Elucidating Which Pairwise Mutations Affect Protein Stability : an Exhaustive Big Data Approach

Nicholas Majeske
Computer Science
Western Washington University
Bellingham, WA
majesken@wwu.edu

Filip Jagodzinski
Computer Science
Western Washington University
Bellingham, WA
filip.jagodzinski@wwu.edu

***Abstract*—The specific sequence of amino acids in a polypeptide chain dictates the three dimensional structure, and hence function, of a protein. Mutagenesis experiments on physical proteins involving amino acid substitutions provide insights enabling pharmaceutical companies to design medicines to combat a variety of debilitating diseases. However such wet lab work is prohibitive, because even studying the effects of a single mutation may require weeks of work. Computational approaches for performing exhaustive screens of the effects of single mutations have been developed, but methods for conducting a systematic, exhaustive screen of the effects of all multiple mutations are not available due to the large number of mutant protein structures that would need to be analyzed. In this work we motivate and demonstrate a proof of concept approach for conducting in silico experiments in which we generate all possible mutant structures with 2 amino acid substitutions for three proteins with 46, 67, and 99 residues; for the largest protein we in silico generate 1,751,211 mutants. We leverage an efficient combinatorial algorithm to assess the effects of the mutations among the mutant protein structures. We also produce heat maps for several mutation metrics to facilitate identifying which pairs of amino acid in a protein have the greatest impact on protein stability based on how those amino acid substitutions affect the protein's flexibility.**

***Keywords*—protein, multiple mutations, big data, visualization**

I. INTRODUCTION

Amino acids and the sequence in which they occur in a protein determine a protein's structure and consequently its function. Even a single amino acid substitution can drastically alter a protein's shape, which may be the cause of a serious disease. For example, Fabry disease is caused by mutations to α-galactosidase, which causes serious cardiac complications.

Wet lab mutagenesis experiments provide definitive insights about the effects of mutations, but such experiments are time and cost prohibitive. Moreover, because there are 20 naturally occurring amino acids, even a small 50-amino acid protein could be mutated to generate 20^{50} mutants. Even if only 1 mutation is allowed, the number of possible variants of a 50-amino acid protein is still $50 \times 20 = 1000$.

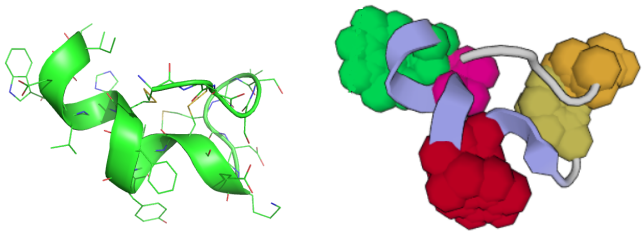
Assessing the effects of mutations is nonetheless an important screening tool needed by pharmaceutical companies that aim to develop medicines to combat serious diseases caused by protein mutants. A variety of computational approaches such as molecular docking algorithms rely on costly all-atom energy calculations to generate and score large sets of feasible protein variants. Because of the huge possible configuration space of a single protein, identifying as energetically or structurally feasible the set of protein mutant candidate structures is a time consuming task, often requiring months of compute time.

In this work we motivate and demonstrate an efficient computational approach enabling an exhaustive screen for assessing which pairwise mutations in a protein have the greatest impact on a protein's stability. We achieve this via the use of an efficient combinatorial algorithm for calculating the flexibility of a protein. **This work is distinguished from others in that it permits an exhaustive screen of all possible pairwise mutations**. The implications of this work are numerous. Analyzing the effects of all possible pair-wise mutations can yield information about allosteric sites. Allostery is the phenomenon by which proteins transmit the effect of binding to a location in the protein far removed from the binding site [1].

II. RELATED WORK

To complement mutation studies performed on physical proteins, computational methods are available. Most of them aim to predict how an amino acid substitution affects a protein's stability. Approaches that rely on energetic analysis as an indicator of the effects of mutations are often unable to predict how mutations far removed from the active site affect a protein's structure. That is because amino acid substitutions at residues not in the active site induce little or no changes to the structure of a protein, else induce limited energetic perturbations [2]. Various modeling and computational methods, including some available via web servers, are available. They strive to predict the effects of mutations, with varying degrees of accuracy. Early work algorithms ranged from those that searched for best side-chain conformations as a measure of the impact of a mutation [3]–[5], to those that relied on heuristic energy functions or database-derived potentials [6], [7]. Others were dependent on sufficiently large datasets of homologous proteins [8]–[10]. More recently, machine learning (ML) approaches, which is a branch of artificial intelligence, exhibit great variety, with some relying on Vector Machine methods [11], [12], while others utilize Random Forest and similar approaches [13], [14]. Among these ML methods, several have high prediction rates of the effects of the mutations, upwards of 70 and 80%.

The energy-, homology-, and ML-based approaches have several drawbacks. All but a few of them permit reasoning



(a) cartoon (b) rigidity analysis

Fig. 1: Rigidity Analysis : Cartoon (a) and Rigidity analysis (b) of PDB file 1edn. In (b), atoms in different rigid clusters are colored by cluster membership and displayed as spheres.

about the effects of single point mutations only [10], [15]–[21]. And those that afford assessing the effects of multiple amino acid substitutions do not permit an exhaustive screen.

In the case where a web server is available, most cannot be used to assess the effects of multiple mutations, in spite of the fact that there is a wide range of diseases associated with proteins with multiple mutations. For example, for HIV-1 protease it has been shown that the median number of mutations in the protease gene which confers drug-associated resistance to protease inhibitors duranavir and tipranavir is twenty-eight [22].

Thus progress has been made in developing software to complement wet lab methods, but many such tools rely on computationally intensive energy calculations, may need a rich data set built up from a variety of experimental methods, which is not available, or permit hypothesis testing of the effect of a single amino acid substitution at a specific residue only. Moreover most approaches have not been extended for use via high-throughput analysis to assess the effect of all possible pairwise mutations.

A. Rigidity Analysis and Rigidity Distance

Rigidity Analysis [23] is a combinatorial technique for identifying the rigid and flexible regions of biomolecules. Figure 1 depicts the cartoon and rigidity analysis results of Protein Data Bank (PDB) file 1edn of Human Endothelin-1. Rigidity analysis, which identifies rigid clusters of atoms, is distinguished from most other methods by being very fast. It does not rely on homologous protein data, nor on costly energy calculations. See [24] for a full explanation of rigidity analysis.

Rigidity analysis was first used to explore the effects of mutations by calculating a rigid cluster's configuration entropy value [25]. Later tools for rigidity-based mutation analysis were developed, but the extent of the types of in silico mutations that they could perform were limited.

III. METHODS

Our compute pipeline, made up of 4 distinct stages (each color coded), is shown in Figure 2. The blue portions of that figure designate the input (either in the form of data or parameters). The orange box designates the scripting routines that identify the set of all possible amino acid sequences representing the mutant structures, while the green components refer to the parallelized mutant generation and rigidity analysis routines. The computation that would be performed by a single CPU is outlined in purple. The parallel rigidity analysis invocations generate rigidity metrics, which are aggregated into a single metric data set (yellow), from which the Mutation Heat Maps are produced. In this section we explain each component of our pipeline.

A. Input

The input to our compute pipeline is a single Protein Data Bank (PDB) file, which contains the $x-$, $y-$ and $z-$ coordinates of the atoms of a protein whose structure has been X-ray crystallography resolved. In the context of our work, a PDB structure file is referred to as the wild type, or non-mutated form, of a protein. We use the PDB file as a template, from which mutants are in silico generated (see Section III-C). Also as input is a parameter k, that specifies the k-wise exhaustive mutations to be performed (Section III-B).

B. Identifying Mutant Set

Because the off-the-shelf software we rely on for generating protein mutants requires as input a sequence of amino acids corresponding to the mutant structure, as a first step our custom scripts systematically enumerate the amino acid sequences of all the possible variants with the k mutations. For example, assume there were only 3 kinds of amino acids, abbreviated `S`, `P` and `G`, and that the wild type protein sequence was `GS`. Then, the following would be the exhaustive list of sequences of mutant structures that have 2 mutations:

- Mutant 1 : `PP`
- Mutant 2 : `PG`
- Mutant 3 : `SG`
- Mutant 4 : `SP`

The general formula for the count of mutants, m, representing the exhaustive set of variants for an n residue protein, where each variant has k distinct amino acid substitutions relative to the wild type, is the following :

$$m = r^k \times \binom{n}{k} \tag{1}$$

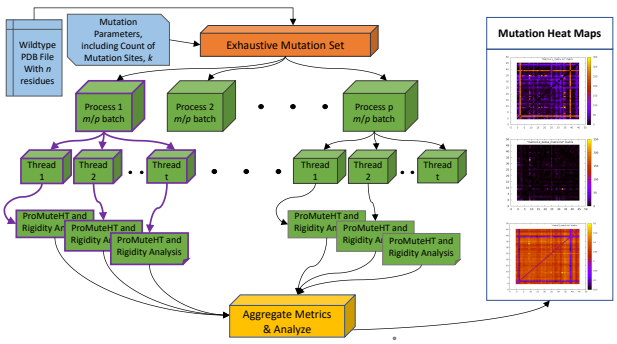


Fig. 2: Compute Pipeline : All pair-wise mutations are generated, rigidity analysis is performed, and the rigidity metrics are used to generated mutation heat maps.

where r refers to how many different amino acids can be substituted at any location in the wild type sequence of residues of a protein. Because there are 20 naturally occurring amino acids, then r has the value 19 in the context of protein structures. For the case study in Section IV, we generated and analyzed all possible mutants with 2 amino acid substitutions for PDB file 1crn, the 46 amino acid protein crambin. Thus, we generated a total of $19^2 \times \binom{46}{2} = 361 \times 1,035 = 373,645$ mutants, each with 2 mutations.

C. Generating Mutants & Rigidity Analysis

To generate the mutant structures from the wild type template, we used the freely available ProMuteHT software [26]. It is a streamlined command-line java, C and python suite of tools capable of in silico mutating any residue to any other amino acid in a PDB protein file. We invoke ProMuteHT via the command line, and each invocation generates an output, mutant, PDB structure file. For our experiments, we used an Intel Core i7-2600 CPU at 3.4GHz, with 8 CPUs and 16GB of memory. We leveraged all 8 cores by spawning multiple processes, each of which was responsible for calculating the rigidity properties of an equally divided portion of the entire mutant set identified by the previous step of our compute pipeline. Run-times range from a few minutes for very small proteins, to hours in the case of PDB file 1hhp.

D. Rigidity Analysis Metrics

To help reason about the effects of mutations, we take an approach that does not rely on propensity tables, costly energy calculations, nor is dependent on homology data. Instead we rely on a fast combinatorial approach for assessing the rigidity of a protein [23], [27] (See Figure 1, Section II-A). In rigidity analysis, atoms and their chemical interactions are used to construct a mechanical model. A graph is constructed from the model, and pebble game algorithms [23] are used to analyze the rigidity of the associated graph. An analysis of the flexibility of the associated graph provides insights about the flexibility of the protein from which the model and associated graphs were constructed.

In this work we compare the rigidity analysis results of the wild type, non-mutated form of a protein, to the rigidity analysis results of a mutant with 2 amino acid substitutions. We do this for all wild type, mutant pairs. We take inspiration from previous studies in which the use of several rigidity-based metrics were demonstrated in helping to discern the effects of mutations. The first metric, the Largest Rigid Cluster (LRC) [28], is a tally of the count of atoms in the largest rigid cluster. Comparing the LRC of the wild type relative to the LRC of a mutant was shown to be a fair indicator of the effect of single mutations, using the hypothesis that a mutant with a smaller LRC than the wild type's LRC has a mutation that is structurally destabilizing. We make use of the LRC rigidity metrics in this work, and refer to any metric that relies on LRC as an **LRC** metric.

A second class of rigidity based metrics which were developed subsequently considered all cluster sizes and their counts – and not only the LRC – in discerning the effect of an in silico mutation [29], [30]. We refer to these metrics as **Rigidity Cluster Bin** metrics. The $RD_{WT \rightarrow mutant}$ rigidity distance metric was developed to assess the impact of an in silico mutation(s) on the stability of a protein :

$$RD_{WT \rightarrow mutant} : \sum_{i=1}^{i=LRC} i \times [WT_i - Mut_i] \qquad (2)$$

where WT refers to Wild Type, Mut refers to mutant, LRC is the size of the Largest Rigid Cluster (in atoms). Each summation term of the $RD_{WT \rightarrow mutant}$ metric calculates the difference in the count of a specific cluster size, i, of the wild type and mutant, and weighs that difference by i.

E. Rigidity Metric Heat Maps

Because of the large count of protein variants that would be produced from an exhaustive mutation screen involving 2 mutations for even a relatively small protein, we were in need of a visualization scheme to facilitate the interpretation of the large output data. We have developed several Rigidity Metric Heat Maps to help identify the residue or pairs of residues, which when mutated, would have the greatest impact on the stability of a protein.

The first of these heat maps considers the Largest Rigid Cluster among all of the possible 361 mutants that can be generated for any pair of 2 amino acids in a protein. There are 361 possible mutants because each amino acid can be mutated to one of the 19 other ones, and thus the possible set of all pairwise mutations for two amino acids is $19^2 = 361$. The general form of an LRC rigidity metric for any pair of 2 amino acids is of the following form:

$$\{mean, max, min\} \sum_{x=1}^{19} \sum_{y=1}^{19} LRC_{WT} - LRC_{Mut_{xy}} \qquad (3)$$

where the summation over x refers to the 19 different mutations that can be performed at the selected first residue, and the summation over y refers to the 19 different mutations that can be performed at the selected second residue. The $mean, max, min$ components refers to the fact that the LRC-based metric for all 361 mutants for any pair of amino acids can take of different forms, such as the average, minimum, difference, etc. of the RD value.

The second class of heat maps considers the $RD_{WT \rightarrow mutant}$ scores for all 361 mutants for a distinct pair of amino acids in a protein. The general form of an $RD_{WT \rightarrow mutant}$ rigidity metric for any pair of 2 amino acids is of the following form:

$$\{mean, max, min\} \sum_{x=1}^{19} \sum_{y=1}^{19} RD_{WT \rightarrow mutant_{xy}} \qquad (4)$$

where the x and y summations refer to the same concept as in the LRC rigidity metric, as do the $mean, max, min$ components.

A sample heat map is shown in Figure 3. Both the $x-$ and $y-$axis values designate amino acids in the chain of residues

in a protein. The color of any one cell in a heat map designates a value for a metric for all of the 361 mutant structures when the residues indicated by the x and y values were exhaustively mutated ($19 \times 19 = 361$). For example, the cell at $x = 2$ and $y = 3$ in Figure 3 is the metric calculated using all 361 mutants for when residues 2 and 3 were exhaustively mutated.

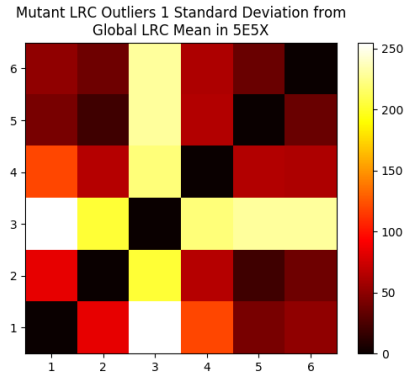


Fig. 3: Sample Heat Map : The color of any one cell specifies the value for a metric for all 361 mutants which were generated by exhaustively mutating the amino acids labeled on the x and y axes.

IV. RESULTS - CASE STUDIES

To demonstrate the utility of our approach in performing an exhaustive screen of all pair-wise mutations in a protein with the aim that we identify pairs of amino acids that have a significant pronounced effect when mutated, we analyze the PDB file 1crn, the 46 amino acid protein crambin, and PDB file 1hhp, the 99 amino acid HIV-1 protease. We also performed an analysis of the 67 residue pdb file 1csp, which is of the crystal structure of the bacillus subtilis major cold shock protein. Because each of the 46 amino acids in 1crn can be mutated to 19 different amino acids, and because all of our mutants had 2 amino acid substitutions, the total count of mutants we generated was 373,635 (see Section III-B), and for 1hhp, 1,751,211 mutants. For 1csp, a total of 798,171 mutants were generated.

A. LRC metric, average $LRC_{WT} - LRC_{Mut}$

For our first heat map visualization of the rigidity metric data, we tallied the average of the $LRC_{WT} - LRC_{Mut}$ score for all 361 mutants for each 2 pairs of amino acids in 1crn. The average scores ranged from approximately 0 to +200, shown in Figure 4. In the case of a large positive number, the LRC_{mut} is far smaller than the LRC_{WT}, which we infer to mean that the mutant is less stable than the wild type. Said differently, mutating two residues that causes a big decrease to the rigidity metric of a protein indicates that those 2 amino acid substitutions are together highly structurally destabilizing because the mutant has far fewer rigid clusters of significant size.

Noteworthy in the heat map in Figure 4 are the residues 10 and 26. The striped bars designate that regardless of which other residue is also mutated along with either of those two, the

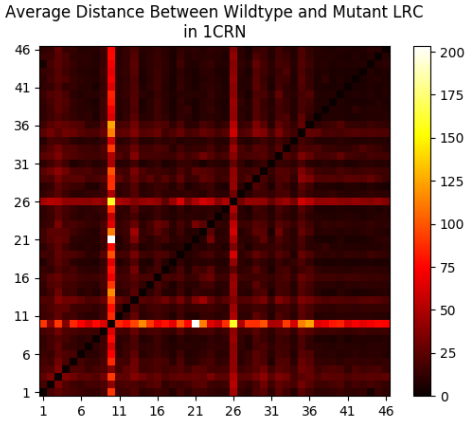


Fig. 4: LRC Heat Map : average $LRC_{WT} - LRC_{Mut}$.

average $LRC_{WT} - LRC_{Mut}$ score for all pairs of mutations is nearing 100 or far more. This is telling that any pair mutations that involve one of those residues will result in a mutant structure that is far less stable than the wild type.

The information gleamed from this heat map might prove of biological significant on many fronts. For example, if a researcher wants to identify which pairs of residues should be mutated to have the greatest chance of destabilizing a protein, then residues 10 and 26 should not be included among the likely candidates of residues.

B. Rigid Cluster Bin Distance, average $RD_{WT \to mutant}$

To assess the utility of a Rigid Cluster Bin Distance in its ability to identify pairs of residues that when mutated had a pronounced effect on the stability of a protein, we generated a heat map analogous to Figure 4, but using binned cluster values instead of LRC scores, for 1crn. In Figure 5, the horizontal bands at residue 11 and 27 appear. This indicates that the averaged RD scores as well as the averaged LRC scores are able to identify those residues that are not good mutation candidates if the goal is to destabilize a protein.

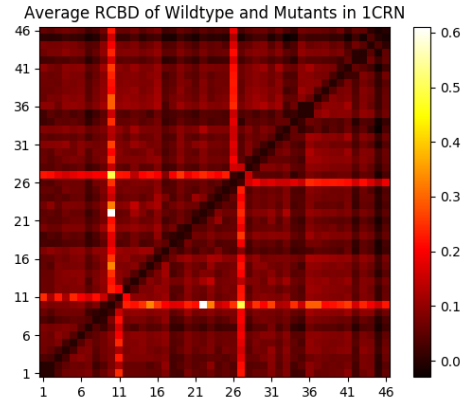


Fig. 5: Rigid Cluster Bin Heat Map : average $RD_{WT \to mutant}$.

C. LRC Metric, $LRC_{WT} - LRC_{Mut}$ Outliers, 1SD+

Because of our large data set (373,635) for PDB file 1crn, any use of averaging of the metric scores might hide outlier values that may exist for any one pair of amino acids that were mutated. To identify outlier individual pairs of amino acids that had a very strong effect on the stability of a protein when mutated, we tallied the number of mutations whose LRC metric is more than one standard deviation from the mean LRC metric over all 373,635 mutants. Heat Map show in Figure 6.

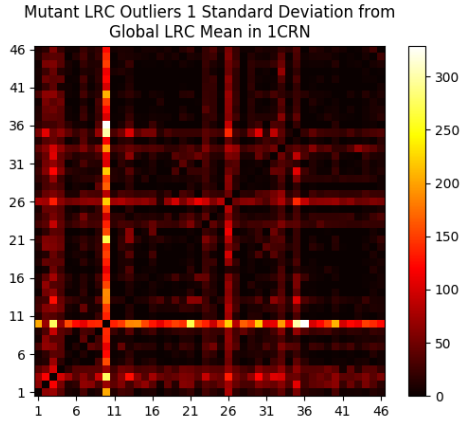


Fig. 6: LRC Metric, $RD_{WT \rightarrow mutant}$ Outliers 1SD+.

In Figure 6, we see that the point at residues 35 and 36 (on the y axis) are bright yellow, indicating that of the 361 possible mutants involving those amino acids, more than 300 have an $LRC_{WT} - LRC_{Mut}$ score that is more than 1 standard deviation from the average LRC among all pairs of mutated residues.

D. LRC Metric, $LRC_{WT} - LRC_{Mut}$ Outliers, 3SD+

To identify those pairs of amino acids that when mutated had a significant, pronounced effect on the structural stability of a protein, we performed a similar analysis described in Section IV-G, but tallied the count of mutants that had $LRC_{WT} - LRC_{Mut}$ scores at least 3 standard deviations from the mean LRC metric for all 373,635 mutants (Figure 7).

In Figure 7, we see several pairs of amino acids, including 3 and 27, that when mutated, have a non-trivial count (upwards of 10 or more) of LRC scores among their 361 mutants that had LRC scores at least 3 standard deviations from the mean. Our approach identified these residues as the most resistant to mutations, because so many of the mutants involving those residues had LRC scores that were excessively high.

E. LRC Metric, $LRC_{WT} - LRC_{Mut}$ Outliers, 1SD-

To identify those pairs of residues that, when mutated, had a pronounced destabilizing effect on the protein, we tallied the count of the 361 mutants for each pair of amino acid substitutions that had an LRC score that was at least 1 standard deviation below the mean of the LRC score for all 373,635 mutants. Heat Map shown in Figure 8.

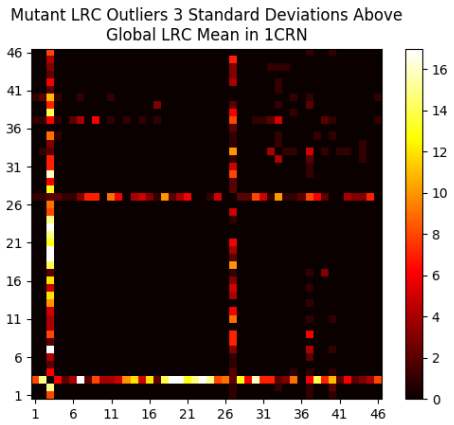


Fig. 7: LRC Metric, $RD_{WT \rightarrow mutant}$ Outliers 3SD+.

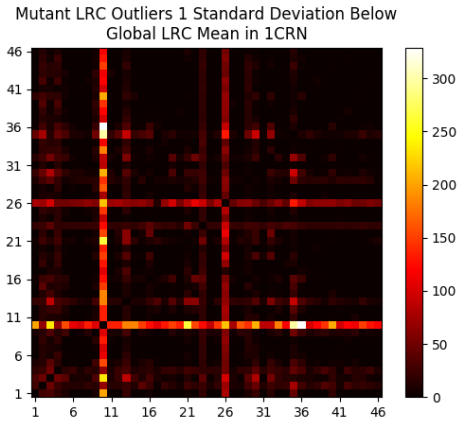


Fig. 8: LRC Metric, $RD_{WT \rightarrow mutant}$ Outliers 1SD-.

In Figure 8, several pairs of amino acids had high counts (nearing 350) of the 361 mutants that had LRC values at least 1 standard deviation below the average LRC across all 373,636 mutants. Specifically, when residues 33 and 8, and 33 and 7, were mutated, most of the resulting mutants had Largest Rigid Cluster far small than the average LRC for all experiments. This indicates that those pairs of residues might be good targets of mutation studies or protein engineering attempts aiming to maximally destabilize a protein.

We also assessed the heat maps for the far larger PDB file 1hhp. Noteworthy in the heat map in Figure 9 are a series of residues in the 20-25 range, as well as residues 34, and 83 and 87. The dark striped bars at those locations designate that even if those residues are mutated, along with any other one, the average $LRC_{WT} - LRC_{Mut}$ score for all pairs of mutations is nearing 0, indicating the the mutation seems to have no effect on the size of the largerst rigid cluster. Interesting, those residues are not involved in the catalytic action performed by the protein, and one might say therefore they are **not** critical, so mutating them would not have any effect on the protein's stability. Indeed that is what we see because there is no change in the mutant's rigidit relative to the wildtype's rigidity when

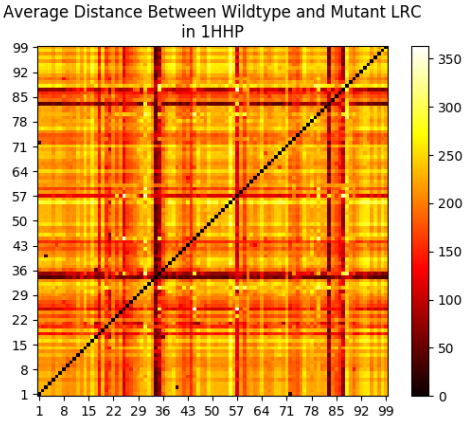


Fig. 9: LRC Heat Map : average $LRC_{WT} - LRC_{Mut}$.

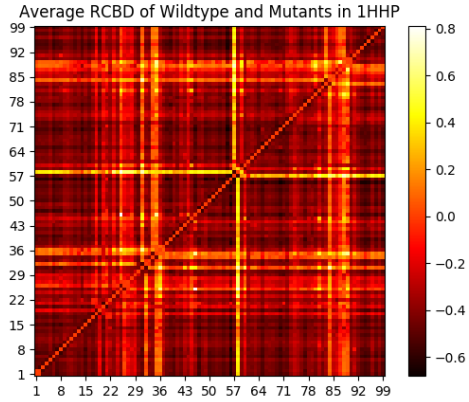


Fig. 10: Rigid Cluster Bin Distance : average $RD_{WT \to mutant}$.

those residues are mutated.

There are bright (nearly white) points in the heatmap in figure 9, which tell the inverse story. Residue 25 (y axis) often is a bright white point, specifying a very high average rigidity distance between the wild type and mutant. This indicates that mutating that residue results in a mutant that is **very** unstable relative to the wildtype because the difference between the wildtype and mutant rigid metric is high. Interestingly, residue 25 is one of only a few residues that are involved in the catalytic activity of the protein, so they play a special, critical, role in the protein's shape and function. The heat map thus shows that when residue 25 is mutated, in many cases the resulting mutant structure is vastly less stable that the wildtype, as might be expected.

F. Rigid Cluster Bin Distance, average $RD_{WT \to mutant}$

To assess the utility of a Rigid Cluster Bin Distance in its ability to identify pairs of residues that when mutated had a pronounced effect on the stability of a protein, we generated a heat map analogous to Figure 9, but using binned cluster values instead of LRC scores. In Figure 10, the brightest yellow bar appears near residue 57. This indicates that the averaged RD scores as well as the averaged LRC scores are able to identify that residues that are not good mutation candidates if the goal is to destabilize a protein.

G. LRC Metric, $LRC_{WT} - LRC_{Mut}$ Outliers, 1SD+

Because of our large data set for PDB file 1hhp, any use of averaging of the metric scores might hide outlier values that may exist for any one pair of amino acids that were mutated. To identify outlier individual pairs of amino acids that had a strong effect on the stability of a protein when mutated, we tallied the number of mutations whose LRC metric is more than one standard deviation from the mean LRC metric over all 1,751,211 mutants. Heat Map show in Figure 11.

In Figure 11, we see that the points at residues 35 (on the x axis) and residue 57 (y axis) is bright yellow, indicating that of the 361 possible mutants involving those amino acids, nearly 350 have an $LRC_{WT} - LRC_{Mut}$ score that is more than 1

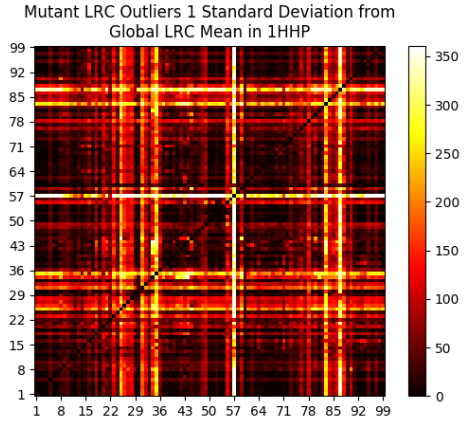


Fig. 11: LRC Metric, $RD_{WT \to mutant}$ Outliers 1SD+.

standard deviation from the average LRC among all pairs of mutated residues.

H. LRC Metric, $LRC_{WT} - LRC_{Mut}$ Outliers, 2SD+

To identify those pairs of amino acids that when mutated had a significant, pronounced stabilizing effect on the structural stability of a protein, we performed a similar analysis described in Section IV-G, but tallied the count of mutants that had $LRC_{WT} - LRC_{Mut}$ scores at least 3 standard deviations from the mean LRC metric for all 1,751,211 mutants (Figure 12).

In Figure 12, we see several pairs of amino acids, including 70 and 32 ,that when mutated, have a non-trivial count (upwards of 10) of LRC scores among their 361 mutants that had LRC scores at least 2 standard deviations from the mean. Our approach identified these residues as the most resistant to mutations, because so many of the mutants involving those residues had LRC scores that were excessively high.

I. LRC Metric, $LRC_{WT} - LRC_{Mut}$ Outliers, 1SD-

To identify those pairs of residues that, when mutated, had a pronounced destabilizing effect on the protein, we tallied

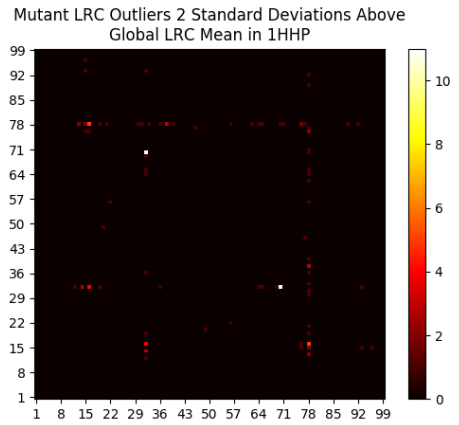


Fig. 12: LRC Metric, $RD_{WT \rightarrow mutant}$ Outliers 2SD+.

the count of the 361 mutants for each pair of amino acid substitutions that had an LRC score that was at least 1 standard deviation below the mean of the LRC score for all 1,751,211 mutants. Heat Map shown in Figure 13.

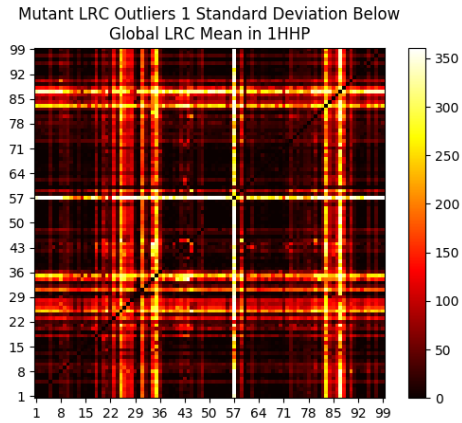


Fig. 13: LRC Metric, $RD_{WT \rightarrow mutant}$ Outliers 1SD-.

In Figure 13, several pairs of amino acids had high counts (nearing 350) of the 361 mutants that had LRC values at least 1 standard deviation below the average LRC across all mutants. Specifically, when residue 57 was mutated, most of the resulting mutants had Largest Rigid Cluster far small than the average LRC for all experiments, resulting in a very high rigidity metric value. This indicates that mutating that residue might be good target of mutation studies or protein engineering attempts aiming to maximally destabilize a protein. Indeed that residue is near a part of the protein commonly referred to as a flap or arm, which is involved in a mechanical motion necessary for the protein's function, and mutating it might reduce the protein's capacity to perform its function.

Lastly, we also analyzed the 67 amino acid protein structure 1csp. Various heatmaps are shown in Figure 14, including average $LRC_{WT} - LRC_{Mut}$, average $RD_{WT \rightarrow mutant}$, $RD_{WT \rightarrow mutant}$ Outliers 1SD+, and $RD_{WT \rightarrow mutant}$ Outliers 1SD-. We include the heatmaps for 1csp to showcase that the pairwise mutation rigidity analysis approach is quite unique to the structure being analyzed. For 1csp, residues 8 and 9, as well as 16 and 17, are identified as having the greatest impact on the structural stability of the protein when mutated in combination with any other residue.

V. CONCLUSIONS & FUTURE WORK

In this work we have motivated the need for, and developed a compute pipeline capable of exhaustively generating in silico all mutants with 2 mutations for a user-specified PDB protein structure file. We have analyzed the flexibility of each mutant using an efficient combinatorial algorithm, and have analyzed the rigidity metrics and fashioned a heat map to enable identifying those pairs of amino acids that have a high impact on the structure of a protein when mutated. To our knowledge, this is the first compute pipeline capable of conducting an in silico exhaustive mutation screen for all pairwise mutations.

We envision a variety of next steps. Firstly, several of the off-the-shelf tools that are components of our pipeline require extensive I/O operations. And although performing a few I/O operations such as reading from or writing to a file is not problematic, doing it hundreds of thousands, or millions of times – as we have done – can greatly increase overall runtimes. To address this, we are refining our pipeline as well as several of the open-source tools that we use to minimize the use of I/O operations. Also, correlation studies are underway, in which we are calculating Pearson Correlation as well as RMSE values for predictions based on the rigidity metrics we have generated, and the predictions they are used to make, against experimental data for the effects of the mutations involving those residues.

REFERENCES

[1] T. Xiao, J. Takagi, B. S. Coller, J.-H. Wang, and T. A. Springer, "Structural basis for allostery in integrins and binding to fibrinogen-mimetic therapeutics," *Nature*, vol. 432, no. 7013, p. 59, 2004.

[2] S. Piana, P. Carloni, and U. Rothlisberger, "Drug resistance in hiv-1 protease: Flexibility-assisted mechanism of compensatory mutations," *Protein Science*, vol. 11, no. 10, pp. 2393–2402, 2002.

[3] R. J. Dunbrack and M. Karplus, "Conformational analysis of the backbone-dependent rotamer preferences of protein sidechains," *Nature Structural Biology*, vol. 1, pp. 334–340, 1994.

[4] J. Janin and S. Wodak, "Conformation of amino acid side-chains in proteins." *J Mol Biol*, vol. 125, no. 3, pp. 357–386, Nov 1978.

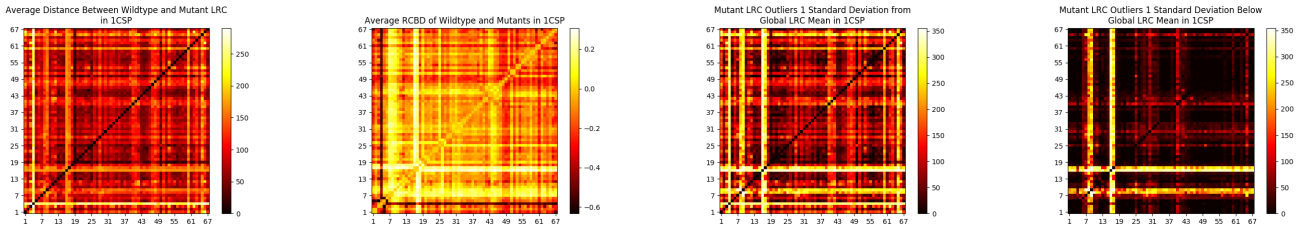
[5] J. Ponder and F. Richards, "Tertiary templates for proteins: Use of packing criteria in the enumeration of allowed sequences for different structural classes," *Journal Molecular Biology*, vol. 193, pp. 775–791, 1987.

[6] D. Gilis and M. Rooman, "Predicting protein stability changes upon mutation using database-dervied potentials: Solvent accessibility determines the importance of local versus non-local interactions along the sequence," *Journal of Molecular Biology*, vol. 272, no. 2, pp. 276–290, 1997.

[7] C. Lee and M. Levitt, "Accurate prediction of the stability and activity effects of site-directed mutagenesis on a protein core," *Nature*, vol. 352, pp. 448–451, 1991.

[8] C. Topham, N. Srinivasan, and T. Blundell, "Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitutions and propensity tables," *Protein Engineering*, vol. 10, no. 1, pp. 7–21, 1997.

[9] J. R. Brender and Y. Zhang, "Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles," *PLoS Comput Biol*, vol. 11, no. 10, p. e1004494, 2015.



(a) (b) (c) (d)

Fig. 14: Heatmaps for exahustive mutations for 1csp. (a) LRC Heat Map : average $LRC_{WT} - LRC_{Mut}$. (b) Rigid Cluster Bin Heat Map : average $RD_{WT \to mutant}$. (c) LRC Metric, $RD_{WT \to mutant}$ Outliers 1SD+. (d) LRC Metric, $RD_{WT \to mutant}$ Outliers 1SD-.

[10] C. Worth, R. Preissner, and L. Blundell, "Sdm-a server for predicting effects of mutations on protein stability and malfunction," *Nucleic Acids Research*, vol. 39, no. Web Server Issue, pp. W215–W222, 2011.

[11] J. Cheng, A. Randall, and P. Baldi, "Prediction of protein stability changes for single-site mutations using support vector machines," *PROTEINS: Structure, Function, and Bioinformatics*, vol. 62, pp. 1125–1132, 2006.

[12] F. Jagodzinski, B. Akbal-Delibas, and N. Haspel, "An evolutionary conservation & rigidity analysis machine learning approach for detecting critical protein residues," in *CSBW (Computational Structural Bioinformatics Workshop), in proc. of ACM-BCB (ACM International conference on Bioinformatics and Computational Biology)*, September 2013, pp. 780–786.

[13] L. Jia, R. Yarlagadda, and C. C. Reed, "Structure based thermostability prediction models for protein single point mutations with machine learning tools," *PloS one*, vol. 10, no. 9, p. e0138022, 2015.

[14] Y. Li and J. Fang, "Prots-rf: a robust model for predicting mutation-induced protein stability changes," *PloS one*, vol. 7, no. 10, p. e47247, 2012.

[15] E. Capriotti, P. Fariselli, and R. Casadio, "A neural-network-based method for predicting protein stability changes upon single point mutations," *Bioinformatics*, vol. 20, Supplemental, pp. i63–i68, 2004.

[16] W. Lee, P. Yue, and Z. Zhang, "Analytical methods for inferring functional effects of single base pair substitutions in human cancers," *Human Genetics*, vol. 126, no. 481-498, 2009.

[17] S. Mooney, "Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis," *Briefings in Bioinformatics*, vol. 6, pp. 44–56, 2005.

[18] S. Henikoff and P. Ng, "Predicting the effects of amnio acid substitutions on protein functions," *Annual Reviews of Genomics Human Genetics*, vol. 7, pp. 61–80, 2006.

[19] S. Teng, E. Michonova-Alexova, and E. Alexov, "Approaches and resources for prediction of the effects of non-synonymous single nucleotide polymorphisms on protein function and interactions," *Current Pharmacology Biotechnology*, vol. 9, pp. 123–133, 2008.

[20] C. Topham, N. Srinivasan, and T. Blundell, "Prediction of the stability of protein mutants based on structural environment-dependent amino acid substitution and propensity tables," *Protein Engineering*, vol. 10, pp. 7–21, 2012.

[21] M. Masso and I. I. Vaisman, "Structure-based prediction of protein activity changes: assessing the impact of single residue replacements," in *Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE*. IEEE, 2011, pp. 3221–3224.

[22] S.-Y. Rhee, J. Taylor, W. J. Fessel, D. Kaufman, W. Towner, P. Troia, P. Ruane, J. Hellinger, V. Shirvani, A. Zolopa, and R. W. Shafer, "Hiv-1 protease mutations and protease inhibitor cross-resistance," *Antimicrobial Agents and Chemotherapy*, vol. 59, no. 8, pp. 4253–4261, 2010.

[23] D. Jacobs, A. Rader, M. Thorpe, and L. Kuhn, "Protein flexibility predictions using graph theory," *Proteins*, vol. 44, pp. 150–165, 2001.

[24] N.Fox, F.Jagodzinski, Y.Li, and I.Streinu, "KINARI-web: A server for protein rigidity analysis," *Nucleic Acids Research*, vol. 39 (Web Server Issue), pp. W177–W183, 2011.

[25] S. Radestock and H. Gohlke, "Exploiting the link between protein rigidity and thermostability for data-driven protein engineering," *Engineering in Life Sciences*, vol. 8, no. 5, pp. 507–522, 2008.

[26] E. Andersson and F. Jagodzinski, "Promuteht: A high throughput compute pipeline for generating protein mutants in silico," in *Proceedings of the 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, ser. ACM-BCB '17. New York, NY, USA: ACM, 2017, pp. 655–660. [Online]. Available: http://doi.acm.org/10.1145/3107411.3116251

[27] N. Fox, F. Jagodzinski, and I. Streinu, "KINARI-Lib: a C++ library for pebble game rigidity analysis of mechanical models," in *Minisymposium on Publicly Available Geometric/Topological Software, Chapel Hill, NC, USA*, June 2012.

[28] F. Jagodzinski, J. Hardy, and I. Streinu, "Using rigidity analysis to probe mutation-induced structural changes in proteins," *Journal of Bioinformatics and Computational Biology*, vol. 10, 2012.

[29] R. Farhoodi, M. Shelbourne, R. Hsieh, N. Haspel, B. Hutchinson, and F. Jagodzinski, "Predicting the effect of point mutations on protein structural stability," in *proc. of ACM-BCB (International Conference on Bioinformatics, Computational Biology and Health Informatics)*, August 2017.

[30] E. Andersson, R. Hsieh, H. Szeto, R. Farhoodi, N. Haspel, and F. Jagodzinski, "Assessing how multiple mutations affect protein stability using rigid cluster size distributions," in *Computational Advances in Bio and Medical Sciences (ICCABS), 2016 IEEE 6th International Conference on*. IEEE, 2016, pp. 1–6.