Large scale analyses of single residue mutations on Residue Interaction Networks

Gummala Sreenivas
MS16009

A dissertation submitted for the partial fulfilment of BS-MS dual degree in Science

Indian Institute of Science Education and Research Mohali
April 2021

Certificate of Examination

This is to certify that the dissertation titled "**Large scale analyses of single residue mutations on Residue Interaction Networks**" submitted by **Mr Gummala Sreenivas** (Reg. No. **MS16009**) for the partial fulfilment of the BS-MS dual degree programme of the Institute, has been examined by the thesis committee duly appointed by the Institute. The committee finds the work done by the candidate satisfactory and recommends that the report be accepted.

Dr Santosh B. Satbhai

Dr Kuljeet Sandhu

Dr Shashi Bhushan Pandit
(Supervisor)

Dated: April 30, 2021

Declaration

The work presented in this dissertation has been carried out by me under the guidance of **Dr Shashi Bhushan Pandit** at the Indian Institute of Science Education and Research Mohali.

This work has not been submitted in part or in full for a degree, a diploma, or a fellowship to any other university or institute. Whenever contributions of others are involved, every effort is made to indicate this clearly, with due acknowledgement of collaborative research and discussions. This thesis is a bonafide record of original work done by me and all sources listed within have been detailed in the bibliography.

Gummala Sreenivas

MS16009

April 30, 2021

In my capacity as the supervisor of the candidate's project work, I certify that the above statements by the candidate are true to the best of my knowledge.

Dr Shashi Bhushan Pandit

(Supervisor)

Acknowledgement

I would like to take this rare opportunity of acknowledging and expressing my gratitude to all those who helped me all along this journey. To the foremost, I would like to express my deepest gratitude to my guide Dr Shashi Bhushan Pandit, Associate Professor of Biological Sciences at Indian Institute of Science Education and Research Mohali, for giving me a chance to work in his lab despite my not so professional programming skills. I am so grateful for this opportunity and I have learned a lot under his guidance. No matter how late or early in the day, he was always available. I always introduced him to my parents and my friends as the most hardworking person that I have ever met. He is one of my greatest inspirations of all time, who reminds me to keep trying. Thanks to the senior members of the lab Deeksha and Paras, whom you can always approach for guidance. I always felt they had answers to all the questions, no matter if it is something related to biology or programming for that matter. Special thanks to Purva and Rupesh, whom I always bothered with all my silly questions, if it was not for them this project would have been really hard for me. They were more than just colleagues in the lab. They were good friends and I always enjoyed their company even after the lab hours. I would also like to thank Deepanshi, Ruby, Shabduli and Teja for their support all along this journey.

Special thanks to Vishal, a physics major at IISER Mohali, for his programming inputs and a unique perspective about my work which helped me save a lot of time and rethink my approach to different problems in the project. It is great to have such a geeky friend.

Sincere thanks to all DBS faculty members, all colleagues and teaching staff for their help.

To the last but not least, I would also like to thank my sister and my parents, who would always guide and support me through various ups and downs in life. I am always grateful for their support and I can’t thank them enough.

I feel very much lucky to have met such an amazing company of unique people all along this untrodden path.

-Sreenivas Gummala

List of figures

List of tables

Notations

ATP	Adenosine triphosphate
CD-HIT	Cluster Database at High Identity with Tolerance
GO	Gene Ontology
mmCIF	macromolecular Crystallographic Information File
MT	Mutant
PCN	Protein Contact Network
PDB	Protein Data Bank
RIN	Residue Interaction Network
RING	Residue Interaction Network Generator
RMSD	Root Mean Square Deviation
SNP	Single Nucleotide Polymorphism
WT	Wild Type

Contents

Abstract

The phenotypic effect of single residue mutation in proteins can vary from no effect to complete loss of function. Such varied effects have intrigued structural biologist and biophysicist, over decades, because it has remained difficult to derive a general guiding structural principle to explain changes in structural properties of proteins due to mutations. Usually, detailed structural effects on mutations are obtained by comparison of experimentally determined tertiary structures of wild-type and mutant proteins. Such previous studies have mostly focused on understanding changes in protein thermodynamic stability, which led to development of modeling or prediction of change in protein stability. Moreover, these have been also employed to understand human diseases as it has been shown that of these are due to mutations involving non-synonymous single-nucleotide polymorphisms (nsSNP). The effect on protein function due to single residue mutation can be attributed to impairment of stability, defective interactions with other biomolecules (ligand/proteins), and residue packing density. Despite many studies on single residue mutant structures, the effect on wild type protein residue interaction network (RIN) due to such mutation, mostly, unexplored. In the present study, we have systematically investigated the effect of mutation by comparing RIN of wild type and mutant proteins. Furthermore, we studied the network perturbation using closeness centrality due to mutation. Through these studies, we have explored whether changes in residue interaction network can explain the functional shift in proteins. The comparison of C-α residue network of wild-type and mutant multi-domain proteins suggested that global network centrality measures remain mostly unchanged on mutation. However, a small subset of proteins showed a large change in global network parameters that could not be correlated to conformational change due to mutation or belonged to specific functional families. Interestingly, local network features show remarkable changes, which seem to propagate in the protein network in some cases to really large spatial distances. Thus, suggesting allosteric effect of mutation. Importantly, this study can provide insights for rational design of protein with a desired feature.

Chapter 1

Introduction

1.1 Protein tertiary structure

Most globular proteins adopt tertiary structure for their function. However, some proteins remain floppy or have no stable structure, which on interaction with their cognate ligand can adopt partial secondary structures. These are referred to as intrinsically disordered proteins (IDP). The individual amino acids, the basic building blocks of proteins, interact to form regular secondary structures elements: helices and the sheets. These arrange in specific orientations having many tertiary contacts between sequentially distant but spatially close residues to form tertiary structures.

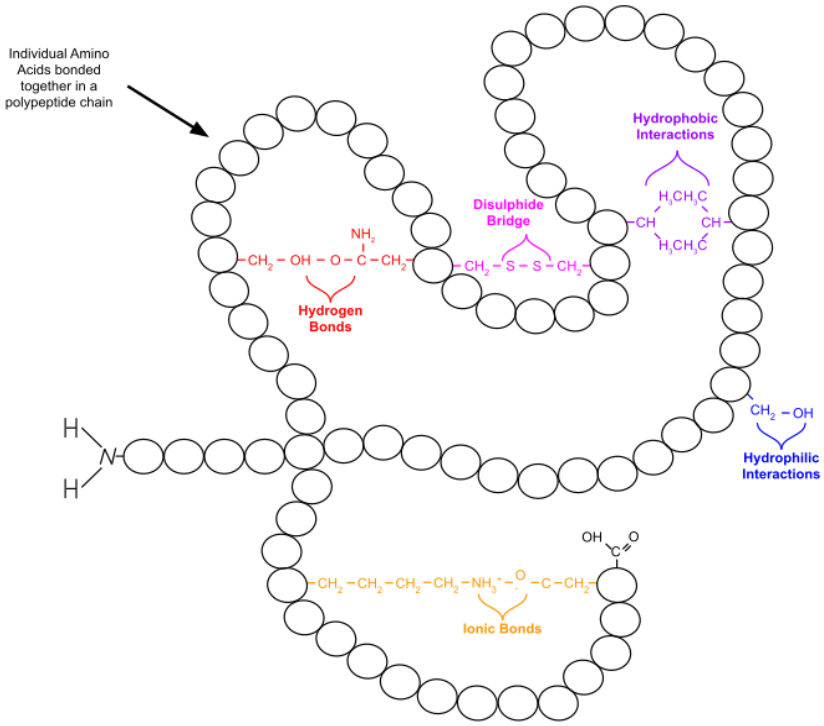


Fig 1.1 Interactions leading to the formation of a tertiary structure

The tertiary structures arrange themselves to form quaternary arrangement. The basic interacting forces, which stabilize protein structures, are: hydrophobic interactions, electrostatic forces, disulfide bonds and hydrogen bonds. Among these hydrogen bonds is an important interacting force as it dictates both specificity and directionality of interaction.

The tertiary contacts interactions enable amino acids, which are otherwise far away in the order of sequence to interact with one another. The number and type of interactions that these amino acids form with one another decide the shape of the protein. The tertiary structure of protein provides a specific shape along with forms multiple clefts/crevices providing sites of interactions with other biomolecules. Therefore, knowledge of tertiary structure can provide insights into molecular details the proteins function and also provide understanding of fundamental forces responsible for protein stability as well as their folding. In general, most globular proteins can fold to form tertiary structures, however, some proteins require assistance of chaperones in folding. Several experimental and computational efforts have been going on for decades to unravel the combination and contribution of various interacting forces in stabilizing or folding of globular proteins.

Since knowledge tertiary structure of proteins is essential for deciphering molecular details of protein function, significant efforts have been invested in experimental determination of protein tertiary structures. Towards this, a structural genomics consortium was established to determine structures of representative proteins from genomic sequences. This has led to a massive increase in experimental structures deposited in Protein Data Bank (PDB). Over past decade, significant progress has been made in the area of structure modeling allowing prediction of structures of proteins sequences even remotely related to protein sequence of known structure. Through these efforts, one wish to leverage sequence-structure-function paradigm to model structure of protein based on experimentally determined structure and using this one can find function of proteins. Even though, this is feasible but the success rate of function prediction has not proven to be remarkable due to complexity in defining molecular function.

Moreover, this also opens up a great potential for directing the genetic engineering of proteins for therapeutic, and the laboratory design of proteins with desirable properties for biotechnological purposes.

1.2 Single residue mutations (SRM)

Protein evolves through the process of mutation and functional divergence. The single residue mutation (SRM), as the name suggests refer to single residue mutant of a protein. In experimental studies, this has been expensively employed to elucidate the function of proteins by creating a single residue mutation and assaying for the function of mutant protein. For instance, alanine scanning mutation method wherein every residue is mutated to Alanine for identifying critical functional residues in protein. Apart from these designed single residue mutations, these are also known to be responsible for many human diseases and many of these mutations occur in population as non-synonymous single-nucleotide polymorphisms (nsSNP). As can be seen, phenotypic effect of mutation has been found to range from no effect to mild effect and a complete loss of function, which leads to diseased state. Apart from these, protein has also been found to exhibit changes in biophysical properties such as increase/decrease in their thermodynamic stability (Anupan banerjee, Levy and Mitra, Journal of Proteome, 2020). Such wide varied effects of mutations have puzzled structural biologists and biophysicist over decades and several studies have been dedicated to understand or decipher role residue spatial position in the protein. From many such studies it has been deduced that mutant protein loses its function because impairment of their stability, defective interactions with other biomolecules (ligand/proteins), and loss in residue packing density (reference: Pandurangan and Blundell, Tools for protein science, 2019; Anupan banerjee, Levy and Mitra, Journal of Proteome, 2020). However, effect of many single residue mutations still remains to be explained. Importantly, insights from studies on single residue mutation can be useful in protein engineering. These mutations have been of great interest to people for several decades, because of their role in diseases, it could aid in drug discovery and their importance in understanding structural and evolutionary studies.

One such example is the study of unnatural amino acids (UAAs) in *Chromobacterium violaceum*. UAAs are chiral amines with high application in drug discovery and synthesis of other valuable compounds. Though bio catalysis, this is also known to synthesize pure UAAs, there was a dire need for non-native substrates and also to understand key amino acids, which are important for the activity of these UAAs. This study required systemic testing on different substrates, which relied heavily on these single residue mutations. A similar approach has been taken while trying to understand the reason for a significant difference in thermostability

between two isoforms of human cytosolic creatine kinase, which helps in maintaining the homeostasis of intracellular ATP level (Ref: Liu, H., Gao, YS., Chen, XJ. et al. A single residue substitution accounts for the significant difference in thermostability between two isoforms of human cytosolic creatine kinase, 2016). As mentioned before, single nucleotide polymorphism (SNP) is substitution of a single nucleotide at a specific position in the genome that is present in a sufficiently large fraction of the population. While SNPs can have a wide range of effects, in general, but if it is a non-synonymous SNP *i.e.*, when present within a coding sequence it leads to an amino acid change or single residue mutation. If the mutation causes a premature stop codon, it can result in the production of a truncated protein product or a near-null phenotype due to nonsense-mediated decay. These SNPs are referred to as non-synonymous SNP or missense mutation, which can have dire consequences of proteins and manifests as diseases in human. SNPs are also reported to occur in several proteins involved in drug transport, metabolism, and receptors ultimately affecting the pharmacokinetic and pharmacodynamic properties of many medications. Several such studies have been reported in the past, highlighting the importance of these single residue mutations.

Usually, detailed structural effects on mutations are obtained by comparison of experimentally determined tertiary structures of wild-type and mutant proteins. However, it is not always possible to experimentally determine tertiary structure of both wild type and mutant proteins. In such cases, computational approaches such modeling of mutant protein followed by their molecular dynamics have been used to study the effect of mutation. In past, most studies have focused on understanding changes in protein thermodynamic stability, which led to development of modeling or prediction of change in protein stability (Ref: Pandurangan and Blundell, Tools for protein science, 2019). Despite many studies on single residue mutant, it remains to find a general structural effect of single residue mutation on protein functions.

1.3 Residue Interaction Networks

Residue interaction network is graph representation of a protein where residues are modelled as nodes and interactions among them as edges. This provides an abstract representation of protein structure amenable to understand many system-level perturbation effects on protein. In past, these have been used to find allosteric paths in proteins, modeling of structural properties

(Ref: Stetz G, Verkhivker GM. Computational Analysis of Residue Interaction Networks and Coevolutionary Relationships in the Hsp70 Chaperones: A Community-Hopping Model of Allosteric Regulation and Communication. PLoS Comput Biol. 2017).

In most studies, usually involves C-alpha based residue interaction network where the C-alpha atom of a given amino acid is used to represent the nodes instead of all the residues present in amino acid. Representing all the atoms of an amino acid is computationally expensive while forming the network and calculating different network attributes could being confusing and intractable. RIN can utilize distances between C-alpha residue to construct a network or involve distances between side-chain to construct side-chain residue interaction network. The way in which edge is decided to form between two nodes in a network defines various types of RIN. This criterion can be used to broadly classify the networks into two types: distance-based and energy-based networks.

a. Distance-based RINs

The RIN, which, primarily, use distances between either two C-alpha atoms or any heavy atom between two residues to define an edge, is called distance-based RINs. Usually, the distance cutoff is set to be 7Å between C-alpha atoms of residues for modelling the proteins is around. It can sometimes be as low as 5Å, depending on the aim of the study. We have used networks constructed with distance-based cutoff of 7Å. Such a network is referred to as distance-based Protein Contact Network (PCN) or C-alpha based networks. Residues that are just adjacent to each other in the protein sequence anyway form a bond when modelled as distance-based PCNs. But that would form several redundant edges. To address this a slight modification is considered while defining an edge, which eliminates edges between sequence neighboring residues i.e., an edge is not defined between residues which are within ±2 or ±3 residues. Such kind of network is called a long-range network, in which no edge is formed with their immediate adjacent residues.

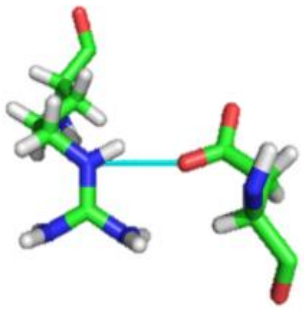
Despite their simplicity, these distance-based networks have their limitations. A value slightly larger than the cutoff is ignored while a value little less than is considered as important for interactions. These networks though might help in getting a basic

understanding, but to get a more robust network energy-based networks had to be considered.

b. Energy-based RINs

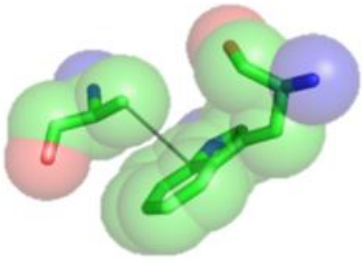
Just as the name suggests the edges within these networks are defined based on the energy-based interactions between them. These interactions apart from the distance criteria also set a wide range of characteristics such as the angle formed among the nodes, donor and acceptor atoms etc., depending on the type of interaction network. The energy-based networks can be broadly classified as under:

I. Hydrogen bond based



- Donor (D) and acceptor (A) atom pair is considered
- The distance, d, between A and D is less or equal to 3.5-5.5 Å
- The angle H-D-A, formed by the hydrogen of the donor, the donor itself and the acceptor atom, is less or equal to 63°
- Donors (according to RING-2.0): main chain NH; Arg NE, NH1, NH2; Asn ND2; His NE2, ND1; Ser OG; Tyr OH; Cys SG; Thr OG1; Gln NE2; Lys NZ; Trp NE1
- Acceptors (according to RING-2.0): Main chain C; Asn OD1; Gln OE1; Met SD; Asp OD1, OD2; Glu OE1, OE2; Ser OG; Thr OG1; His ND1; Tyr OH

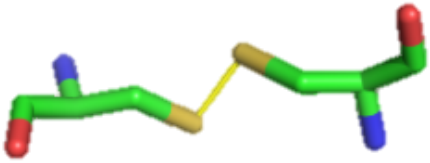
II. Van der Waals Interactions



- Distance between the surface of two atoms is measured
- The distance threshold is 0.5-0.8 Å
- Only carbon-carbon or carbon-sulfur atom pairs are considered for a valid VDW interaction

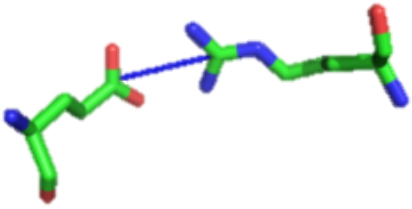
- Van der Waals radii are taken into consideration while calculating the distance
- Another special type of VDW involving N or O atoms and carbon is considered
- In particular only Glutamine NE2, OE1 and Asparagine ND2, OD1 are allowed

III. Disulfide bridges



- Covalent bonds
- Found with a very constant distance
- A bond is defined if the distance between SG atoms of cysteine is lower or equal to 2.5-3.0 Å

IV. Salt bridges (IONIC)



- Occur between residues with opposite charges
- Ionic interaction is possible if the distance between the mass centers of the charged groups are 4.0-5.0 Å

Other kinds of energy-based interactions include π-π stacking, π cation, coulomb interactions etc. The type of interaction that is formed depends on the question we are trying to address. In some cases, we can consider all the above-mentioned interactions together by constructing a weighted network in which the edge weight of a given edge, between any two residues, is decided by the interactions between those two residues. These networks though seem a bit complicated, give a more realistic understanding of the interactions that are occurring among the residues within a protein. Any small disturbance to this complex net of interactions could have a huge effect on proteins structural and functional properties. Considering the applications of single residue mutations as discussed in the above section, modelling them as RINs could give new insights into the protein behavior.

1.4 Objective

In general, it has been of great interest for scientists over decades to analyze the structural changes or functional changes accompanying residue mutations. Previous studies have mostly

focused on studying structural changes accompanying these single residue mutations and strong emphases have been on modeling biophysical property of thermodynamic stability. The mutations can severely affect the underlying residue interactions; however, this has not been paid much attention, except in cases when the protein is known to exhibit allostery and residue is an allosteric effector site. Recent studies have been reported addressing these issues but are mostly limited to specific sets of proteins (Ref: Srivastava A, Birari V, Sinha S. Small Conformational Changes Underlie Evolution of Resistance to NNRTI in HIV Reverse Transcriptase. Biophys J. 2020 May 19). Owing to increase in number of protein structures in PDB, it provides a unique opportunity to systematically perform a large-scale study on understanding the network perturbation of residue mutations.

The main goal of these studies is to explore the effects of single residue mutation on the structure and function of a protein using network-based representation of proteins. Through this study, we have addressed following questions: does mutation lead to global/local changes in protein networks? and is it possible to employ network properties to explain the change in functional shift of proteins due to mutation? In order to address these questions, we have performed a large-scale systematic comparison of various network parameters derived from RIN of wild (WT) and mutant (MT) proteins and correlated to general effect of residue mutation on protein function.

Chapter 2

Methodology

2.1 Dataset construction

To study effect of mutation on the network proteins, we initially considered a multi-domain protein dataset with ~80,000 PDB structures. This consists of protein structures having resolution less than 2.5 Å. We need to construct a dataset of pairs of structures having one wild type (WT) and other as single residue mutant (MT). For constructing such a dataset, first we mapped all possible protein structures in PDB database to the 'Uniprot' sequence identifiers or Uniprot ID either relying on SIFT database or by performing pairwise sequence alignment between uniprot sequence and 'atom' record sequences. Subsequently, we identified all Uniprot IDs having more than one PDB structures associated with it. These were further processed to identify WT or MT sequence by performing aligned of uniprot sequence to the PDB 'atom' record sequence. Since, in experimental determination of protein structures it is possible to have missing atom coordinates for some residues, we considered residue alignment overlaps of more than 90% i.e., the overlapping sequence should constitute a minimum of 90 per cent of the entire sequence. Further, we defined WT structures as having 100% identity with uniprot sequence and MT as those having changes in a single residue. This resulted in WT-MT pair for various uniprot IDs. We also filtered the dataset using 90% redundancy to construct a non-redundant dataset of proteins. Thus, resulting dataset consists of protein pairs of wild type (WT) and single residue mutant (MT), wherein the mutant protein differs from WT at a single residue position. After addressing other issues in dataset construction, we identified ~20,000 WT-MT pairs. The pairs, though might represent single residue mutant pairs, they need not necessarily be mutants because these sequences could have been poor fitting of residues, insertions/deletions, cloning artefact etc. Therefore, for a subset of pairs we manually curated the dataset to identify mutant protein by confirming it with other sources such as PDB (from 'author entry'), UniProt and literature. Various filtering parameters in several stages of dataset generation are described in detail later in this method section.

We used this dataset to build C-alpha based networks using network attributes obtained from RING executable file (which were further on visualized using Plotly library) and different network properties were calculated using NetowrkX, both at the global and the local levels.

2.1.1 Multi-domain proteins

In the present study, we have used multi-domain proteins for our study. This consisted of 80,637 PDB structures having resolution less than 2.5 Å and these span various diverse functional proteins. We considered multi-domain proteins for our study to consider the effect of domains in modeling the network perturbations. These proteins structures obtained from PDB were mapped to Uniprot IDs relying on SIFT database or performing alignment with uniprot sequence, which is provided in mmCIF mapping. Since different chains of a PDB usually tend to have similar sequences, therefore one among the chains belonging to the same PDB was randomly picked within the above-formed groups.

2.1.2 Construction of non-redundant dataset

It is widely known that PDB database is biased for certain proteins; therefore, we would like to reduce the effect of redundancy in the dataset. Moreover, over-representation of proteins in dataset can adversely bias the observation. We chose redundancy cut-off 90% sequence identity between Uniprot sequences after mapping of protein structures to keep the dataset enriched for deriving useful observations. In order to construct non-redundant dataset, we used CD-HIT (Ref: Weizhong Li, Adam Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics, 1 July 2006) clustering with a clustering parameter of 0.9.

1. The clustering of the above UniProt IDs was performed at 90% sequence identity using CD-HIT command:

   command: "cd-hit -i <input_fasta_file> -c 0.9 -g 1 -sc 1 -sf 1 -o <output_file>"

2. The representative UniProt ID of each cluster was picked.

3. The PDBs of the UniProt IDs belonging to a cluster were all grouped together and uniprot having the maximum number of pairs were considered for further studies.
4. The entire dataset of 80,637 different types of proteins is now broadly classified into 489 groups (i.e., 489 representative UniProt IDs).

2.1.3 Identification of single residue mutant structure

Before constructing WT-MT pairs, we identified a PDB structure as WT or MT based on sequence identity between uniprot sequence and 'atom' record sequence of PDB. For this, we performed pairwise sequence alignment between them and classified a structure as WT if it has 100% sequence identity to uniprot sequence otherwise, it is classified as MT structure, importantly, it should involve single residue mutation or change in single residue position. It is known in solving structures that it is difficult to model side chains of Asparagine, Glutamine and in many instances, residues are modeled as 'Alanine' if electron density can resolve until Cβ atom. In order to reduce these structural determination artifacts, we did not consider if the mutations involve any of these residues. This reduced the number of pairs in our dataset; however, it helped in minimizing erroneous mutants in the dataset.

2.1.4 Confirming of mutant structure with the author

A mutation observed in the above case, can be due to any of the various types of conflict such as alignment errors, cloning artifact, insertions, deletions etc. In this study, we considered engineered mutants.

1. The above mutants were only considered for further analysis, only if they are mutants as confirmed by the author.
2. The above information was processed from mmCIF files using Biopython.
3. Further, the mmCIF information was also checked for the wild type proteins
4. Only the mutants were considered for further analysis.

2.1.5 Construction of WT-MT pairs

Subsequent to classifying structures as WT or MT, we constructed pairs of WT and MT considering at least one residue change between WT and MT. This cut-off is considered to reduce effect on structures due to additional domain or structures. Since there can be many

missing residues in a protein structure, having additional regions can affect the network properties. Therefore, we considered 90% overlapping region between WT and MT structures and overhanging residues should not be more than 50 residues. Thus, we constructed a dataset that resembles the conditions similar to single-point mutants having 19,758 unique WT-MT pairs grouped under 489 representative UniProt IDs.

2.2 RIN generation

In this study, we constructed residue interaction network of WT/MT proteins to investigate effect of mutation on network features. Even though structural changes due to mutation can be studied using other approaches, protein network provides computationally less expensive and a system-level effect of changes. Despite its simplicity, it can in turn give much better insights given the development in the field for several decades. The abstraction of protein structure as network has immensely helped in modeling allostery in proteins. As mentioned before, a residue interaction network (RIN), is a graph of three-dimensional representation of a protein structure as a set of nodes (residues) with their connections (edges).

2.2.1 Nodes and Edges

1. Since the same PDB can be present in multiple WT-MT pairs formed above, to avoid repetition, all the unique PDB IDs were collected.
2. The above PDBs were run through RING software (Ref: Martin, A.J.M., Vidotto, M., Boscariol, F., Di Domenico, T., Walsh, I., Tosatto, S.C.E. RING: networking interacting residues, evolutionary information and energetics in protein structures. 2011), which was obtained from authors, to generate RIN
3. Commands to generate Nodes and Edges respectively:
   <ringpath> -i <pdb path> -N <output_path> -n ca --all_edges --get_iac --no_energy
   <ringpath> -i <pdb path> -E <output_path> -n ca --all_edges --get_iac --no_energy
4. C-alpha atoms of each amino acid residue is chosen as the node.
5. Long-range contacts were formed with a distance cutoff of 7Å.
6. Thus, a Protein Contact Network (PCN) or a Residue Interaction Network (RIN) is generated for each of the above proteins.

2.2.2 Shells for perturbation studies

To understand the network perturbation or percolation of changes in edges through the three-dimensional structure in their spatial orientation of residues because of mutation, we divided structures into various non-overlapping spherical shells. The objective is to study how changes in one shell can affect changes in distantly located shells from the origin of perturbation. Thus, it can provide how changes due to mutation affects spatially distantly located residues within a protein. For this purpose, spherical shells of 5Å were constructed with the source of disturbance (the mutant residue) as the center and their long-range effects were studied. A disturbance in a shell is defined to occur, if any residue in a given shell showed an absolute change in their centrality (betweenness) value between WT and MT more than 60% with respect to WT betweenness value. It can be either increase/decrease in betweenness parameter in MT with respect to WT.

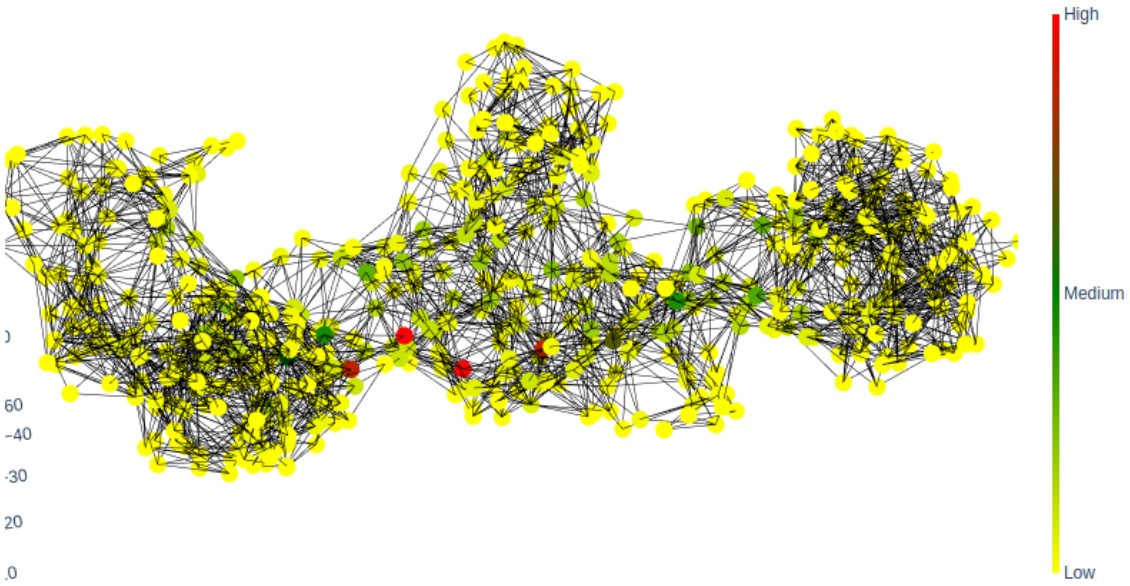
2.2.3 NetworkX for analysis

We used NetworkX software (Ref: Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman, 2008) to calculate various global/local (residue wise) network properties of RIN. Further, these above properties were visualized in matplotlib, wherever required.

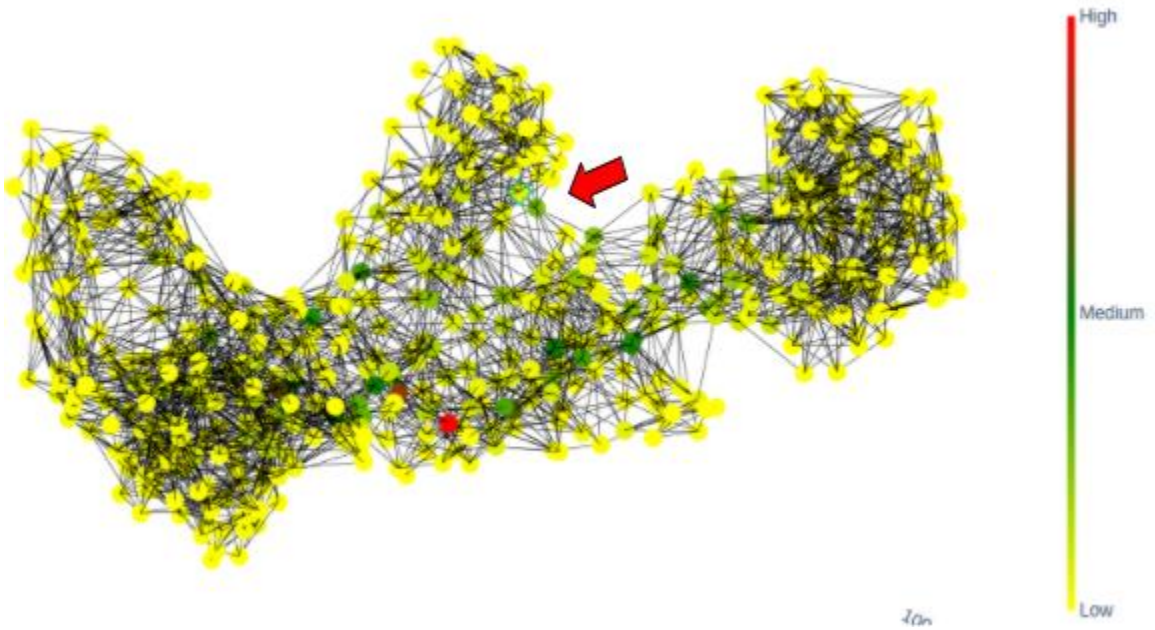
2.2.4 Plotly for Visualization

Visualizing the network parameters could provide a better understanding and intuition of different properties. Therefore, a visualization tool specifically for this project was built in which the nodes are colored based on their centrality values and the mutant residue is also highlighted.

E.g., Visualizing the effect of mutant residue on the betweenness centrality values of the residues in HIV Type 1 Reverse Transcriptase (3hvtA_wt, 1hniA_mt)



(a) 3hvtA - WT protein Betweenness Centrality Values



(b) 1hniA - MT protein Betweenness Centrality Values

Fig 2.1 Plotly for visualizing the effect of mutation on the overall values of the protein network

2.3 Metrics of network comparison

Throughout the study, network comparisons are primarily carried to understand the effect of the mutations at the level of residues and the level of protein. For this, we studied local and global network properties of proteins. Since, we are trying to understand the effect of the mutation and how these residues are communicating with one another in protein structure, we computed following network properties.

2.3.1 Centrality measures

Table 2.1 Centrality measures and their importance

Network Property	**Feature**
Closeness Centrality	How central/closer a node is, to all other nodes
Betweenness Centrality	The probability that communication from source(s) to target(t) will go through that node
Degree Centrality	No. of edges connected to a particular node
Eigenvector Centrality	Importance of a node in the network w.r.t its neighbors, based on the connectivity pattern

The above properties are calculated at the residue level. Calculating the above properties helps us in understanding the changes at the residue level. To study the changes at the global level, the average of a given network property can be considered.

2.3.2 The Delta Parameter

We modeled each protein structure as RIN and various network properties (listed in section 2.3.1) can be calculated for each protein network. In order to facilitate quantitative comparison of global network features, we defined an absolute difference between WT and MT feature, which we referred to as ‘Delta parameter’. This essentially captures changes in average global

features, therefore, defines changes between WT and MT networks. Assuming that this would follow a normal distribution, we identified set of WT-MT pairs lying 1σ away from mean showing either high or low delta values. The high and low delta value shows increase and decrease in average network features in MT protein network. We used this delta parameter to characterize protein structures, which show large variation in these values and explored whether these can correspond to special protein functional class.

Chapter 3

Results and Discussion

3.1 Dataset

3.1.1 Sequence identity distribution of WT-MT pairs

As described in methods sections, we constructed a dataset of 19,758 WT-MT pairs from non-redundant at 90% sequence identity of 489 uniprot sequences. First, we computed sequence overlapping between WT-MT by performing pairwise sequence alignment. In general, the WT-MT pairs overlap >90% suggesting that pairs have high coverage with each other (Figure 3.1). Such high coverage dataset can be used for structural comparison or network properties because additional residues in either WT/MT protein structure would minimally affect structural/network features. We do not observe any pair showing 100% overlap, which shows that protein structures have missing residues and justifies our reasoning to relax overlap criteria that allows increase in WT-MT pairs for the present study.

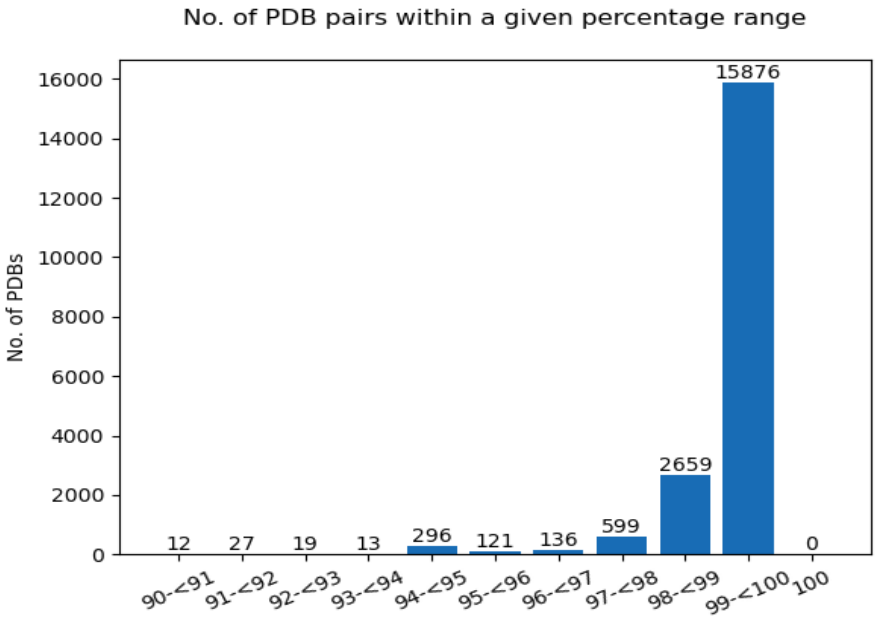


Fig 3.1 Distribution of WT-MT pairs for different percentages of sequence overlap

3.1.2 Distribution of delta Parameter

In order to characterize changes in feature of mutant RIN, we first find difference in global network properties using a delta parameter. This is computed as the difference in average network centrality measures between WT and MT protein structure networks (see methods). This will provide whether a residue mutation leads to large or small change in global network properties. The distribution of delta parameter for betweenness, closeness, eigen and degree centrality are shown in Figure 3.2. The mean and standard deviation of these are tabulated in Table 3.1. It is evident from the figure, more than 60% of WT-MT pairs show no or negligible change in these centrality measures. Moreover, average is 0.0 and standard deviation is not large. Therefore, mutation of a residue, in general, does not lead to a large variation in the global network features.

Next, we analyzed pairs of WT-MT showing unusual high/low delta values by considering pairs, which lie at least 1 σ away from the mean. The number of pairs could be anywhere between 1500-4500, depending on the type of centrality measure that is being considered. We analyzed such pairs having unusual high/low delta values to find whether these perform any specific molecular function or these values are associated with large conformational changes, using Root Mean Square Deviation (RMSD) as its measure.

Table 3.1 Mean and standard deviation of different centrality measures

Centrality Measure	**Mean ± 1 sigma**
Betweenness Centrality	0.0 ± 0.00046
Closeness Centrality	0.0 ± 0.00674
Eigen Centrality	0.0 ± 0.00152
Degree Centrality	0.0 ± 0.34553

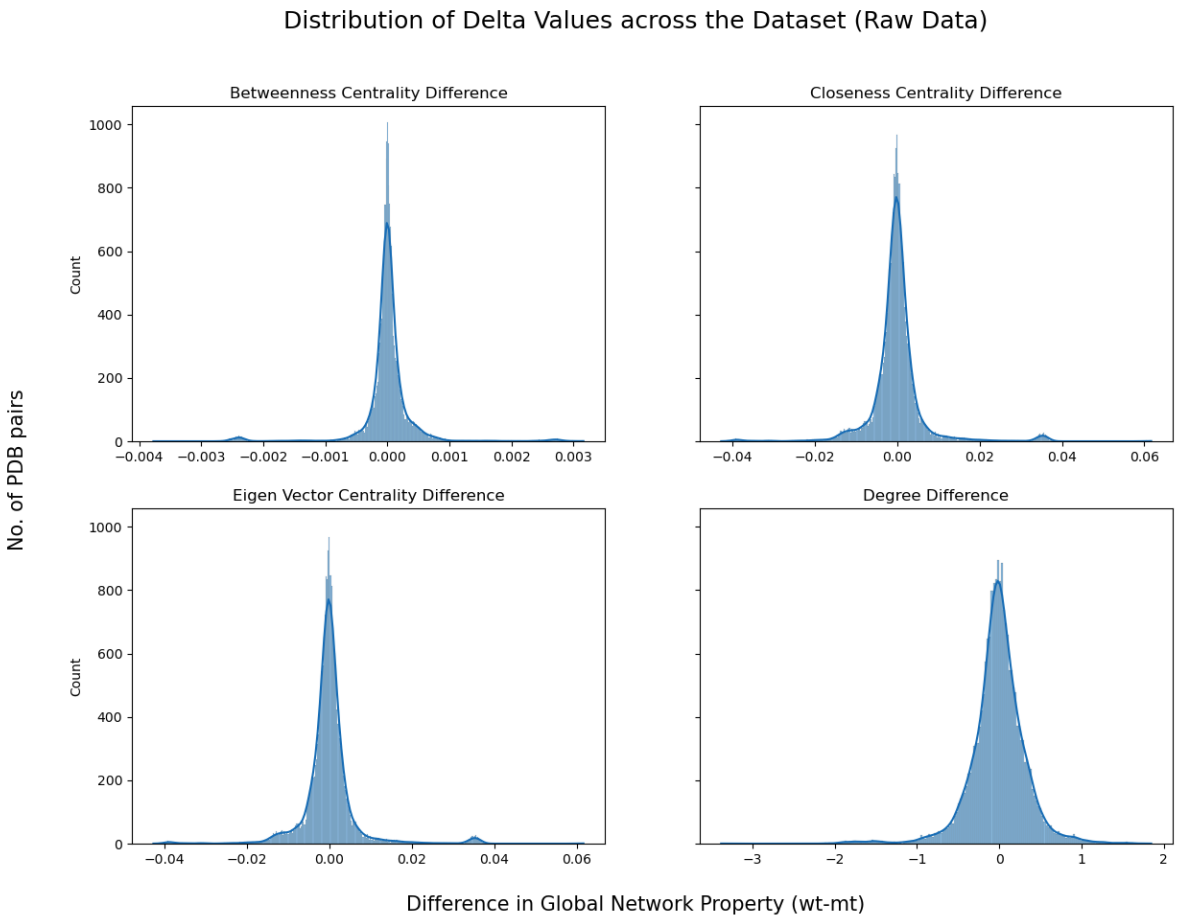


Fig 3.2 Distribution of the delta parameter for different centrality measures

3.2 Correlation of network features with change in conformation

The change in network properties could arise from conformational change in the protein structure, which is measured using RMSD between WT and MT protein structure. We classified mutants as 'engineered' mutants (TRUE), which are author confirmed mutants and others as potential mutants (FALSE), yet not confirmed from literature. We could not find any correlation between any network centrality measure and RMSD and the same is evident from the figure 3.3. This is observed for both types of mutants whether author confirmed or not. It is worthwhile to note in Figure 3.3 that no change in global parameters change can be observed even for WT-MT pairs, which have large conformational change as measured by RMSD. Therefore, conformational changes do not significantly affect average network feature

suggesting the interaction dynamics is preserved despite movement in region of protein structures.

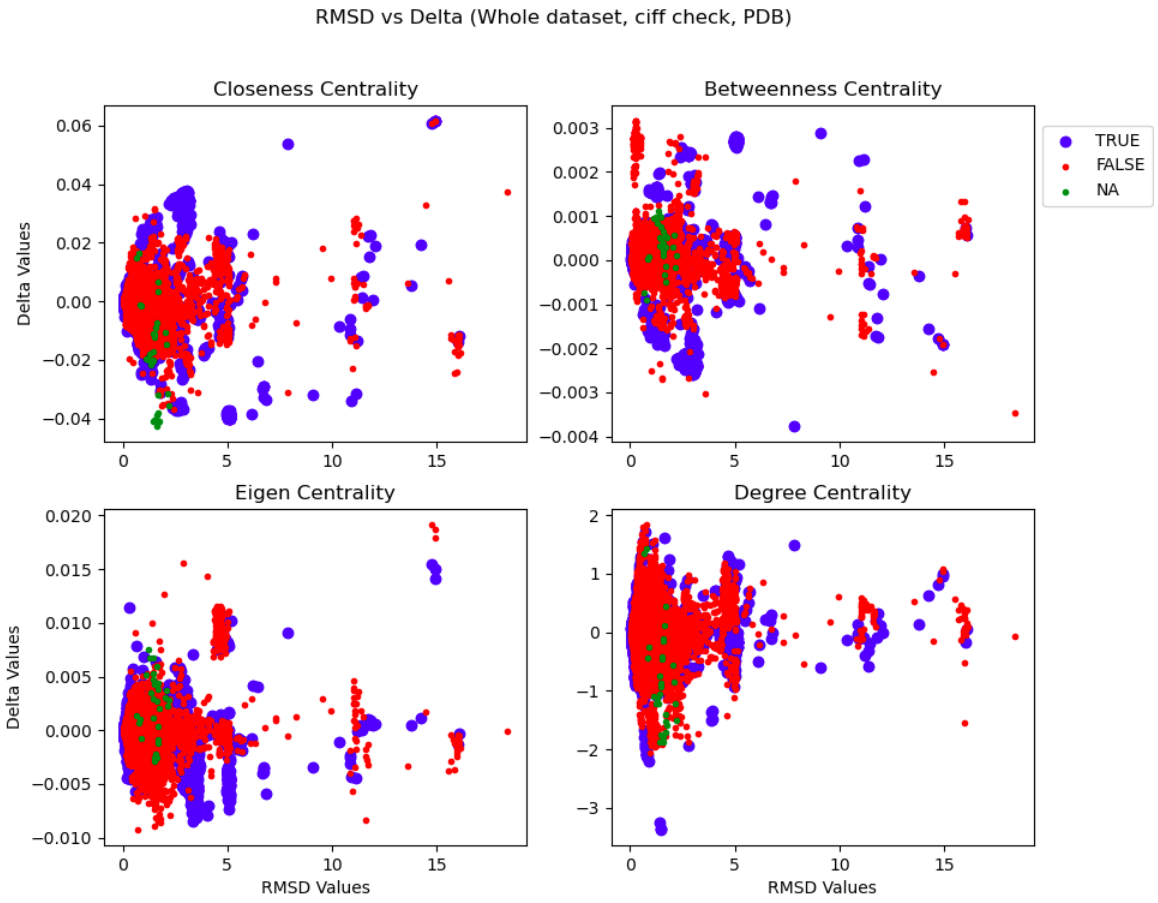


Fig 3.3 Correlation of RMSD and delta values for the entire dataset. TRUE are engineered mutants and FALSE are other cases

It is quite likely that correlation may be missed in such scatter plot, we took TRUE cases of WT-MT pairs and binned them with a bin size 1Å RMSD to plotted box plot to observe any correlation between delta parameter and RMSD. As can be seen in figure 3.4, there is no obvious association of RMSD and difference in centrality measures. Notably, a large RMSD does not lead to change in global network feature and also low structural change cannot ensure that mutant proteins does not undergo significant change in global network properties. Therefore, these analyses suggests that usually mutation does not affect the global network properties, however, a change in these features should be interpreted with complementary

analyses. Importantly, these changes are completely unaffected by conformational change in proteins because of single residue mutation.

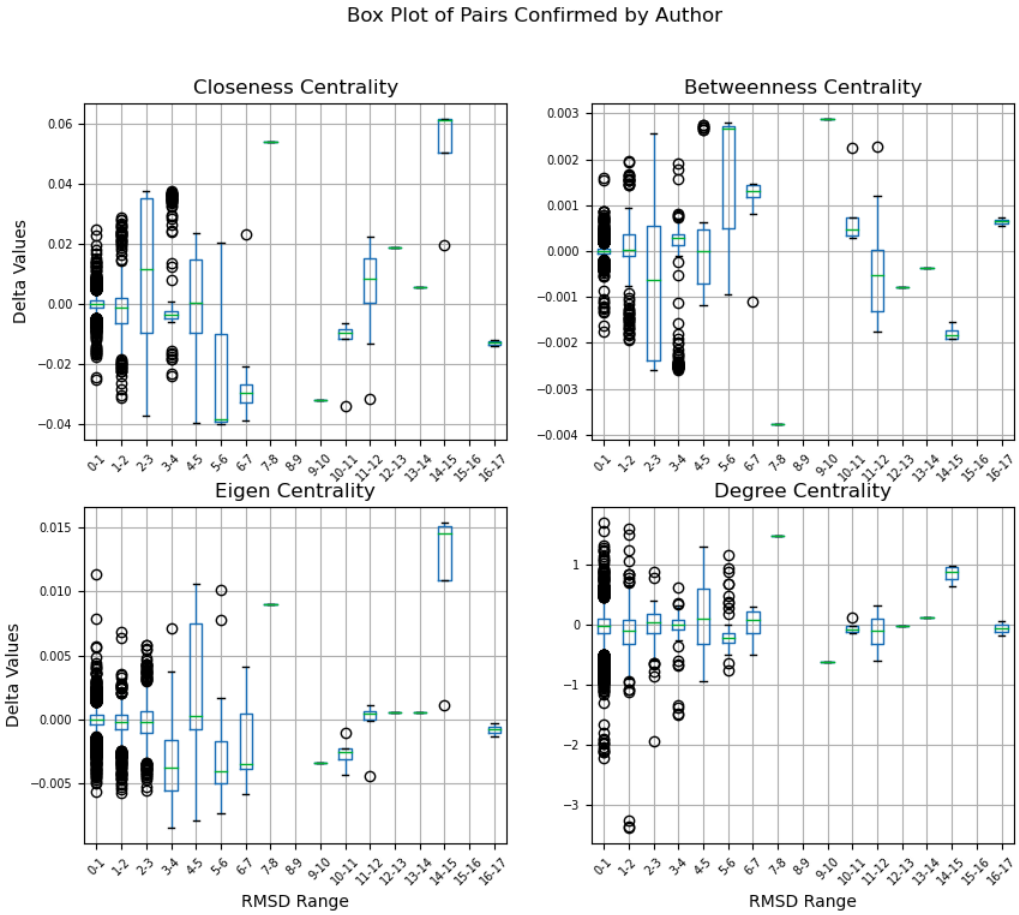


Fig 3.4 Correlation between pairwise RMSD and the Delta parameter. The plot is binned box plot of RMSD binned with 1 Å

3.3 Correlation with the function

3.3.1 Mutation data from different sources

As shown before, global network properties does not undergo any significant change due to single residue mutation. There were some pairs of WT-MT proteins, which showed variation

more than 1 σ away from mean of delta value. We investigated whether these proteins are involved in specific function using Gene Ontology (GO) based function enrichment. The objective is to identify distinguishable functions associated with WT-MT pairs having delta at least more than 1σ of delta distribution for any centrality measure. As described before, we confirmed mutation in protein structures as given by authors in PDB file (mmCIF) as well as Uniprot database. The distribution of author confirmed (TRUE) mutant (MT) proteins pairs either from PDB or Uniprot is shown in Figure 3.5 As can be seen, Uniprot has poor documentation on mutation data. Therefore, for rest other analyses we relied on TRUE cases derived from PDB.

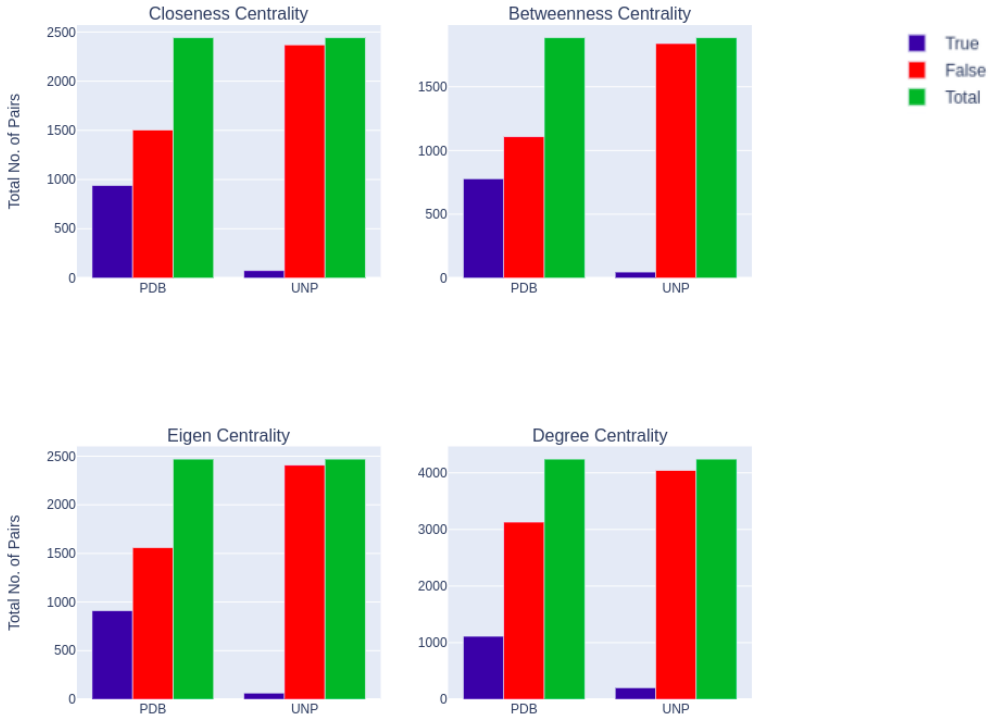
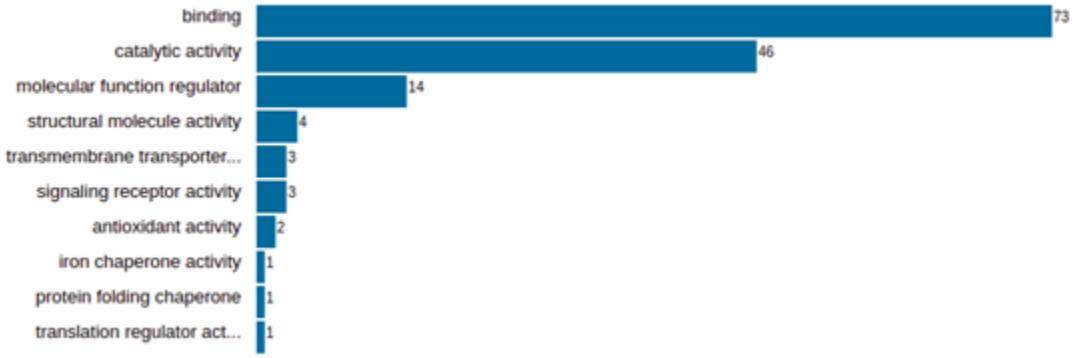


Fig 3.5 Distribution of WT-MT pairs having delta at least 1σ or more and also classifies these as author confirmed (TRUE) and others potential mutant structures (FALSE) based on information from PDB/Uniprot

3.3.2 GO functional annotation

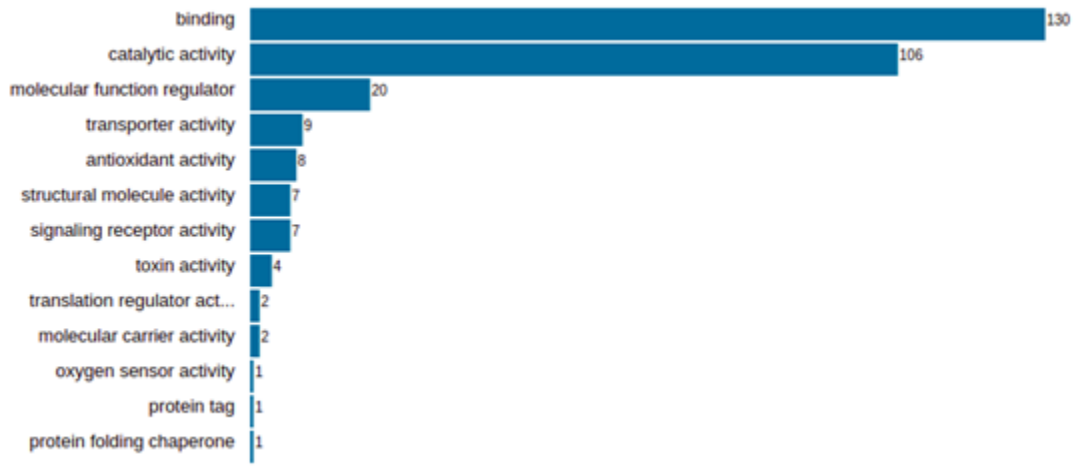
Gene Ontology (GO) provides semantic based ontologies describing functions of proteins that are consistent across genes. GO can be described at the level of biological processes, cellular components, and molecular function. Since GO can facilitate structured function comparison

among proteins, we examined whether WT-MT pairs showing large absolute delta values (at least 1σ) are associated with any specific GO entry. This can assist whether specific functional class of proteins is more prone to large variation in network properties, for instance do allosteric proteins. In this study, we considered only author confirmed WT-MT pairs having large change in any centrality measure and for these find enriched molecular function GO terms associated with them. First, we performed this analysis for each centrality measure and finally, combined all WT-MT entries from various measures. As can be seen in Figure 3.6, binding and catalytic function are found to be enriched in all cases. These are also commonly observed in WT-MT pairs having no change in delta value.

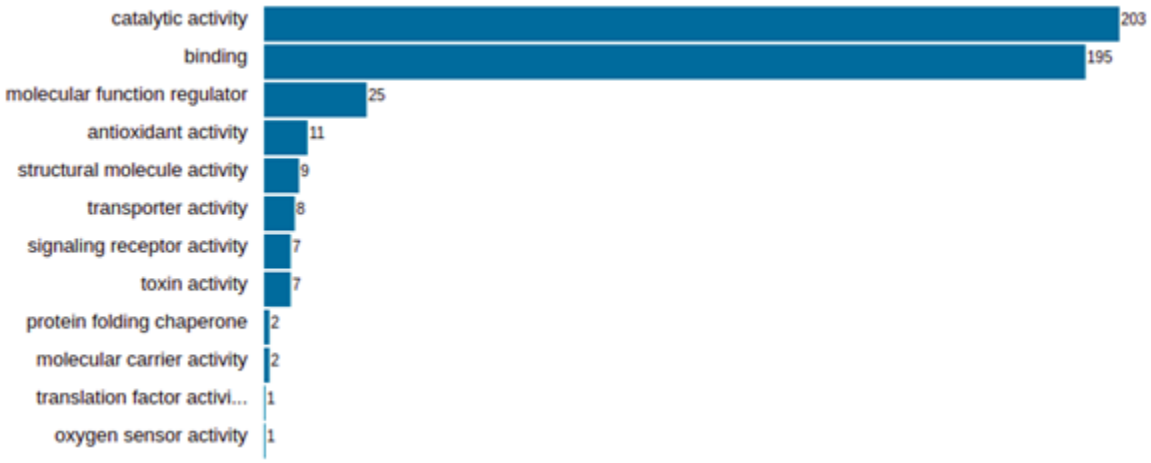


(a) Betweenness Centrality

(b) Closeness Centrality



(c) Eigen Centrality



(d) Degree Centrality

Fig 3.6 GO functional enrichment of WT-MT pairs having delta values at least 1σ from mean of distribution.

Figure 3.7, shows results for all WT-MT pairs having large delta in any centrality measure and as can be seen the results as observed before. Thus, suggesting that there no specific function could be associated with proteins showing large change in delta parameter.

Fig 3.7 GO functional classification of common WT-MT pairs having delta at least more than 1σ from mean of distribution

The results showed that the global network properties do not seem to show any correlation with the structural and functional properties of the proteins. To analyze, the effect of mutation on other residues within the protein, local perturbation studies were done.

3.4 Local network effect of mutation

3.4.1 Determination of Δbetweenness cut-off

Since mutation can perturb the local residue packing and interactions, we analyzed whether residue level network features can capture perturbations caused due to mutation. Further, we investigated whether it can propagate through the RIN. The latter is interesting property as such perturbations are commonly used to define allostery and identify allosteric communication paths. The comparison of various node property (degree, betweenness, closeness) showed that these change in MT protein in comparison to WT. Among these, betweenness centrality is important feature as it provides significance of a residue in shortest paths in a given network. In order to study propagation of perturbation through RIN, we computed residue betweenness for WT and MT and calculated absolute percentage difference between them as ΔBC with respect to WT as given in equation below:

$$\Delta BC = \frac{|BC_{WT} - BC_{MT}|}{BC_{WT}}$$

Since RIN is a 2-dimension representation of protein, the propagation in such a network does not provide information communication between spatially distant residues. For this, we empirically draw concentric spheres separated by 5Å and having center at the site of mutation.

In such representation of concentric spheres, changes in last sphere (shell) would represent changes in spatially distant residues.

Now, we have definition to find change in residue betweenness (ΔBC) and a way to find spatial location of such change. With this, first we determined a suitable cut-off to define a significant change in residue betweenness. First, we find all WT-MT pairs of proteins having at least one residue undergoing change in this centrality measure until the last shell. Then, we computed (ΔBC) for all residues for these WT-MT pairs. Figure 3.8 shows the fraction of WT-MT pairs showing change until last shell plotted with various cut-off values ΔBC expressed in percentage. As expected, if we consider small change in betweenness (ΔBC <10%), we can observe a large number of WT-MT pairs (~80%) show change until the last shell. From this curve, we decided to select a stringent cut-off of 60% to defined significant change in residue local property of betweenness. Even though it may miss some WT-MT pairs, it will probably identify mutations, which are able to make long range effects on proteins.

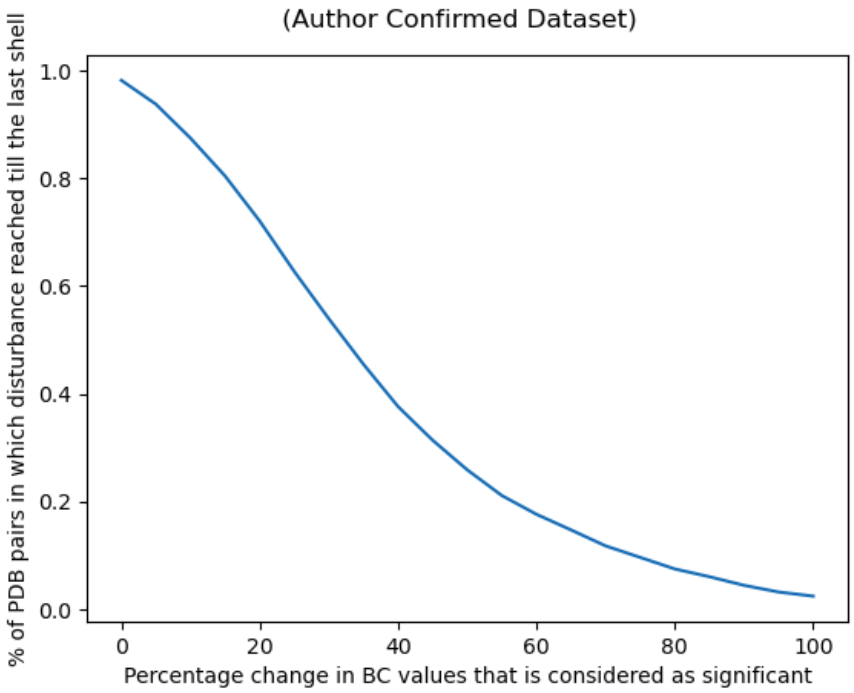


Fig 3.8 Percentage of PDB pairs vs Percentage disturbance

Apart from this, we also looked at the ΔBC changes propagated to various shells (Figure 3.9). As can be seen at low values (0 or 10%), changes can be seen until the last shell and with increase in ΔBC cut-off, the maximal effect can be seen reduced from last shell till lower fraction of shells and at suggested cut-off we can find ~1200 pairs which show change until last shell.

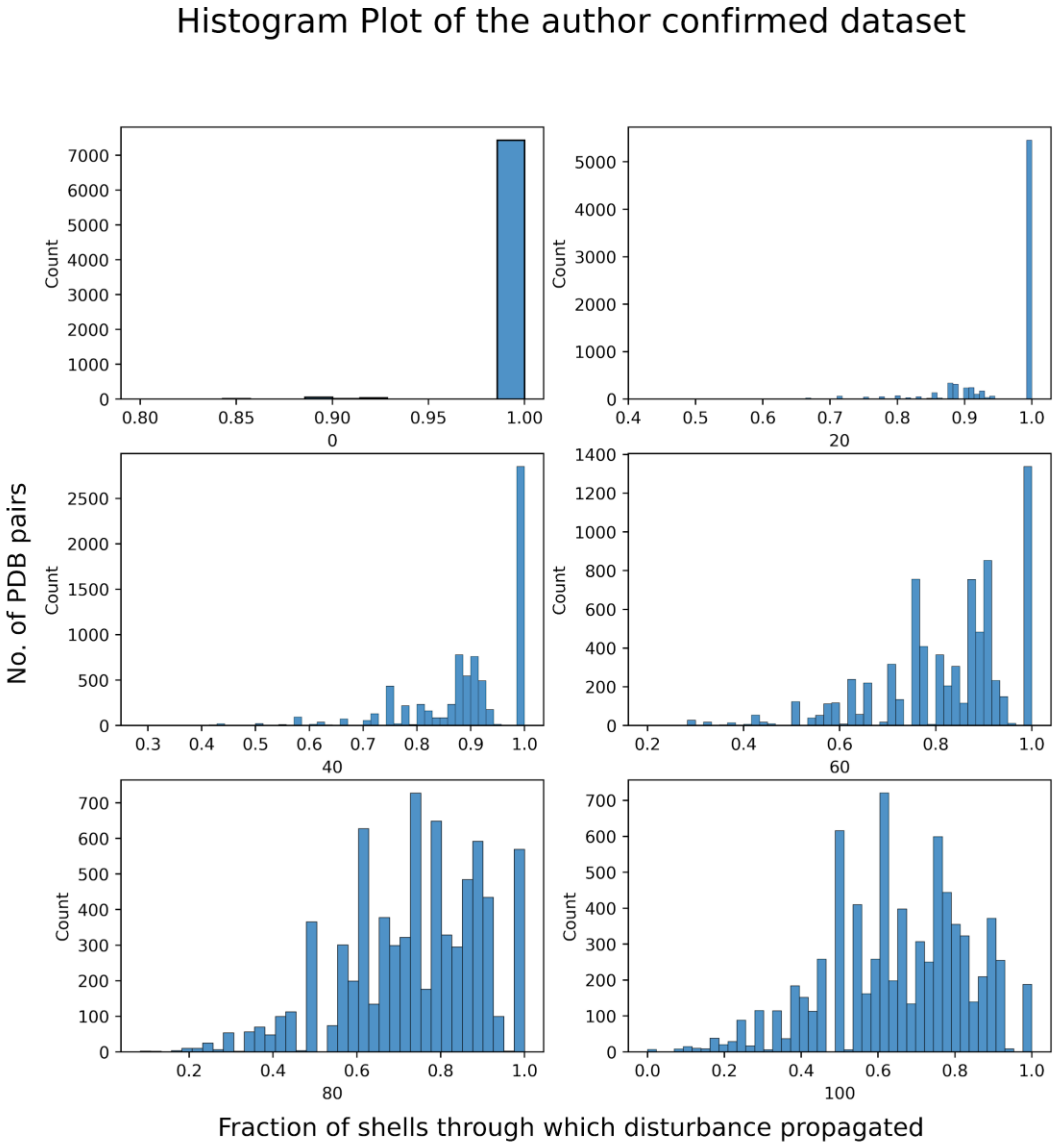


Fig 3.9 Distribution of WT-MT pairs for different percentage cutoffs

When the cutoff is set at 60%, there are several pairs in which the disturbance caused by the mutant residue within a protein network has propagated until different fractions of shells. This

suggests that mutation is not affecting far of residues. For further analysis, we considered WT-MT pairs which have changes at >90% shells (i.e., those in which the disturbance is observed in more than 90% of the shells). In addition, we also removed proteins not undergoing large conformational change (<2Å).

Using the above-mentioned criteria, we found enrichment of uniprot IDs to find which WT-MT pairs contributes the maximum to changes in betweenness (Figure 3.10). Most of the highly represented UniProt IDs belonged to the enzyme category. As can be seen, dipeptidyl peptidase 4 is among the top ranked protein contributing to large change in residue level local network properties.

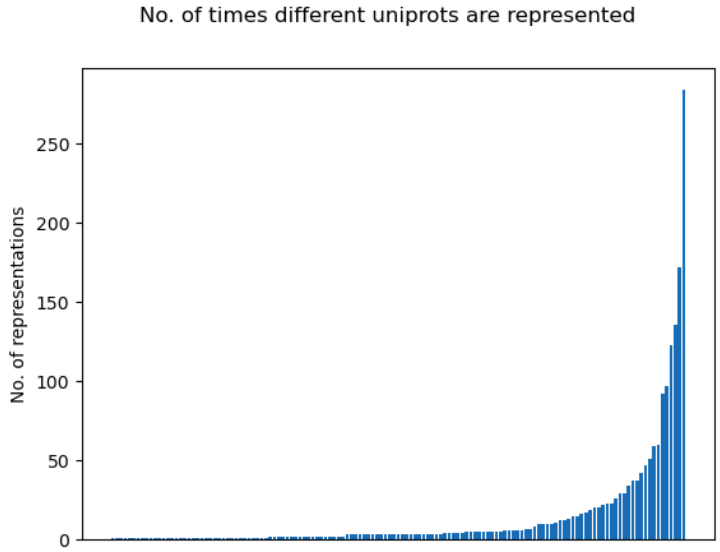


Fig 3.10 Representation of different UniProt IDs for the above dataset (author confirmed - high Betweenness value and low RMSD)

UnpID	#Representations	Protein Name
P27487	284	Dipeptidyl peptidase 4
P00742	172	Coagulation factor X
P0C0Y9	136	Reaction center protein M chain
Q54727	123	Sialidase B
Q97W02	97	DNA polymerase IV
P19120	92	Heat shock cognate 71 kDa protein
Q05769	60	Prostaglandin G/H synthase 2
P0C0Y8	59	Reaction center protein L chain
P00509	51	Aspartate aminotransferase

Highly represented UniProt IDs (more than 50 representations)

3.4.2 Local perturbation in DPP4

Dipeptidyl-peptidase 4 (DPP4) is a glycoprotein of 110 kDa, which is ubiquitously expressed on the surface of a variety of cells. This exopeptidase selectively cleaves N-terminal dipeptides from a variety of substrates, including cytokines, growth factors, neuropeptides, and the incretin hormones (Ref: Röhrborn D, Wronkowitz N, Eckel J. DPP4 in Diabetes. Front Immunol. 2015; 6:386. Published 2015 Jul 27. doi:10.3389/fimmu.2015.00386). Interestingly, this protein is important drug target in treating diabetes. The catalytic important residues are 630, 708 and 740. Given that single residue mutation in many residues can cause a long-range propagation suggests that this protein must show allosteric behavior. In our study, we find the mutation in residue id 39 (SER), leads to change in residue betweenness located in active site, which is ~43 Å from the site of mutant residue, which the origin of propagation in the network. In terms of shell distance, these two sites are about 8 shells apart, indicating the long-range effect these mutants can have within the protein network. To get a better understanding one such instance was further visualized with the above built Plotly visualization tool.

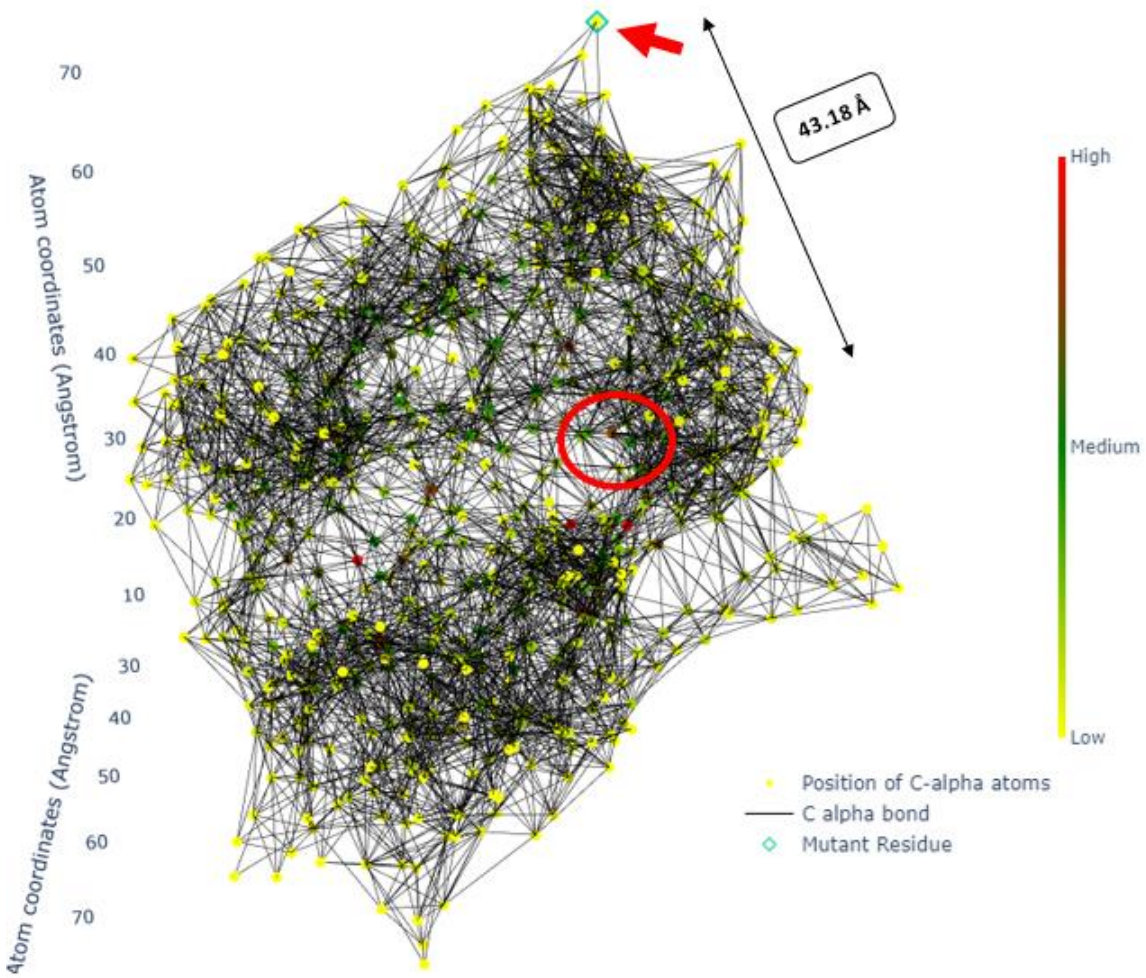


Fig 3.11 Long-range effect of mutation on active site residues in DPP4 (3eioB_wt, 2qt9A_mt)

Conclusions

The present serve as preliminary study to investigate effect on RIN due to mutation. We studied changes in global (average centrality) as well as local (residue level) on RIN constructed using C-alpha distance criteria. For this analysis, we constructed a dataset of ~20,000 WT-MT multidomain proteins. In general, we did not observe large changes in the global network properties. Upon examining proteins having relatively large changes in these features could not be associated with any specific functional feature of proteins. Though it might sound surprising and counter-intuitive the above C-alpha based network analysis suggests that a change in network properties as quantified by the Delta parameter does not correlate with the pairwise RMSD.

Subsequently, we analyzed local (residue/node level) change in betweenness and their propagation to spatially distant residues. This showed that for many proteins we could observe change in betweenness >60% to residues lying spatially far away from the site of mutation. For example, in DPP4, we could identify residue undergoing change in betweenness ~43Å away from the site of mutation. Thus, suggesting that analyses of local network feature can reveal interesting characteristics of protein interaction network. These can be correlated with phenotypic effect of mutation to shed light on role of network communication in proteins.

Future work

We need to analyze residue interaction networks constructed using other features such as based on energy before we finalize the proper RIN to employ for understanding network perturbation propagation in proteins. The local perturbation studies also suggest a possible scope for allostery, which is worthwhile to pursue for defining a general communication architecture in proteins.

Bibliography

- W.T. Godbey, "An Introduction to Biotechnology. The Science, Technology and Medical Applications." Elsevier Inc, 2015. Chapter 2- Proteins: 9-33, DOI: https://doi.org/10.1016/C2013-0-18161-5, ISBN: 978-1-907568-28-2
- Joseph Feher. "Quantitative Human Physiology. An Introduction." Elsevier Inc, 2017. 2.3- Protein Structure: 130-141, ISBN: 978-0-12-800883-6
- Vella, F. (1992), Introduction to protein structure. By C Branden and J Tooze. pp 302. garland publishing, New York. 1991 ISBN 0–8513–0270–3 (pbk). Biochem. Educ., 20: 121-122. https://doi.org/10.1016/0307-4412(92)90129-A
- Larry R.Engelking," Textbook of Veterinary Physiological Chemistry." Elsevier Inc, 2015. Chapter 4- Protein Structure: 18-25, DOI: https://doi.org/10.1016/C2010-0-66047-0, ISBN: 978-0-12-391909-0
- Almahboub, S.A., Narancic, T., Fayne, D. et al. Single point mutations reveal amino acid residues important for Chromobacterium violaceum transaminase activity in the production of unnatural amino acids. Sci Rep 8, 17397 (2018). https://doi.org/10.1038/s41598-018-35688-7
- Liu, H., Gao, YS., Chen, XJ. et al. A single residue substitution accounts for the significant difference in thermostability between two isoforms of human cytosolic creatine kinase. Sci Rep 6, 21191 (2016). https://doi.org/10.1038/srep21191
- Robert F, Pelletier J. Exploring the Impact of Single-Nucleotide Polymorphisms on Translation. Front Genet. 2018;9:507. Published 2018 Oct 30. doi:10.3389/fgene.2018.00507
- National Institutes of Health. National Human Genome Research Institute
- Piovesan, D., Minervini, G., Tosatto, S.C.E. The RING-2.0 web server for high-quality residue interaction networks. 2016. Nucleic Acids Research 44(W1), W367-74. doi:10.1093/nar/gkw315
- Martin, A.J.M., Vidotto, M., Boscariol, F., Di Domenico, T., Walsh, I., Tosatto, S.C.E. RING: networking interacting residues, evolutionary information and energetics in protein structures. 2011. Bioinformatics 27, 2003–2005. doi:10.1093/bioinformatics/btr191

- Stetz G, Verkhivker GM. Computational Analysis of Residue Interaction Networks and Coevolutionary Relationships in the Hsp70 Chaperones: A Community-Hopping Model of Allosteric Regulation and Communication. PLoS Comput Biol. 2017 Jan 17;13(1):e1005299. doi: 10.1371/journal.pcbi.1005299. PMID: 28095400; PMCID: PMC5240922.
- Contreras-Riquelme S, Garate JA, Perez-Acle T, Martin AJM. RIP-MD: a tool to study residue interaction networks in protein molecular dynamics. *PeerJ*. 2018;6:e5998. Published 2018 Dec 7. doi:10.7717/peerj.5998
- The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021, Nucleic Acids Research, Volume 49, Issue D1, 8 January 2021, Pages D480–D489, https://doi.org/10.1093/nar/gkaa1100
- The Protein Data Bank H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne (2000) Nucleic Acids Research, 28: 235-242. doi:10.1093/nar/28.1.235
- Weizhong Li, Adam Godzik, Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences, Bioinformatics, Volume 22, Issue 13, 1 July 2006, Pages 1658–1659, https://doi.org/10.1093/bioinformatics/btl158
- Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. Nat Struct Mol Biol 10, 980 (2003). https://doi.org/10.1038/nsb1203-980
- Cock PA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B and de Hoon MJL (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics, 25, 1422-1423
- Aric A. Hagberg, Daniel A. Schult and Pieter J. Swart, "Exploring network structure, dynamics, and function using NetworkX", in Proceedings of the 7th Python in Science Conference (SciPy2008), Gäel Varoquaux, Travis Vaught, and Jarrod Millman (Eds), (Pasadena, CA USA), pp. 11–15, Aug 2008
- Jennifer Golbeck, "Analyzing the Social Web." Elsevier Inc, 2013. Chapter 3- Network Structure and Measures: 25-44. DOI: https://doi.org/10.1016/C2012-0-00171-8; ISBN: 978-0-12-405531-5

- Jennifer Golbeck. “Analyzing the Social Web.” Elsevier Inc, 2013. Chapter 2- Nodes, Edges, and Network Measures: 9-23. DOI: https://doi.org/10.1016/C2012-0-00171-8; ISBN: 978-0-12-405531-5
- Author: Plotly Technologies Inc. Title: Collaborative data science Publisher: Plotly Technologies Inc. Place of publication: Montréal, QC Date of publication: 2015 URL: https://plot.ly
- J. D. Hunter, "Matplotlib: A 2D Graphics Environment", Computing in Science & Engineering, vol. 9, no. 3, pp. 90-95, 2007
- Srivastava A, Birari V, Sinha S. Small Conformational Changes Underlie Evolution of Resistance to NNRTI in HIV Reverse Transcriptase. Biophys J. 2020 May 19;118(10):2489-2501. DOI: 10.1016/j.bpj.2020.04.008. Epub 2020 Apr 16. PMID: 32348721; PMCID: PMC7231925.
- Ashburner et al. Gene ontology: a tool for the unification of biology. Nat Genet. May 2000;25(1):25-9.
- The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res. Jan 2021;49(D1):D325-D334
- Röhrborn D, Wronkowitz N, Eckel J. DPP4 in Diabetes. *Front Immunol*. 2015;6:386. Published 2015 Jul 27. doi:10.3389/fimmu.2015.00386