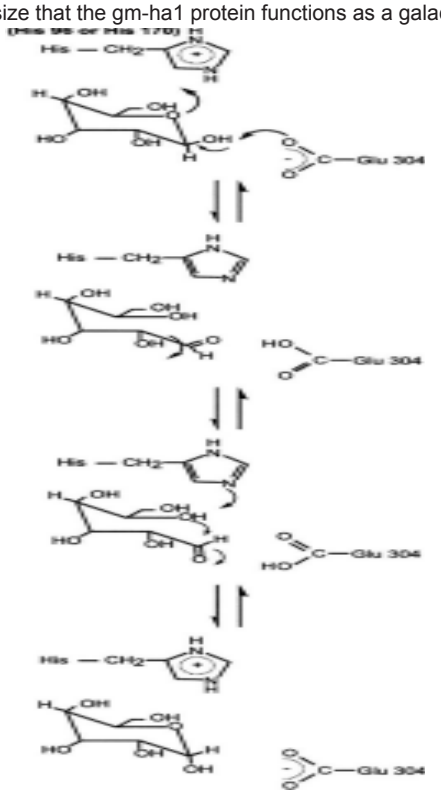
4RNL-A (gm-ha1) Structure and Sequence Shows Homologies to Galactose Mutarotase Enzymes

Haaris Ahmed, Bomby Ahuja, Dr. William Conrad, Leslie Gonzales, Kayenath Khan, Nathaniel Kregar, Karina Mora, Mathieu Norcross, Tate Rosenhagen, Max Smith, Sergio Sosa, Anthony Sullivan, and George Vladimirov
Lake Forest College
Lake Forest, Illinois 60045

Abstract
Gm-ha1 is a protein of unknown function with the PDB ID: 4RNL. This protein is found in *Streptomyces platensis*, which is a bacteria species that produces two highly effective antibiotics. Despite the protein being of unknown function, it is hypothesized that this protein functions as a galactose mutarotase. Galactose mutarotase, also known as an aldose-1-epimerase, is an enzyme that commits the first step in the metabolism of galactose. In order to investigate the function of gm-ha1, various bioinformatic methods were used. ProMOL revealed that gm-ha1 had a homologous active site to PDB: 1snz, which is a human galactose mutarotase enzyme. Moreover, Pfam showed that gm-ha1 had conserved residues with the aldose-1-epimerase family and likely had a similar sequence to other proteins in that family. Further, Dali showed gm-ha1 had a high level of global alignment with other aldose-1-epimerase enzymes. Besides, Autodock and PyMOL showed that NAD was most likely an important ligand in relation to gm-ha1 enzyme catalysis. Lastly, purification was performed on a different protein (c8orf32) and confirmed the presence of purified protein in the gel. However, the kinetics experiment of the same protein proved that the data was invalid and could not be used in this study. The bioinformatic data obtained in this study do not reject the hypothesis that gm-ha1's function as a galactose mutarotase enzyme. This was expected according to the data of the PDB profile of gm-ha1 (Tan et al. 2014). If the function of gm-ha1 is confirmed, it can give more knowledge of the *Streptomyces platensis* bacteria strain which can ultimately result in increased production of important antibiotics.

Introduction
The galactose mutarotase enzyme is a common protein found in both prokaryotes and eukaryotes. It is essential for the metabolism of galactose as it converts beta-D-galactose to alpha-D-galactose, which is the first step in normal galactose metabolism. This interchange can occur spontaneously in pure water *in vitro*, but organisms require the enzyme to perform this conversion *in vivo* because this reaction requires catalysis (Bouffard et al. 1994). In *Escherichia Coli*, galactose mutarotase is coded in the *gal* operon. This operon is induced by D-galactose, which indicates that the galactose mutarotase has some importance when D-galactose is present (Lee et al. 2008). Galactose mutarotase is essential for many prokaryotic cells as galactose is an abundant sugar and an excellent source of energy for bacteria. The PDB ID: 4RNL protein is found in *Streptomyces platensis* and is proposed to be a galactose mutarotase. However, this has not been confirmed as its function is unknown (Tan et al. 2014). It may prove beneficial to understand the function of the proteins in *Streptomyces platensis* as this species has a very important use in society. *Streptomyces platensis* is responsible for producing two key antibiotics, platensimycin and platencin. These two antibiotics have been shown to be effective against bacteria strains, such as MRSA and Streptococcus pneumoniae, that have developed resistance to many antibiotics (Falzone et al. 2017). Some researchers experimented and found a way for *Streptomyces platensis* to overproduce these two antibiotics tenfold the normal amount (Smanski et al. 2009). Learning how to overproduce antibiotics is a huge benefit to the medical field as antibiotics are in high demand, especially potent ones such as platensimycin and platencin (Smanski et al. 2009). A study that endeavored to determine the nutrients that increase *Streptomyces platensis* growth used a medium with high levels of glucose and lactose, which had growth of the bacteria on it. This indicates that *Streptomyces platensis* most likely utilizes galactose in its metabolism and growth. Therefore, determining the function of this protein may prove beneficial for providing media for culturing *Streptomyces platensis*. The protein was not named, so in this manuscript, the protein will be called gm-ha1. However, this study focused on the 4RNL-A chain for all sections of this report, so this name will be used when necessary. The original gm-ha1 study used *E. coli* which is one of the most common bacteria used for biochemical studies. One experiment had cloned the mutarotase gene of *Acinetobacter calcoaceticus*, which is another type of bacteria. This was performed by a complicated process of purifying the proteins, creating probes for the gene, cloning the gene, inserting the gene into a plasmid, and inserting the plasmids into *E. coli* (Gatz et al. 1986). In theory, this same process should be able to be performed on *Streptomyces platensis* to produce the same resulting *E. coli* with the mutarotase gene present. Another study that examined the mechanisms of galactose mutarotase had identified the amino acids and their positions that are crucial in performing the mutarotase functions. For example, it identified that Glu 304 and His 170 are the key peptides for catalysis (Thoden et al. 2003). The sequence of the 4RNL structure is listed in the PDB, so perhaps the sequence can be examined to see if the key peptides listed in this study are present in the same location in gm-ha1. We hypothesize that the gm-ha1 protein functions as a galactose mutarotase enzyme.



Schematic 1. This is the catalytic mechanism of galactose mutarotase. It is therefore the proposed mechanism of the gm-ha1 protein. This image shows the conversion of beta-D-galactose to alpha-D-galactose. This schematic was obtained from (Thoden et al. 2003).

Methods and Materials
PyMOL/ProMOL
ProMOL was used for examining the active site homologs for protein gm-ha1 as a plugin for PyMOL. 4RNL-A was examined by using the Motif Finder in ProMOL and searching"4RNL" in the query box. The template libraries selected were A set and P set. A list of structural homologs came up and the protein that had the lowest RMSD value was selected because this was the closest active site structural match. Important data were obtained such as the Levenshtein distance, EC class, and the RMSD value. This allowed a closer look at the homology of the two proteins because a Levenshtein distance of zero means that all of the residues of the active site homolog are found in the 4RNL-A. A low RMSD value means that the distance between the alpha and beta carbons are very close between the two structures. These values were all considered in finding the closest active site alignment for gm-ha1. The full protocol is located in the student guide in the Labarchives notebook at mynotebook.labarchives.com.
BLAST and Pfam
BLAST and Pfam were used to compare proteins with similar sequences to gm-ha1 to further the examination of the function of these proteins. For BLAST, the graphic summary and alignment were looked at and for Pfam, the HMM logo and the alignments were used. For BLAST, the FASTA sequences of the POIs were searched in Protein-Protein BLAST to find the top "hits" of the proteins, or the proteins that had the best alignment scores. The E-values and percent identities of the top three hits for each POI were recorded. These matches can provide insight into the function of the POIs as the three matches have the function of the protein in the name, so this allows easy comparisons. A SmartBLAST search was used to compare the POIs to proteins in other organisms and the E-values and percent identities were

also compared to determine whether the other proteins were good matches. For Pfam, the sequence search option was used. This was performed by pasting in the FASTA sequence for the POIs. This search gave the superfamily the POIs most likely belonged to, which was selected to examine the HMM (Hidden Markov Model) logo graph. The HMM logo was used to determine highly conserved amino acids, as it is a statistical algorithm that predicts the sequence based on previous data, which can be compared to the alignment of the sequences of the POIs. Peptides that appear large on the HMM logo would be highly conserved, so a small section would be chosen in the HMM logo to compare amino acids residues. In the search results page of Pfam, there is an option to hide/show alignment, so this would be selected to show the alignment, and the area selected from the HMM logo would be searched to determine whether the POI sequence showed these highly conserved amino acids. This is supporting evidence that the POIs belong to the families that they were listed in. The full protocol is located in the student guide in the Labarchives notebook at mynotebook.labarchives.com.

DALI

Since protein structure is key to function, Dali was used to visualize and quantify the global alignments of the POIs and backbone structural homologs to give insight into the function of the POIs. The PDB search was used by entering the PDB ID and chain letter of the POIs. This search was performed for "4RNL-A". When the search was finished, the "matches against full PDB" option was selected to observe the alignments against the entire POI structure. This option was selected because it compares the structure of the entire backbone of the protein minus the side chains. To view the structure comparisons, the "3D superimposition" was selected and the cartoon image option was used for the screenshots in the table (Table 1). The active site residues were found on PDB and a few peptides in sequence were used to compile a search of the POI sequence active site against the structural homologs in order to determine whether the active sites were conserved among the homologs. These results, alongside the z-score, RMSD, the length of the alignment (Lali), Nres, and % ID were all used to determine whether the homologs were actually good fits compared to the POI. Alignment and fitness can be determined through analyzing z-scores and Lali values, where greater values of either measure reflect greater alignment. Examining the functions of the good-fit homologs gives insight into the function of the POIs despite the fact that the side chains were not included in this search. The three proteins that were compared to gm-ha1 were 1SNZ-B, 1NSS-A, and 1SO0-A. The full protocol is located in the student guide in the Labarchives notebook at mynotebook.labarchives.com.

Autodock

This lab was started by first downloading a .pdb file of the POIs and loading the molecules in PyRX. These molecules were made to be macromolecules, which converts the file into a .pdbqt, which is required for Autodock. The ligands were found by searching the EC class of the POIs (5.1.3.3) in PDB and selecting the unique ligands that were larger in size than single molecules or very small structures. The five specific ligands chosen for this lab were 4QQ, MID, NAD, MIT, and MKY. These ligands were downloaded as one "ideal SDF" file as this is compatible with PyRX. This file was imported to PyRX, and the ligands were all minimized and were converted to .pdbqt files to correctly autodock the ligands to the protein. The ligands were docked using Vina Wizard in PyRX by first clicking start under Vina Wizard. The ligands were selected, and the forward option was selected. Then, "analyze results" was selected to provide a list of the ligands and their RMSD values to determine the best fit mode for each ligand, which was mode 0 for all ligands. The ligand NAD_A_352 was selected for visualization because it had a high binding affinity. The protein with the docked ligand was saved as a .pdb file to import it into PyMOL to compare the location of the ligand binding site to the active sites of the POIs. In PyMOL, the visualization settings were set to lines for the 4RNL-A structure and sticks for the ligand. The surface setting for 4RNL-A did not work because the ligand could not be seen. To compare the active site and ligand binding site, the 4RNL-A had everything hidden except for the two active site residues and the ligand, which were both as sticks. This visualization allowed for easy comparison of ligand binding site and the active sites to determine if the ligands may play a role in catalysis of the reaction or is a substrate. The full protocol is located in the student guide in the Labarchives notebook at mynotebook.labarchives.com.

Protein Purification

Column purification is used to isolate a purified protein of interest, which allows the protein to be manipulated for future experiments. The purification used in this experiment followed the Hook™ G-Biosciences protein purification spin column protocol. The bacteria were lysed with Bacterial PE LB™ and *PE LB*™-Lysozyme so that the 6X His tagged proteins were accessible. The tagged proteins were purified using immobilized metal affinity chromatography (IMAC). 0.4 mL of immobilized metal affinity resin is added to the protein lysate. Then, the resin is transferred to a spin column, where the His tagged proteins were washed twice with Tris-NaCl and eluted three times with the imidazole buffer. The bacteria before lysis, the protein lysate, the flow through, the two washes, and the three elutions were all run through a gel to determine if the protein of interest was correctly isolated and purified. The full protocol is located in the student guide in the Labarchives notebook at mynotebook.labarchives.com, and these methods were performed by Dr. Will Conrad.

Protein Kinetics

First, the POI elution samples were dialyzed into a 50 mM sodium phosphate buffer of neutral pH (7.6) using the Pierce 3kDa MWCO centrifugal filter concentrator. It was spin-concentrated three times using a 20 mL buffer per 1 mL elution. Protein abundance was measured using absorbance at 280 nm and 1 cm path length. This was converted in mg/mL using Beer's Law. Each well was filled with 160 uL of 50 mM sodium phosphate buffer, X uL (0-20) of 10 mM para nitrophenol acetate (PNPA) and Y uL of acetonitrile so that X and Y equal 20 uL total. Baseline absorbance was read at 405 nm to determine product concentration without enzymatic activity. Then, 20 uL of enzyme was added to each well and absorbance was read every 41 seconds at 405 nm. The full protocol is located in the student guide in the Labarchives notebook at mynotebook.labarchives.com and these methods were performed by Dr. Conrad.

Results

ProMOL search indicates active site homolog

Active site homologs give key insights into protein function since protein activity takes place in the active site. The active site of gm-ha1 was searched in ProMOL to find and compare active site homologs of gm-ha1. Figure 1 shows the closest active site alignment where there is almost complete overlap between gm-ha1 and 1SNZ. It had a very low RMSD value in ProMOL of 0.12 for the alpha and beta carbons in the structure and it had the same EC as gm-ha1. The 1SNZ active site is homologous in structure to the active site of gm-ha1. 1SNZ is a human galactose mutarotase enzyme, which indicates that gm-ha1 is homologous to a human galactose mutarotase.

Pfam sequence comparison shows conserved residues in aldose epimerase protein family

While local active site alignment is key to protein function, examining the protein sequence is also an important aspect of determining protein function. Comparing a protein sequence to a family of proteins can indicate the function of said protein may be similar to the protein family. The sequence of gm-ha1 was searched in Pfam and the HMM logo graph was used to compare the gm-ha1 sequence with the predicted sequence of the HMM logo graph. Since the suggested function of gm-ha1 is a galactose mutarotase enzyme, it would be expected that gm-ha1 would be a part of the aldose 1-epimerase family, as the galactose mutarotase enzyme is a type of aldose 1-epimerase. In Figure 2, the sequences in the bottom image show that the HG residues are 5 peptides before the uncertainty line and the W residue is 3 Peptides after the uncertainty, so the highly probably amino acids in the aldose 1-epimerase family are conserved in the correct placements. This indicates that the sequence of gm-ha1 is homologous to proteins in the aldose 1-epimerase family.

DALI data shows good global alignment with Aldose 1-Epimerase enzymes

While sequence comparisons are important in investigating protein function, global structure is arguably the most important aspect of protein function. DALI shows global alignment of proteins both visually and statistically, which is a powerful tool in the investigation of proteins of unknown function. The data from Pfam gives the indication that gm-ha1 will most likely match closest to aldose 1-epimerase and galactose mutarotase enzymes. The data from DALI confirms this as the three closest matches were all aldose 1-epimerase and galactose mutarotase enzymes. The image capture from Table 1 shows very close 3D alignment between the gm-ha1 among the three proteins and this is confirmed with 1SNZ-B and 1So0-A both showing the highest z-score of 53.6 and a very low RMSD of 1.0. Both active site residues were also conserved in all three homologs, so gm-ha1 shows very close global structural alignment for aldose 1-epimerase enzymes.

Autodock and PyMOL shows NAD ligand binding near the active site

Ligands with high binding affinities near the active site indicate mol-

ecules that are most likely essential for enzyme catalysis. PyRX was used to determine the binding affinity of the ligands to find tightly binding ligands as these typically play an important role in the function of an enzyme. PyRX gave numerical and visual data when searching the ligands and docking them to the POIs. Autodock revealed that the NAD_A_352 ligand was bound in the 4RNL-A structure with a high binding affinity. This high binding affinity indicates that this ligand may play an important role with gm-ha1, such as a substrate or a cofactor. PyMOL was used to visualize the binding site of the ligand with the protein structure and compare the location of the ligand binding site with the active site residues. The protein and the ligand were imported to PyMOL, and the active site residues were highlighted. Since the ligand has a high binding affinity, it is expected that the ligand most likely plays a role in the enzyme (gm-ha1) catalysis. Therefore, it is expected that the ligand should be bound near the active site. Figure 3 shows that the ligand is bound very closely to the active residues which further indicates that the ligand plays a role in the enzyme function. Therefore, the ligand is most likely a substrate or cofactor that supports the enzymatic reaction.

Nickel His purification yielded purified c8orf32 protein

Moving away from bioinformatics, proteins of interest were purified to obtain the pure protein. In order to determine the function of a specific protein, it must be purified and isolated. The POI was purified using the HOOK 6X His Protein Spin Purification Kit. The molecular weight of the c8orf32 protein with the Maltose Binding Protein tag was 65,971 kDa. The black circle in Fig 4 shows the protein of interest in the gel. These lanes were the three elutions, which is where the protein is expected to be present. The protein bands are also between the 75 kDa and 50 kDa markers, which is expected because the weight was 65,971kDa. The elutions contain the protein of interest with some other bands that are most likely impurities. There are also some impurities present in the final wash, so there were most likely still some impurities still present in the column. In ideal condition, there should have been one band of pure protein present. Therefore, the c8orf32 protein was present and purified in the elution, however, there were some impurities present in the gel.

Protein kinetics data indicates invalid elution sample of c8orf92

Continuing on after purification, eluted protein samples were tested to determine if the protein is present. Protein kinetics data is used to determine if an eluted sample has a specific protein of interest by monitoring whether it catalyzes a reaction and shows product formation. This was performed using absorbance vs time, then plotting the data with Michaelis-Menten and Lineweaver-Burk plots. The Michaelis-Menten graph (Fig 5) failed to show a burst phase as the data was linear from 0-10 on the x-axis. The data then showed a peculiar curve and a possible plateau, but not likely valid data. The Lineweaver-Burk plot (Fig 5) showed a negative y-intercept, which means that the data is invalid. It is invalid because a negative y-intercept of -0.0171 which would yield a negative Vmax. This would lead to other variables, such as Km and Kcat being negative, which is not possible. Therefore, this enzyme data is not valid and cannot be used for protein c8orf32.

Discussion and Conclusion

This study coincided with the expected results for each bioinformatic section that examined gm-ha1. ProMOL revealed that the A chain of gm-ha1 had a homologous active site to 1snz, which is a human galactose mutarotase. This supports the hypothesis that gm-ha1 is a galactose mutarotase. The active site is where the enzyme catalysis takes place, so having a homologous active site indicates a very similar, if not the same reaction taking place. However, only having a similar active site does not confirm a relation between the homologies. For the sequence of gm-ha1, the HMM logo graph from Pfam highlighted that the gm-ha1 protein sequence had conserved some common residues in the same location as proteins of the aldose-1-epimerase family. This also coincides with the hypothesis of gm-ha1's function as a galactose mutarotase because some common residues in the sequence are the same. However, this section only examined three residues among an entire polypeptide sequence, so this is very little data in comparison to the entire sequence. Dali had compared the global alignment of gm-ha1 and showed that the closest matches were all aldose-1-epimerase enzymes. This also supports that gm-ha1 is a galactose mutarotase enzyme because its overall structure is most similar to aldose-1-epimerase enzymes. Since structure determines protein function, this is sound support for gm-ha1's hypothesized function. Autodock was used to find ligands that bound to the A chain of gm-ha1 and NAD was found to be a high affinity ligand. It was confirmed in PyMOL that NAD also binds very close to the active sites, so NAD is most likely a key ligand in the catalysis of gm-ha1. The purification and kinetics experiments were not of gm-ha1, but of different proteins. However, the proteins used in this study could be substituted with gm-ha1 to obtain the corresponding data for this POI. The purification experiment confirmed the presence of purified c8orf32 protein in the gel, but the kinetics experiment data was invalid and could not be used in this study. This may have been due to impurities present in the protein elution, which could have hindered the protein activity. Since the purification and kinetics experiments were not able to be performed for gm-ha1, this would be a reasonable future path to take in determining the function of gm-ha1. With purifying the protein, one experiment that could be performed is to put an eluted sample of the protein in a solution of gel that contains beta-D-galactose. I hypothesize that gm-ha1 is a galactose mutarotase enzyme, so there should be alpha-D-galactose product forming if my hypothesis is correct. I believe that this would be a conclusive study on the function of gm-ha1 when combined with the bioinformatic data provided by this study. For the bioinformatics, only the A chain was used for searches in this study, so repeating the same bioinformatics for the other three chains would help confirm the results of this study. This study was performed remotely due to COVID-19, so the gm-ha1 protein was never physically experimented on. All of the experiments were bioinformatic, or a different protein was experimented on by Dr. Conrad. This limits the scope of this study as the protein was never physically worked with. However, this study provides multiple bioinformatic sources that support the hypothesis of gm-ha1's function. This study was aiming to better understand a protein in the *Streptomyces platensis* bacteria species, as this is an important bacteria strain in producing antibiotics. It is still inconclusive whether the gm-ha1 protein is a galactose mutarotase, but the data from this study gives both structural and sequential indications that gm-ha1 is a galactose mutarotase enzyme.

Figures

ProMOL depicts good active site alignment with 1SNZ motif

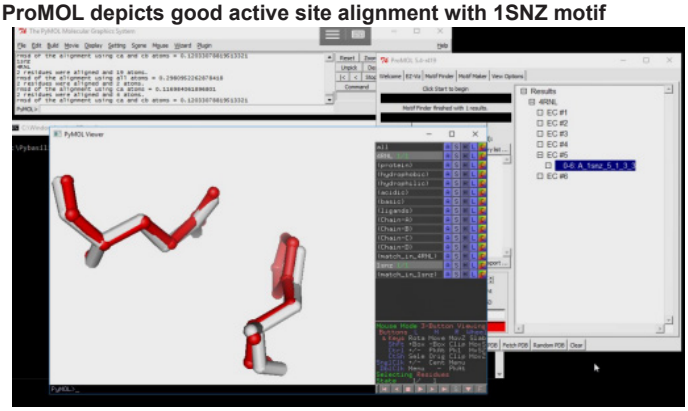


Figure 1. This image is a screenshot from ProMOL that shows the active site structural alignment of 4RNL-A (red) and the active site homolog motif, 1SNZ (white). The image on the right shows the EC of 1SNZ (5.1.3.3) as well as the Levenshtein distance (0-6).

Pfam shows conserved amino acid residues of gm-ha1 among the aldose epimerase protein family

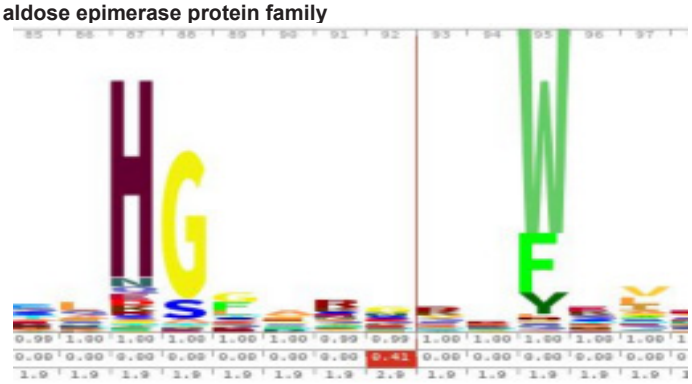
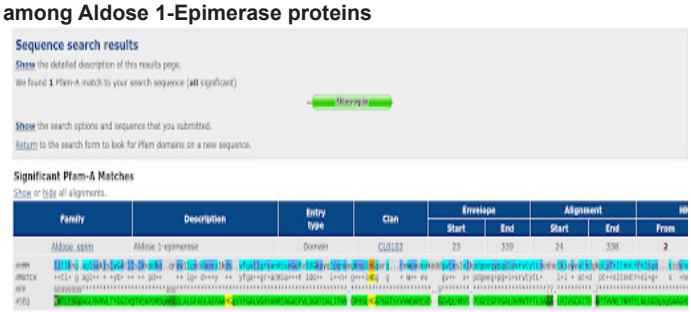
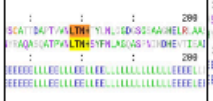
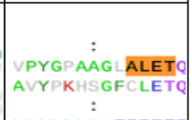


Figure 2. The top image is of the HMM logo from Pfam for the aldose epimerase family (PF01263) where the larger the letter, the higher probability that the amino acid residue is present at that location. The red line indicates uncertainty in the sequence. The bottom image compares the sequences of the HMM and the sequence of gm-ha1. The highlighted section was the result of searching for the consecutive HG peptides as seen in the HMM logo.

Dali shows good global alignment and conserved active residues among Aldose 1-Epimerase proteins



PyMOL shows NAD ligand in 4RNL-A near active residues

Chain/ Description	1SNZ-B Aldose 1-Epimerase	1NSS-A Galactose Mutarotase	1So0-A Aldose 1-Epimerase
Image Capture (Cartoon with Structure Conservation View)			
Active Site: Histidine (H)-182 LTNH			
Active Site: Glutamate (E)-308 ALET			
# of active site match	2/2	2/2	2/2
Good Fit? (Y/N)	Yes	Yes	Yes
Z-Score	53.6	44.3	53.6
RMSD	1.0	1.6	1.0
LALI (length of alignment)	333	319	335
NRES (Number of residues)	342	339	344
%ID	44%	29%	44%

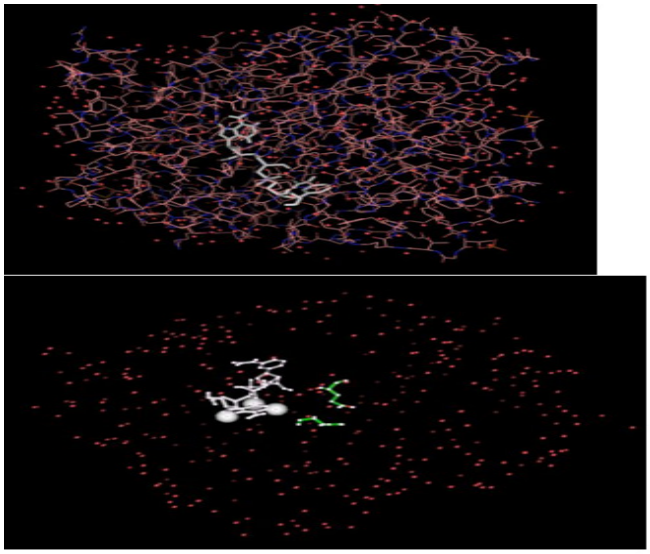


Figure 3. These images were screenshots from PyMOL after importing the ligand from Autodock. The left image shows where the NAD ligand is bound in the 4RNL-A structure where the 4RNL is shown in red and blue lines and the NAD ligand is shown in white sticks. The image on the right shows the same ligand, but the 4RNL-A structure is hidden, except for the two active sites residues of gm-ha1, shown in green. The top residue is glutamate, and the bottom is histidine.

Nickel His column purification shows presence of purified POI

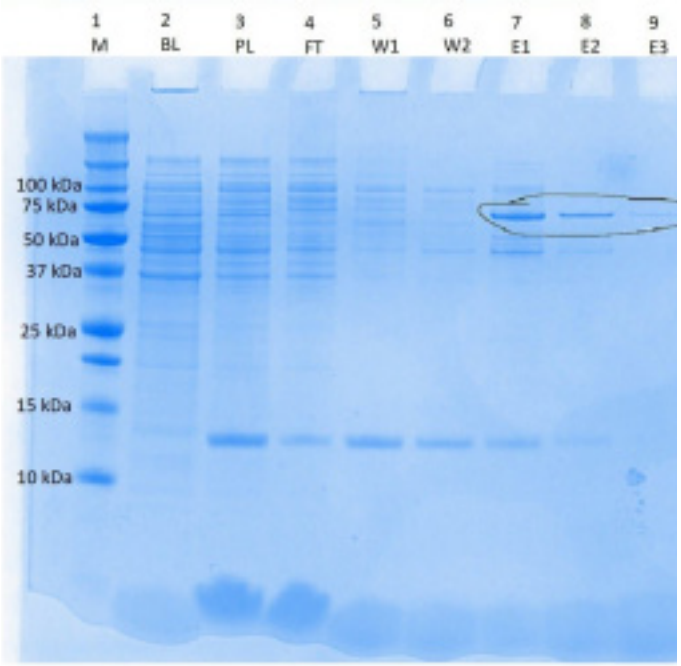


Figure 4. The proteins in this gel were purified using the HOOK 6X His Protein Spin Purification Kit. The first lane is the marker with relevant weights labeled. The second lane was the bacteria before lysis. Lane 3 was the protein lysate and lane 4 was the flow through. Lanes 5 and 6 were the two washes. Lanes 7, 8, and 9 were the elutions, and the dark bands in the circle are c8orf32.

Protein kinetics plots for c8orf32 protein show invalid data for

c8orf32 kinetics

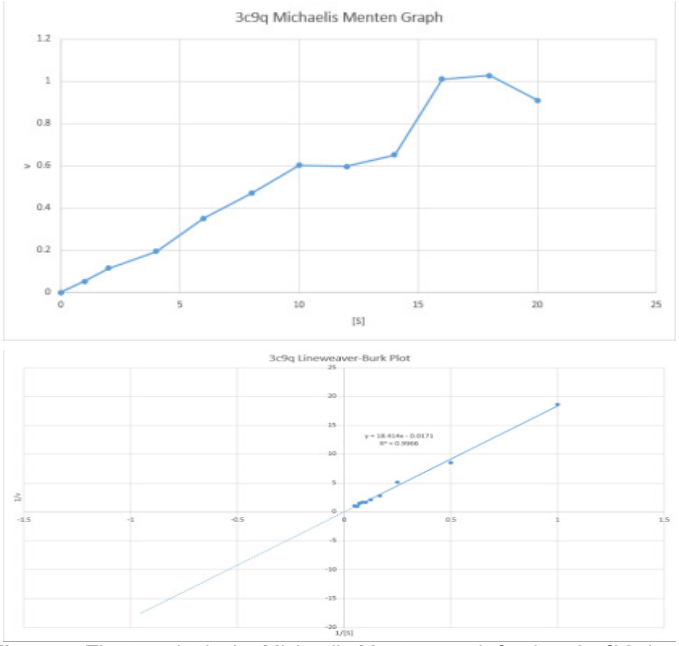


Figure 5. The top plot is the Michaelis-Menten graph for the c8orf32 data, generated using Excel. It depicts the velocity vs substrate concentration from 0-20 uL of substrate. The bottom plot is the Lineweaver-Burk plot for the same c8orf92 data from the Michaelis-Menten graph. The equation of the best-fit line of the data is shown in the graph above the trendline. Inverse velocity vs inverse substrate concentrations were the axes in this graph.