Angleton, TX 77515

Style manual designation: APA Style

This thesis was typed by Logan E. Sheffield