Determining the Underlying Distributions of Change in Free Energy Change for Pathogenic and Benign Protein Mutations

Jorden Thompson[1] (Honors Capstone), Jaie Woodard[1,2], Sriram Chandrasekaran[1]

1. Department of Biomedical Engineering, University of Michigan

2. Department of Computational Medicine and Bioinformatics, University of Michigan

Abstract

A mutation in a patient's genome can affect a protein in that patient's body, resulting in either no change in the health of the patient or a disease experienced by the patient. Assigning terminology, the mutations can therefore be referred to as benign or pathogenic, respectively. When these benign or pathogenic mutations occur, there is an associated change in change in free energy (ΔΔG) when the protein folds, which essentially means the act of the protein folding can become more or less stabilizing. The questions we were interested in are the following: are pathogenic protein mutations stabilizing or destabilizing when compared to benign protein mutations and is there a difference between ΔΔG distributions for benign and pathogenic mutations. In order to analyze the distribution of the ΔΔG's, we looked at both data from a previous study and data obtained from an extensive literature search for pathogenic mutations found in patients who exhibit a disease. We found that there appears to be a statistical difference between the distribution of benign ΔΔG's and pathogenic ΔΔG's when organizing proteins by general function and that pathogenic mutations appear to be more destabilizing than benign mutations. Furthermore, pathogenic distributions appear better described by two gaussians, or a bimodal distribution, whereas benign distributions are adequately described by a single gaussian. Pathogenic distributions also appear to have greater range and variance. While the causes are not yet entirely understood, these results can play a role in understanding what, if any, role ΔΔG has on the pathogenicity of a mutation and could be one day used alongside other methods to generate a model that can help predict the pathogenicity of an arbitrary mutation.

Introduction

Many diseases that are prevalent in society are caused by protein mutations. Protein mutations begin at the DNA level when a nucleotide that helps encode a specific protein is affected. There are regions of DNA called exons and some of these regions contain the genetic code required for cells to produce whatever proteins they may need to properly function. In order for a cell to be able to produce a protein, the protein encoding section of a cell's DNA must first be translated into RNA, and then the RNA sequence can be converted into a chain of amino acids that will fold into a structured protein. Both the DNA and RNA sequences are made up of nucleotides, and, when coding for a protein, each group of three nucleotides specifies a different amino acid (the building blocks of proteins). There are various things that can happen to these nucleotides, namely a nucleotide can be deleted, added, or swapped out for a different nucleotide, all of which are considered DNA mutations. These three specifically mentioned mutations are known as deletions, insertions, and base substitutions, respectively.

Depending on the type of mutation that occurs and if the mutations occur in a protein encoding region, there can be downstream effects on the proteins produced from that region of the DNA. There is the chance that the mutation can have no effect, meaning that changing a nucleotide did not change the amino acid called for at that point in the sequence. This type of mutation is called a silent mutation, meaning the mutation would go unnoticed to the patient. However, there is also the chance that the mutation can affect the structure and/or stability of the protein by changing one or

multiple amino acids that are called for by the groups of three nucleotides in the DNA sequence. These are called missense mutations. If the DNA mutation that occurred causes one of the groups of nucleotides in the encoding region to call for a stop codon, then protein production would prematurely end resulting in an incomplete protein. It is important to note that in the case of deletion or insertion DNA mutations an entire chain of amino acids can be changed because inserting or deleting a single nucleotide causes the entire coding region to be offset. However, base substitution DNA mutations only cause a single point amino acid missense mutation since only one nucleotide is being altered, meaning there is no offset. Our study focused solely on single point amino acid missense mutations.

It may be important to understand that there are multiple reasons that these mutations may be present in a patient. Primarily, these mutations can be passed genetically. Patients with a pattern of protein mutation specific diseases in their family history either have those diseases or are at a genetic predisposition for those diseases. Secondly, these mutations can actually just occur by random chance. This happens during the DNA replication process. Although these replication errors do not have a high probability of occurring (a single error occurs, on average, every 100,000 nucleotides), when considering the amount of DNA replication that occurs in the human body they become unavoidable [1]. While the human body has checks in place to account for and correct almost all of the errors that may occur during the replication process, some of these errors fall through the cracks and can result in disease. Thirdly, radiation or chemical factors, known as mutagens, can interact with DNA and physically change its structure. These mutagens can come from many different places, including the environment, food, or viruses, to name a few. These are typically the causes that come to mind when people think of diseases such as cancer. Similarly to the DNA replication errors, the human body has systems to help repair DNA when breakage or mutation occurs from mutagens, but not all of them are caught.

As mentioned, these mutations in the DNA can cause diseases because changing the DNA nucleotide sequence can change the amino acids that are called for when the DNA is read to build a protein. If the amino acids that make up a protein change, then it is possible that the protein will physically not be able to fold the same or have a different stability. This is because different amino acids are composed of different atoms and are of different sizes, meaning that the site of change would not fit as the normal sequence did and would not interact with the rest of the chain as the normal sequence did. This process is depicted in Figure 1. The way a protein folds largely determines its function since it is the orientation

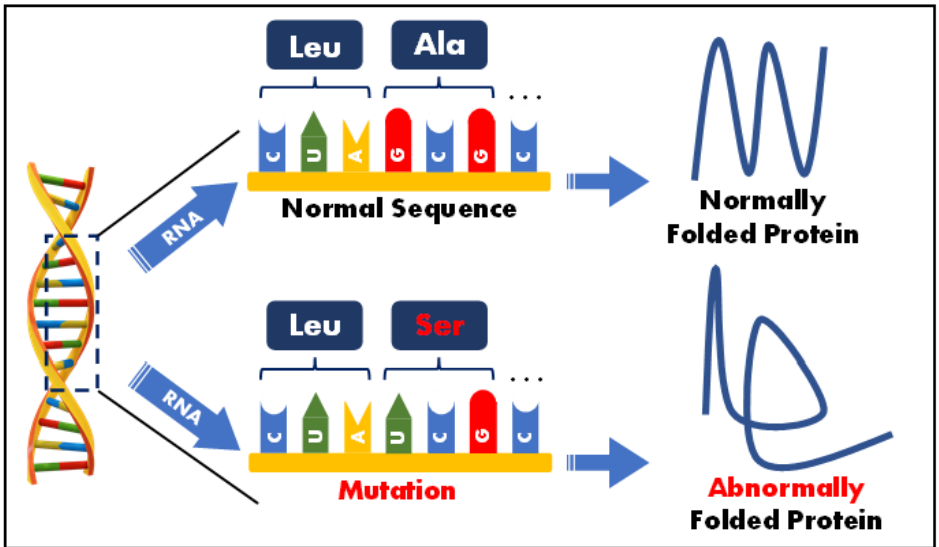

Figure 1: A simplified schematic depicting an example of a DNA base substitution mutation and the downstream effects that mutation can have on RNA translated from that DNA sequence and further protein generation from that RNA. In the normal nucleotide sequence, the fourth position in the DNA translates to a guanine (G) in RNA which results in the second amino acid shown to be alanine (Ala). In this hypothetical scenario, the resulting protein folds normally. In the mutated nucleotide sequence, the fourth position in the DNA has become a thymine (T) (translating to an uracil (U) in RNA) which causes the resulting amino acid to become a serine (Ser). Changing the sequence of the protein causes it to fold differently. Note that in this scenario the mutation is not a result of the translation to RNA but is occurring at the DNA level.

and location of the various atoms in the protein that determine how the protein interacts with the environment and other molecules around it. Thus, if the protein is no longer folded the same, there is a chance that it can no longer properly interact with its targets, which is the cause of many diseases. Mutations in the DNA that cause proteins to change such that a disease occurs are known as pathogenic mutations. There are, of course, many mutations that can occur in the DNA that are not severe enough to change a protein enough to cause diseases. These mutations are known as benign mutations.

A chain of amino acids has an associated free energy, meaning that in the unfolded state a protein has a certain potential to do work. For our purposes, this is really just to say that an unfolded protein has a certain amount of energy associated with it. As with any system, a protein folding is an attempt to minimize enthalpy (heat) and maximize entropy (disorder). The enthalpy is typically decreased because as a protein folds new bonds and interactions are made between the atoms in the amino acid chain, and when these bonds are made energy (typically in the form of heat) is released. It is entirely possible that in order for a bond to be made energy is required, but this would not occur spontaneously and is beyond the scope of this introduction. Furthermore, as these new bonds and interactions are made, water molecules that were previously bonded to the amino acid chain are displaced and released which increases the overall entropy of the system. Both the resulting change in enthalpy and change in entropy are summarized as an overall change in free energy, referred to specifically as a change in Gibbs free energy and shorthanded as ΔG. If ΔG is negative, then the reaction is spontaneous because overall energy was released. This must be true because for ΔG to be negative it means that whatever system we are looking at started with more energy than it ended up with. Energy cannot be created nor destroyed, so energy was released from the system. Furthermore, something having a lower energy is more stable and more favorable than having a higher energy, which is why a negative ΔG is said to be spontaneous. If the new state has less free energy than the old state (meaning ΔG is negative), the system will naturally move in the direction of that state (the energy minimizing state). To summarize, if ΔG is negative then the reaction is said to be spontaneous and stabilizing. From similar logic, if ΔG is positive the reaction is said to be non-spontaneous and destabilizing since it means energy must have been put into the system and the system is at a higher energy state. The discussion on the thermodynamics behind the folding of a protein can be much more in depth, but for the understanding of this study nothing more than what has been discussed needs to be understood.

We have thus introduced that an unfolded protein has an associated energy, and this energy changes when a protein folds. Additionally, a folded protein is typically of lower energy and more stable than an unfolded protein. We have also established that when a mutation occurs the result can be that the protein no longer folds in the same way. What this means is that when a mutation occurs there can be a change in the change in energy (a $\Delta\Delta G$) that a protein experiences when folding since the protein may no longer fold the same. More concisely, when a mutation occurs there is a $\Delta\Delta G$ when the mutated protein folds compared to the wild type protein folding. A $\Delta\Delta G$ value is calculated by subtracted the energy change from the initial fold, known as the wild type, from the energy change from the new fold that results after the mutation occurs. Depending on the exact details of the mutation, this $\Delta\Delta G$ can either be positive, negative, or zero. Recall, as mentioned earlier, that energy changes from protein folding are typically negative since they occur spontaneously. If the $\Delta\Delta G$ is zero, that just means that from an energetic standpoint nothing has changed with the fold, i.e., there is no difference between the mutated fold and the wild-type fold. However, if the $\Delta\Delta G$ is positive, then that means that the mutated fold is less negative than the wild type, meaning that the mutated fold is less stable than the wild type. Thus, a positive $\Delta\Delta G$ indicates that a mutation was destabilizing. Following a similar logic, if a $\Delta\Delta G$ is negative it means that the mutated fold is more negative than the wild type. Thus, a negative $\Delta\Delta G$ indicates that a mutation was stabilizing. Another thing to consider is that the change in free energy, ΔG, can also be thought of as a measure of the ratio of the folded and unfolded populations of the protein [2]. Thus, if the free energy of a protein changes the relative populations of folded to

unfolded proteins to each other will change. Specifically, if ΔΔG is negative the population of the folded protein to the unfolded protein will increase, and the opposite is true for a negative ΔΔG [2]. Either of these situations could be bad for the human body and result in disease.

The questions that we are interested in are twofold. Firstly, are pathogenic mutations stabilizing or destabilizing when compared to benign mutations? Secondly, is there a statistical difference between the underlying ΔΔG distributions of pathogenic and benign mutations? The data obtained from analyzing these questions could potentially, and hopefully, be used to create a model that can predict pathogenicity for a given protein and mutation pair. Knowing the answers to or thinking about these questions can be especially helpful in fields that are currently working with protein sequences. Such fields include gene therapy, medicine design for protein therapeutics, various tissue engineering or biomaterial fields, or protein modeling in other research settings. In all of these applications, having access to a model that can help predict pathogenicity of a protein mutation for a given protein would be very beneficial. It can also potentially lead to the discovery of new diseases and give researchers a new lens with which to develop treatments for various diseases or analyze protein mutations.

Methods

The change in energy that a protein experiences when folding can be computationally predicted. There are various programs available for this, including a program called FoldX and a relatively novel program developed by the Yang Zhang Lab of the University of Michigan, EvoEF2 [3], [4]. For this project, EvoEF2 was utilized to conduct protein folding modeling but some data from a previous study that utilized FoldX was also analyzed. However, this will be discussed later. For now, it is important to understand that these programs allow the input of a file containing information about a protein, including its structure, for example. Alongside the protein, the programs can also take in a specific mutation at a specific location in the protein. In order to calculate a ΔΔG, the change in energy of the wild-type protein is first calculated, and then mutations can be simulated and new changes in energy can be calculated. Then, as mentioned earlier, the ΔΔG for a given mutation can be calculated by subtracting the change in energy that occurs in the wild-type protein fold from the change in energy that occurs after the mutation has been simulated. Note that the units for ΔΔG are kilocalories per mole.

This process was done for a large number of pathogenic and benign mutations in a previous study by the Yang Zhang Lab, and the results were compiled into a database called ADDRESS [5]. ADDRESS contained ΔΔG calculations both from EvoEF and FoldX. The first part of this study involved analyzing the ΔΔG calculations from ADDRESS. This analysis was conducted by first organizing the data by function by utilizing GOnet [6]. In particular, GOnet organizes the proteins by function, and it is important to note that a particular protein may fall into more than one category on GOnet. Once the data was organized by gene function, the pathogenic data points and benign data points for each function were plotted as separate histogram distributions and boxplots. Next, gaussian curves of degrees one and two were fit for each distribution and various statistical tests were performed for each distribution.

The datapoints compiled in ADDRESS were obtained from the UniProt Humsavar database (version 2020_04) [5], [7]. Thus, it was of interest to us to gather pathogenic mutations reported in literature from patients with diseases and determine their ΔΔG to compare them to benign mutations of the same protein. In particular, the literature search was framed around specific protein function classes, and data points were gathered from patients with diseases that stem from mutations in a protein that belong to one of the protein function classes of interest. The specific diseases were not as important for this study as was the specific protein, its class, and the mutation. For each mutation found, the information that was gathered included the protein that the mutation occurred in, the position in the amino acid sequence that the mutation occurred in, what the amino acid at that position is in a wild-type sample, and what the amino acid at that position mutates to. For this study, two classes were chosen based on the ADDRESS results, transporter proteins and methyltransferase activity

proteins, and both of these classes were analyzed separately.

This literature search was conducted from multiple angles. First, a preliminary search was conducted on Google Scholar and PubMed using key words like "pathogenic", "mutation", or "disease" as well as words relevant to the class of protein of interest, such as "transporter proteins", "channels", or "methyltransferase". In addition to this, sometimes specific proteins of the interested class were singled out and searched for in an attempt to pull more relevant papers, such as glucose transporters or solute carriers. To supplement this literature search, a search was also done on UniProt for the desired class of protein [8]. UniProt then allows reviewed variants of a specific protein in the class to be viewed. Conducting a portion of our search through UniProt was advantageous as it allowed for a relatively large number of pathogenic mutations in literature to be viewed while summarizing the mutation information and providing a link to the paper that it was reported.

Once a certain number of pathogenic mutations were obtained for each desired function class (the threshold for this study was at least 100 mutations), the lists of proteins were cross referenced to gnomAD, from which benign mutations for each protein were obtained [9]. Next, for proteins that were in Recon3D the PDB was obtained from there [10], [11]. Recon3D is a genome scale human metabolic network reconstruction, and it also contains protein files for a good number of proteins in the human body. If Recon3D did not contain a sequence file for a desired protein, then the PDB was obtained from UniProt or RCSB through UniProt [12]. The location from where each PDB was obtained was recorded as well as the method the PDB was generated, i.e., experimental or modeled.

As mutations were discovered through the literature search, the data was stored in a CSV file. A script was written in MatLab that could read and validate the information from the CSV file. This was done by having MatLab open the PDB file and ensure that the protein was aligned by comparing it to an alignment file obtained for that protein. Then, for each mutation for that protein script verified that the amino acid at the mutation position was the correct initial amino acid. Any mutation that had a different initial amino acid than what was in our PDB file for that protein was removed from analysis. Then, the script called EvoEF on the PDB file before any mutations to calculate the wild type change in energy from folding. Next, the script called EvoEF to simulate the mutation and stored the change in energy of the protein folding after the mutation occurred. Next, for each mutation a $\Delta\Delta G$ value was calculated by subtracting the wild type change in free energy for the protein from the change in free energy after the mutation had occurred. These values were stored and used to generate the $\Delta\Delta G$ distribution analysis. For each protein class, a histogram and boxplot was plotted for both its benign and pathogenic $\Delta\Delta G$ distributions and various statistical tests were run on the distributions. Additionally, as with the ADDRESS analysis, gaussian curves of modalities one and two were fit to the distributions. For context, modality refers to the number of peaks in a graph, but more generally this can be understood as monomodal gaussians being a normal, single gaussian distribution and bimodal gaussians being a summation of two gaussian distributions. In simple terms, something with a bimodal distribution just has two peaks.

In order to fit the gaussian curves to the distributions in both the ADDRESS analysis and the literature search analysis, two scripts were written in MatLab that could handle plotting monomodal gaussian distributions and bimodal gaussian distributions. These scripts mostly contained built-in MatLab functionality for curve fitting. For this study, the method used for fitting was the nonlinear method of least squares. In order to give the curve fitting scripts a list of points to fit, points were estimated using the distribution histograms. More specifically, a list of x-coordinates was generated by averaging the x-location of the two edges of each bin and a list of y-coordinates was generated by taking the height of each bin, noting that a bin is a bar in a histogram.

These x and y-coordinates were fed into the scripts to generate both monomodal and bimodal fits and accompanying R-values which were analyzed to determine fit quality. Namely, the increase in R-value from monomodal to bimodal distribution was used as a measure of which distribution best

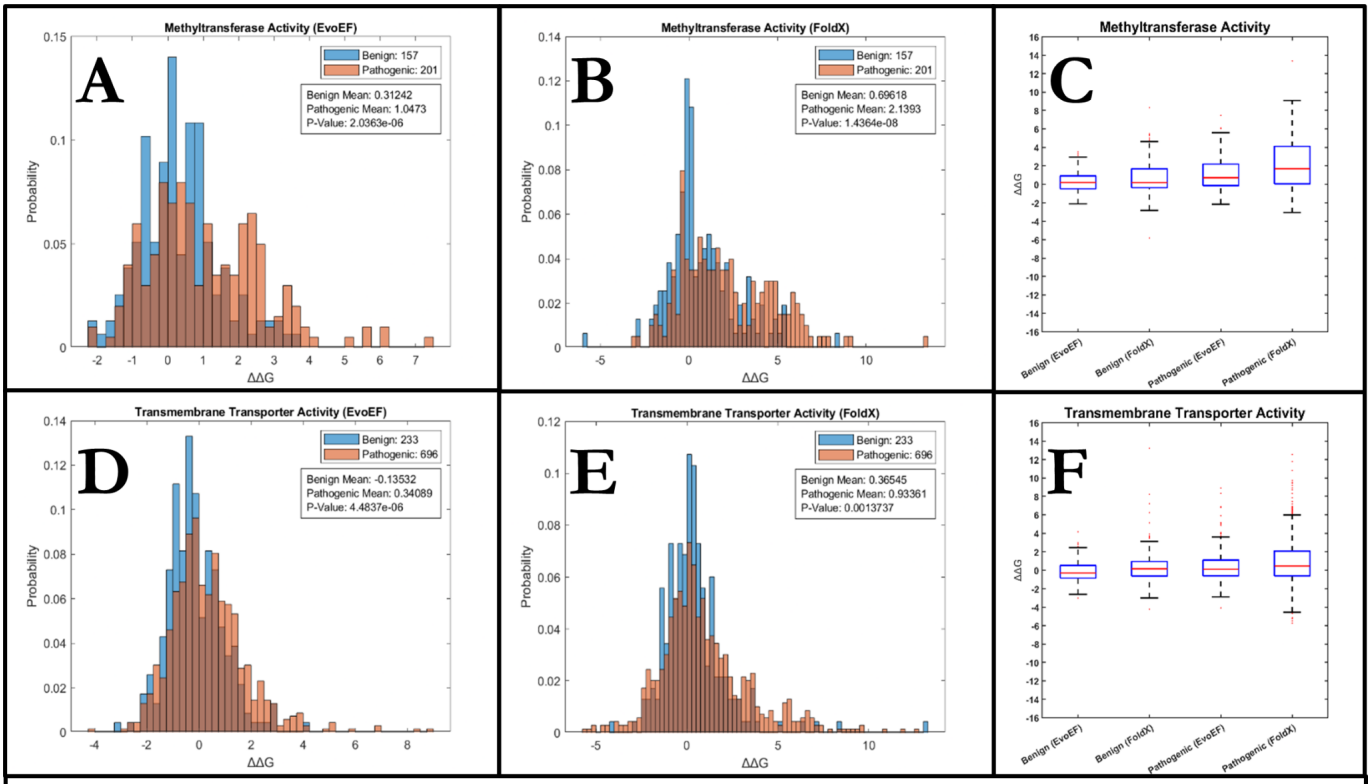

Figure 2: Plots containing histograms for two gene function categories from both EvoEF and FoldX and boxplots. The top-right corner of each histogram subplot contains the number of data points for each histogram as well as the means of the histograms and the p-value generated from a statistical test between the two distributions. The boxplots, from left to right, contain benign EvoEF distributions, benign FoldX distributions, pathogenic EvoEF distributions, and pathogenic FoldX distributions. In particular, the plots are **A)** genes relating to methyltransferase activity with calculations from EvoEF, **B)** genes relating to methyltransferase activity with calculations from FoldX, **C)** boxplots of all methyltransferase data, **D)** genes relating to transmembrane transporter activity with calculations from EvoEF, **E)** genes relating to transmembrane transporter activity with calculations from FoldX, and **F)** boxplots of all transmembrane transporter activity.

benefited from a bimodal distribution. Once all the data was collected, histograms, statistical tests, and gaussian fits were analyzed to determine if any patterns or differences were discernible.

Results

From the ADDRESS database, 21,251 protein mutations were analyzed. These mutations were then organized into 37 different gene function categories, such as ion binding, transmembrane transporter activity, and methyltransferase activity, for example. The full list of gene function categories can be found in Table 1 or Table 2. For each of these gene functions, both EvoEF energy calculations and FoldX energy calculations were examined. Furthermore, for each gene function, a benign mutation histogram and a pathogenic mutation histogram (and boxplots) were created for each type of energy calculation, which means there were over 100 plots created. Some representative plots can be seen in Figure 2, which will be discussed later. The basic statistical measures gathered from the histograms and box plots are summarized in Table 1.

All tests were performed for both the EvoEF data points and the FoldX data points, but for the purpose of this paper and because the results between data sets were very similar, from this point on we will only show representative images from the EvoEF data sets and include some discussion on the FoldX data sets in *Discussion*. Once the distributions had been created and plotted, we began fitting monomodal and bimodal gaussians to each distribution. Representative plots can be seen in Figure 3.

As mentioned, something of interest to us was to

Table 1: A table summarizing the statistical data relevant to the benign and pathogenic distributions of each of the gene function categories. Statistical measures include number of mutations, mean, standard deviation and range of the distribution. At the bottom of the table, averages, maximums, and minimums are noted for across all gene functions, and the number of genes in which the pathogenic distribution value is greater than the benign distribution value is noted as well.

Gene Function Category	Statistical Measurements							
	Mutations		Mean		STDev		Range	
	Ben.	Path.	Ben.	Path.	Ben.	Path.	Ben.	Path.
ATPase Activity	282	**761**	0.043	**0.571**	1.135	**1.574**	12.290	**18.590**
Cytoskeletal Protein Binding	495	**1088**	-0.001	**0.463**	1.011	**1.310**	**13.520**	10.870
DNA Binding	822	**1708**	0.118	**0.588**	1.148	**1.566**	17.550	**26.740**
DNA-Binding Transcription Factor Activity	184	**958**	0.066	**0.521**	**1.368**	1.314	14.180	**14.590**
Enzyme Binding	1544	**3396**	0.028	**0.636**	1.171	**1.630**	16.930	**22.750**
Enzyme Regulator Activity	670	**872**	-0.090	**0.570**	1.202	**1.499**	**15.880**	13.550
GTPase Activity	72	**427**	0.145	**0.441**	1.025	**1.509**	4.420	**13.760**
Helicase Activity	75	**126**	-0.209	**0.585**	1.297	**1.738**	10.200	**15.730**
Histone Binding	87	**91**	0.272	**0.646**	1.396	**1.580**	**11.510**	8.830
Hydrolase (Carbon-Nitrogen) Activity	99	**180**	-0.019	**1.046**	1.024	**2.076**	5.370	**15.530**
Hydrolase (Glycosyl Bonds) Activity	153	**968**	0.244	**1.092**	1.240	**1.963**	7.580	**27.470**
Ion Binding	4144	**8940**	0.086	**0.724**	1.142	**1.652**	17.800	**30.130**
Isomerase Activity	121	**485**	-0.067	**0.777**	1.087	**1.584**	6.300	**12.620**
Kinase Activity	839	**1344**	0.097	**0.753**	1.190	**1.508**	**17.690**	14.690
Ligase Activity	96	**369**	-0.091	**0.785**	1.002	**1.649**	6.000	**12.160**
Lipid Binding	606	**1328**	-0.030	**0.748**	1.026	**1.475**	9.410	**14.890**
Lyase Activity	179	**484**	0.190	**0.749**	1.150	**1.504**	6.750	**15.030**
Methyltransferase Activity	157	**201**	0.312	**1.047**	1.103	**1.638**	5.640	**9.650**
Nuclease Activity	**173**	163	0.074	**0.398**	0.981	**1.428**	7.070	**7.920**
Nucleotidyltransferase Activity	85	**266**	0.254	**0.610**	1.198	**1.485**	7.240	**10.730**
Oxidoreductase Activity	1293	**1959**	0.193	**0.742**	1.117	**1.594**	10.010	**15.760**
Peptidase Activity	456	**903**	-0.043	**0.910**	1.223	**1.799**	13.840	**17.470**
Phosphatase Activity	130	**215**	0.146	**0.934**	1.333	**2.160**	11.250	**17.000**
Protein Binding, Bridging	99	**339**	0.096	**0.649**	0.944	**1.611**	4.600	**12.560**
RNA Binding	469	**887**	0.081	**0.519**	1.031	**1.645**	6.270	**21.020**
Structural Constituent of Ribosome	2	**8**	0.170	**1.355**	0.156	**1.993**	0.220	**5.750**
Structural Molecule Activity	240	**658**	0.047	**0.818**	0.958	**1.679**	5.840	**13.830**
Transcription Factor Binding	250	**940**	0.052	**0.605**	0.930	**1.394**	5.090	**22.690**
Transferase (Acyl Groups) Activity	107	**185**	0.017	**0.850**	0.937	**1.473**	4.070	**9.570**
Transferase (Alkyl, Aryl, Methyl Groups) Activity	55	**217**	0.093	**0.688**	1.039	**1.321**	7.310	**9.300**
Transferase (Glycosyl Groups) Activity	114	**386**	0.163	**0.847**	1.107	**1.620**	7.360	**13.410**
Translation Factor (RNA Binding) Activity	**11**	4	**0.625**	0.268	1.175	**2.142**	4.030	**5.030**
Transmembrane Transporter Activity	233	**696**	-0.135	**0.341**	1.049	**1.453**	7.190	**12.990**
Ubiquitin-like Protein Binding	45	**99**	0.158	**0.286**	0.924	**1.280**	4.870	**6.850**
Unfolded Protein Binding	**64**	31	0.094	**1.250**	0.880	**1.627**	3.980	**7.760**
mRNA Binding	32	**168**	-0.128	**0.432**	0.792	**1.354**	3.650	**10.500**
rRNA Binding	15	**16**	**0.668**	0.089	0.768	**1.662**	2.550	**6.240**
Average	392	861	0.100	0.685	1.061	1.608	8.526	14.161
Maximum	4144	8940	0.668	1.355	1.396	2.160	17.800	30.130
Minimum	2	4	-0.209	0.089	0.156	1.280	0.220	5.030
Number of Pathogenic Greater than Benign	**34**		**35**		**36**		**33**	
Number of Null Hypothesis Rejected	N/A		N/A		N/A		N/A	

Table 2: A table summarizing the gaussian fitting and t-test data relevant to the benign and pathogenic distributions of each of the gene function categories. Gaussian fitting data includes monomodal and bimodal R-values as well as R-value increase for each distribution for each gene function and t-test data includes the p-value and the decision. At the bottom of the table, averages, maximums, and minimums are noted for across all gene functions, and both the number of genes in which the pathogenic distribution value is greater than the benign distribution value and the number of null hypotheses rejected are noted as well.

Gene Function Category	Gaussian Fit Results						T-Test Results	
	Single R-Value		Double R-Value		R-Value Increase		Decision	P-Value
	Ben.	Path.	Ben.	Path.	Ben.	Path.		
ATPase Activity	0.955	0.966	0.957	0.979	0.002	**0.013**	Reject	3.06E-07
Cytoskeletal Protein Binding	0.960	0.959	0.968	0.985	0.009	**0.026**	Reject	4.06E-12
DNA Binding	0.979	0.978	0.982	0.995	0.003	**0.016**	Reject	2.52E-14
DNA-Binding Transcription Factor Activity	0.873	0.976	0.876	0.985	0.003	**0.010**	Reject	2.11E-05
Enzyme Binding	0.983	0.976	0.990	0.988	0.007	**0.011**	Reject	4.21E-39
Enzyme Regulator Activity	0.971	0.954	0.992	0.985	0.021	**0.031**	Reject	3.85E-20
GTPase Activity	0.000	0.951	0.000	0.972	0.000	**0.021**	**Accept**	**1.10E-01**
Helicase Activity	0.834	0.842	0.839	0.850	0.004	**0.008**	Reject	7.32E-04
Histone Binding	0.814	0.725	0.814	0.771	0.000	**0.046**	**Accept**	**9.65E-02**
Hydrolase (Carbon-Nitrogen) Activity	0.723	0.766	0.831	0.781	**0.107**	0.015	Reject	2.67E-06
Hydrolase (Glycosyl Bonds) Activity	0.835	0.977	0.845	0.981	**0.011**	0.004	Reject	2.57E-07
Ion Binding	0.989	0.978	0.995	0.996	0.006	**0.018**	Reject	5.03E-110
Isomerase Activity	0.767	0.933	0.770	0.940	0.003	**0.007**	Reject	4.36E-08
Kinase Activity	0.975	0.935	0.996	0.982	0.021	**0.048**	Reject	4.76E-26
Ligase Activity	0.816	0.865	0.821	0.949	0.006	**0.084**	Reject	9.47E-07
Lipid Binding	0.964	0.950	0.974	0.976	0.010	**0.026**	Reject	7.29E-31
Lyase Activity	0.924	0.945	0.947	0.970	0.023	**0.025**	Reject	7.55E-06
Methyltransferase Activity	0.739	0.745	0.749	0.831	0.010	**0.087**	Reject	2.04E-06
Nuclease Activity	0.966	0.818	0.968	0.851	0.003	**0.033**	Reject	1.53E-02
Nucleotidyltransferase Activity	0.684	0.910	0.813	0.927	**0.129**	0.017	Reject	4.52E-02
Oxidoreductase Activity	0.977	0.977	0.986	0.987	0.010	**0.010**	Reject	1.33E-26
Peptidase Activity	0.977	0.938	0.978	0.978	0.001	**0.040**	Reject	1.71E-23
Phosphatase Activity	0.814	0.903	0.861	0.922	**0.047**	0.019	Reject	2.07E-04
Protein Binding, Bridging	0.650	0.854	0.782	0.909	**0.132**	0.055	Reject	1.24E-03
RNA Binding	0.943	0.976	0.980	0.987	**0.037**	0.012	Reject	1.74E-07
Structural Constituent of Ribosome	0.000	0.000	0.000	0.000	0.000	0.000	**Accept**	**4.45E-01**
Structural Molecule Activity	0.878	0.943	0.943	0.970	0.065	**0.027**	Reject	3.26E-11
Transcription Factor Binding	0.893	0.976	0.901	0.985	0.008	**0.009**	Reject	4.17E-09
Transferase (Acyl Groups) Activity	0.605	0.900	0.605	0.914	0.000	**0.014**	Reject	2.76E-07
Transferase (Alkyl, Aryl, Methyl Groups) Activity	0.000	0.848	0.000	0.952	0.000	**0.104**	Reject	2.10E-03
Transferase (Glycosyl Groups) Activity	0.843	0.909	0.843	0.933	0.001	**0.024**	Reject	2.83E-05
Translation Factor (RNA Binding) Activity	0.000	0.000	0.000	0.000	0.000	0.000	**Accept**	**6.81E-01**
Transmembrane Transporter Activity	0.883	0.952	0.887	0.971	0.004	**0.019**	Reject	4.48E-06
Ubiquitin-like Protein Binding	0.000	0.658	0.000	0.658	0.000	**0.000**	**Accept**	**5.50E-01**
Unfolded Protein Binding	0.000	0.000	0.000	0.000	0.000	0.000	Reject	1.97E-05
mRNA Binding	0.000	0.857	0.000	0.875	0.000	**0.018**	Reject	2.46E-02
rRNA Binding	0.000	0.000	0.000	0.000	0.000	0.000	**Accept**	**2.28E-01**
Average	0.681	0.806	0.700	0.831	0.018	0.024	N/A	5.94E-02
Maximum	0.989	0.978	0.996	0.996	0.132	0.104	N/A	6.81E-01
Minimum	0.000	0.000	0.000	0.000	0.000	0.000	N/A	5.03E-110
Number of Pathogenic Greater than Benign	**23**		**26**		**26**		N/A	
Number of Null Hypothesis Rejected	N/A		N/A		N/A		**31**	

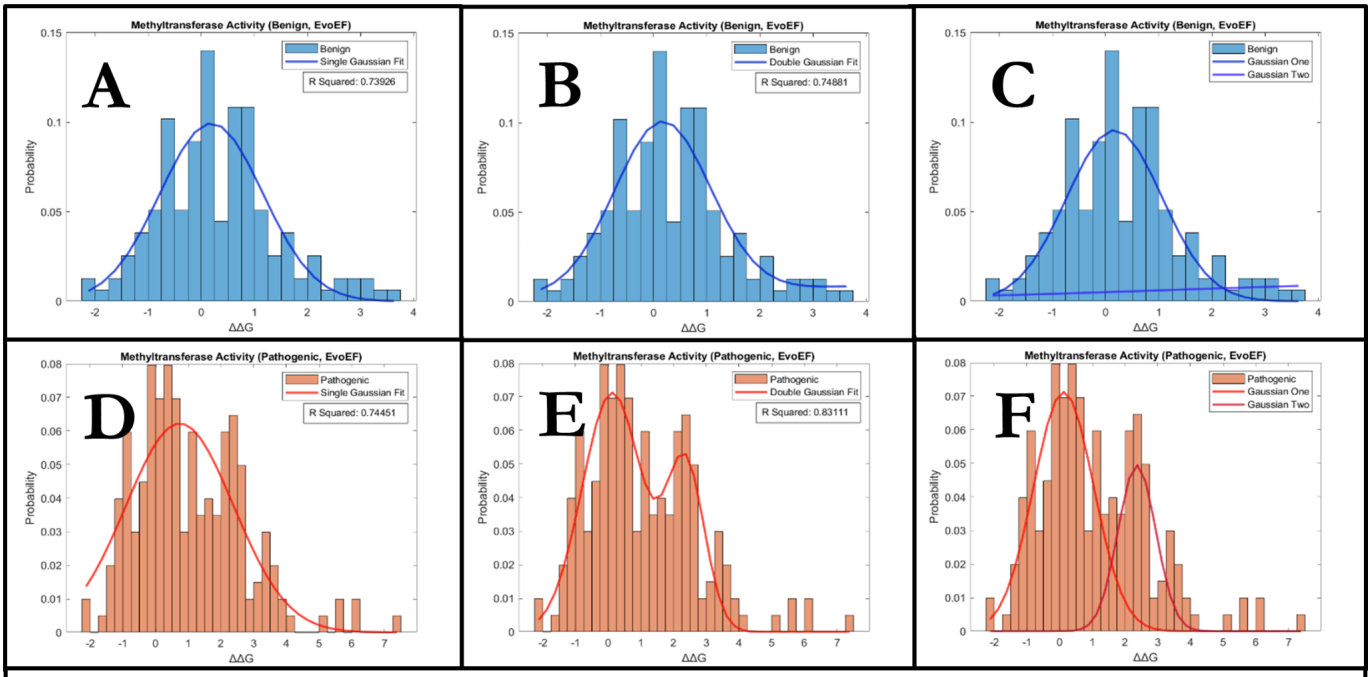

Figure 3: Plots showing a gaussian fit example for methyltransferase activity. Note, that monomodal gaussian is the same as single gaussian and bimodal gaussian is the same as double gaussian in this context. R-squared values for each non-deconstructed fit are shown in the top right of each sub-plot. In particular, the plots show for methyltransferase activity genes and data from EvoEF **A)** a single gaussian fit to the benign data, **B)** a double gaussian fit to the benign data, **C)** a deconstructed double gaussian fit to the benign data, **D)** a single gaussian fit to the pathogenic data, **E)** a double gaussian fit to the pathogenic data, and **F)** a deconstructed double gaussian fit to the pathogenic data.

determine if there were any differences in which gaussian model better described the benign distributions versus the pathogenic distributions. To get an idea of this, we looked at each distribution and determined the increase that occurred when switching from a monomodal distribution model to a bimodal distribution model. These results are shown in Table 2. In addition, t-tests were conducted between pathogenic and benign distributions for each gene function category, and the results can also be seen in Table 2.

Next, the results of the literature search were obtained. As mentioned, we looked at two gene categories: transporter proteins and proteins relating to methyltransferase activity. Transporter

Table 3: A table summarizing the statistical data relevant to the benign and pathogenic distributions of each of the gene function categories from the literature search. For reference, the same categories from ADDRESS are included as well. Statistical measures include number of mutations, mean, standard deviation and range of the distribution. At the bottom of the table, averages, maximums, and minimums are noted for across all gene categories.

Gene Function Category	Statistical Measurements							
	Mutations		Mean		STDev		Range	
	Ben.	Path.	Ben.	Path.	Ben.	Path.	Ben.	Path.
Transporter Proteins (Lit Search)	**3766**	262	0.219	**0.623**	0.987	**1.431**	8.510	**11.380**
Transporter Proteins (ADDRESS)	233	**696**	-0.135	**0.341**	1.049	**1.453**	7.190	**12.990**
Methyltransferase Activity (Lit Search)	**2113**	150	0.208	**0.758**	0.950	**1.268**	**9.890**	6.400
Methyltransferase Activity (ADDRESS)	157	**201**	0.312	**1.047**	1.103	**1.638**	5.640	**9.650**
Average	1567	327	0.151	0.692	1.022	1.448	7.808	10.105
Maximum	3766	696	0.312	1.047	1.103	1.638	9.890	12.990
Minimum	157	150	-0.135	0.341	0.950	1.268	5.640	6.400

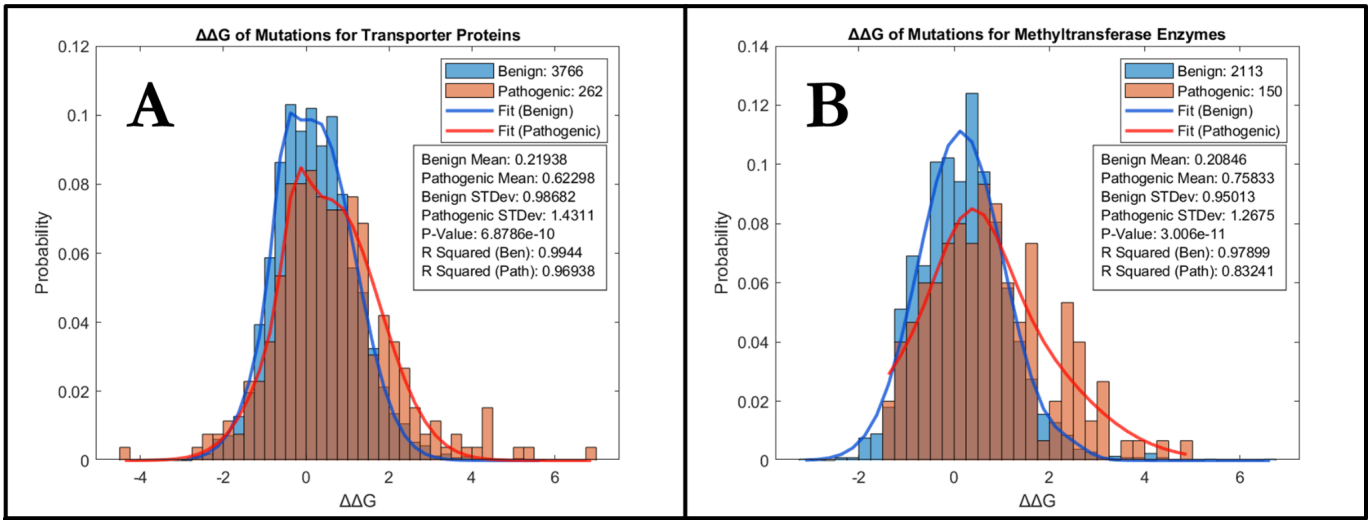

Figure 4: Histogram plots showing the distributions obtained from the literature search with bimodal gaussians fit to the distributions. The top right of each plot also contains information such as number of mutations, distribution means, distribution standard deviations, t-test p-value, and r-squared values from the fit. In particular, the plots show **A)** distributions from the transporter protein literature search and **B)** distributions from the methyltransferase enzyme/activity literature search.

proteins were chosen because they were of interest to us, and methyltransferase activity related proteins were chosen because they had strong pathogenic bimodal representation from ADDRESS (Figure 3). Additional proteins could've been chosen and likely will be in the future. For transporter proteins, we looked at 18 different genes; across those 18 genes, we identified 262 pathogenic mutations and 3,766 benign mutations. For methyltransferase activity related proteins, we looked at 10 genes; across those 10 genes, we identified 150 pathogenic mutations and 2,113 benign mutations. As previously mentioned, finding benign mutations was much more trivial than pathogenic ones since gnomAD already has tables of benign mutations available. The resulting histograms (from the process defined in *Methods*) can be seen in Figure 4. In addition, a summary of

Table 4: A table summarizing the gaussian fitting and t-test data relevant to the benign and pathogenic distributions of each of the gene function categories from the literature. In addition, t-test results are also included from comparing the literature search distributions to the corresponding ADDRESS distributions. Gaussian fitting data includes monomodal and bimodal R-values as well as R-value increase for each distribution for each gene function and t-test data includes the p-value and the decision. At the bottom of the table, averages, maximums, and minimums are noted for across all gene functions.

Gene Function Category	Gaussian Fit Results						T-Test Results	
	Single R-Value		Double R-Value		R-Value Increase		Decision	P-Value
	Ben.	Path.	Ben.	Path.	Ben.	Path.		
Transporter Proteins (Lit Search)	0.984	0.954	0.994	0.969	0.010	**0.016**	Reject	6.88E-10
Transporter Proteins (ADDRESS)	0.883	0.952	0.887	0.971	0.004	**0.019**	Reject	4.48E-06
Transporter Proteins Benign (Lit vs ADDRESS)	N/A	N/A	N/A	N/A	N/A	N/A	Reject	1.19E-07
Transporter Proteins Pathogenic (Lit vs ADDRESS)	N/A	N/A	N/A	N/A	N/A	N/A	Reject	7.29E-03
Methyltransferase Activity (Lit Search)	0.979	0.812	0.979	0.832	0.000	**0.020**	Reject	3.01E-11
Methyltransferase Activity (ADDRESS)	0.739	0.745	0.749	0.831	0.010	**0.087**	Reject	2.04E-06
Methyltransferase Activity Benign (Lit vs ADDRESS)	N/A	N/A	N/A	N/A	N/A	N/A	**Accept**	**1.91E-01**
Methyltransferase Activity Pathogenic (Lit vs ADDRESS)	N/A	N/A	N/A	N/A	N/A	N/A	**Accept**	**7.34E-02**
Average	0.896	0.866	0.902	0.901	0.006	0.035	N/A	3.40E-02
Maximum	0.984	0.954	0.994	0.971	0.010	0.087	N/A	1.91E-01
Minimum	0.739	0.745	0.749	0.831	0.000	0.016	N/A	3.01E-11

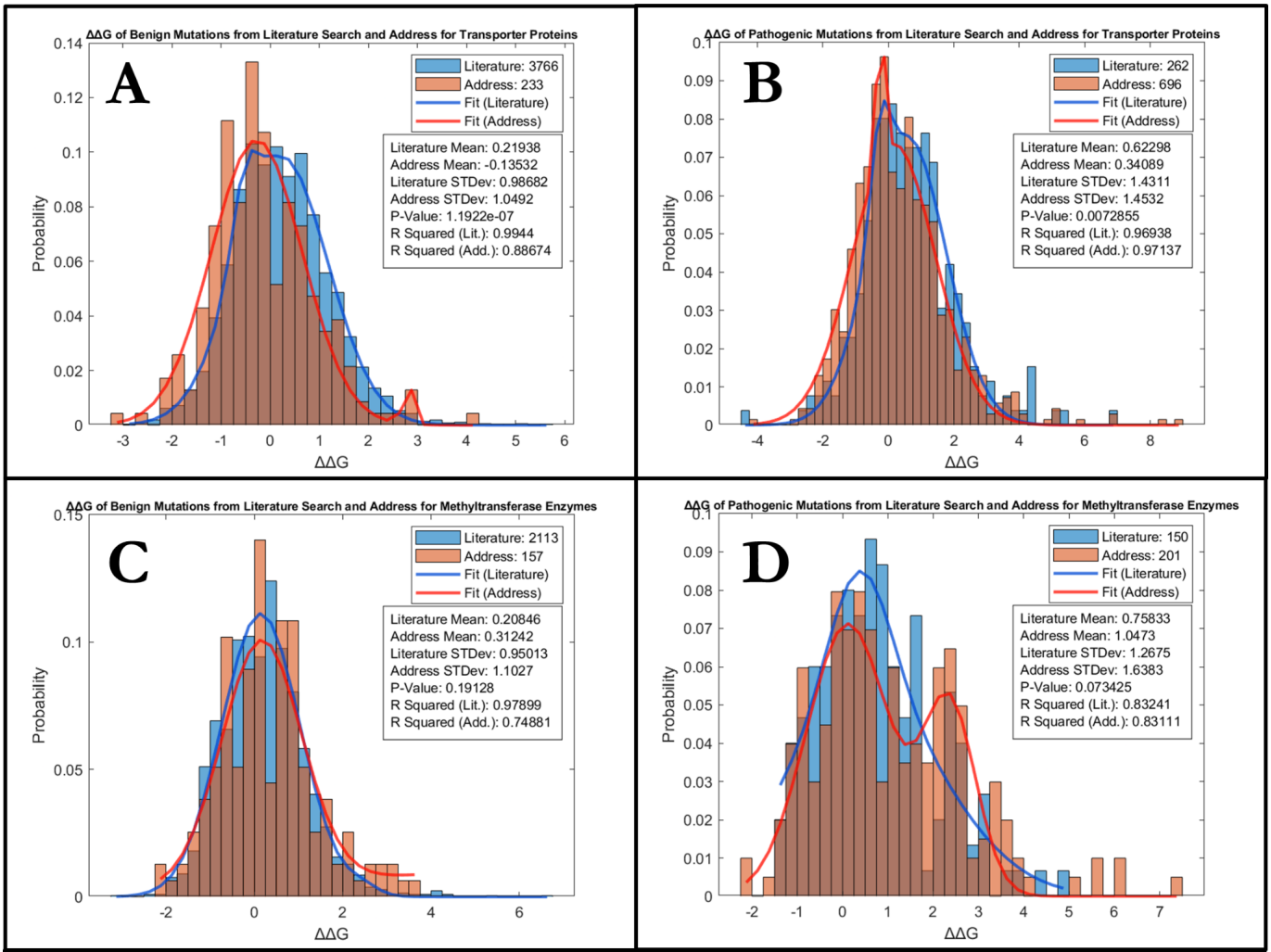

Figure 5: Plots showing the distributions from the literature search with the corresponding distributions from ADDRESS. The top right corner of each plot contains information such as number of mutations, distribution mean, distribution standard deviation, t-test p-value, and r-squared values from the fit. In particular the plots shown are **A)** the benign distributions from the literature search and ADDRESS for transporter proteins, **B)** the pathogenic distributions from the literature search and ADDRESS for transporter proteins, **C)** the benign distributions from the literature search and ADDRESS for methyltransferase enzyme proteins, and **D)** the pathogenic distributions from the literature search and ADDRESS for methyltransferase enzyme proteins.

the statistical quantities relevant to the distributions can be seen in Table 3.

Just as was done with the ADDRESS analysis, a gaussian fit analysis was conducted in order to get an understanding of the difference in monomodal and bimodal fit accuracy for the distributions. The resulting quantities from this analysis can be seen in Table 4.

Finally, we compared the distributions obtained from the literature search to the corresponding category that was in ADDRESS in order to try and get some kind of idea as to if they statistically come from the same distribution. We plotted the histograms for both categories from both sources by pathogenic and benign mutations. These plots can be seen in Figure 5. Furthermore, we conducted a t-test between the corresponding distributions and those results are summarized in Table 4.

Discussion

One of the first things that we noticed was that the pathogenic distributions appeared to be slightly more positive than their benign counterparts. This was true for 35 out of 37 gene categories from the EvoEF data from ADDRESS. The results from FoldX were very similar. In addition, both of the

gene categories from the literature search exhibited this behavior as well. Furthermore, none of the pathogenic distributions had a mean that was below zero. This suggests that pathogenic mutations are more destabilizing than their benign counterparts. Benign distributions appear to be more about zero, with some distributions being destabilizing on average and some being stabilizing on average. However, the key observation that we took from this was that the pathogenic distributions appear to be more destabilizing.

Another thing that we noticed is that the pathogenic distributions appeared to have a greater standard deviation than their benign counterparts. This was true for 36 out of the 37 gene categories from ADDRESS and both of the gene categories from the literature. It also appeared that the pathogenic distributions had a greater spread, with 33 out of 37 categories from ADDRESS and both categories from the literature search showing this. This can suggest that pathogenic distributions can have more extreme outliers, which can mean that pathogenic mutations have the potential to be extremely destabilizing. Something else of interest to note is that FoldX pathogenic distributions appeared to have the greatest standard deviation. Although we did not confirm this rigorously, it is an important observation nonetheless as it brings to our attention that computational estimation methods can play a role in the distributions generated. However, even in FoldX's case, 36 out of 37 gene categories showed that pathogenic distributions had greater standard deviation.

One of the main things that we sought to determine was whether there was a statistical difference between the pathogenic and benign distributions for a protein category. The above discussion on mean and standard deviation suggests that they are, in fact, statistically different, however we wanted to show this further with a t-test. With a p-value threshold of 0.05, 31 out of the 37 ADDRESS gene categories and both of the literature search categories had statistically different pathogenic and benign distributions. Some distributions were more different than others, with the p-values ranging from 1.3E-2 to 5.0E-110. No p-value adjustments were made, so it is important to remember that there may be a couple of false rejections, however the fact remains that a majority of the distributions would still be statistically different even if the p-value was corrected for multiple comparisons. The FoldX data also had 31 out of 37 gene categories marked as statistically different, but interestingly two of the categories marked as different by EvoEF were not by FoldX and vice versa which brings into the light the need to fully understand the computational difference between the two programs to better understand any discrepancies between the results of the two programs.

Another goal was to see if the pathogenic and benign distributions were better modeled by different modalities of gaussian fits. From the EvoEF data from ADDRESS, it did appear that in the case of some protein categories the pathogenic distribution was better described by a bimodal gaussian fit whereas the benign distribution was already accurately described by a monomodal gaussian fit. A prime example of this was shown in Figure 3. The way we tried to quantify this was by analyzing the improvement in r-squared values when going from a monomodal distribution to a bimodal distribution. The results for this from EvoEF ADDRESS can be seen in Table 2. Note that this analysis was not conducted for the FoldX data. From this table, we can see that for 26 of the protein categories the pathogenic distributions saw a greater increase in r-squared value when going from the monomodal fit to the bimodal fit when compared to the benign distributions. However, some of these greater increases were very minimal, so it is more interesting to see the cases where the pathogenic distribution benefited by a good amount and the benign distribution benefited very little. We did not define numbers that met this requirement, but some categories that we believe exhibit this behavior are ligase activity, methyltransferase activity, nuclease activity, peptidase activity, some transferase activity, and transporter proteins. Visually, this can be seen in Figures 2 and 3. In Figure 2, the pathogenic distributions of both categories appear to have a second "spike" which is not as prominent in the benign distribution. This behavior is more closely examined in Figure 3 for methyltransferase activity, which shows that in the case of the benign bimodal distribution a second peak was not even detected. Our hypothesis for why this is occurring is that pathogenic mutations could have multiple different

routes of causing disease. The distributions could be bimodal because one group of mutations could be causing diseases by destabilizing the protein, causing a secondary peak at higher ΔΔG, whereas the other proteins cause diseases via other avenues, such as disrupting the protein fold or altering the proteins ability to bind to its targets. Verification of this hypothesis will require further testing.

The results from the literature search distributions were less clear on the matter. Mathematically, both categories saw that the pathogenic distributions benefited from the bimodal distribution greater than the benign distributions did. However, the fits that were generated for the distributions did not correctly pick up the second, more positive spike that was shown in some of the distributions from ADDRESS. Interestingly enough, we can visually see the makings of this second peak in both of the pathogenic distributions (at about the 2.5 kilocalories per mole mark), so we believe that the lack of having statistical representation of these second peaks could be due to the fact that we simply need more data points for the pathogenic distributions, because in the case of the each of the literature search categories the pathogenic distribution had less data points than their ADDRESS counterparts. Another reason that our fit was unable to detect the second peak could be that we just need to try additional gaussian fit techniques or perhaps adjust the bin width settings of the histogram.

Finally, we wanted to compare the distributions that we obtained from our literature search to the distributions that we obtained from ADDRESS. It is important to note that we did not check for redundancy between the ADDRESS mutation pool and the literature search mutation pool, which should be done in the future. Interestingly enough, when conducting a t-test between the methyltransferase distributions, with a p-value threshold of 0.05, the null hypothesis was actually accepted for both the benign and pathogenic distributions. As shown in Figure 5, the distributions from the methyltransferase activity look incredibly similar. Even though the gaussian fits do not line up (especially in the case of the pathogenic distribution), the makings of the histogram look remarkably similar. Thus, we can conclude that at the very least separate distributions of mutations relating to methyltransferase appear to be very similar. The results from doing this with the transporter protein categories were less clear on this since the t-test suggested that they came from different distributions. However, the p-value from the pathogenic distributions was relatively high and visually the distributions look quite similar, so it appears that they may be similar, but we were not able to show this concretely. Discrepancies could come from the largely differing number of data points between the two distributions.

Conclusion and Future Directions

Generally, we found that compared to benign mutations, pathogenic mutations appear to be more destabilizing and have greater variance. Furthermore, we found that the underlying ΔΔGs distributions from pathogenic and benign mutation do appear to be statistically different. In addition, for some categories of protein function, pathogenic mutation ΔΔG distributions appear to be better described by a bimodal fit whereas this phenomenon is not seen as clearly with the benign mutation ΔΔG distributions. Finally, there appears to be some correlation of ΔΔG distributions by protein category from different sources, i.e. from ADDRESS and an independent literature search, suggesting that for a given protein function the mutation ΔΔG distribution may have an expected underlying distribution.

Moving forward, it is of great importance to try and identify why this bimodal pathogenic distribution is occurring. It may involve conducting an analysis in which we try to identify specific mutations that have a specific effect in the protein, such as destabilizing the protein, disrupting its fold, disrupting its binding, etc., and analyzing the contributions those sub-categories of mutations contribute to the overall pathogenic ΔΔG distribution for that protein category. As mentioned in the discussion, some of the results are not as clear as desired, so it would certainly be worthwhile to run this analysis on additional data sets and to increase the size of the literature search data sets. Working with more data may be able to bring out stronger correlations or results. It may also be necessary to take a closer look at the underlying differences between different computational approaches to estimating ΔΔG since

we saw that while generally EvoEF and FoldX agreed there were some discrepancies in the fine details. Understanding why this may be occurring and furthermore understanding how using the different computational models can affect the resulting distributions is an important distinction to make if these observations are to be applied to real world models and applications. Other things that we could look into doing were previously mentioned, such as trying different fitting estimation techniques, changing the histogram bin width settings, and checking for redundancies between the two data collections (ADDRESS versus the literature search).

In the future, it would also be interesting to run the same kind of analysis but instead do multiple mutations at a time. In real life, it is often the case that a disease results from multiple mutations. Said in another way, when patients with diseases are genotyped, it is often the case that the patient has multiple mutations, and sometimes it is unknown which mutation has what effect as far as causing a disease goes. Furthermore, sometimes a combination of mutations is required to get the protein to fold and/or behave a certain way. For example, sometimes one mutation can cause a large destabilizing effect and a following mutation can cause a large stabilizing effect. So, from a $\Delta\Delta G$ standpoint these mutations "cancel" each other out. However, if those mutations are believed to be associated with a disease, then it may be important to understand both their individual roles and their combinatory effects. Thus, only running mutations one at a time could be a potential limitation of this analysis. However, sometimes it is possible for a single mutation to cause a protein's binding affinity to a specific molecule, ligand, nucleic acid, or other protein, so a future study could include this aspect in the analysis.

This study has shown that there does appear to be a difference in how benign and pathogenic mutations affect a protein from an energy standpoint. These differences may be enough to one day allow researchers to better understand and predict the pathogenicity of certain protein mutations and potential treatments. Although this study does not provide concrete models to be used by researchers, it can act as the foundation upon which future studies and models can be built.

References

[1] L. Pray, "Errors in DNA Replication." http://www.nature.com/scitable/topicpage/dna-replication-and-causes-of-mutation-409 (accessed Mar. 04, 2023).

[2] B. Hess, "Free Energy Change,"

[3] "FoldX." https://foldxsuite.crg.eu/ (accessed Mar. 12, 2023).

[4] X. Huang, R. Pearce, and Y. Zhang, "EvoEF2: accurate and fast energy function for computational protein design," Bioinformatics, vol. 36, no. 4, pp. 1135–1142, Feb. 2020, doi: 10.1093/bioinformatics/btz740.

[5] J. Woodard, C. Zhang, and Y. Zhang, "ADDRESS: A Database of Disease-associated Human Variants Incorporating Protein Structure and Folding Stabilities," J. Mol. Biol., vol. 433, no. 11, p. 166840, May 2021, doi: 10.1016/j.jmb.2021.166840.

[6] "GOnet." https://tools.dice-database.org/GOnet/ (accessed Mar. 26, 2023).

[7] "Humsuvar Database." https://www.uniprot.org/help?query=humsavar (accessed Mar. 26, 2023).

[8] "UniProt." https://www.uniprot.org/ (accessed Mar. 26, 2023).

[9] "gnomAD." https://gnomad.broadinstitute.org/ (accessed Mar. 26, 2023).

[10] E. Brunk et al., "Recon3D enables a three-dimensional view of gene variation in human metabolism," Nat. Biotechnol., vol. 36, no. 3, Art. no. 3, Mar. 2018, doi: 10.1038/nbt.4072.

[11] "Recon3D." Systems Biology Research Group, Dec. 01, 2022. Accessed: Mar. 26, 2023. [Online]. Available: https://github.com/SBRG/Recon3D

[12] RCSB Protein Data Bank, "RCSB PDB." https://www.rcsb.org/ (accessed Mar. 26, 2023).