PROTEIN DYNAMICS IN DELETERIOUS AND COMPENSATORY MUTATIONS

by

Yiğit Kutlu

B.S., Chemical Engineering, Boğaziçi University, 2016

Submitted to the Institute for Graduate Studies in
Science and Engineering in partial fulfillment of
the requirements for the degree of
Master of Science

Graduate Program in Chemical Engineering

Boğaziçi University

2020

ACKNOWLEDGEMENTS

First and foremost, I would like to express my sincere gratitude to my thesis supervisor Prof. Türkan Haliloğlu for her guidance, support and patiance. It was a great experience to work with her and I have learned a lot from her knowledge, expertise and wisdom. I would never achieve such success in my studies without her motivation, continuous support and understanding.

I wish to express my sincere appreciations for the members of the thesis committee, Assoc. Prof. Burak Alakent and Asst. Prof. Özge Kürkçüoğlu Levitas, for accepting to be a member of the thesis committee, devoting their valuable time to read and comment on my thesis.

I would like to thank Prof. Nir Ben-Tal and Rachel Kolodny who provided me the themes dataset and contributed to my study generously.

I would like to give my deepest thanks to Özge Duman, Zeynep Erge Akbaş Buz and Burçin Acar for their continuous support, patience and friendship, especially to Özge Duman for being there whenever I need. Also, I would like to thank other members of the PRC family; Gökçehan Kara and Büşra Özgüney for their friendship.

I also thank my friends Mert Doğruer, Çağatay Göler and Utku Özcan for their frienships and emotional support during my thesis period.

Finally, I wish to thank my family for their endless love and support.

It was a great opportunity for me that my research has been supported by Boğaziçi University B.A.P. (13460) and Betil Fund.

ABSTRACT

PROTEIN DYNAMICS IN DELETERIOUS AND COMPENSATORY MUTATIONS

Mutations are associated with many diseases (cancer, Alzheimer's disease etc.). These disease-causing mutations are called deleterious mutations and their effects can be corrected by other mutations, i.e. compensatory mutations. Understanding the underlying dynamics of deleterious and compensatory mutations is of high importance for the treatment of diseases and drug design. To that end, the relationship between evolutionary conserved/reused segments and dynamic domains as well as the dynamic determinants of deleterious and compensatory mutations are investigated in this work. Themes; reused segments (~35-200 amino acids) among proteins with high sequence similarity, are correlated with the dynamic domains unveiled with Gaussian Network Model (GNM) analysis. The correlation between themes and dynamic domains evaluated with Adjusted Mutual Information and Standardized Mutual Information measures is found to be statistically significant. GNM based perturbation analysis revealed that the highest response to perturbations occur at the terminal points of themes. In order to investigate the dynamic properties of deleterious mutations, a dataset consisting of proteins that have been the subject to deep sequencing studies is created. In order to mimic the effect of mutation, a perturbation is placed in the GNM algorithm. The effect of this perturbation on the dynamics of the structures is examined according to the changes in the total fluctuation profiles and eigenvalues. Residues with high mutation sensitivity are found to be the residues that cause distinct change in protein dynamics upon perturbation and this relationship is statistically significant. For compensatory mutation studies, deleterious mutations of tumor suppressor protein p53 and their compensatory mutations are used. The results revealed that deleterious and compensatory mutations are correlated in slow modes of motion and demonstrate the importance of coevolution, hinge points and allosteric interaction in slow modes of motion for these mutations. Furthermore, the dynamic information about deleterious and compensatory mutations reported in this study will be a guide for further studies on the prediction of deleterious and compensatory mutations.

ÖZET

ZARARLI VE TELAFİ EDİCİ MUTASYONLARDA PROTEİN DİNAMİĞİ

Mutasyonlar bir çok hastalık (kanser, alzheimer hastalığı vs.) ile ilintilidir. Hastalıklara sebep olan mutasyonlara zararlı mutasyonlar denir ve bu mutasyonların hastalık yapıcı etkileri telafi edici mutasyonlar ile düzeltilebilir. Zararlı ve telafi edici mutasyonların altında yatan dinamiği anlamak, mutasyonlar ile ilgili hastalıkların tedavisi ve ilaç dizaynı için yüksek öneme sahiptir. Bu amaçla, evrimsel korunan/yeniden kullanılan bölümler ile dinamik parçalar arasındaki ilişki ile zararlı ve telafi edici mutasyonların dinamik belirleyicileri incelenmiştir. Yüksek sekans benzerliğine sahip ve proteinler arasında yeniden kullanılan bölümler olan temaların, Gauss Ağ Modeli analizi ile ortaya çıkarılan dinamik parçalar ile ilişkili olduğu ortaya çıkmıştır. Temaların dinamik parçalar ile olan korelasyonu, Adjusted Mutual Information ve Standardized Mutual Informartion ölçüleriyle değerlendirilip bu korelasyonun istatistiksel olarak önemli olduğu sonucuna varılmıştır. GNM tabanlı pertürbasyon analizi ile pertürbasyona en çok tepki veren bölgelerin temalraın uç noktaları olduğu görülmüştür. Zararlı mutasyonların dinamik özelliklerinin incelenmesi amacıyla derin sekanslama çalışmalarına konu olmuş proteinlerden oluşan veri seti oluşturulmuştur. Mutasyon etkisini taklit etmesi için GNM algoritmasının içine bir pertürbasyon yerleştirilip, bu pertürbasyonun yapıların dinamiğine etkisi, toplam dalgalanma miktarlarındaki ve özdeğerlerindeki değişimlerine göre incelenmiştir. Mutasyon hassasiyeti yüksek olan rezidülerin, perturbasyon sonrası protein dinamiğinde en çok değişime sebep olan rezidüler ile ilişkili olduğu gözlemlenmiş ve bu ilişki istatistiksel olarak incelenmiştir. Telafi edici mutasyon çalışmaları için, tümör süpresör protein p53'ün zararlı mutasyonları ve bu mutasyonların telafi edici mutasyonları kullanılmıştır. Sonuçlar, zararlı mutasyonların ve telafi edici mutasyonların yavaş hareket modlarında bir korelasyonda olduğunu ve bu mutasyonlar için yavaş hareket modlarında birlikte evrimin, menteşe noktalarının ve allosterik etkileşimin önemini göstermektedir. Ayrıca, bu çalışmada bildirilen zararlı ve telafi edici mutasyonlar hakkında dinamik bilgi, zararlı ve telafi edici mutasyonların tahmini üzerine daha fazla çalışma için bir rehber olacaktır.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF SYMBOLS

Cα	Alpha Carbon
k	Mode
k_b	Boltzmann constant
n	Number of residues
R_i	Position vector of residue i
T	Absolute temperature
u_k	kth eigenvector
u^*	Perturbed eigenvector
U	Matrix of eigenvectors
U^*	Perturbed matrix of eigenvectors
X	Sample Mean
Z	Z-score
Å	Angstrom
γ	Force constant of the Hookean pairwise potential
λ_k	kth eigenvalue
λ^*	Perturbed eigenvalue
μ	Population mean
σ	Standard deviation
Γ	Kirchhoff (or connectivity) matrix
Γ^*	Perturbed Kirchhoff (or connectivity) matrix
ΔR_i	Fluctuation of residue i
Λ	Diagonal matrix of eigenvalues
Λ^*	Perturbed diagonal matrix of eigenvalues

LIST OF ACRONYMS/ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
AMI	Adjusted mutual information
cdf	Cummulative distribution function
DNA	Deoxyribonucleic acid
N/A	Not available
NGS	Next generation sequencing
mRNA	Messenger ribonucleic acid
PDB	Protein Data Bank
RNA	Ribonucleic acid
SMI	Standardized mutual information

1. INTRODUCTION

Replication of DNA is a highly accurate process that consists many control and repair mechanisms. Nevertheless, it is not perfect, and the DNA is susceptible to mutations in each replication. There are also some outside factors (chemicals, radiation, etc.) that causes mutations. Occurrence of mutations in the DNA has wide range of consequences due to changes in the amino acid sequence of proteins. They may have no effect at all or may result in gaining or losing of a function or can be catastrophic by preventing the proper folding (Yue *et al.,* 2005; Studer *et al.,* 2013; Gao *et al.,* 2015). Many diseases, like cancer, sickle cell anemia, Alzheimer's disease, are linked with mutations. These mutations which cause loss of function or diseases (loss of fitness) are called deleterious mutations (Omenn, 2010; Nesse *et al.,* 2010). The negative effects of deleterious mutations can be rescued by another mutation or mutations. These mutations, which restore the loss of fitness are called compensatory mutations (Kimura, 1985; Poon *et al.,* 2005).

Compensatory mutations are first introduced by Kimura. The term defined as ‘compensatory neutral mutations’ for two mutations at different sites, which has a deleterious effect individually but together have a neutral or beneficial effect on overall fitness. The latter work was about molecular evolution aspect of compensatory mutations, which suggested that a model on compensatory mutations will provide a better understanding the mechanism of evolution at molecular level (Kimura, 1985).

Understanding the underlying mechanisms of compensatory mutations gives insight about mutations, diseases and curing of these diseases. As a case in point, the tumor suppressor protein p53 has many mutations that cause loss of fitness in the protein and prevents it from binding to DNA. It is the main reason of many cancer related diseases. The tumor suppressor protein p53 also has many compensatory mutations that enables the protein to bind to DNA again. These rescue mutations are keystones to many cancer related drug design projects (Bullock and Fersht, 2001; Chen *et al.,* 2010). In addition, compensatory mutations are also observed in pathogens. Pathogens use compensatory mutations as a mechanism to gain drug resistance. There are many studies aiming to study the

compensatory mutations and thus preventing pathogens from gaining drug resistance in fast rates (Maisnier-Patin and Anderson, 2004; Anderson, 2012).

Mutations especially disease-causing mutations have been studied both experimentally and computationally (Merz *et al.*, 2010; Baresic *et al.*, 2011). Understanding disease mechanisms is an important subject for the drug design efforts. The dynamical behavior of the deleterious and compensatory mutations can particularly be important to explain the allostery in this phenomenon and also disclose the determinants of the global compensatory mutations that are not only specific to the deleterious mutations.

In the first part of the thesis, relationship between evolutionary conserved/reused segments and dynamic domains is investigated (Nepomnyachiy *et al.*, 2014; Nepomnyachiy *et al.*, 2017). In a recent study, a systematic survey is conducted in order to examine reuse in protein space. The study focuses on shared segments of 35 to 200 amino acids among proteins. Those segments called as "themes" are not identical, but they suggested to share high sequence similarity. Themes are suggested to differ from classical definition of domains. The main difference between themes and domains is that an amino acid can only belong to a single domain whereas it may be shared by any number of themes. This difference is suggested to expose reuse in full capacity. An amino acid can belong to a long theme which is shared by some closely related proteins and a shorter theme which is shared by many remotely related proteins. The study revealed many themes but did not provide any hints about the biological function of these evolutionary footprints. In the thesis, we examine possible links between the themes, which have been detected based on sequence similarity alone, and protein structural dynamics, as revealed by elastic network analysis. In particular, we use the Gaussian Network Model (GNM), where the protein is represented as a collection of interaction sites, corresponding to its amino acids, with springs between these that are sufficiently close to each other in 3D space (Bahar *et al.*, 1997; Haliloglu *et al.*, 1997; Emekli *et al.*, 2008). The analysis decomposes protein dynamics into modes of motion, each of which comprises rigid parts that move in opposite directions around hinges. The main objective of the analysis is to explore a correlation between the rigid parts and the themes, to suggest a link between protein dynamics and evolution. Moreover, GNM based Dynamic Perturbation analysis, again within the GNM approximation, will provide a further information for the correlation between protein dynamics and the themes.

As for the second part of the thesis, the dynamic determinants of deleterious mutations and their related compensatory mutations are investigated. The main premise here is that there is an interrelation between sequence, structure and function of proteins; thus, the dynamic perspective of the initial and compensatory mutations would plausibly disclose the main dynamic traits that could be important for the design of compensatory mutations/perturbations. For deleterious mutation analysis, a dataset is created by using the information gained from deep sequencing studies (Adkar et al.,2012; McLaughlin et al., 2012; Melamed et al., 2013; Firnberg et al., 2014; Kitzman et al., 2015; Mavor et al., 2016; Bandaru et al., 2017). For the compensatory mutations, p53 is selected as a case structure since p53 is the only protein which has crystal structures with its deleterious and its related compensatory mutations available on Protein Data Bank (Berman *et al.,* 2000). In order to gain dynamical information about the selected structures in deleterious and compensatory mutation datasets, GNM analysis is performed on the selected structures. The purpose of the GNM analysis is to observe correlations between hinge residues and mutation sites. For the compensatory mutations, hinge analysis is made for deleterious and their related compensatory mutations. Additionally, 2D cross correlation maps are produced for the available crystal structures.

Mode perturbation analysis is also made as a part of GNM analysis in both first and second part of the thesis. The purpose of the mode perturbation analysis is to mimic the effects of the mutations by perturbing the selected residue in the structure. This could be done in two different ways, increasing or decreasing the strain. Effects of perturbation on the whole structure will be analyzed as total fluctuation changes compared to unperturbed structure. Fluctuation changes are representing the entropy changes. The capacity to change the entropy of each residue aimed to be explored by mode perturbation analysis. The effect of perturbations is also examined as perturbations' capacity to alter the eigenvalues. Mechanistically key residues will be identified from perturbation analysis and their relationship with deleterious mutation sites will be explored.

The obtained information from GNM and GNM based mode perturbation analysis is believed to be a keystone for the prediction of deleterious and compensatory mutation sites. The knowledge gained from the dynamical behavior of the deleterious and compensatory

mutations will provide a better understanding of diseases and can be crucial in drug design projects.

2. MATERIALS AND METHODS

2.1. Themes Dataset

The themes were taken from the themes database of reuse in proteins (Nepomnyachiy *et al.,* 2017), which contains segments of at least 35 residues of similar sequence and structure (themes) that reoccur in different proteins. Of these, we examined themes taken from five proteins that correspond to two very different architectures: all-beta and all-alpha. The histone-binding protein CAF1 (PDB ID: 2XYI) and its close homologue WD repeat-containing protein 5 (WDR5, PDB ID: 3EMH) are both seven-blade beta-propellers, while ZYG-9 (PDB ID: 2OF3), protein phosphatase PP2A (PDB ID: 1B3U) and 26S proteasome subunit Rpn2 (PDB ID: 4ADY) are alpha-helix bundles. The themes detected in each of the five proteins are listed in Tables A1-3.

As an example, the 27 themes detected in the histone-binding protein CAF1 (2XYI) are presented in Figure 2.1; their length vary from 35 to 101 amino acids. Some of the themes can be observed in multiple positions along the protein; an additional index is added to these (e.g., 14815-1 through 14815-4).

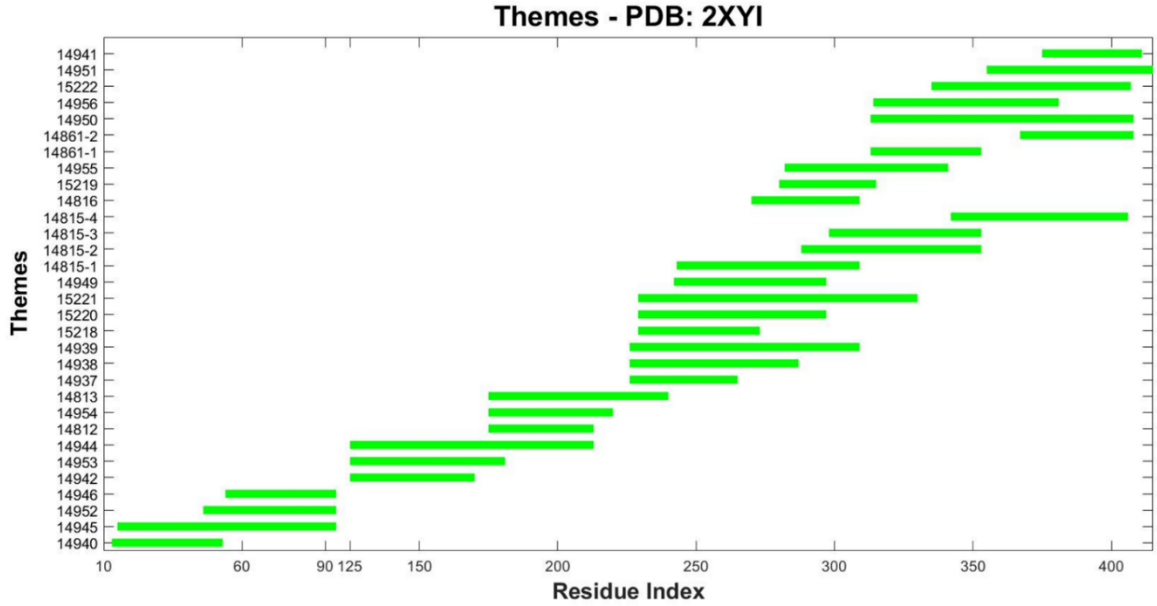


Figure 2.1. The themes detected in the 2XYI propeller. The x-axis indicates the amino acid positions of the themes, and the y-axis indicates their index (e.g. 14941,…14815-1, 14815-2… 14940).

2.2. Deep Sequencing Dataset

Deep sequencing, which is also known as next generation sequencing (NGS), high-throughput sequencing and massively parallel sequencing, is a new sequencing method which had a great impact on genomics. Deep sequencing method allows researchers to sequence an entire human genome in a single day whereas with the later method, Sanger sequencing, it would take almost a decade to do the same task. In deep sequencing or NGS, small fragments of DNA sequenced multiple times and in parallel. This process provides highly accurate data and provides an insight into DNA variations. Variations in the DNA involves substitutions, insertions or deletions. The full spectrum of genomic variation data can be obtained from deep sequencing data directly in a single experiment without the need of dedicated assays. Deep sequencing can be applied to whole genome or to small number genes. (Behjati & Tarpey, 2013; Goldman & Domschke, 2014)

In this work, a dataset which contains proteins that were analyzed with deep sequencing method, is produced. Each residue on the protein is changed to all remaining 19 amino acid with deep sequencing method. That provides a whole genome for a specific

protein. On those experiments, a fitness score is calculated for each amino acid change. According to those fitness scores, mutation sensitive residues are identified for each protein. The dataset of proteins which are analyzed with deep sequencing method and their corresponding studies are given in Table 2.1.

Table 2.1. Deep sequencing dataset.

PDB ID	Molecule	Reference
1BE9	PSD95-PDZ Domain	McLaughlin *et al.,* 2012
2VUB	CcdB Dimer	Adkar *et al.*,2012
3VUB	CcdB Monomer	Adkar *et al.*,2012
1D66	GAL4	Kitzman *et al.,* 2015
1CVJ	PAB1-RRM2 Domain	Melamed *et al.,* 2013
1UBQ	Ubiquitin	Mavor *et al.,* 2016
1XPB	TEM1 β-lactamase	Frinberg *et al.*, 2014
3K8Y	GTPase H-Ras	Bandaru *et al.*, 2017

2.2.1. PSD95-PDZ domain

PDZ domains are one of the most common protein interaction domains, which recognizes sequence motifs at the C-termini of target proteins. PDZ domains share similar topology and usually are around 80-100 residues long. PDZ domains play important role in multiple biological processes like transport, ion channel signaling and formation of supramolecular signaling complexes. (Harris & Lim, 2001; Lee & Zheng, 2010). The PDZ domain of PSD-95 is mostly involved in synapse formation. (Cui *et al.,* 2007; Toto *et al.,* 2016)

Mutation sensitive residues of PSD-95 PDZ domain utilized here are obtained from a study, which uses deep sequencing method on PSD-95 PDZ domain (McLaughlin *et al.,* 2012).

2.2.2. CcdB

The CcdB protein is the F plasmid carried bacterial toxin of the CcdA/CcdB toxin-antitoxin system. Main responsibility of CcdB is regulating cell death when CcdB's action is not prevented by its antitoxin CcdA. The target of CcdB is DNA gyrase subunit GyrA. CcdB forms a complex with GyrA which poisons the gyrase-DNA complex and blocks the passage of polymerases. This action causes double-strand breakage of the DNA thus cell death. (Bernard and Couturier, 1992; Bernard *et al.*, 1993; Bahassi *et al.*, 1999; Madl *et al.*, 2006)

Mutation sensitive residues of CcdB used here is obtained from a study about CcdB based on deep sequencing method (Adkar *et al.*,2012).

2.2.3. GAL4

GAL4 transcription factor is a positive regulatory protein of the yeast. GAL4 binds to UAS upstream activation sequence in the DNA which enables the transcription of the genes required for galactose metabolism GAL1, GAL7 and GAL10. GAL4 activity is inhibited in the presence of GAL80 and uninhibited in presence of galactose. (Guarente et al. 1982; Ginger *et al.*, 1985)

Due to its unique mechanism with upstream activating region UAS, it is widely used in the gene expression and function studies in organisms such as Drosophila. (Duffy 2002)

The DNA-binding domain of GAL4 (residues 2-65) is used in a deep sequencing study. Each GAL4 DNA binding domain residue is replaced by 19 other amino acid or by a stop codon. (Kitzman *et al.*, 2015). Mutation sensitivity of DNA-binding domain of GAL4 is obtained from the data collected in the mentioned study.

2.2.4. PAB1

Polyadenylate-binding protein (PAB) is an RNA binding protein which binds to the poly A tails of mRNAs. Almost all eukaryotic mRNAs have poly A tails at their 3' ends.

Main functions of PAB protein are initialization of translation and regulation of mRNAs in the cell by controlling their decay rates. (Kühn and Wahle, 2004)

PAB1 is the major poly (A) binding protein of the yeast Saccharomyces cerevisiae, which contains four RNA recognition motifs (RRM). The second RRM domain (RRM2) is revealed to be essentially responsible for poly(A) binding. (Deardorff and Sachs, 1997)

From a deep sequencing study primarily focuses on RRM2 domain of PAB1, mutation sensitive residues of the RRM2 domain is obtained (Melamed *et al.,* 2013).

2.2.5. Ubiquitin

Ubiquitin is a regulatory protein which is found in almost all eukaryotic cells. It is one of the most conserved proteins. Ubiquitin is mainly responsible from degradation of proteins by binding them. This process is called as ubiquitylation. (Varshavsky 2001)

Mutation sensitive residues of Ubiquitin is obtained from a deep sequencing study (Mavor *et al.,* 2016).

2.2.6. TEM1 β-lactamase

TEM-1 β-lactamase is an enzyme produced by bacteria, which provides resistance to antibiotics that contains β-lactam rings such as penicillin. β-lactamase hydrolyses the β-lactam ring of the antibiotics thus deactivates the antibacterial properties of the antibiotic. Since this process is crucial for new antibiotic design projects, the TEM1 β-lactamase is one of the best-studied antibiotic resistance enzymes. (Salverda *et al.,* 2010)

Mutation sensitivity information of residues of TEM-1 β-lactamase is obtained from a deep sequencing study which focuses on TEM1 β-lactamase (Frinberg *et al.,* 2014).

2.2.7. GTPase H-Ras

GTPase H-Ras is a small GTPase protein of the Ras family. Alternates between an inactive form bound to GDP and an active form bound to GTP. Activated by a guanine nucleotide-exchange factor (GEF) and inactivated by a GTPase-activating protein (GAP) (Williams, 2003). GTPase H-Ras is an oncogene protein. Mutations of GTPase H-Ras are implicated in a variety of human tumors (Cox and Der 2010).

Mutation sensitivity of residues for GTPase H-Ras are obtained from a deep sequencing study about GTPase H-Ras (Bandaru *et al.*, 2017).

In order to obtain uniformity among structures, 20 residues with highest functional cost of mutation are selected as mutation sensitive residues. So, for each case, we identified 20 residues as mutation sensitive residues according to the experimental results.

2.3.Gaussian Network Model

Gaussian Network Model (GNM) assumes folded state of proteins as three-dimensional elastic network in which interactions between residues close to each other are replaced by linear springs. The junctions are defined with Cα atoms in the protein and these atoms endure Gaussian distributed fluctuations. According to the model, the equilibrium correlation between fluctuations of two α carbons i and j is given by

$$\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle = \left(\frac{3k_b T}{\gamma}\right)[\boldsymbol{\Gamma}^{-1}]_{ij} \tag{2.1}$$

where Γ is a symmetric matrix known as Kirchhoff (connectivity) matrix, the subscript ij identifies the ijth element of the matrix, Ri is the position vector of the ith α-carbon. T and k_b are absolute temperature and the Boltzmann constant respectively and γ is the force constant of the Hookean pairwise potential which represents the interactions between the residues in the folded structure. The elements of Γ are given by

$$\Gamma_{ij} = \begin{cases} -1 \; if \; i \neq j \; and \; \boldsymbol{R}_{ij} \leq r_c, \\ 0 \; if \; i \neq j \; and \; \boldsymbol{R}_{ij} > r_c, \\ -\sum_{i,i\neq j} \boldsymbol{\Gamma}_{ij} \; if \; i = j \; . \end{cases} \quad (2.2)$$

where r_c is the cutoff separation defining the range of interaction of non-bonded α-carbons and R_{ij} is the distance between the ith and jth C_α atoms. The inverse of Γ may be written in terms of U and Λ. U is an orthogonal matrix whose columns u_i are the eigenvectors of Γ and Λ is the diagonal matrix of the eigenvalues λ_i of Γ.

$$\boldsymbol{\Gamma}^{-1} = \boldsymbol{U}(\boldsymbol{\Lambda}^{-1})\boldsymbol{U}^T \quad (2.3)$$

So, it is possible to decompose Γ^{-1} as the sum of contributions from individual modes and thus the correlation of the fluctuations of ith and jth C_α atoms can be expressed as the sum of the contribution of individual modes as

$$\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle = \left(\frac{3k_bT}{\gamma}\right) [\boldsymbol{U}(\boldsymbol{\Lambda}^{-1})\boldsymbol{U}^T]_{ij} = \left(\frac{3k_bT}{\gamma}\right) \sum_{k=2}^{n} \left[\boldsymbol{\lambda}_k{}^{-1} \boldsymbol{u}_k \boldsymbol{u}_k{}^T\right]_{ij} \quad (2.4)$$

where k is the kth vibrational mode. Since the first eigenvalue of Γ is equal to zero, first mode is not included in the summation. (Haliloglu *et al.*, 1997)

2.4. Dynamic Segments and Dynamic Domains

In each individual mode of motion, the dynamic parts are identified by the cooperativity of residue fluctuations; residues fluctuate either in the same direction, i.e., positively correlated, or opposite, i.e., negatively correlated (Emekli *et al.*, 2008). Accordingly, each mode includes two dynamic parts that move in opposite directions around hinges.

The dynamic parts can be projected onto the amino acid sequence of the protein, where a continuous stretch of amino acids of the same dynamic part is called ‘dynamic segment’

(see, e.g., the blue and red stripes in Figure 2.2). To avoid minor fluctuations, short fragments containing less than 15 residues are merged with the neighboring longer dynamic segment. The merged parts are called “dynamic domains”.

In Figure 2.2, the lower bars represent the raw results (dynamic segments) and upper bars represent the dynamic domains. For representative purpose only two GNM modes are shown: The slowest and third slowest. The dynamic segments of each mode are marked in blue and red along the protein sequence (lower bars). The dynamic domains (upper bars) are filtered versions of the dynamic segments, after smoothing using a 15 amino acids window. Each of the red and blue segments (after smoothing) is considered a dynamic domain.

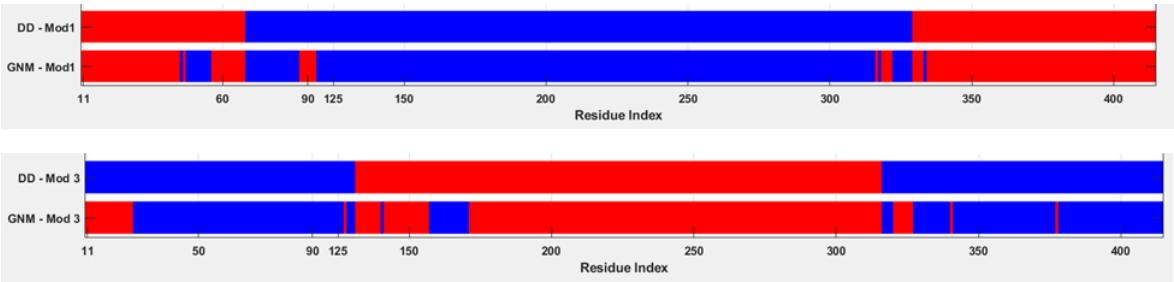


Figure 2.2. Dynamic segments and dynamic domains in the 2XYI propeller.

2.5. Mode Perturbation Analysis

In the mode perturbation analysis, a perturbation implemented in the GNM algorithm. This perturbation is given as adding a stiffness or loosening the interactions of one selected residue, which is considered as to mimic a mutation on that residue. Effect of that perturbation on the fluctuation of the residues is compared to unperturbed state.

Perturbation is implemented as stated below.

$$\Gamma^*_{ij,p} = \begin{cases} \boldsymbol{changed\ value\ if\ i \neq j\ and\ R}_{ij} \leq r_c, \\ 0\ if\ i \neq j and\ \boldsymbol{R}_{ij} > r_c, \\ 0\ if\ k = i, i \neq j and\ \boldsymbol{R}_{ij} > r_c, \\ -\sum_{i,i\neq j} \boldsymbol{\Gamma}_{ij}\ if\ i = j. \end{cases} \quad (2.5)$$

$$\langle \Delta \boldsymbol{R}_i \cdot \Delta \boldsymbol{R}_j \rangle^*_p = \left(\frac{3k_bT}{\gamma}\right) \left[\boldsymbol{U}^* (\boldsymbol{\Lambda}^{*-1}) \boldsymbol{U}^{*T}\right]_{ij} = \left(\frac{3k_bT}{\gamma}\right) \sum_{k=2}^{n} \left[{\boldsymbol{\lambda}^*}_k^{-1} \boldsymbol{u}^*_k {\boldsymbol{u}^*}_k^T\right]_{ij} \quad (2.6)$$

where p is the perturbed residue. This perturbation will be repeated for each residue i and a new connectivity matrix Γ^* will be obtained. This approach will give new eigenvectors, u^*, and new eigenvalues, λ^*, for each perturbation. Perturbed dynamics of eigenvalues and eigenvectors will be compared to the unperturbed state.

The analysis made with different cutoff values for GNM (7Å, 10 Å, etc.) and with different stiffness or loosening constants. The force constant in the connectivity matrix of GNM is "-1". By changing the constant as stated in Equation 2.5, stiffening or loosening effects of perturbation are implemented to the system, respectively, as example values "-1.2" for stiffening and "-0.8" for loosening.

PDZ structure 1BE9 is considered as sample structure and all various combinations of perturbations are tried on that structure. Also, the difference in fluctuations of perturbed and unperturbed states is formulated differently. In comparing the differences in residue fluctuations between the perturbed and unperturbed states, the following forms are used; actual value and absolute value of cumulative fluctuation difference of all residues.

Change in eigenvectors upon perturbation is calculated as fluctuation difference on whole structure upon perturbation. Steps of the calculation upon perturbing residue p is as following. First, eigenvectors are normalized according to eigenvalues of related mode and then mean squared fluctuations of each residue calculated for selected number of modes. Calculation of total fluctuation of residue i is given in Equation 2.7 for the unperturbed state and in Equation 2.8 for the perturbed states.

$$Total\ Fluctuation\ (i) = \sum_{k=1}^{\#\ of\ Modes} \boldsymbol{U}\ (i,k) * \boldsymbol{U}\ (i,k) / \boldsymbol{\Lambda}(k) \tag{2.7}$$

$$Total\ Fluctuation^{*}\ (i) = \sum_{k=1}^{\#\ of\ Modes} \boldsymbol{U}^{*}\ (i,k) * \boldsymbol{U}^{*}\ (i,k) / \boldsymbol{\Lambda}^{*}(k) \tag{2.8}$$

Fluctuation difference on whole structure upon perturbing residue p is calculated as the summation of each residue's total fluctuation differences of the perturbed states and the unperturbed state. Calculation of the fluctuation difference on whole structure upon perturbing residue p is given in Equation 2.9 for actual difference and Equation 2.10 for absolute difference, where n is number of residues.

$$\Delta \mathrm{U}\ (p) = \sum_{i}^{n} Total\ Fluctuation^{*}\ (i) - Total\ Fluctuation\ (i) \tag{2.9}$$

$$|\Delta \mathrm{U}\ (p)| = \sum_{i}^{n} |Total\ Fluctuation^{*}\ (i) - Total\ Fluctuation\ (i)| \tag{2.10}$$

The effect of perturbation on eigenvalues is also investigated. Total change in eigenvalues upon perturbing residue p is given in Equation 2.11.

$$\Delta\lambda(p) = \sum_{k=1}^{\#\ of\ Modes} \lambda^{*}(k) - \lambda\ (k) \tag{2.11}$$

The optimum set up for GNM mode perturbation analysis is selected as 10 Å cutoff value and "-1.2" as force constant in the connectivity matrix to mimic stiffening effect. Additionally, actual value of the fluctuation difference is used to analyze GNM mode perturbation, however, absolute value results are kept as back-up results.

On the mode perturbation analysis, fluctuation difference, which is considered as entropy change on the system, is considered as determinant factor. Residues which cause higher fluctuation difference on the structure upon perturbation are expected to be mutation sensitive residues.

2.6. Adjusted Mutual Information and Standardized Mutual Information

Mutual Information (MI) is a commonly used measure for clusters' comparison. Consider two random variables x and y with a joint probability mass function p(x, y) and marginal probability mass functions p(x) and p(y). Mutual information (x;y) is the relative entropy between the joint distribution p(x,y) and the product distribution p(x) p(y) (Cover and Thomas, 1991).

Here we use two related measures: Adjusted Mutual Information (AMI) and Standardized Mutual Information (SMI).

AMI is the normalized variant of mutual information. AMI ranges between 1, when the two partitions are identical, and 0, when the mutual information between two partitions equals to the value expected by chance alone (Vinh *et al.,* 2010).

SMI is obtained by probabilistic adjustment for chance on mutual information, it is simply the standardized form of mutual information. The value of SMI is the number of standard deviations the mutual information is from the mean, under a null distribution of random clustering solutions with fixed marginal (Romano *et al.,* 2014). Thus, an SMI value of 50 signifies that the mutual information is 50 standard deviations away from the mean of random clustering solutions.

Consider two types of clusterings from a dataset consisting of N records, A and B. Let the data clustered in k clusters of size a_i for each cluster i = 1,…., k, in A and let the data clustered in l clusters of size b_j for each cluster j = 1,…,l, in B. The number of records shared between clusters i and j are expressed as n_{ij}. The overlap between the two clustering can be represented in matrix form by the $k \times l$ contingency table as represented in Table 2.2.

Table 2.2. *k* x *l* contingency table of the overlaps between two clusterings.

		B				
		$\mathbf{b_1}$	…	$\mathbf{b_j}$	…	$\mathbf{b_l}$
A	$\mathbf{a_1}$	$\mathbf{n_{11}}$	…	.	…	$\mathbf{n_{1l}}$
	…	…		…		.
	$\mathbf{a_i}$	.		$\mathbf{n_{ij}}$		.
	…	…		…		.
	$\mathbf{a_k}$	$\mathbf{n_{k1}}$	…	.	…	$\mathbf{n_{kl}}$

Calculation of MI related analysis is performed as follows:

$$H(\mathrm{A}) = -\sum_{i=1}^{k} \frac{a_i}{N} \log \frac{a_i}{N} \qquad entropy\ of\ A \tag{2.12}$$

$$H(\mathrm{A},\mathrm{B}) = -\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{n_{ij}}{N} \log \frac{n_{ij}}{N} \quad joint\ entropy\ of\ A\ and\ B; \tag{2.13}$$

$$H(\mathrm{A}|\mathrm{B}) = -\sum_{i=1}^{k}\sum_{j=1}^{l} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{b_j/N} \quad conditional\ entropy\ of\ A\ given\ B; \tag{2.14}$$

$$MI(\mathrm{A},\mathrm{B}) = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{n_{ij}}{N} \log \frac{n_{ij}/N}{a_i b_j/N^2} \quad mutual\ information\ of\ A\ and\ B \tag{2.15}$$

$$MI(\mathrm{A},\mathrm{B}) = H(\mathrm{A}) - H(\mathrm{A}|\mathrm{B}) = H(\mathrm{A}) + H(\mathrm{B}) - H(\mathrm{A},\mathrm{B}) \tag{2.16}$$

$$E\{MI(\mathrm{U},\mathrm{V})\} = \sum_{i=1}^{k}\sum_{j=1}^{l}\sum_{n_{ij}=\max(a_i+b_j-N,0)}^{\min(a_i,b_j)} \left(\frac{n_{ij}}{N}\right.\cdot\cdot$$
$$\left.\cdot\cdot\log\left(\frac{N.n_{ij}}{a_i b_j}\right)\frac{a_i!\,b_j!\,(N-a_i)!\,(N-b_j)!}{N!\,n_{ij}!\,(a_i-n_{ij})!\,(b_j-n_{ij})!\,(N-a_i-b_j+n_{ij})!}\right) \tag{2.17}$$

$E\{MI(\mathrm{A},\mathrm{B})\}$ is the expected mutual information;

$$AMI(\mathrm{A},\mathrm{B}) = \frac{MI(\mathrm{A},\mathrm{B}) - E\{MI(\mathrm{A},\mathrm{B})\}}{\sqrt{H(\mathrm{A})\cdot H(\mathrm{B})} - E\{MI(\mathrm{A},\mathrm{B})\}} \tag{2.18}$$

$$SMI(\mathrm{A},\mathrm{B}) = \frac{MI(\mathrm{A},\mathrm{B}) - E\{MI(\mathrm{A},\mathrm{B})\}}{\sqrt{Var\{MI(\mathrm{A},\mathrm{B})\}}} \tag{2.19}$$

where $Var\{MI(\mathrm{A},\mathrm{B})\}$ is the variance of mutual information.

To carry out the MI analysis between two clusterings, A and B, both need to be partitions of the same data, and thus need to have the same length in total (Romano *et al.*, 2014).

In our case, we compare the dynamic domains defined in the slow modes of motion with the themes. Dynamic domains define all residues in a protein's structure, so in order to do the MI analysis, themes need to define or cover maximum possible number of residues in a structure. There are some gaps and overlaps between themes. With different cutoff values for the overlaps and the gaps, we identify sequence of themes that maximally covers the whole structure. The cutoff is taken as 3 and 5 residues on overlapping regions and 8, 10 and 15 residues for gaps between themes.

Thus, in total we consider three versions of decomposition of the protein into themes, 3 residue overlap and 8 residue gap, 5 residue overlap and 10 residue gap, and 5 residue overlap and 15 residue gap. We generate these sequences of the themes with a Monte Carlo like algorithm which enables us to produce all possible combinations with the given restrictions.

The SMI and AMI computations in Equations 2.18 and 2.19 are performed using the MATLAB code provided by Romano et al. (Romano *et al.*, 2014).

2.7. Statistical Significance Analysis for Perturbation

In order to interpret the results of perturbation analysis, a statistical significance analysis is performed. Local/global minimum positions are considered when analyzing the results of actual fluctuation difference values between perturbed and unperturbed states. These residues are considered to be the ones which decrease the global entropy of the structure most among their neighbors when the perturbation is added as stiffness to the system. Additionally, for the eigenvalue analysis, local/global maximum positions are considered. Those residues are the ones which change eigenvalues most among their neighbor residues.

Local minimum or maximum positions are calculated according to their preceding and subsequent residues. If a position has lowered the fluctuation more than preceding and subsequent residues, that position is considered as local minimum for fluctuation difference results. In addition, for the eigenvalue difference results, if a position has increased total eigenvalue more than preceding and subsequent residues, that position is considered as local maximum. Local minimum positions are obtained with the Equation 2.20.

$$\begin{gathered} if\ \ \Delta\mathrm{U}\,(i)\ <\ \Delta\mathrm{U}\,(i-1)\ \ \&\ \ \Delta\mathrm{U}\,(i)\ <\ \Delta\mathrm{U}\,(i+1) \\ i = local\ minimum \end{gathered} \tag{2.20}$$

Local maximum positions are obtained with the Equation 2.21.

$$if\ \ \Delta\lambda\,(i) > \ \Delta\lambda\,(i-1)\ \ \&\ \ \Delta\lambda\,(i) > \ \Delta\lambda\,(i+1)$$
$$i = local\ maximum \tag{2.21}$$

After identifying the local minimum and maximum positions, relationship between mutation sensitive residues and these minimum and maximum positions is investigated. The number of mutation sensitive residues which overlap with the local minimum positions on fluctuation difference results is obtained. Following, random sampling method is used in order to decide whether the relationship between mutation sensitive residues and these minimum and maximum positions is statistically significant or not.

In random sampling method, twenty residues are selected randomly. Number of residues which overlap with local minimum positions is calculated and this procedure is repeated for 10000 times. Population mean, μ, and standard deviation, σ, are obtained from the generated population. Additionally, the mean distance of mutation sensitive residues from local minimum positions is calculated. Population mean and standard deviation for distance analysis are also gathered from the population that generated from random sampling procedure.

Next step in the statistical significance analysis is to obtain z-score from population mean and standard deviation. Z-score gives how many standard deviations does sample mean (X) away from the population mean and can be calculated with Equation 2.22.

$$Z = \frac{(X - \mu)}{\sigma} \tag{2.22}$$

X: sample mean
μ: Population mean (generated with random sampling)
σ: Standard deviation (generated with random sampling).

The final step of statistical significance analysis is to obtain p-value by using z-score. P-value gives the probability of obtaining sample mean, X, or higher (if left tail X or lower) in the population generated from random sampling. P-value can be calculated with

Equation 2.23, where tcdf is the cumulative distribution function (cdf**)** of the t distribution at the value z using the corresponding degrees of freedom, df.

$$P\ value = 1 - tcdf(z, df) \tag{2.23}$$

3. RESULTS AND DISCUSSION

3.1. Themes and Dynamic Domains

To demonstrate the correlation between the structural dynamics and themes we start with in depth analysis of selected cases with very different architectures: Two homologous propeller structures, and three alpha helix bundles. We consider only the number of slowest modes that are responsible to sharper decay from the slowest to fastest in the eigenvalue distribution; contributing relatively significant to the global dynamics. For the cases studied here, this number of modes change from four to seven slowest modes, which approximates the significant part of the dynamic spectrum.

3.1.1. All-beta architecture: Propellers

We start with two homologous proteins, sharing 27.8% sequence identity and the same seven-blades propeller architecture (superimposition RMSD of 1.83 Å): Histone-binding protein CAF1 (PDB ID: 2XYI) and WD repeat-containing protein 5 (PDB ID: 3EMH).

Figure 3.1 displays the predicted dynamic segments in the seven slowest modes of motion (x-axis) in 2XYI, as well as the themes observed in this protein (results for each mode given in Figures A1-A7. Each mode of motion constitutes two dynamic parts that move in opposite directions. The sites at the intersection between the dynamic segments (in the bar below the figure) are the positions where the structure flexes. Each of the three slowest modes of motion features three dynamic segments, which are relatively long. The higher modes have more dynamic segments, which are typically shorter. Within a given mode, each amino acid can belong to one dynamic segment only. However, when considering all the slowest modes, the same amino acid belongs to many segments. In this respect, the modes of motion are intertwined, just like the themes.

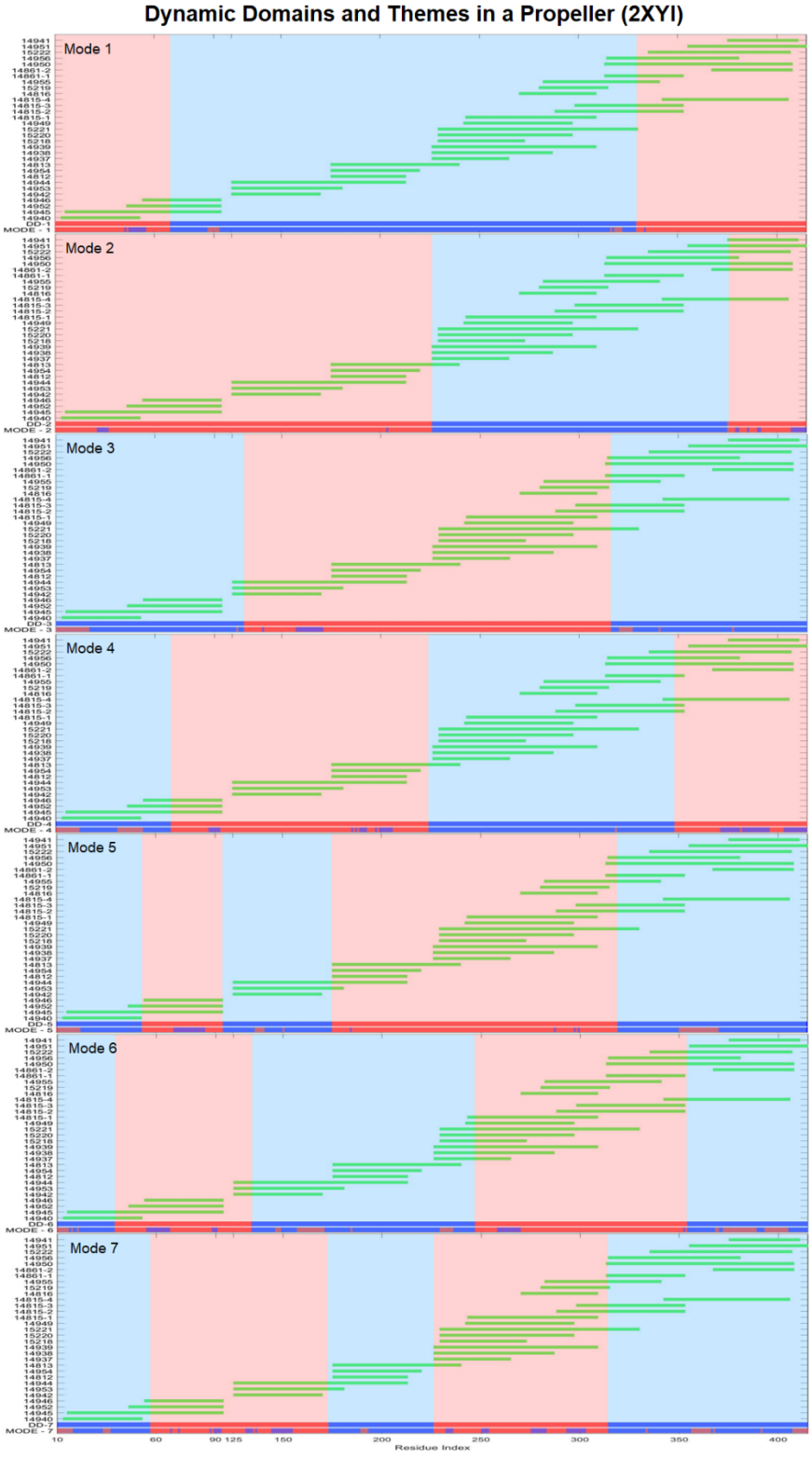


Figure 3.1. Correlation between structural dynamics and commonly used sequences. The dynamic domains of the 2XYI propeller in each of the seven slowest modes, are marked along the x-axes. The recurring sequences, 'themes' are marked along the y-axes and their positions are highlighted in green.

Alignment shows good correlation between the dynamic domains and the themes of 2XYI; most of the dynamic domains align with the themes, either alone or in combination with other themes. While a few short dynamic domains may constitute a theme, several themes may also constitute a dynamic domain. As it can be seen, the dynamic domains (as well as the themes) may be part of different dynamic domains for different modes of motion. Based on the motion defined by each mode, they can be recruited or integrated in multiple ways. This suggests a possible hierarchal rewiring of the themes to facilitate functional motions.

Next, we analyze 3EMH, another propeller structure. This protein shares many themes with 2XYI; a list of the shared themes is given in supplementary Table A2. Reassuringly, the dynamic domains of 3EMH align with the shared themes either alone or in combination with each other. Interestingly, sometimes the equivalent themes of the proteins match with dynamic domains from different modes of motion. For example, theme 14815 with a dynamic domain from the slowest mode for 2XYI, and a dynamic domain of the fifth slowest mode of 3EMH (Figure 3.2). In a slightly more complicated example, a combination of themes 14813 and 14815 aligns with a dynamic domain of the fourth slowest mode of 2XYI, and with a dynamic domain of the second slowest mode of 3EMH (Figure 3.3). Dynamic domains for each mode of 3EMH with the associated themes are given in Figures A8-15.

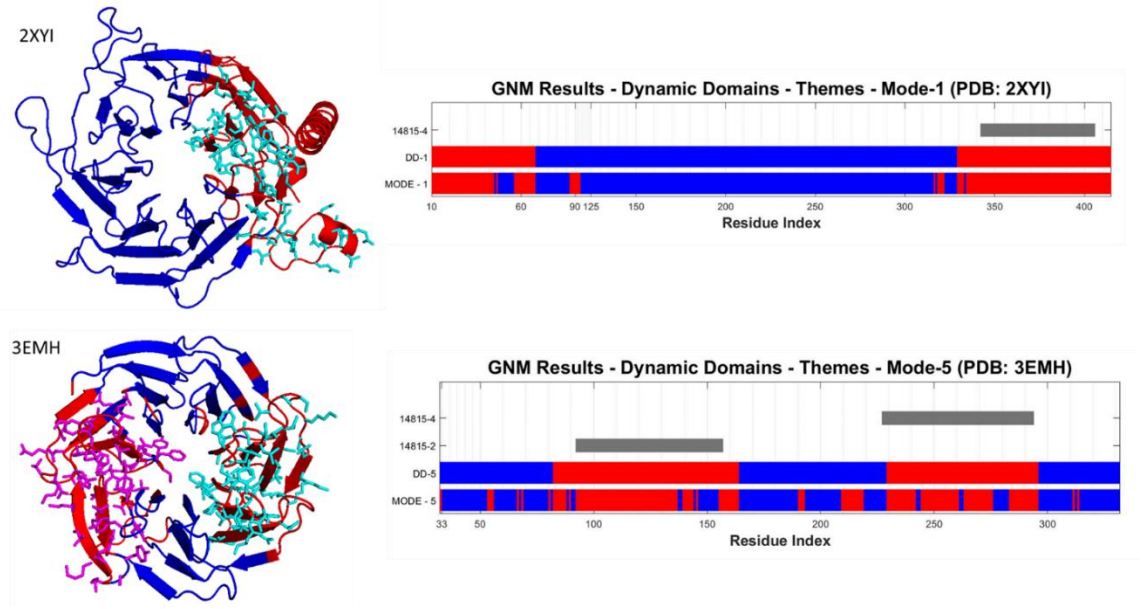


Figure 3.2. Theme 14815 corresponds to dynamic domains in 2XYI and 3EMH. The theme is shown as cyan or purple side chain.

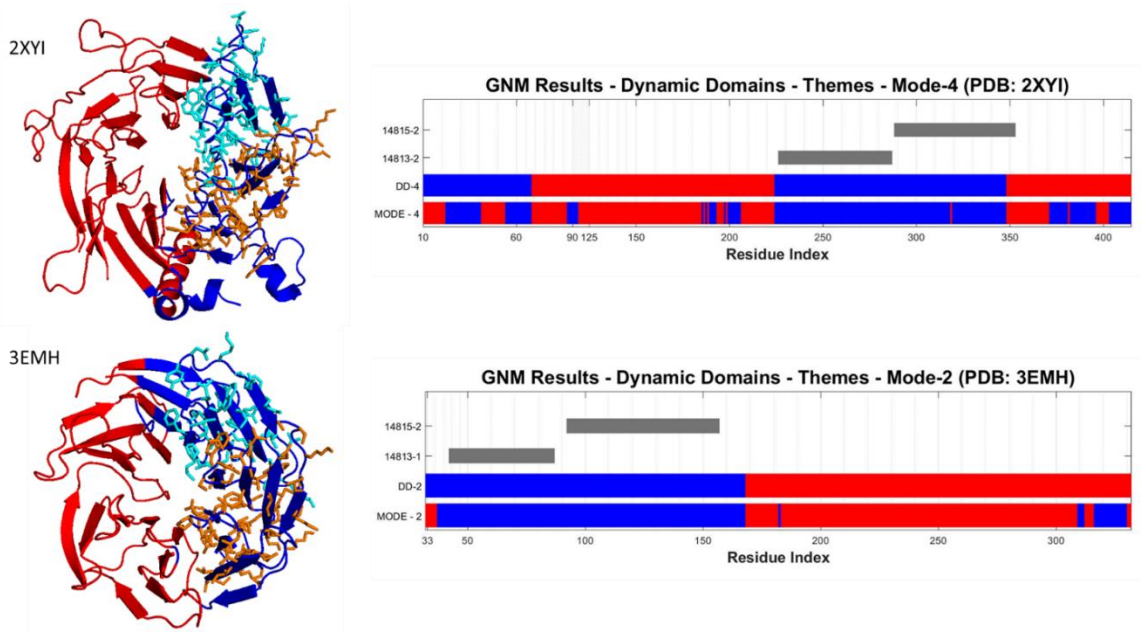


Figure 3.3. Theme 14813 and 14815 correspond to dynamic domains in 2XYI and 3EMH. The themes are shown as ribbon side chains in cyan for 14813 and orange for 14815.

3.1.2. All alpha architectures: alpha helix bundles

We examine the correlation between dynamic domains and the themes also in three all-alpha proteins (sequence identity from BLAST below 50.00%; 2OF3-1B3U (50.00%), 2OF3-4ADY (26.67%), 4ADY-1B3U (26.19%)). All three are alpha helix bundles, but they are architecturally different from each other. Their themes – shared and non-shared – and their positions on the structures are given in supplementary Table A3. The dynamic domains in a group of slowest modes of all three proteins with the themes are presented in supplementary Figures A3, Figures A4, and Figures A5, respectively. Green themes represent the shared themes between all three structures, magenta colored themes are the themes shared between 1B3U and 4ADY (cyan colored themes are the non-shared themes). It is observed that hinge points are mostly located at the edges of the given themes and the themes mostly correlate with the dynamic domains as anticipated. As an example, common theme c180-36 (and c180-8) defines a specific dynamic domain in one of the slowest modes of 1B3U, while at the same time defining specific dynamic domains in different modes of the other two structures 2OF3 and 4ADY (Figure 3.4). The same can be observed for theme c180-19, which is represented for all three cases in Figure 3.5. As observed, these themes align with the dynamic domains almost perfectly.

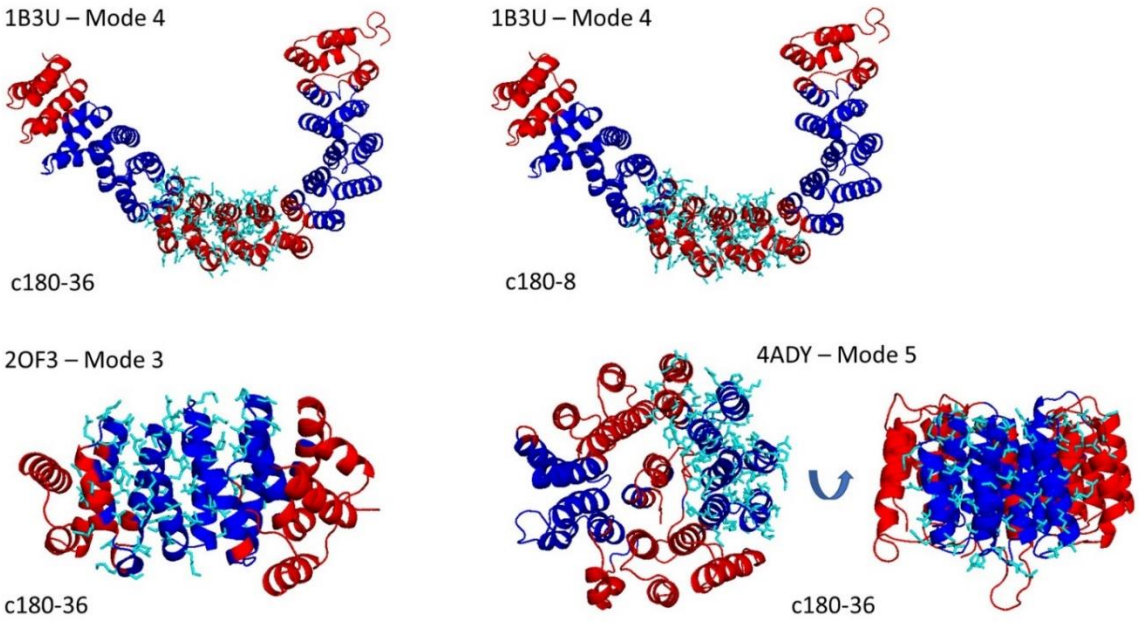


Figure 3.4. Theme c180-36 with dynamics domains of 1B3U, 2OF3 and 4ADY. C180-36 and C180-8 differ only by a few residues. Themes are represented as cyan side chains. Blue and red are the dynamic domains in a given mode (indicated in each).

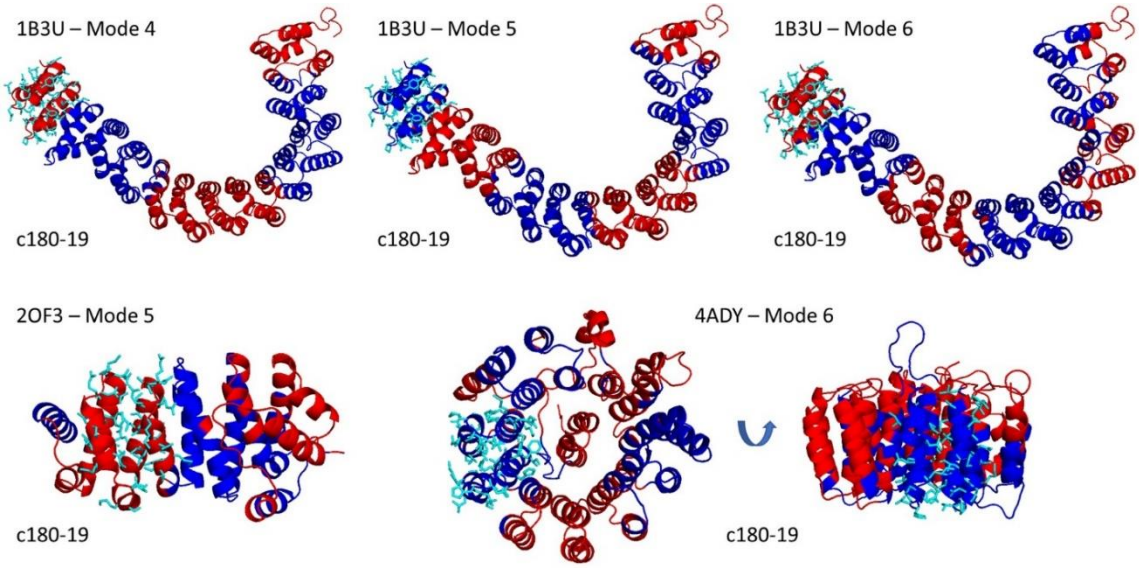


Figure 3.5. Theme c180-19 with dynamic domains of 1B3U, 2OF3 and 4ADY. Themes are represented as cyan side chains. Blue and red are the dynamic domains in a given mode (indicated in each panel).

In the next section, we examine whether the correlation between the dynamic domains and themes is statistically significant.

3.1.3. Mutual Information Analysis

The themes and dynamic domains are fundamentally different entities, which complicates their comparison. In particular, the dynamic domains of each mode always cover the whole structure, but the themes typically do not. In addition, the dynamic domains of a given mode do not overlap, while the themes sometimes do. To minimize the gaps and overlaps, various combinations of themes are randomly combined as presented in Materials & Methods. We start with detailed description of the analysis of the propeller structure of 2XYI.

The list of combinations for the initial restriction of 2XYI is provided in Table A4. The AMI and SMI values between the predicted dynamic domains in the seven slowest modes for all theme combinations are computed. For combinations with initial restriction (3 residue overlap, 8 residue gap), the AMI values are observed to be in the range between 0.464 and 0.913, and the SMI values are observed to be in the range between 90.30 and 168.8, results for each GNM mode are given in Tables 3.1 and 3.2, respectively. For combinations with other restrictions, mutual information results for each GNM mode are given in Tables A5-A8, which appear with the range of 0.431 to 0.913 for AMI and 89.90 to 168.8 for SMI. The results indicate that there is a strong statistical correlation between the dynamic domains and themes, being particularly strong in slow modes three to seven. The bar representation of the best AMI scored theme combination and related dynamic segments is given as an example in Figure 3.6.

Table 3.1. AMI values for the correlation between the dynamic domains of each of the seven slowest GNM modes of 2XYI and theme combinations, filtered with 3 residues overlap and 8 residues gap limit.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5	Mode-6	Mode-7
Minimum	0.464	0.518	0.626	0.586	0.738	0.647	0.719
Maximum	0.558	0.658	0.762	0.743	0.893	0.788	0.913
Average	0.511	0.596	0.699	0.673	0.817	0.724	0.817

Table 3.2. SMI values for the correlation between the dynamic domains of each of the seven slowest GNM modes of 2XYI and theme combinations, filtered with 3 residues overlap and 8 residues gap limit.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5	Mode-6	Mode-7
Minimum	90.30	105.3	135.5	119.5	128.5	112.3	134.9
Maximum	109.9	137.7	179.6	142.7	163.5	144.7	168.8
Average	101.9	120.2	154.2	132.2	147.0	130.3	149.6

Figure 3.6. The theme combination that was assigned the highest AMI score of correlation with the dynamic domains of the seventh slowest mode. (AMI = 0.9126)

The same restriction for combination of themes also applied to 3EMH. For 3EMH, the AMI values are observed to be in the range between 0.438 and 0.897, and the SMI values are in the range between 82.53 and 132.5 (Table 3.3 and Table 3.4). It should be noted here that the results of the slow modes 3, 6 and 7 are indecisive, due to not being able to observe dynamics segments larger than 15 residues. As described in Materials & Methods section, the short dynamic segments are joined to next larger dynamics segment in defining dynamics domains.

Table 3.3. AMI values for the correlation between the dynamic domains of each of the seven slowest GNM modes of 3EMH and theme combinations, filtered with 3 residues overlap and 8 residues gap limit.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5	Mode-6	Mode-7
Minimum	0.533	0.438	N/A	0.613	0.657	N/A	N/A
Maximum	0.661	0.644	N/A	0.761	0.897	N/A	N/A
Average	0.593	0.574	N/A	0.656	0.785	N/A	N/A

Table 3.4. SMI values for the correlation between the dynamic domains of each of the seven slowest GNM modes of 3EMH and theme combinations, filtered with 3 residues overlap and 8 residues gap limit.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5	Mode-6	Mode-7
Minimum	82.53	85.30	N/A	89.75	98.08	N/A	N/A
Maximum	118.5	135.3	N/A	123.9	132.5	N/A	N/A
Average	94.66	110.9	N/A	99.79	110.6	N/A	N/A

3.1.4. Perturbation Analysis

To further examine the dynamics, we conducted perturbation analysis within the GNM framework. A perturbation to a residue (each and every one) here means to increase its interaction force constant, and then estimate its effect on the fluctuations of all other residues. One can then compare the fluctuation differences (perturbed-unperturbed) versus perturbed residue or versus affected residue. Some residues have high plausibility to affect the fluctuation profiles of other residues, and at the same time these residues are being affected at the most by the perturbation of others. These are considered the points of structural dissection.

One can think of various measures of the effects of a perturbation. We examined some and concluded that they are similar. This is demonstrated with the 2XYI propeller structure. Figure 9A displays the absolute cumulative fluctuation change upon perturbation (averaged over the seven slowest GNM modes) vs. perturbed residue. Figure 9B shows the resulting structural dissection, while Figure 9C displays the themes detected in this protein.

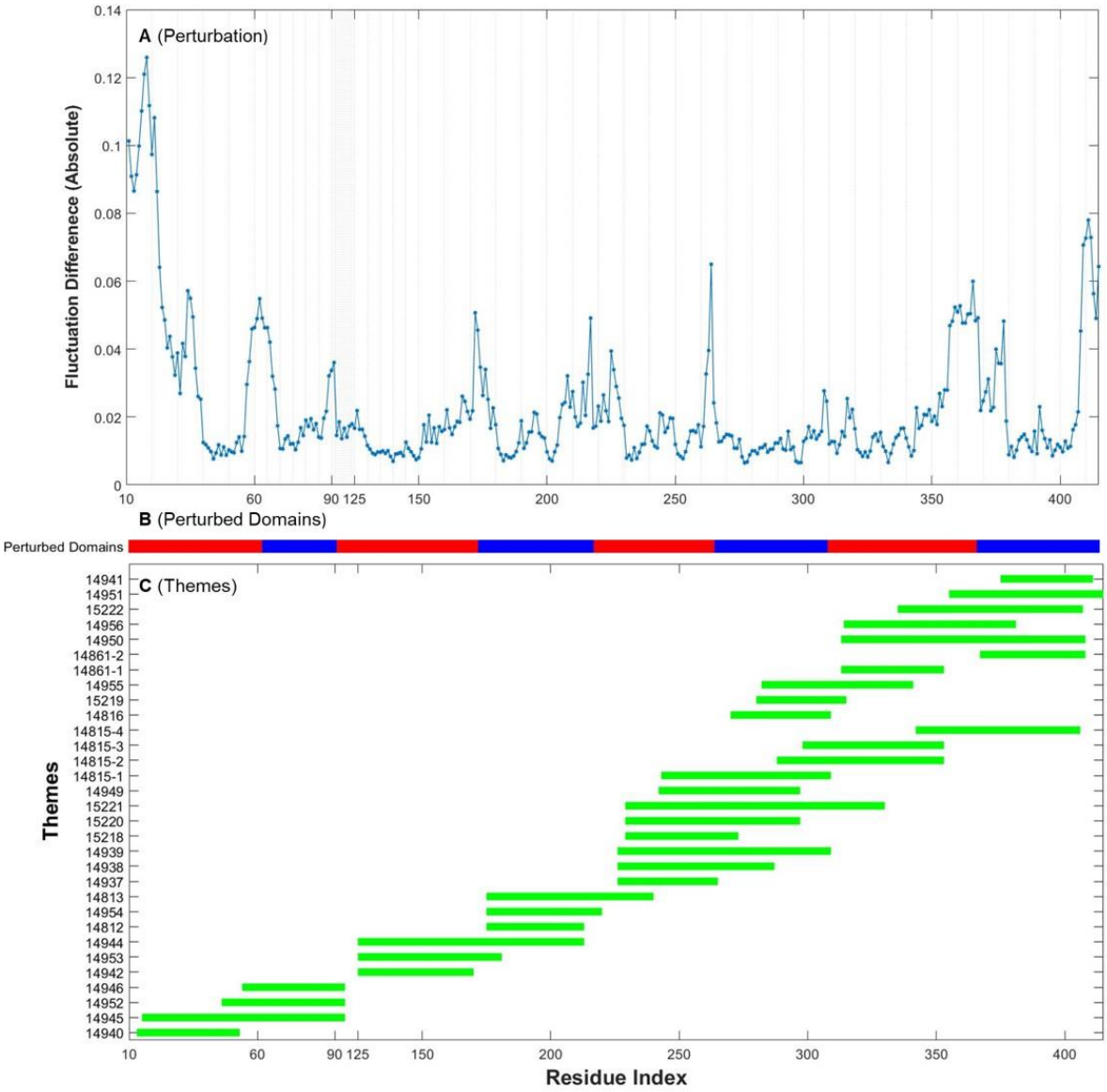


Figure 3.7. Absolute cumulative fluctuation changes as response to perturbation on each residue of 2XYI (A) with the segments/parts defined (B) in comparison with the themes layout (C).

Residues with highest response to perturbations (local/global peaks in Figure 9A) are at theme boundaries. The statistical significance of the correlation between structural dissection by most highly perturbing/perturbed positions and theme boundaries is assessed by the SMI and AMI analysis. AMI values are in the range between 0.688 and 0.887 with an average of 0.774 and SMI values are in the range between 109.3 and 134.3 with an average of 116.9 (Tables 3.5 & 3.6), indicating significant correlation between dynamic perturbation profile and themes alignment in the structure; residues that highly affect or are highly affected by dynamic perturbations tend to be at the boundaries between themes.

Table 3.5. AMI results for the correlation of the structural segments by the perturbations and various theme combinations (2XYI).

	3 Residue Overlap 8 Residue Gap Limit	**5 Residue Overlap 10 Residue Gap Limit**	**5 Residue Overlap 15 Residue Gap Limit**
Minimum AMI	0.737	0.688	0.688
Maximum AMI	0.844	0.844	0.887
Average AMI	0.784	0.764	0.774

Table 3.6. SMI results for the correlation of the structural segments by the perturbations and various theme combinations (2XYI).

	3 Residue Overlap 8 Residue Gap Limit	**5 Residue Overlap 10 Residue Gap Limit**	**5 Residue Overlap 15 Residue Gap Limit**
Minimum SMI	118.2	110.2	109.3
Maximum SMI	127.6	127.6	134.3
Average SMI	121.3	117.4	116.9

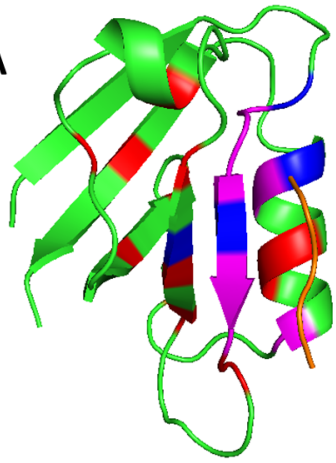
The highest effect of perturbations is expected in residues that are key for the global dynamics of the protein. That they align with the boundaries of themes disclose an entropic view for the themes as building blocks that are reused in the protein worlds, which might have implications for the evolutionary process.

3.2. Deleterious Mutations – GNM Mode Perturbation Analysis

3.2.1. PSD95-PDZ Domain

Mutation sensitivity of residues are obtained with respect to ligand binding affinity. The ligand CRIPT is quantitatively linked to the expression of enhanced green fluorescent protein (eGFP). eGFP levels are measured and compared to wild type in order to get the average functional cost of each amino acid substitution over all positions (McLaughlin et al., 2012). The average functional cost is used to determine the mutation sensitivity of residues. Twenty residues with highest mutation sensitivity and residues on binding site to PDZ's ligand are given in Figure 3.8 as both their positions on 3D structure (Figure 3.8A) and as a list with their residue index (Figure 3.8B).

A



B

Mutation Sensitive Residues	Binding Site Residues
323 324 325 327 328 329 330 336 338 341 347 353 359 362 367 **372** 375 376 **379** 388	322 **323 324 325 326 327 328** 339 372 **379** 380

Figure 3.8. Mutation sensitive residues and residues on binding site for PDZ. **A.** 3D structure of PSD PDZ domain (PDB ID: 1BE9). (**Red:** Mutation sensitive residues. **Blue:** Binding site residues. **Magenta:** Mutation Sensitive Residues which are also located on binding site.) **B.** List of mutation sensitive residues and residues on binding site for PDZ. (**Bold:** Mutation sensitive residues which are also located on binding site.)

PDZ core domain (Residues from 311 to 393) with its ligand is used for GNM perturbation analysis (PDB ID: 1BE9). Fluctuation difference with respect to unperturbed state is calculated for three and ten slowest GNM modes and for all GNM modes. The results

are given in Figures 3.9 for three mods, 3.10 for ten modes and 3.11 for all modes with 10 Å cutoff value and "-1.2" as force constant in the connectivity matrix. Twenty residues of highest functional cost to mutations and known binding site residues are marked.

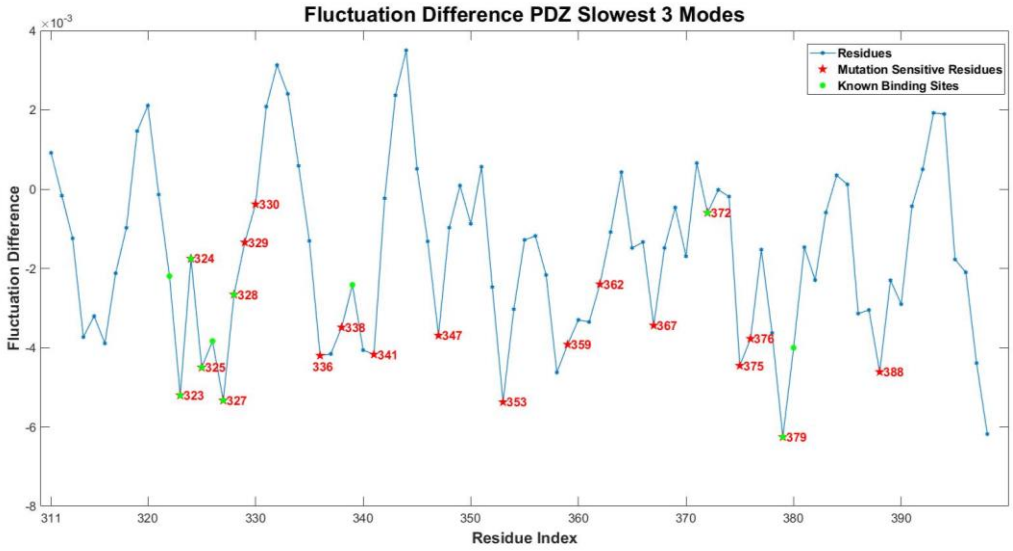


Figure 3.9. PDZ perturbation analysis - Fluctuation difference vs residue index in three slowest GNM modes.

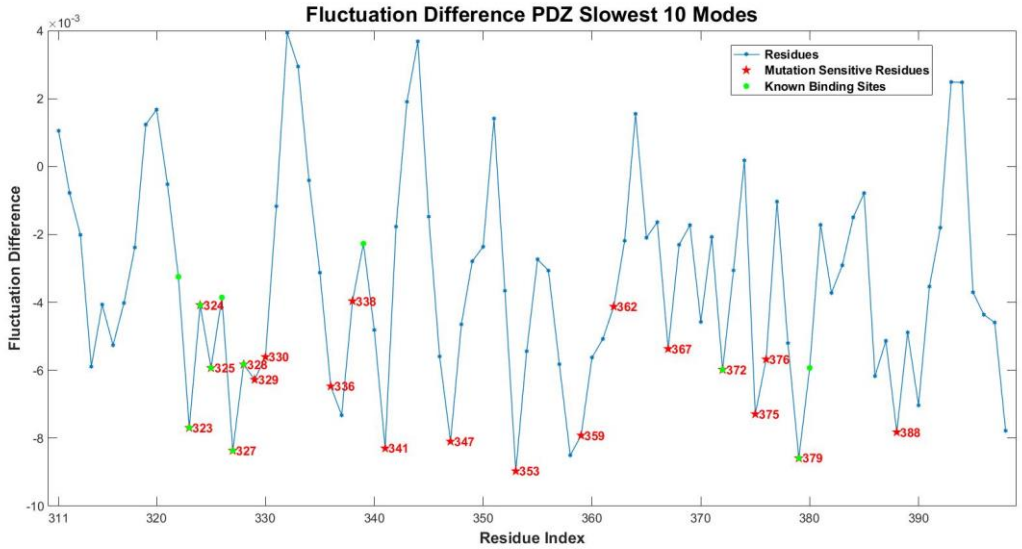


Figure 3.10. PDZ perturbation analysis - Fluctuation difference vs residue index in ten slowest GNM modes.

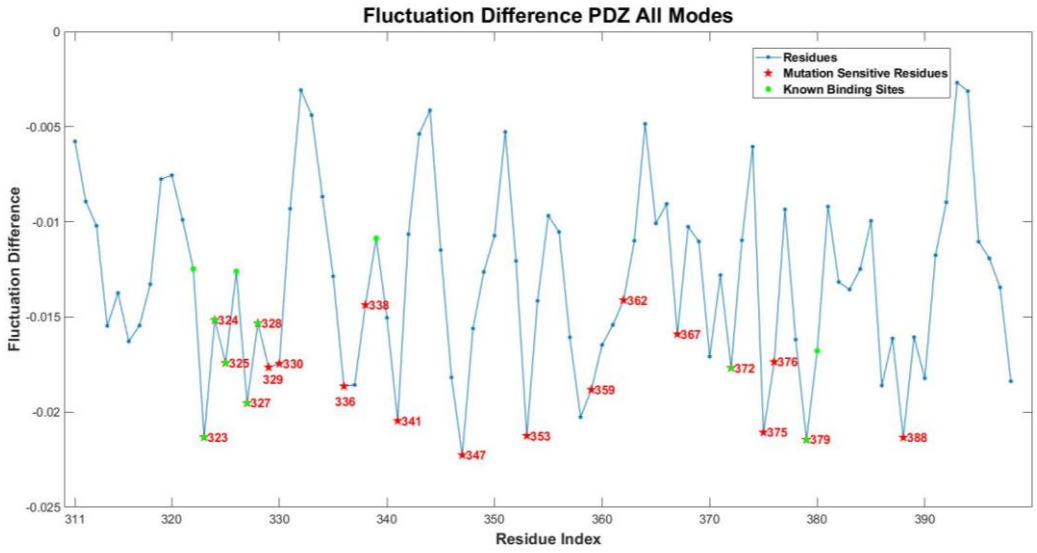
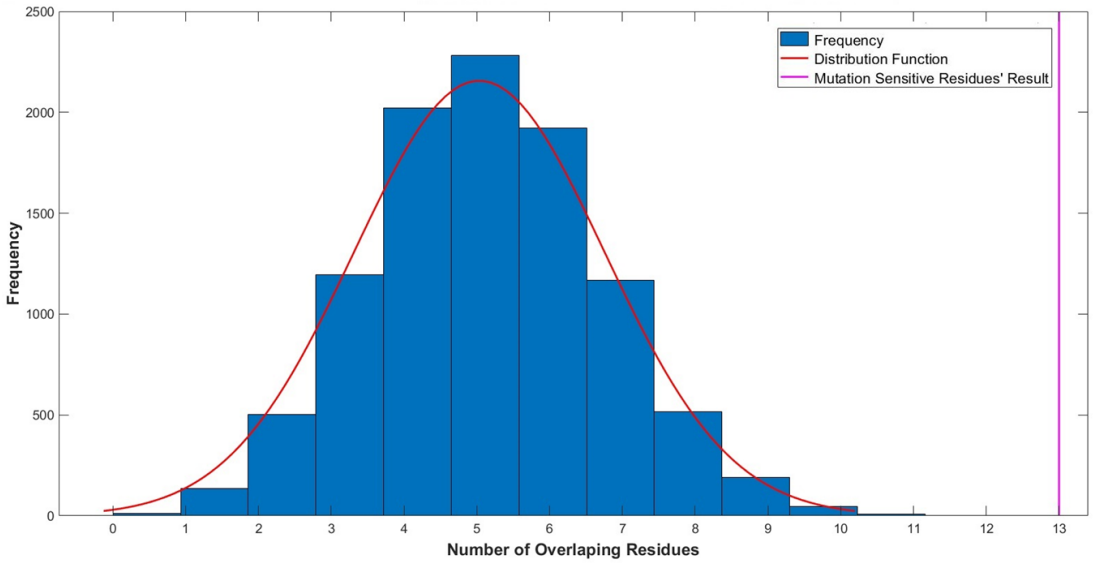


Figure 3.11. PDZ perturbation analysis - Fluctuation difference vs residue index in all modes.

The results showed that mutation sensitive residues are found to be at local/global minimums of the GNM perturbation analysis on fluctuation difference which displays the maximum decrease in the overall mean squared fluctuations. These results indicate that mutation sensitive residues are the ones which alters the global entropy as proportion to fluctuations of the structure most among their neighbors. In order to estimate the significance of the mutation sensitivities of residues to be correlated with those positions, a statistical significance analysis which is explained in detail in Materials & Methods section, is performed. Analysis is performed for both number of residues overlapping with the local minimum positions and mean distance of mutation sensitive residues to local minimum positions (C-alpha atoms). Distance analysis is carried out in order to consider neighboring effect and residue positions in space. Histogram and distribution function for number of residues overlapping with the local minimum positions are given in Figure 3.12 and mean distance to local minimum positions in Figure 3.13, together with detailed results and p-values obtained from the statistical significance analysis. Statistical significance is made for only all modes analysis.

A

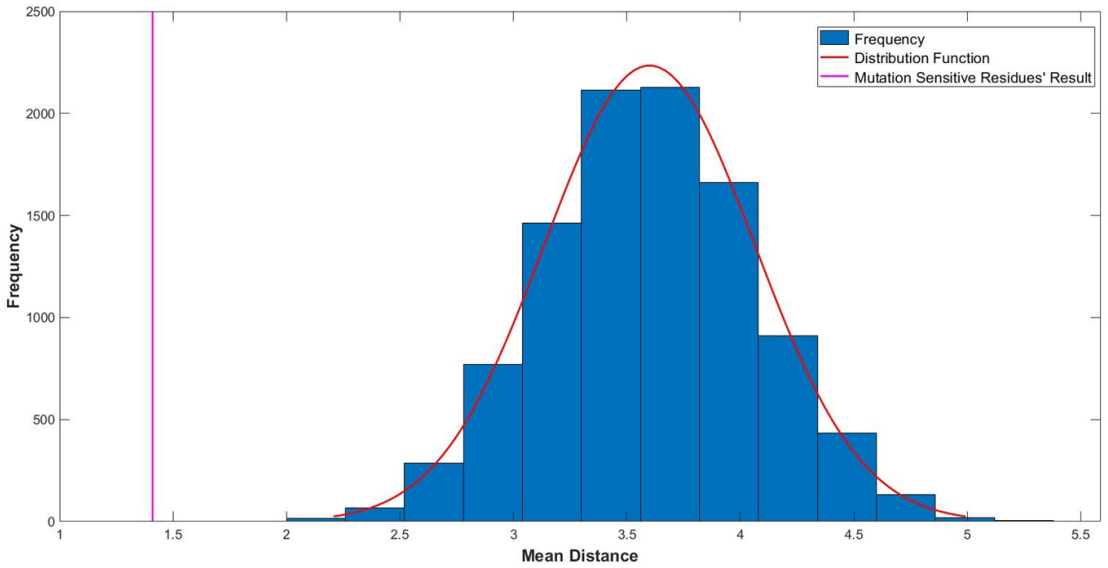


B

# of Mutation Sensitive Residues Located on Local Minimums	Mean # of Residues Located on Local Minimums (random sampling)	Standard Deviation	P-value
13	5.06	1.72	1.91×10^{-6}

Figure 3.12. PDZ results - Histogram and distribution function for number of residues overlapping with local minimum positions.

A



B

Mean Distance of Mutation Sensitive Residues to Local Minimums (C-alpha)	Mean Distance of Random Selected Residues to Local Minimums (Random sampling)	Standard Deviation	P-value
1.41 Å	3.59 Å	0.47 Å	1.67×10^{-6}

Figure 3.13. PDZ results - Histogram and distribution function for mean distance to local minimum positions (C-alpha).

The residues that are located on the local minimum positions are obtained with the constraint mentioned in the Materials & Methods sections. Twenty-one residues are identified as local minimum positions. Those residues are obtained as residues 314, 316, 323, 325, 327, 329, 336, 341, 347, 353, 358, 365, 367, 370, 372, 375, 379, 383, 386 and 388 (Figure 3.11).

Overall, thirteen mutation sensitive residues (323, 325, 327, 329, 336, 341, 347, 353, 367, 372, 375, 379 and 388) out of twenty are found out to be in the local minimum positions. The population mean from random sampling obtained as 5.06 residues with a standard deviation of 1.72. When a one-tailed t-test is applied to those results, the p-value is obtained as 1.91 x 10-6 which indicates a highly significant correlation between the mutation sensitive residues and local minimum positions.

Distance analysis is made for only C-alpha atoms. The mean distance of mutation sensitive residues to local minimums is calculated as 1.41 Å. From random sampling, the population mean distance of randomly selected residues to local minimums is obtained as 3.59 Å with a standard deviation of 0.47 Å. A one-tailed t-test is applied to the results, the p-value is calculated as 1.67 x 10-6.

The effect of perturbation on the distribution of eigenvalues is also analyzed as stated in Materials & Methods section. When a perturbation applied to residues, a change in eigenvalues of GNM occurs. Total change in eigenvalues upon perturbing each residue is calculated and the results are given in Figure 3.14.

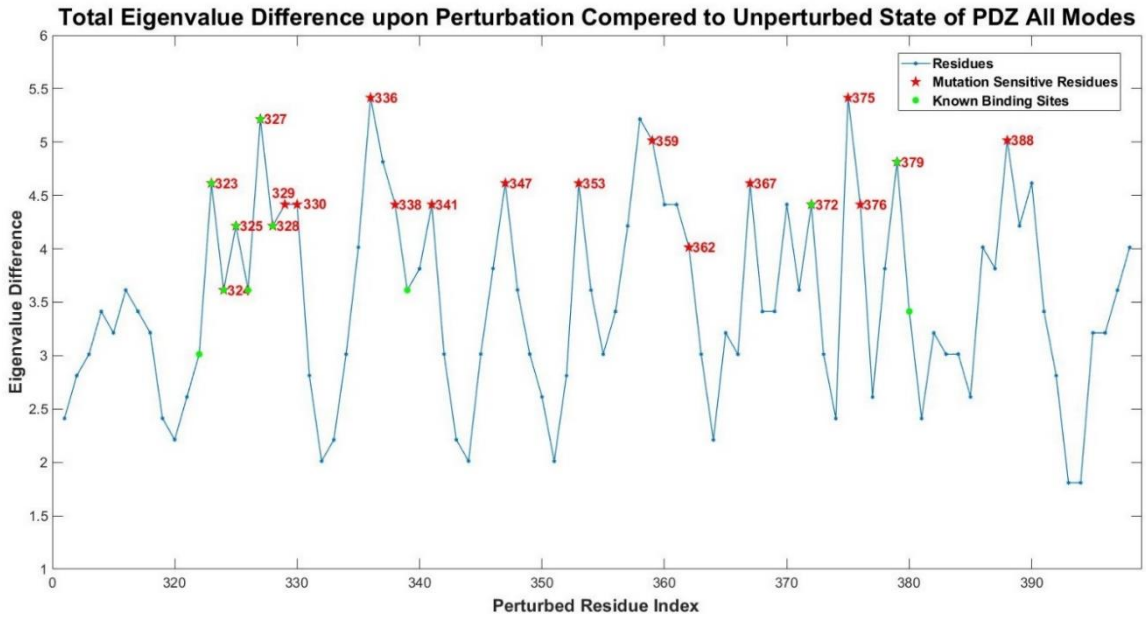


Figure 3.14. PDZ perturbation analysis - Eigenvalue difference vs residue index in all GNM modes.

It is observed that the mutation sensitive residues are located on the local maximum points on the eigenvalue difference graph. These residues are the ones that have the capacity to change the eigenvalues of the GNM with the perturbations on their local positions.

Statistical significance analysis is also carried out for the significance of obtaining mutation sensitive residues on local maximum positions on eigenvalue difference results. The analysis is made for both number of residues overlapping with the local minimum positions and mean distance to local minimum positions (C-alpha atoms), as performed in fluctuation difference analysis.

Twenty-one residues are identified as local maximum positions that are determined with the procedure mentioned in the Materials and Methods section. Those residues are 314, 316, 323, 325, 327, 330, 336, 341, 347, 353, 358, 365, 367, 370, 372, 375, 379, 382, 386, 388 and 390. Thirteen of the mutation sensitive residues (323, 325, 327, 330, 336, 341, 347, 353, 367, 372, 375, 379 and 388) are found out to be located at these maximum points. Random sampling is obtained with the method mentioned in the Materials and Methods section just as in the fluctuation difference analysis. The p-value for overlapping residues is calculated as 1.43 x 10-6 and for the distance analysis, the p-value is calculated as 1.80 x 10-6. P-value results support that there is a strong correlation between the residue's capacity to change eigenvalue magnitude and distribution upon perturbation and residue's mutation sensitivity. Detailed results for each analysis can be observed in Table 3.7.

Table 3.7. Statistical significance results for eigenvalue difference analysis of PDZ.

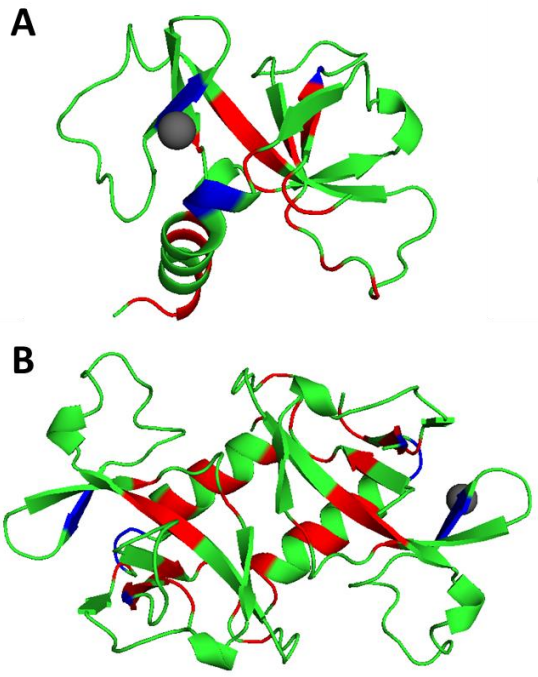
Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	13	5.05	1.70	1.43×10^{-6}
Mean Distance to Local Maximum Positions (Cα)	1.41	3.58	0.47	1.80×10^{-6}

3.2.2. CcdB

When considering the mutation sensitivity, active/inactive phenotype data is used. WT CcdB shows an active phenotype and kills cells when CcdB is inactive cells survive. So, survivability of cells observed for mutation sensitivity of residues on CcdB (Adkar et al.,2012). Both monomer and dimer structures of CcdB are used for analysis (PDB ID:

3VUB is used for CcdB Monomer structure. PDB ID:2VUB Chains A and B are used for CcdB Dimer structure).

Twenty residues with highest mutation sensitivity and residues on known binding sites with their positions on 3D structure are given in Figure 3.15.



C

Mutation Sensitive Residues	Binding Site Residues
2 3 5 6 18 20 22 24 34 35 36 63 68 83 90 93 94 98 99 100	**Metal Binding :** 7 54 55 85 86

Figure 3.15. Mutation sensitive residues and residues on binding site for CcdB. **A.** 3D structure of CcdB monomer (PDB ID: 3VUB). **B.** 3D structure of CcdB dimer (PDB ID: 2VUB). **(Red:** Mutation sensitive residues. **Blue:** Binding site residues) **C.** List of mutation sensitive residues and residues on binding site for CcdB

Fluctuation difference with respect to the unperturbed state is calculated for slowest three and ten GNM modes and for all GNM modes for CCDB case too. The results are given in Figures 3.16-3.18 for monomer and Figures 3.19-3.21 for dimer case with 10 Å cutoff value and "-1.2" as force constant in the connectivity matrix. Twenty residues of highest functional cost to mutations and known binding site residues are marked.

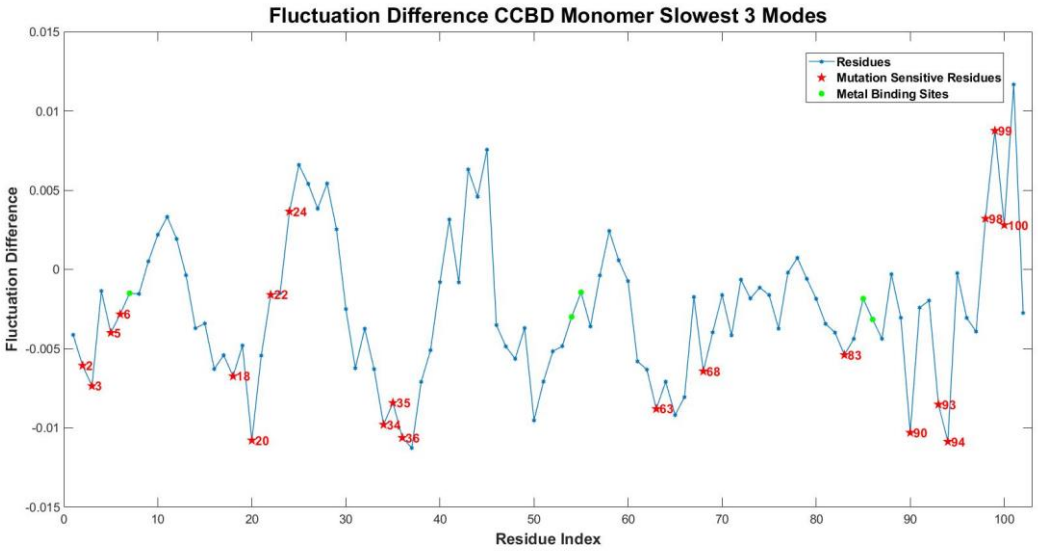


Figure 3.16. CcdB monomer perturbation analysis - Fluctuation difference vs residue index in three slowest GNM modes. (PDB ID: 3VUB)

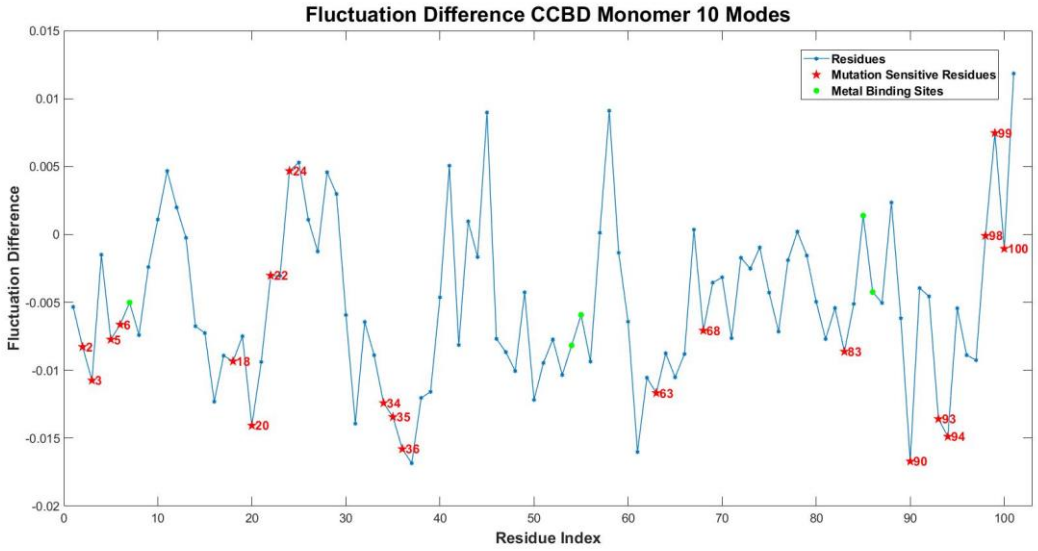


Figure 3.17. CcdB monomer perturbation analysis - Fluctuation difference vs residue index in ten slowest GNM modes. (PDB ID: 3VUB)

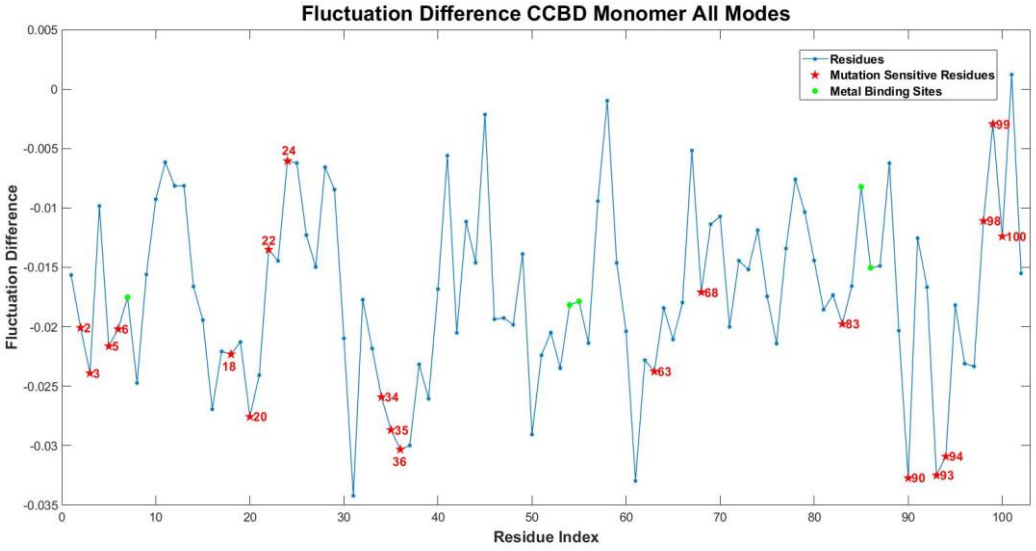


Figure 3.18. CcdB monomer perturbation analysis - fluctuation difference vs residue index in all GNM modes. (PDB ID: 3VUB)

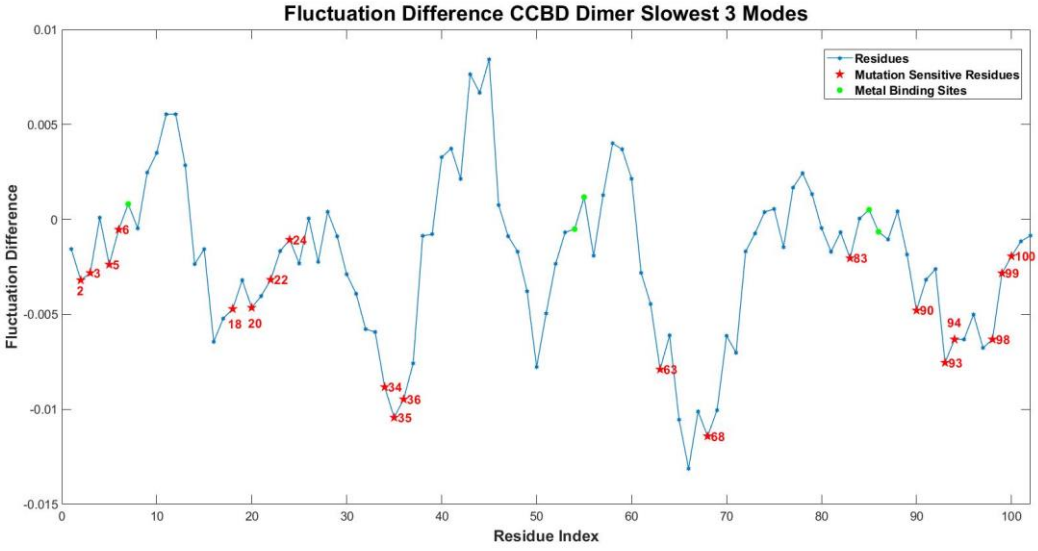


Figure 3.19. CcdB dimer perturbation analysis - Fluctuation difference vs residue index in three slowest GNM modes. (PDB ID: 2VUB)

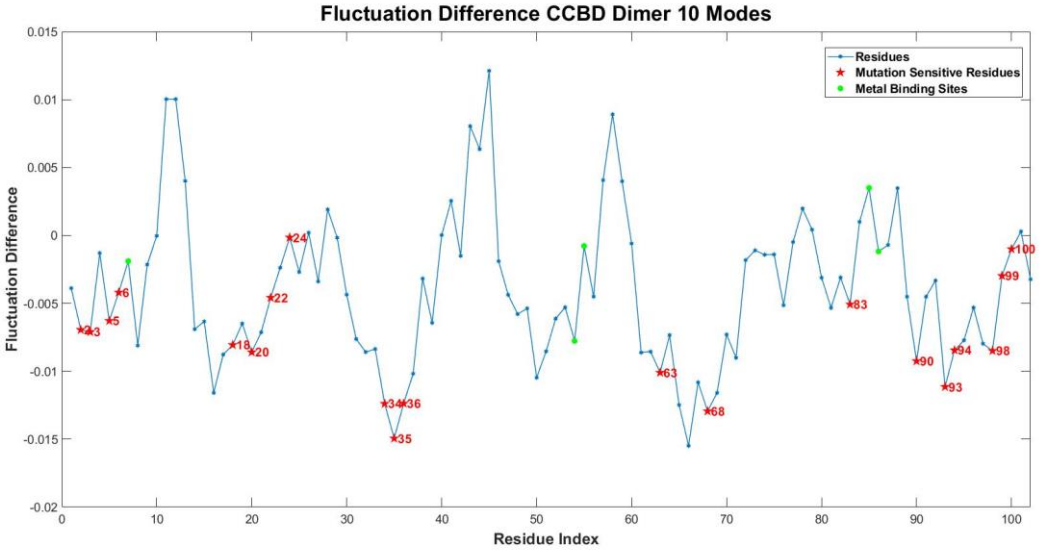


Figure 3.20. CcdB dimer perturbation analysis - Fluctuation difference vs residue index in ten slowest GNM modes. (PDB ID: 2VUB)

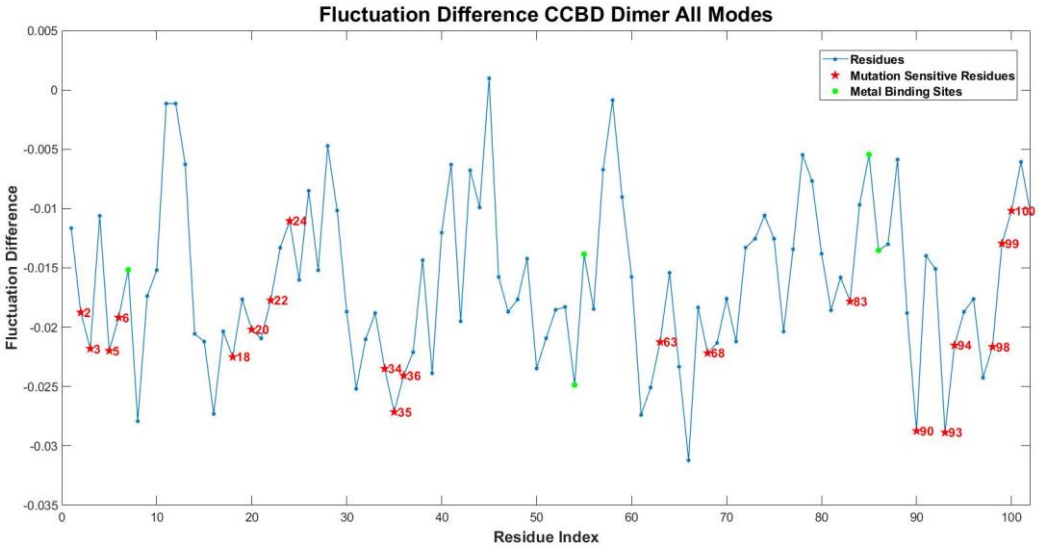


Figure 3.21. CcdB dimer perturbation analysis - Fluctuation difference vs residue index in all GNM modes. (PDB ID: 2VUB)

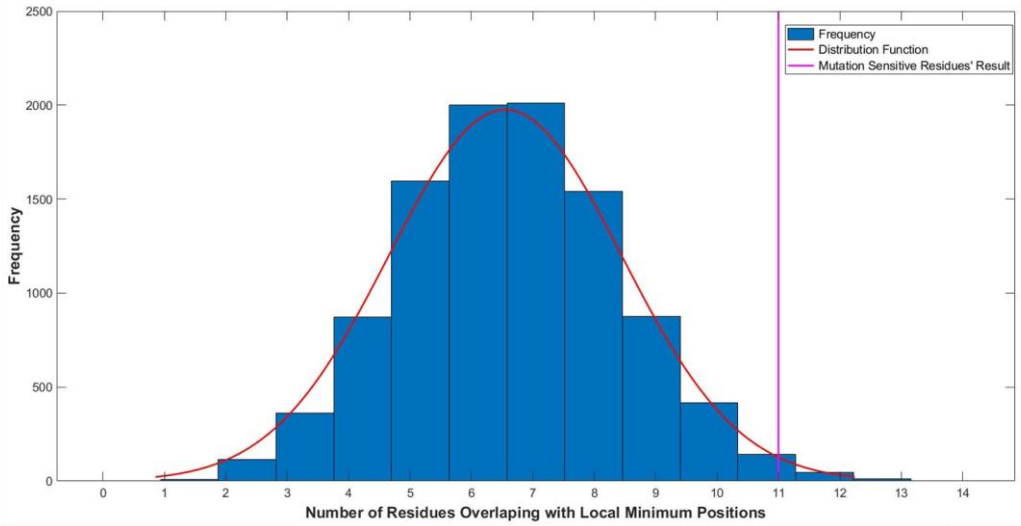
It is observed that mutation sensitive residues are located on or near the local minimum position. A statistical significance analysis is also carried out for CcdB.

Local minimum positions of fluctuation difference analysis are identified both for monomer and dimer structures of CcdB with constraint mentioned in the Materials &

Methods section. 33 residues are identified as local minimums for monomer case and 28 residues are identified as local minimums for dimer case. Random sampling procedure is applied to CcdB case and statistical significance analysis is made.

Histograms and distribution functions are given for monomer case and detailed results obtained from the statistical significance analysis for both cases are given in Figure 3.22 for the number of residues overlapping with the local minimum positions and Figure 3.23 for mean distance to local minimum positions. Statistical significance is made for only all modes analysis.

A

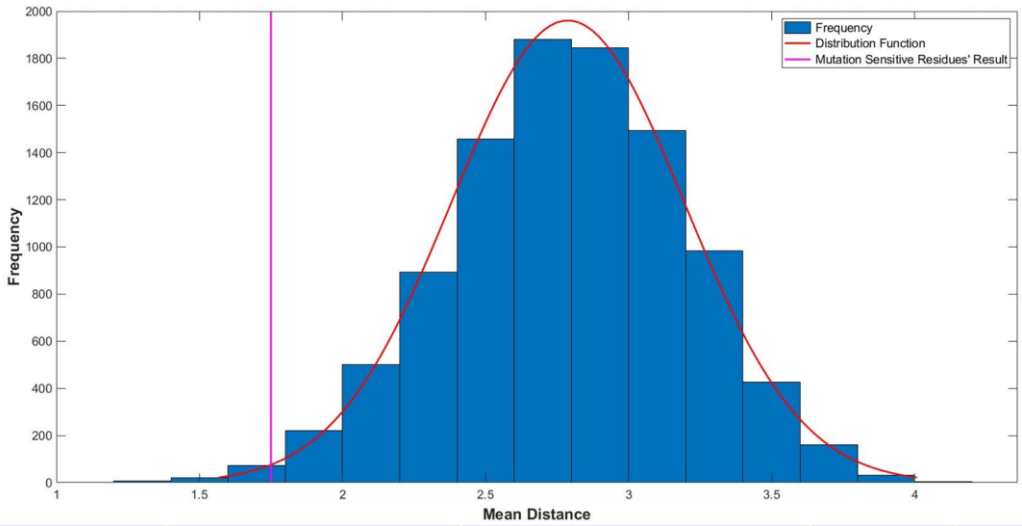


B

Structure	# of Mutation Sensitive Residues Located on Local minimums	Mean # of Residues Located on Local Minimums (random sampling)	Standard Deviation	P-value
CcbD (Monomer)	11	6.55	1.90	0.0095
CcbD (Dimer)	8	5.54	1.81	0.087

Figure 3.22. CcdB results - Histogram and distribution function for number of residues overlapping with local minimum positions.

A



B

Structure	Mean Distance of Mutation Sensitive Residues to Local minimums (C-alpha)	Mean Distance of Random Selected Residues to Local Minimums (random sampling)	Standard Deviation	P-value
CcbD (Monomer)	1.75 Å	2.78 Å	0.41 Å	0.0057
CcbD (Dimer)	2.62 Å	3.33 Å	0.47 Å	0.067

Figure 3.23. CcdB results - Histogram and distribution function for mean distance to local minimum positions.

For monomer case eleven out of twenty mutation sensitive residues are found out to be located on the local minimum positions of the fluctuation difference analysis. The p-value is calculated as 0.0095 for CcdB monomer case. Whereas for the dimer case, eight of the mutation sensitive residues are observed to be located on the local minimum positions of the fluctuation difference analysis. The p-value for the dimer CcdB case is calculated as 0.087.

The effect of perturbation on eigenvalues is analyzed for CcdB. Eigenvalue difference as a result of perturbation for each residue is given in Figure 3.24 for monomer and Figure 3.25 for dimer CcdB.

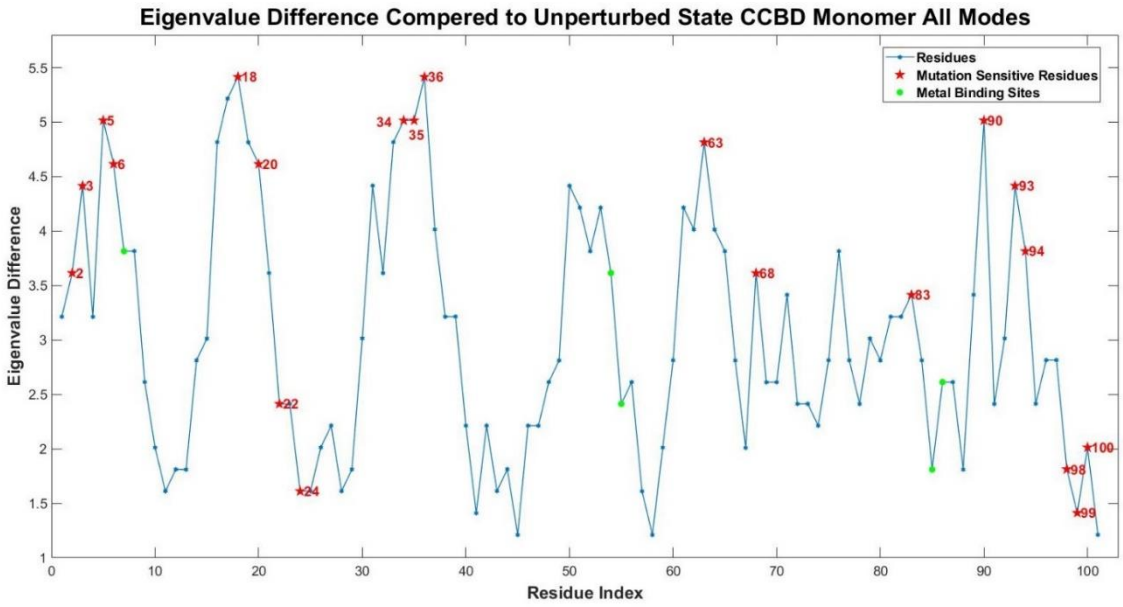


Figure 3.24. CcdB monomer perturbation analysis - Eigenvalue difference vs residue index in all GNM modes. (PDB ID: 3VUB)

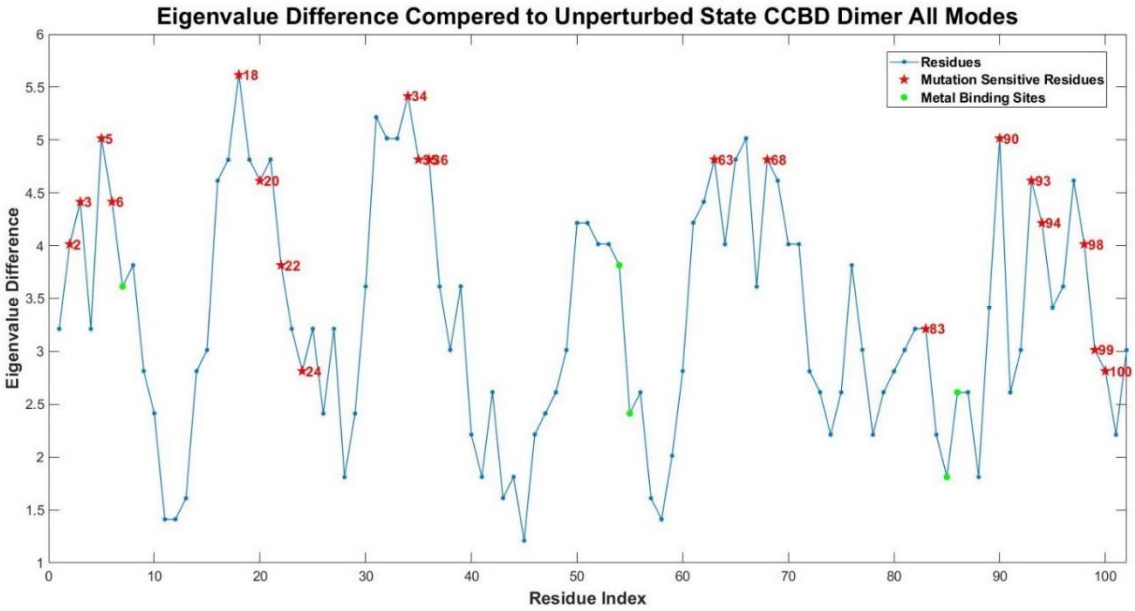


Figure 3.25. CcdB dimer perturbation analysis - Eigenvalue difference vs residue index in all GNM modes. (PDB ID: 2VUB)

It is observed that the mutation sensitive residues are located on the local maximum points on the eigenvalue difference profiles. These residues are the ones that have the capacity to change the eigenvalues of the GNM most in their local positions.

Statistical significance analysis is applied to eigenvalue differences. Maximum points are identified for each case. In monomer, twenty-eight residues are identified as a local maximum point. Ten out of twenty mutation sensitive residues are located on those maximum points. In dimer case, 26 residues are identified as local maximum point and 9 out of 20 mutation sensitive residues are appear at those local maximums. Just as in the fluctuation difference analysis, analysis is made for both number of residues overlapping with the local minimum positions and mean distance to local minimum positions (C-alpha atoms) for monomer and dimer CcdB. Detailed results for each analysis are given in Table 3.8 for monomer CcdB and Table 3.9 for dimer CcdB.

Table 3.8. Statistical significance results for eigenvalue difference analysis of CcdB monomer.

Analysis	**Sample Mean**	**Population Mean**	**Standard Deviation**	**P-Value**
# of overlapping Residues	10	5.57	1.79	0.007
Mean Distance to Local Maximum Positions (Cα)	2.11	3.11	0.41	0.008

Table 3.9. Statistical significance results for eigenvalue difference analysis of CcdB dimer.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	9	5.15	1.77	0.015
Mean Distance to Local Maximum Positions (Cα)	2.31	3.46	0.47	0.007

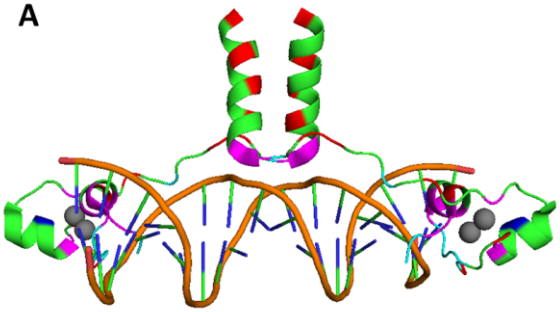
The p-value for the number of overlapping residues is calculated as 0.007 for monomer structure and 0.015 for dimer structure. Whereas for distance analysis, the p-value for C-alpha distance is calculated as 0.008 for monomer structure and 0.007 for dimer structure. The results indicate the correlation between residue's capacity to change eigenvalues and mutation sensitivity.

3.2.3. GAL4

In the deep sequencing study regarding GAL4 DNA binding domain, conditions are arranged in such a way that if mutant GAL4 can bind to DNA and activates HIS3 expression, cells survive. However, if GAL4 binding to DNA fails, cells cannot survive. In this way, the fitness scores of each multination on GAL4's residues are obtained (Kitzman et al., 2015). Mutation sensitivity of each residue is determined from the fitness scores gathered from deep sequencing study.

Analysis of GAL4 was carried out using A and B chains with DNA structure (PDB ID: 1D66). For analysis, P, C4' and C2 atoms of DNA is considered as C-alpha (backbone) atoms. Twenty residues with highest mutation sensitivity and residues on known binding sites are presented on the structure in Figure 3.26.

A



B

Mutation Sensitive Residues	Binding Site Residues
10 11 12 **14 15 17 18** 22 26 **28 38** 41 46 47 **50 51** 54 57 60 64	**DNA:** 8 9 **10** 15 **17 18** 19 20 21 23 43 49 **50 51** **Metal: 11 14 28** 31 **38**

Figure 3.26. Mutation sensitive residues and residues on binding site for GAL4. **A.** 3D structure of GAL4 (PDB ID: 1D66) (**Red:** Mutation sensitive residues. **Blue:** Binding site residues. **Magenta:** Mutation sensitive residues which are also located on binding site.) **B.** List of mutation sensitive residues and residues on binding site for CcdB. (**Bold:** Mutation sensitive residues which are also located on binding site.)

Fluctuation difference with respect to the unperturbed state is calculated for three and ten slowest modes and for all GNM modes of GAL4. The results for all mode analysis are given for with 10 Å cutoff value and "-1.2" as force constant in the connectivity matrix in Figure 3.27. The results for three and ten slowest mode analysis can be found in Figures B1 and B2, respectively. Twenty residues of the highest functional cost to the mutations and known binding site residues are marked.

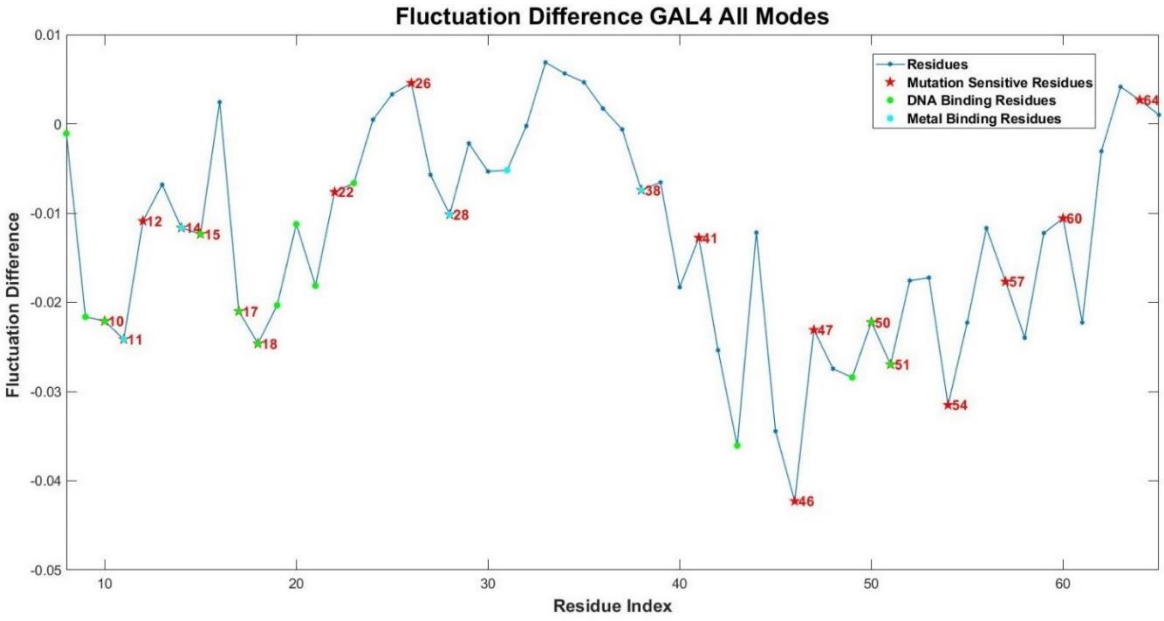


Figure 3.27. GAL4 perturbation analysis - Fluctuation difference vs residue index in all GNM modes. (PDB ID: 1D66)

It is observed that eight of the mutation sensitive residues are located at the local minimum positions of fluctuation difference analysis. Eigenvalue analysis is also performed here, and the results are given in Figure 3.28. Seven out of twenty mutation sensitive residues are located at the local maximum positions.

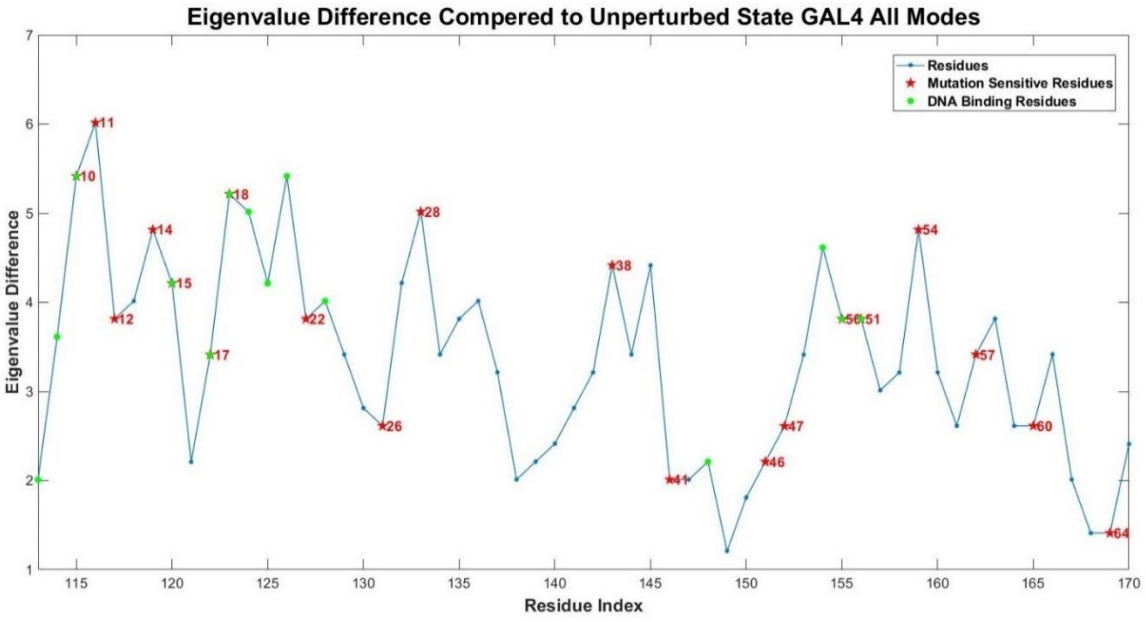


Figure 3.28. GAL4 perturbation analysis - Eigenvalue difference vs residue index in all GNM modes. (PDB ID: 1D66)

Statistical significance analysis is made regarding the number of overlapping residues and C-alpha distance to local minimum/maximum positions for both fluctuation difference and eigenvalue difference analysis with the same method mentioned in the Materials & Methods section. The results are listed in Table 3.10 and Table 3.11.

Table 3.10. Statistical significance results for fluctuation difference analysis of GAL4.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	8	5.27	1.63	0.047
Mean Distance to Local Minimum Positions (Cα)	2.53	3.40	0.425	0.020

Table 3.11. Statistical significance results for eigenvalue difference analysis of GAL4.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	7	5.26	1.61	0.139
Mean Distance to Local Maximum Positions (Cα)	3.06	3.38	0.413	0.219

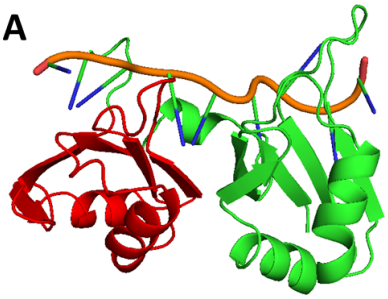
The p-value for the number of overlapping residues is calculated as 0.047, whereas for distance analysis, p-value concerning mean distances is calculated as 0.020 for fluctuation differences. For eigenvalue differences, the p-value for the number of overlapping residues to local maximum positions is calculated as 0.139 and the p-value for mean distance to local maximum positions is calculated as 0.219. It is observed that the p-value results of eigenvalue differences are lower than the p-value results of fluctuation differences. Fluctuation analysis shows the correlation between residues that have the capacity to change entropy of the structure and residues with high mutation sensitivity. On the other hand, the confidence interval for correlation between residue's capacity to change eigenvalues and mutation sensitivity is observed to be low for GAL4 than other structures studied in this thesis.

3.2.4. PAB1 RRM2 Domain

Deep mutational data is based on the binding affinity PAB-1 to RNA (Melamed et al., 2013).

RRM2 domain is obtained from 1CVJ chain A, residues between 99-173. RRM2 domain on PAB1, 20 residues with highest mutation sensitivity on RMM2 domain and residues on known binding sites are represented on 3D structure with the list of residues in Figure 3.29.

A



B

C

Mutation Sensitive Residues	Binding Site Residues
100 101 **102** 103 **104** 106 110 115 122 123 125 127 129 141 **142** 143 **144** 145 146 155	**RNA : 100 102 104** 105 127 138 139 140 **142 144** 169 172 173

Figure 3.29. Mutation sensitive residues and residues on binding site for PAB1. **A.** 3D structure of PAB1 with poly-A tail of RNA (PDB ID: 1CVJ). **Red:** RRM2 domain **B.** 3D structure of RRM2 domain **(Red:** Mutation sensitive residues. **Blue:** Binding site residues.) **C.** List of mutation sensitive residues and residues on binding site for PAB1 RRM2 domain. (**Bold:** Mutation sensitive residues which are also located on binding site.)

Fluctuation difference analysis and eigenvalue difference analysis is carried out for both PAB-1 with RNA and RRM2 domain only. Since the mutation sensitivity analysis is on RRM2 domain only, the results are given to cover only that domain. The results for fluctuation difference of all mode analysis is given in Figure 3.30, three and ten slowest mode analysis are respectively given in Figures B3 and B4. The results for the eigenvalue difference analysis is given in Figure 3.31.

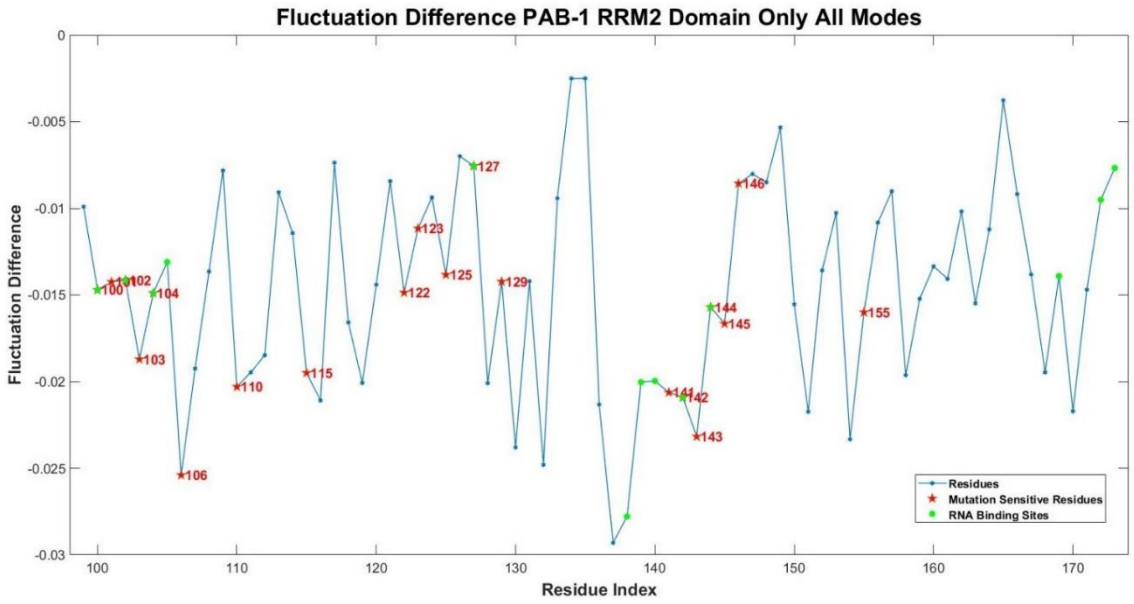


Figure 3.30. PAB1 monomer RRM2 domain perturbation analysis - Fluctuation difference vs residue index in all GNM modes. (PDB ID: 1CVJ)

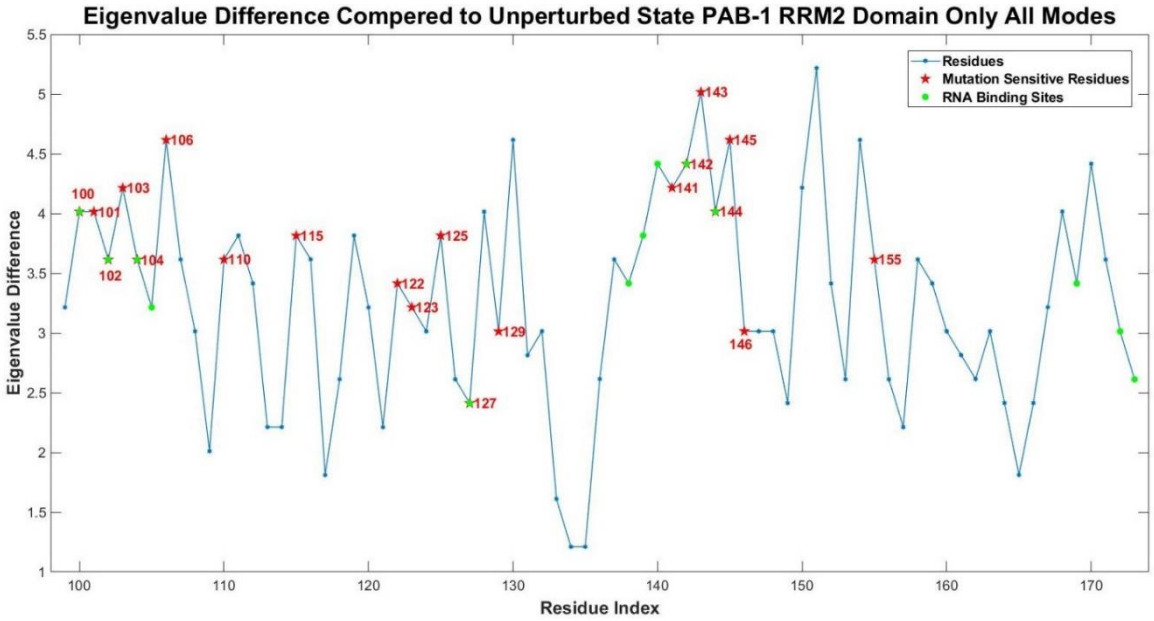


Figure 3.31. PAB1 monomer RRM2 domain perturbation analysis - Eigenvalue difference vs residue index in all GNM modes. (PDB ID: 1CVJ)

It is observed that for both fluctuation differences and eigenvalue differences, eight of the mutation sensitive residues are located at local minimum/maximum positions.

Statistical significance analysis is performed regarding the number of overlapping residues and C-alpha distance to local minimum/maximum positions for both fluctuation differences and eigenvalue differences as stated in Materials & Methods. Statistical significance analysis is made only to RRM2 domain (isolated) results. The results of statistical significance analysis are listed in Table 3.12 for fluctuation differences and Table 3.13 for eigenvalue differences.

Table 3.12. Statistical significance results for fluctuation difference analysis of PAB1 RRM2 domain.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	8	5.68	1.75	0.093
Mean Distance to Local Minimum Positions (Cα)	2.28	3.28	0.50	0.023

Table 3.13. Statistical significance results for eigenvalue difference analysis of PAB1 RRM2 domain.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	8	5.64	1.73	0.086
Mean Distance to Local Maximum Positions (Cα)	2.27	3.34	0.50	0.015

For PAB1 RRM2 domain, the p-value for the number of overlapping residues is calculated as 0.093, whereas p-value considering the mean distance to local minimum positions is calculated as 0.023 for fluctuation differences. For eigenvalue differences, the p-value for the number of overlapping residues to local maximum positions is calculated as 0.086 and the p-value of the mean distance to local maximum positions is calculated as 0.015. It is observed that the p-value of distance analysis gives better results than overlapping residue analysis which may be explained by the fact that mutation sensitive residues of PAB1 RRM2 domain are consecutive residues.

3.2.5. Ubiquitin

Mutation sensitivity of each residue is obtained from the fitness of Ubiquitin mutants in the presence of dimethyl sulfoxide which was provided in deep sequencing data. For the ubiquitin analysis, PDB ID 1UBQ is used. Tail of the structure (residues 73, 74, 75, 76) is not included in the calculations. Mutation sensitive residues on 3D ubiquitin structure and the list of those residues are given in Figure 3.32.

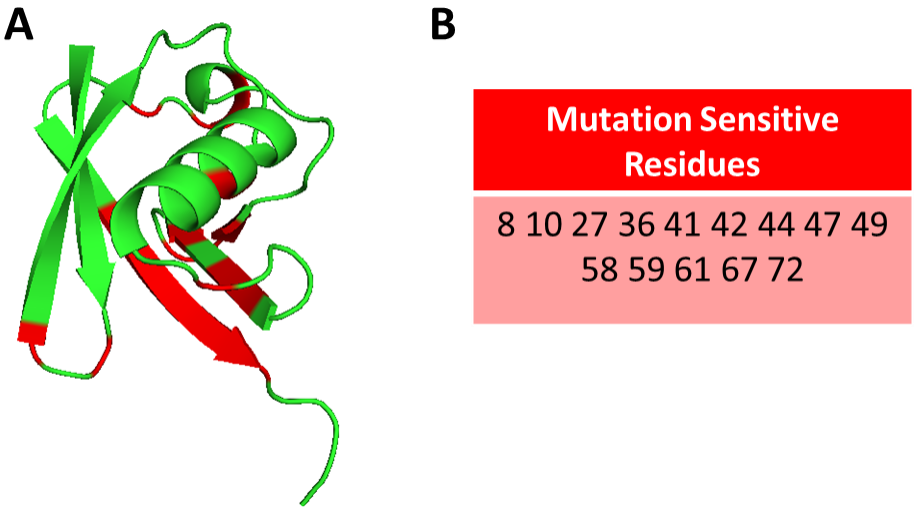


Figure 3.32. Mutation sensitive residues and residues on binding site for Ubiquitin. **A.** 3D structure of Ubiquitin. (PDB ID: 1UBQ) (**Red:** Mutation sensitive residues.) **B.** List of mutation sensitive residues for Ubiquitin.

Fluctuation difference and eigenvalue difference analysis are performed for Ubiquitin. The results for fluctuation difference analysis based on all modes of motion are given in Figure 3.33 and the results for eigenvalue difference analysis are given in Figure 3.34. Three and ten slowest mode results for fluctuation difference can be observed in Figures B5 and B6, respectively.

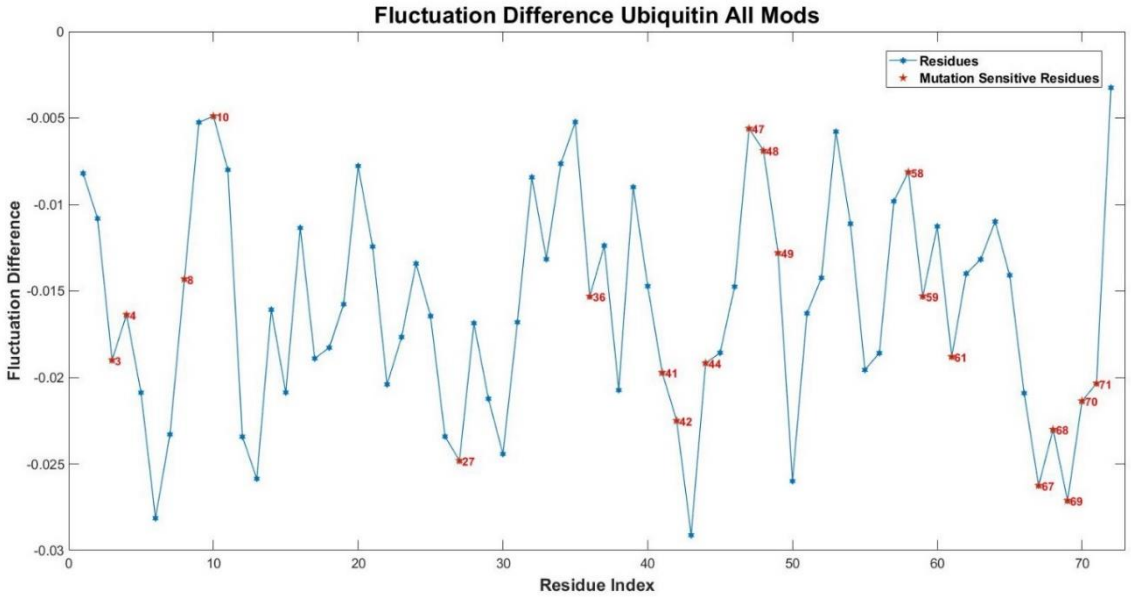


Figure 3.33. Ubiquitin perturbation analysis - Fluctuation difference vs residue index in all GNM modes. (PDB ID: 1UBQ)

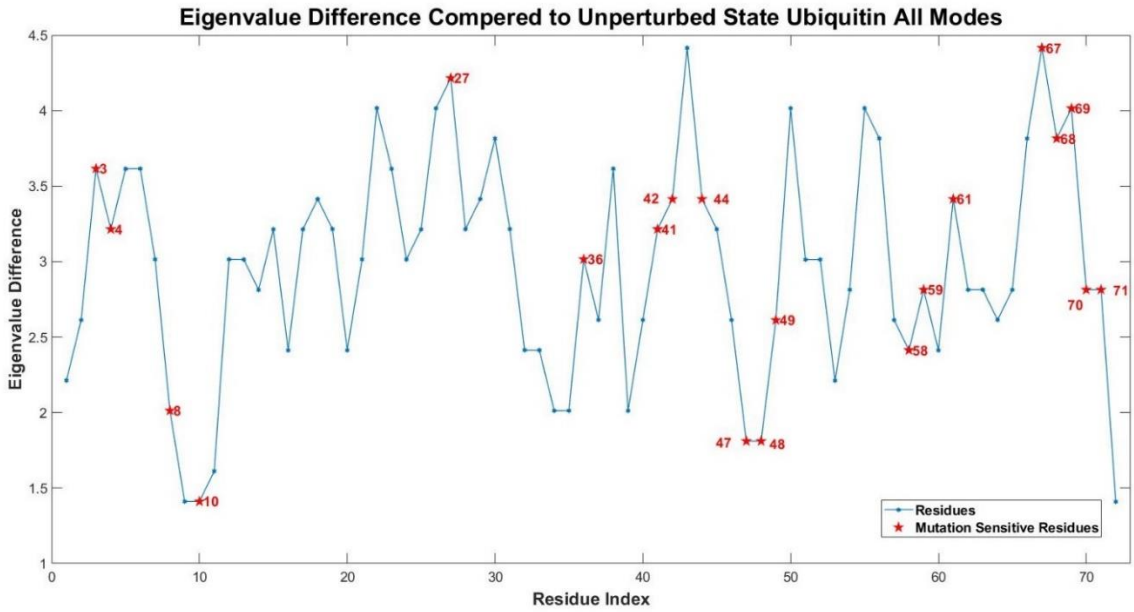


Figure 3.34. Ubiquitin perturbation analysis - Eigenvalue difference vs residue index in all GNM modes. (PDB ID: 1UBQ)

It is observed that seven of the mutation sensitive residues out of twenty residues are located at local minimum positions in the fluctuation difference profile. For eigenvalue

difference analysis, eight of the mutation sensitive residues are observed to be located at local maximum positions. Statistical significance analysis is made accordingly. The results of the fluctuation difference analysis based on all modes of motion are given in Table 3.14 and of the eigenvalue difference analysis in Table 3.15.

Table 3.14. Statistical significance results for fluctuation difference analysis of Ubiquitin in all modes of motion.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	7	4.72	1.64	0.082
Mean Distance to Local Minimum Positions (Cα)	3.38	4.20	0.65	0.104

Table 3.15. Statistical significance results for eigenvalue difference analysis of Ubiquitin.

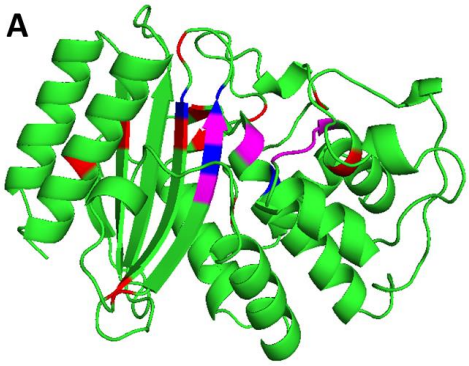
Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	8	4.99	1.69	0.038
Mean Distance to Local Maximum Positions (Cα)	3.02	3.78	0.54	0.079

In the ubiquitin case, the p-value of the number of overlapping residues is calculated as 0.082, whereas for distance analysis, the p-value concerning mean distances is calculated as 0.104 for fluctuation differences. For eigenvalue differences, the p-value of the number of overlapping residues to local maximum positions is calculated as 0.038 and the p-value of the mean distance to local maximum positions is calculated as 0.079. It is observed that p-value results of eigenvalue differences are better than p-value results of fluctuation differences. The results support the correlation between mutation sensitive residues and residues with high capacity to alter the structure’s entropy upon perturbation.

3.2.6. TEM1 β-Lactamase

Twenty mutation sensitive residues are determined from fitness landscape of β-Lactamase in the deep sequencing data (Frinberg *et al.,* 2014). These mutation sensitive residues and binding residues of β-lactamase are represented on 3D structure with list of residues in Figure 3.35.

A



B

Mutation Sensitive Residues	Binding Site Residues
45 66 **70** 73 **130 131 132** 134 157 166 179 181 229 **234 236** 242 245 251 282	**Ligand: 70** 73 **130 131 132 234** 235 **236** 237 244

Figure 3.35. Mutation sensitive residues and residues on binding site for TEM1 β-Lactamase. **A.** 3D structure of TEM1 β-Lactamase. (PDB ID: 1XPB) (**Red:** mutation sensitive residues. **Blue:** Binding site residues. **Magenta:** Mutation sensitive residues which are also located on binding site.) **B.** List of mutation sensitive residues and residues on binding site for TEM1 β-Lactamase. (**Bold:** Mutation sensitive residues which are also located on binding site.)

Fluctuation difference and eigenvalue difference analysis are made for β-Lactamase The results of fluctuation difference analysis based on all modes of motion are given in Figure 3.36 and the results of eigenvalue difference analysis are given in Figure 3.37. Three and ten slowest mode results of fluctuation difference can be observed in Figures B7 and B8, respectively.

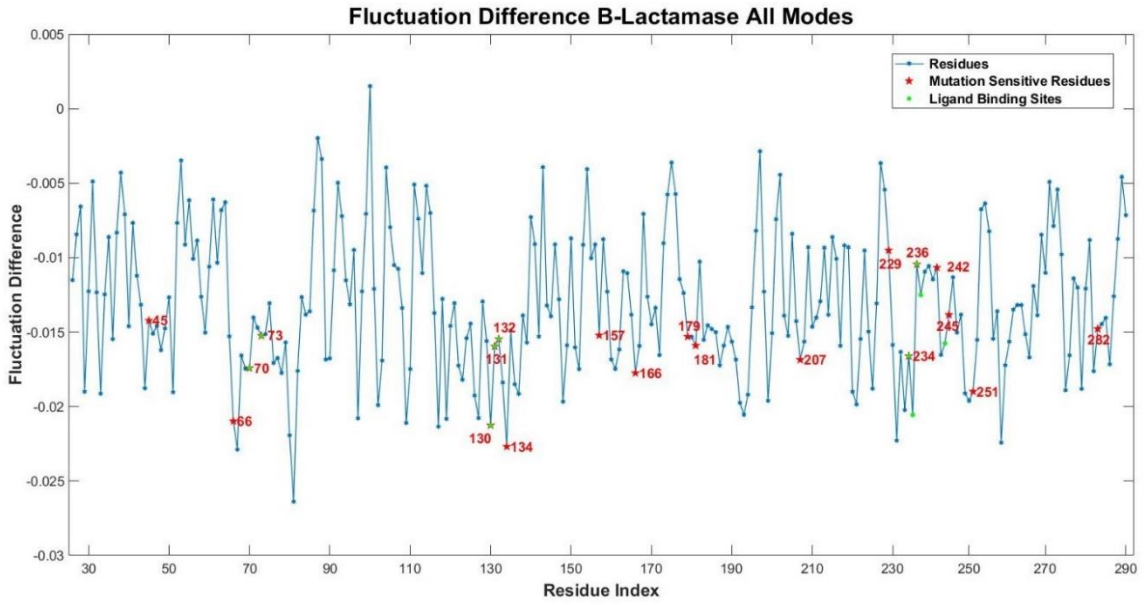


Figure 3.36. β-Lactamase perturbation analysis - Fluctuation difference vs residue index in all GNM modes. (PDB ID: 1XPB)

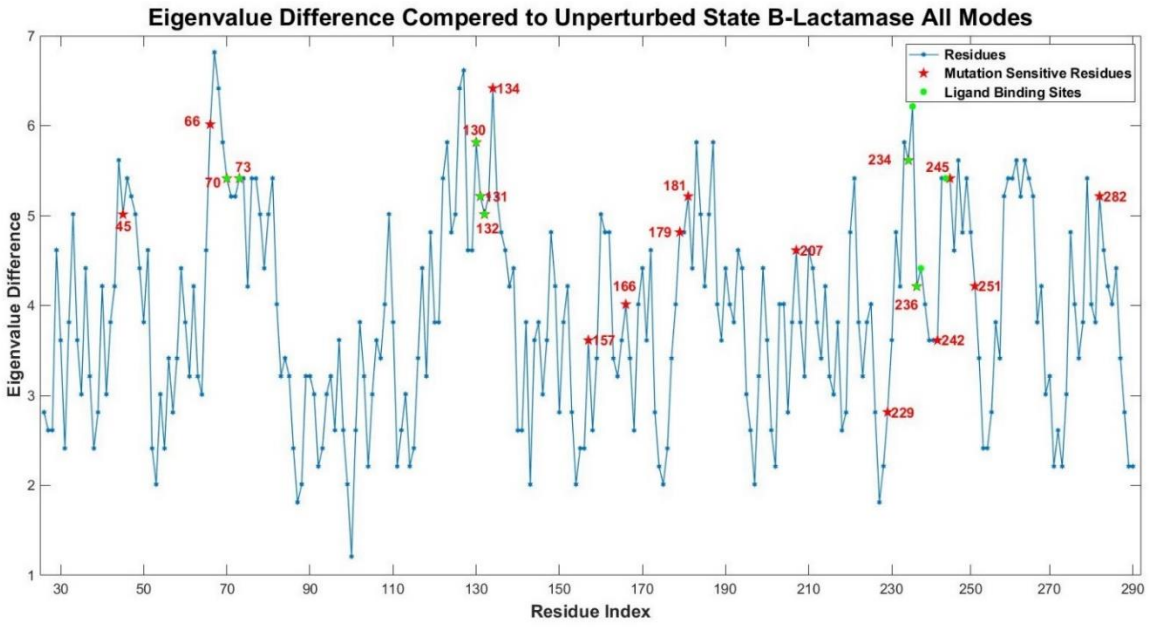


Figure 3.37. β-Lactamase perturbation analysis - Eigenvalue difference vs residue index in all GNM modes. (PDB ID: 1XPB)

Eight of the mutation sensitive residues are observed to be located at the local minimum positions for fluctuation difference analysis and nine of the mutation sensitive

residues are located at the local maximum positions for eigenvalue difference analysis. Statistical significance analysis is made based on these results. The results for statistical significance analysis are listed in Table 3.16 for fluctuation difference analysis and Table 3.17 for eigenvalue difference analysis.

Table 3.16. Statistical significance results for fluctuation difference analysis of β-Lactamase in all modes of motion.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	8	5.67	1.92	0.113
Mean Distance to Local Minimum Positions (Cα)	2.67	3.20	0.48	0.134

Table 3.17. Statistical significance results for eigenvalue difference analysis of β-Lactamase.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	9	5.46	1.92	0.033
Mean Distance to Local Maximum Positions (Cα)	2.44	3.20	0.46	0.049

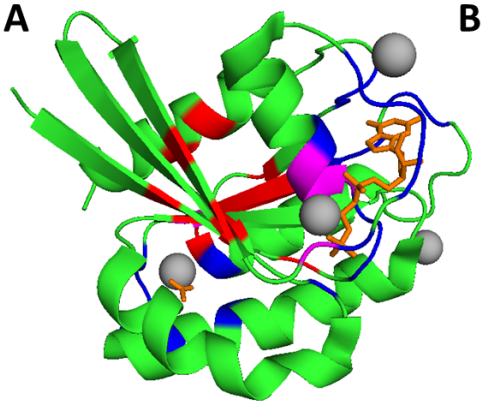
In the β-Lactamase case, for fluctuation differences, the p-value for the number of overlapping residues is calculated as 0.113 and the p-value concerning mean distances is calculated as 0.134. In the case of eigenvalue differences, the p-value of the number of overlapping residues to local maximum positions is calculated as 0.033 and the p-value of the mean distance to local maximum positions is calculated as 0.049. The results indicate that for eigenvalue differences, there is a correlation between mutation sensitive residues and residues with high capacity to alter the eigenvalues. On the other hand, the confidence interval for the correlation between residue's capacity to change entropy of the structure and

mutation sensitivity is observed to be low for β-Lactamase than other structures studied in this thesis.

3.2.7. H-Ras GTPase

In the experiment regarding H-Ras GTPase, H-Ras GTPase is coupled to the transcription of an antibiotic resistance factor. Residue mutation sensitivity of H-Ras GTPase evaluated through their effect on bacterial growth in the presence of an antibiotic (Bandaru *et al.*, 2017). For H-Ras GTPase analysis, PDB ID 3K8Y is used as crystal structure. Twenty residues with highest mutation sensitivity and residues on known binding sites on 3D structure are represented in Figure 3.38.

A



B

Mutation Sensitive Residues	Binding Site Residues
6 10 **15 16 17 35** 38 40 57 78 **111** 112 113 114 115 119 138 141 156 159	12 13 14 **15 16 17** 18 28 29 30 31 32 34 **35** 60 97 101 106 107 108 109 **111** 116 117 119 120 137 145 146 147

Figure 3.38. Mutation sensitive residues and residues on binding site for H-Ras GTPase. **A.** 3D structure of H-Ras GTPase. (PDB ID: 3K8Y) (**Red:** Mutation sensitive residues. **Blue:** Binding site residues. **Magenta:** mutation sensitive residues which are also located on binding site.) **B.** List of mutation sensitive residues and residues on binding site for H-Ras GTPase (**Bold:** Mutation sensitive residues which are also located on binding site.)

Fluctuation difference and eigenvalue difference analysis are made for H-Ras GTPase. The results of fluctuation difference analysis based on all modes of motions are given in Figure 3.39 and the results of eigenvalue difference analysis are given in Figure 3.40. The results based on the slowest three and ten modes of motion for fluctuation difference can be observed in Figures B9 and B10, respectively.

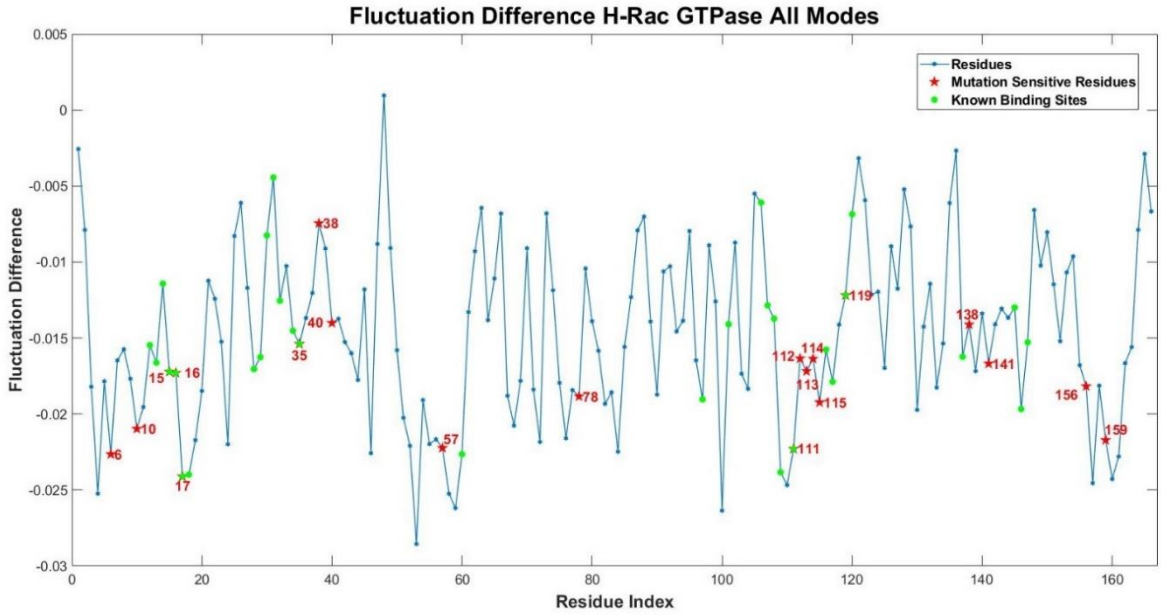


Figure 3.39. H-Ras GTPase perturbation analysis - Fluctuation difference vs residue index in all GNM modes. (PDB ID: 3K8Y)

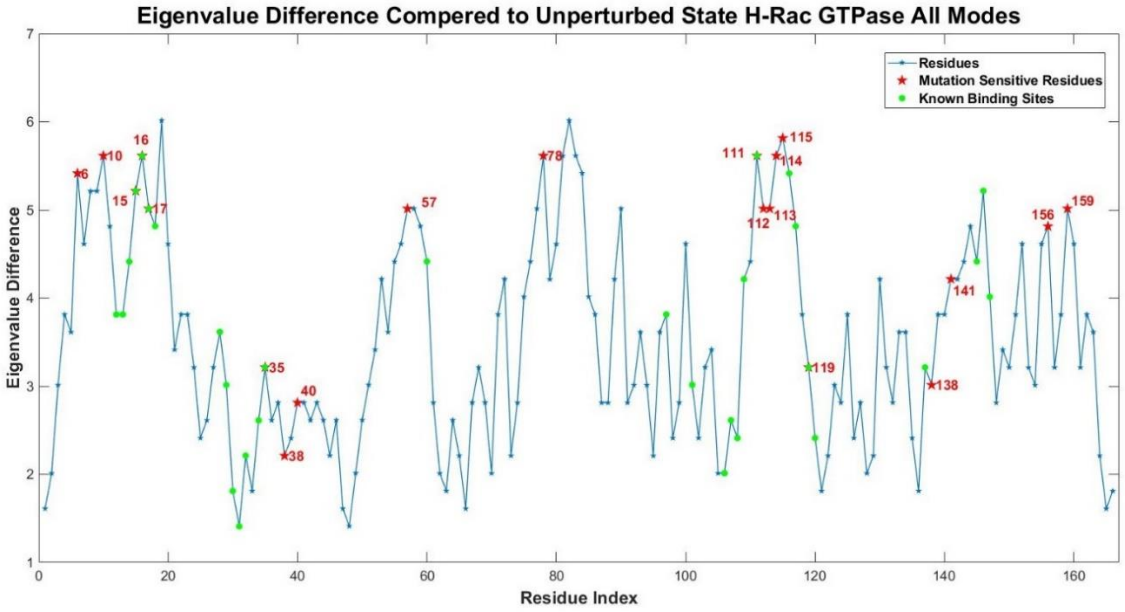


Figure 3.40. H-Ras GTPase perturbation analysis - Eigenvalue difference vs residue index in all GNM Modes. (PDB ID: 3K8Y)

It is observed that nine of the twenty mutation sensitive residues are located at the local minimum points on the fluctuation difference results. For eigenvalue difference analysis, ten

of the mutation sensitive residues are observed to overlap with the local maximum positions. Statistical significance analysis is made according to observations on the fluctuation difference and eigenvalue difference analysis. The results for the statistical significance analysis are given in Table 3.18 for fluctuation difference analysis and Table 3.19 for eigenvalue difference analysis.

Table 3.18. Statistical significance results for fluctuation difference analysis of H-Ras GTPase.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	9	5.52	1.88	0.032
Mean Distance to Local Minimum Positions (Cα)	2.53	3.23	0.46	0.064

Table 3.19. Statistical significance results for eigenvalue difference analysis of H-Ras GTPase.

Analysis	Sample Mean	Population Mean	Standard Deviation	P-Value
# of overlapping Residues	10	5.32	1.86	0.006
Mean Distance to Local Maximum Positions (Cα)	2.16	3.32	0.47	0.007

In the analysis of H-Ras GTPase, the p-value of the number of overlapping residues is calculated as 0.032 and the p-value concerning mean distances is calculated as 0.064 for fluctuation differences. In the case of eigenvalue differences, the p-value of the number of overlapping residues to local maximum positions is calculated as 0.006 and the p-value of the mean distance to local maximum positions is calculated as 0.007. The results of H-Ras GTPase indicate a significant correlation between the mutation sensitive residues and local minimum positions. Moreover, p-values concerning eigenvalue differences support that there is a strong correlation between the residue's capacity to change eigenvalue magnitude and distribution upon perturbation and residue's mutation sensitivity.

Summary tables are produced for the calculated p-values for each analysis in order to observe the p-value results for each structure together. P-value results for each structure regarding fluctuation difference analysis are given in Table 3.20, and eigenvalue difference in Table 3.21.

Table 3.20. Calculated p-values for each structure regarding fluctuation difference analysis.

Structure	**P-Value Overlapping Residue**	**P-Value Distance (Ca)**
PSD95-PDZ domain	1.91×10^{-6}	1.67×10^{-6}
CcdB monomer	0.0095	0.0057
CcdB dimer	0.087	0.067
GAL4	0.047	0.020
PAB1-RRM2	0.093	0.023
Ubiquitin	0.082	0.104
TEM1-β -lactamase	0.113	0.134
H-Ras GTPase	0.032	0.064

Table 3.21. Calculated p-values for each structure regarding eigenvalue difference analysis.

Structure	**P-Value Overlapping Residue**	**P-Value Distance (Ca)**
PSD95-PDZ domain	1.43×10^{-6}	1.80×10^{-6}
CcdB monomer	0.007	0.008
CcdB dimer	0.015	0.007
GAL4	0.139	0.219
PAB1-RRM2	0.086	0.015
Ubiquitin	0.038	0.079
TEM1-β -lactamase	0.033	0.049
H-Ras GTPase	0.006	0.007

3.3. Compensatory Mutations

GNM analysis is done for the core domain of tumor suppressor protein p53 chain A whose structure is obtained from PDB ID: 1TSR chain A. In order to investigate the compensatory mutation sites of tumor suppressor protein p53, a dataset is created for deleterious mutation sites and their related compensatory mutation sites from the literature (Brachmann et al., 1998; Inga and Resnick, 2001; Baroni et al., 2004; Danziger et al., 2007; Danziger et al., 2009). Deleterious mutation sites and their related compensatory mutation sites are given in Table 3.22.

Table 3.22. Deleterious mutations of p53 with known compensatory mutation sites.

Deleterious Mutation Site	Compensatory Mutation Site(s)
141	235 235+239
143	268
152	100 101 103 106
157	235 235+239
158	100 100+104 201 207 224 235 235+239
163	235+239 233+235+239
173	228+239+240 235+239
177	122
205	235 228+239 235+239
220	235 235+239 235+240
244	123
245	113 114+123+172+189 231 123 123+189 230+239 231 235+239
249	118+168 122+124+168 123+168 139+168+239 168+231 277+235+239 235+239
252	122
272	235+239
273	123+240 178+240 183+240 224+240 228+229+235+239+240 233+240 235+239 240
286	235 235+239

Underlying dynamic properties of compensatory mutations and the allosteric mechanism is tried to be revealed by GNM analysis. Deleterious and compensatory mutation sites could be related to the same global mode or may involve in different global modes. Each deleterious mutation in the dataset is investigated in the scope of the slowest five modes of GNM to observe whether they are related to hinge residues of any mode or not. If a deleterious mutation site is found out to be on the hinge axis of these modes of motion, its compensatory mutation sites are investigated in the same scope. Most of the deleterious

mutation sites are observed to be on the same hinge axis of the same GNM slow mode with their corresponding compensatory mutation sites. Deleterious mutations (magenta) and their related compensatory mutations (cyan) are represented in related GNM mode results with hinge sites (orange) in Figures 3.41 – 3.46, following their order in Table 3.21.

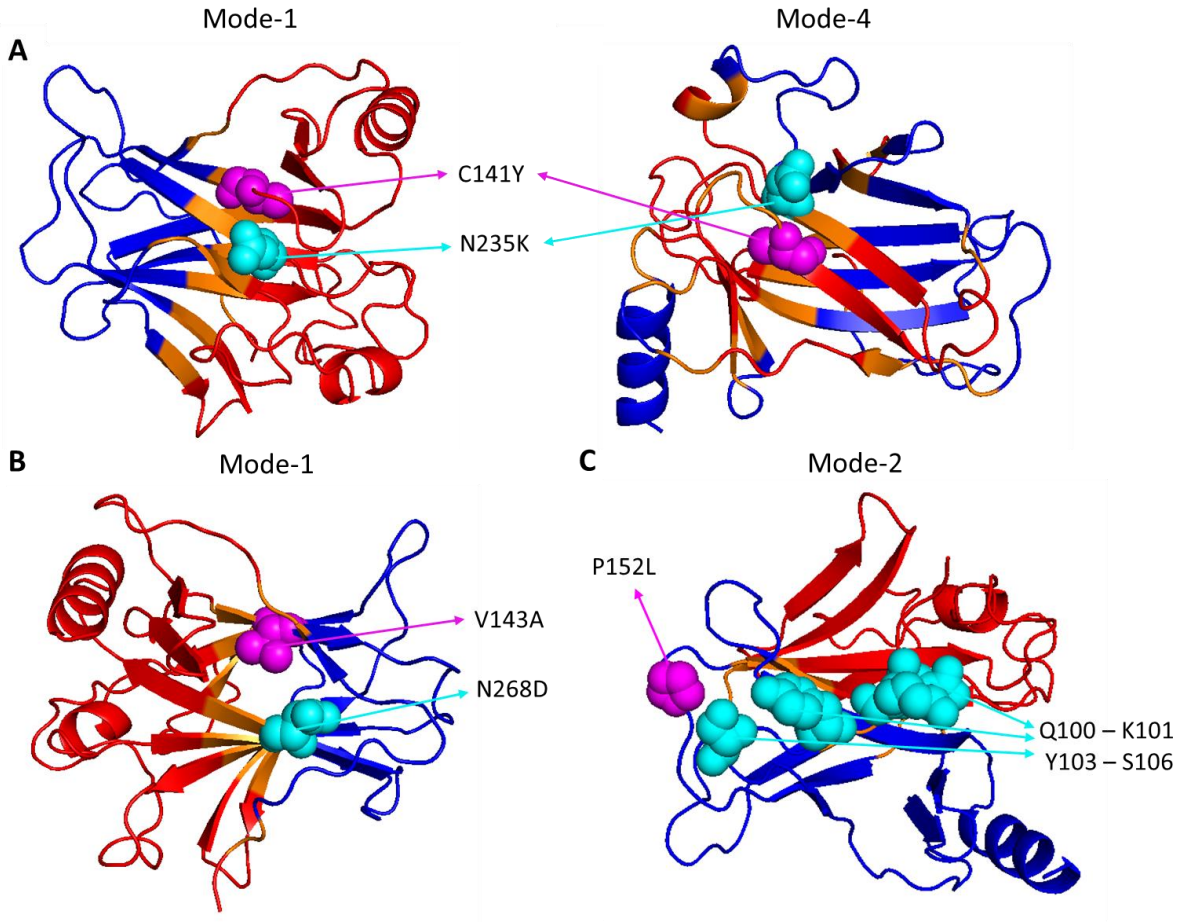


Figure 3.41. Deleterious mutation sites 141, 143 and 152 with their related compensatory mutations. **A.** Deleterious mutation site 141 with its compensatory mutation 235 represented in the slowest and fourth slowest GNM mode. **B.** Deleterious mutation site 143 with its compensatory mutation 268 represented in the slowest GNM mode. **C.** Deleterious mutation site 152 with its compensatory mutations 100, 101, 103 and 106 represented in the second slowest GNM mode.

In Figure 3.41A, it is observed that deleterious mutation site 141 and its compensatory mutation site 235 are hinge residues of the slowest and forth slowest mode of GNM. It is suggested that the mutation on residue 268 compensates the mutation on residue 143. It can

be observed in Figure 3.41B that both residues are hinge residues of the slowest GNM mode. Additionally, residue 152 suggested to have compensatory mutations on residues 100, 101, 103 and 106. It is observed in Figure 3.41C that those compensatory mutation sites and deleterious mutation 152 lie on the hinge axis of the second slowest mode of GNM.

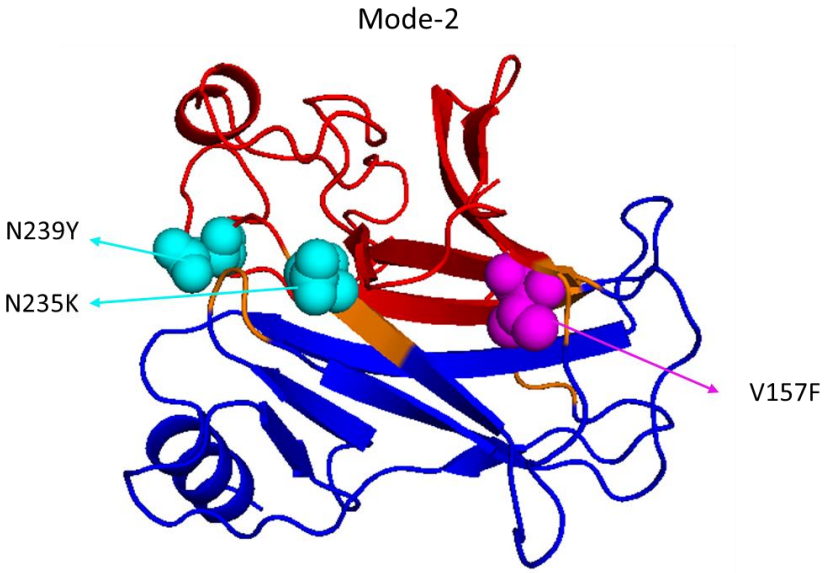


Figure 3.42. Deleterious mutation site 157 with its related compensatory mutations 235 and 239 represented in second slowest GNM mode.

Similarly, it is observed in Figure 3.42 that deleterious mutation site 157 and its compensatory mutation sites 235 and 239 align at the hinge axis of the second slowest mode of GNM.

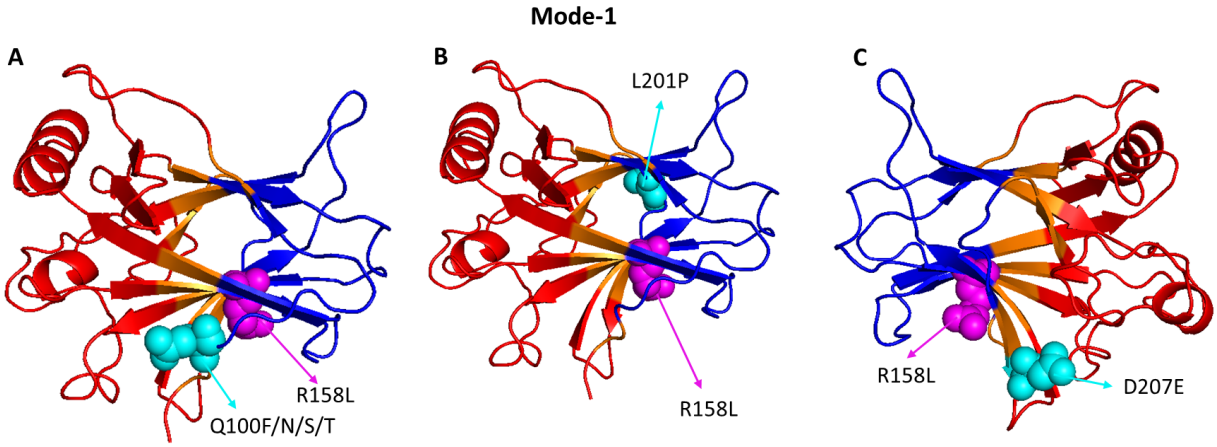


Figure 3.43. Deleterious mutation site 158 with its related compensatory mutations 100, 201 and 207. **A.** Residue 158 with its compensatory mutation site 100 in slowest GNM mode. **B.** Residue 158 with its compensatory mutation site 201 in slowest GNM mode. **C.** Residue 158 with its compensatory mutation site 207 in slowest GNM mode.

Deleterious mutation site 158 has many compensatory mutation sites. In Figure 3.43, its compensatory mutations 100, 201 and 207 are represented with residue 158 in the slowest GNM mode. It can be stated that they all align on the hinge axis of the slowest GNM mode.

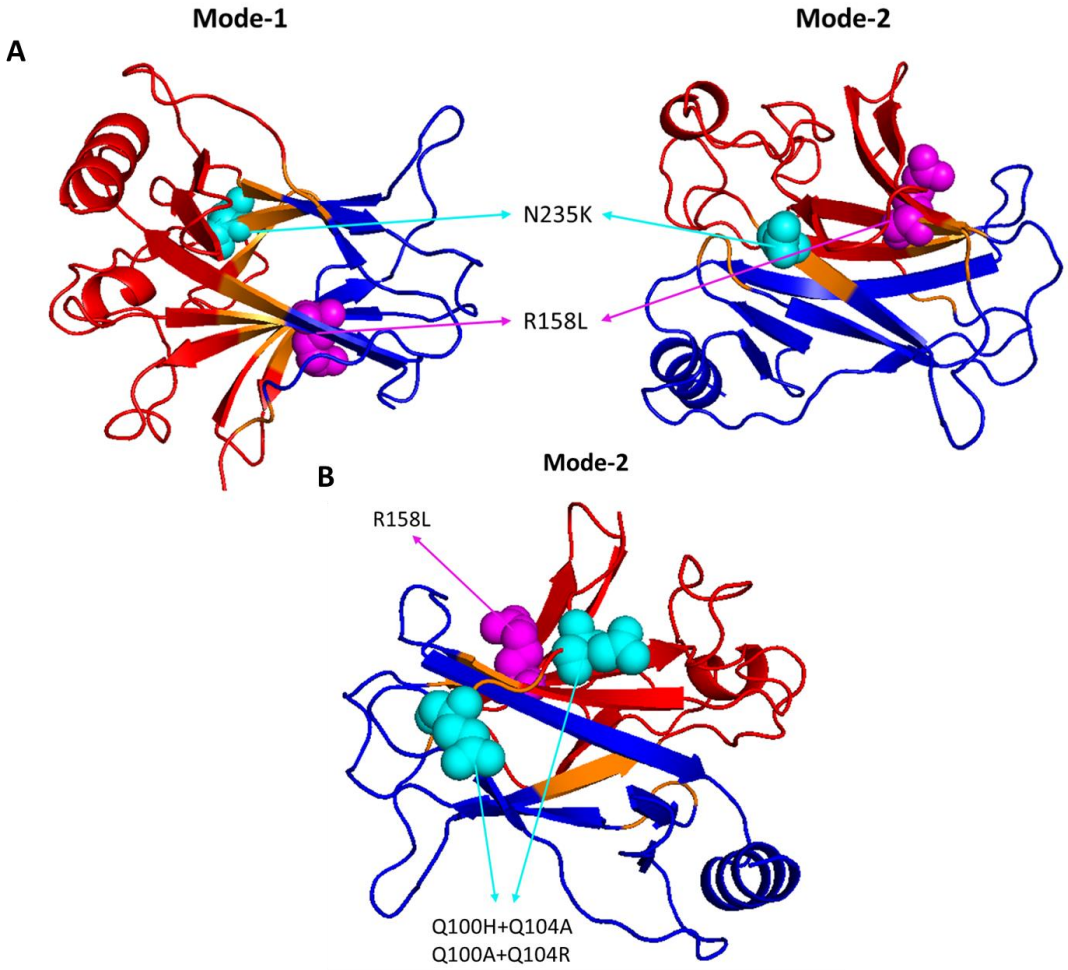


Figure 3.44. Deleterious mutation site 158 with its related compensatory mutations 235, 100 and 104. **A.** Residue 158 with its compensatory mutation site 235 in the slowest and second slowest GNM mode. **B.** Residue 158 with its compensatory mutations 100 and 104 together in second slowest GNM mode.

For the compensatory mutation site 235, it is possible to see that they align both in the slowest and second slowest GNM mode hinge axis as presented in Figure 3.44A. Additionally, residues 100 and 104 are suggested to compensate the deleterious mutation 158 together. It is seen that deleterious mutation 158 and its compensatory mutation sites 100 and 104 align at the hinge axis of the second slowest GNM mode as represented in Figure 3.44B.

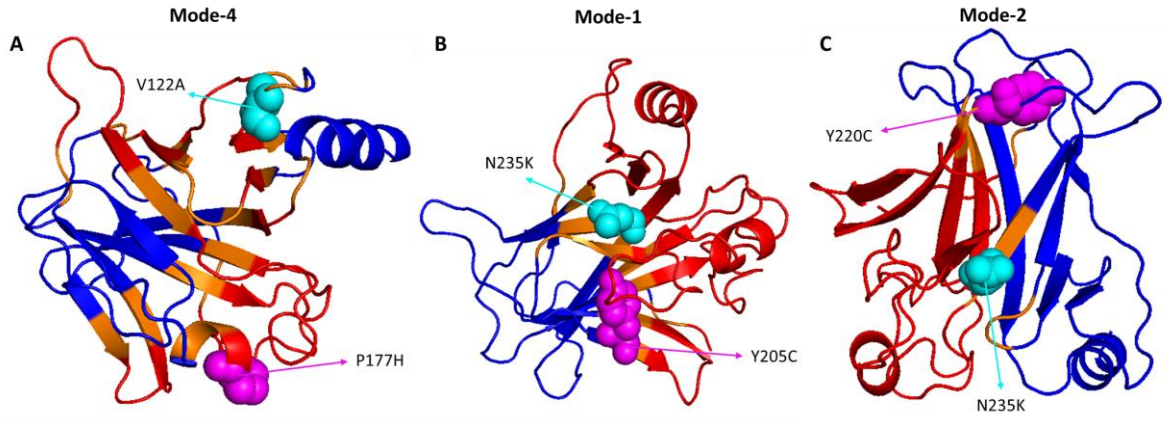


Figure 3.45. Deleterious mutation sites 177, 205 and 220 with their related compensatory mutations. **A.** Deleterious mutation site 177 with its compensatory mutation 122 represented in fourth slowest GNM mode. **B.** Deleterious mutation site 205 with its compensatory mutation 235 represented in GNM slow mode 1. **C.** Deleterious mutation site 220 with its compensatory mutation 235 represented in second slowest GNM mode.

Residue 122 is suggested to compensate the mutation at residue 177. They are observed to be on the hinge axis of the fourth slowest GNM mode (Figure 3.45A). It is suggested that deleterious mutation at residue 205 can be rescued by mutation at residue 235. These residues are found out to be the either hinge residues or hinge neighbor residues for the slowest GNM mode (Figure 3.45B). Additionally, it is noted that deleterious mutation site 272 and its compensatory mutation sites 235 are on the hinge axis of the fourth slowest mode of GNM (Figure 3.45C).

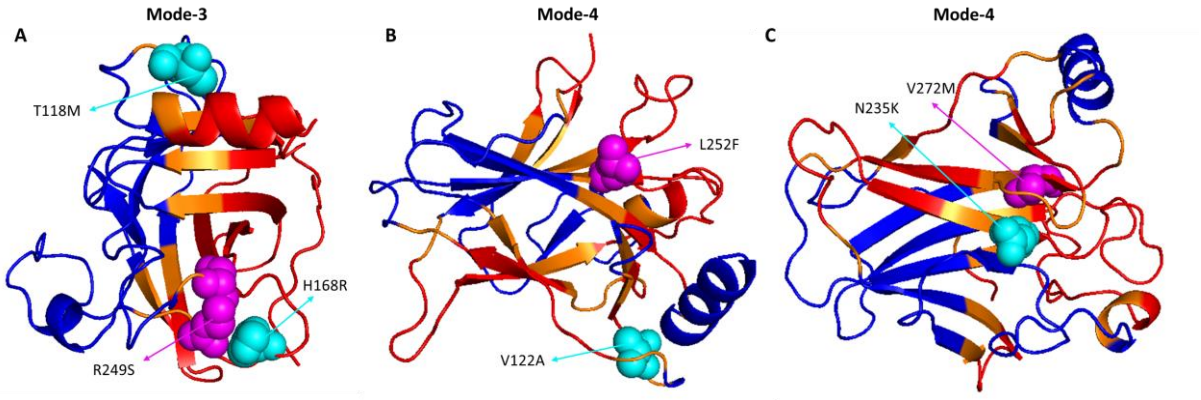


Figure 3.46. Deleterious mutation sites 249, 252 and 272 with their related compensatory mutations. **A.** Deleterious mutation site 249 with its compensatory mutations 116 and 168 represented in the third slowest GNM mode. **B.** Deleterious mutation site 252 with its compensatory mutation 122 represented in the fourth slowest GNM mode. **C.** Deleterious mutation site 272 with its compensatory mutation 235 represented in the fourth slowest GNM mode.

Residues 118 and 168 are suggested to compensate the deleterious mutation at residue 249 together. Figure 3.46A. represents the second slowest mode and mentioned residues. It is seen that these residues align at the second slowest GNM mode hinge axis. Residue 122 is suggested to compensate the deleterious effect of the mutation on residue 252. Figure 3.46B represents these residues and the fourth slowest GNM mode. It is observed that these residues are at the hinges of the fourth slowest mode of GNM. In Figure 3.46C, it is seen that deleterious mutation site 272 and its compensatory mutation sites 235 are at the hinge axis of the fourth slowest mode of GNM.

The compensatory mutation dataset consists of seventeen different deleterious mutation sites. Eleven of these deleterious mutation sites are observed to be on the same GNM mode hinge axis with their related compensatory mutations. Although residues 173 and 273 are hinge residues of the third slowest GNM mode, their suggested compensatory mutations are not on the same hinge axis. On the other hand, the suggested compensatory mutations of residues 173 and 273 are on the hinge axis of the different GNM modes which suggests that dynamic properties that are perturbed by those deleterious mutations can be

rescued by altering the other modes of motions. Residue 163 is a hinge residue for p53 dimer structure, and its compensatory mutations are also hinge residues of other GNM modes just like residues 173 and 273. Residues 244, 245, 268 are not related to the slowest five modes of GNM. Residue 235 which is suggested to be a compensatory mutation for 12 of the 17 deleterious mutations in the dataset, is a hinge residue for multiple GNM modes. Residue 235's contribution to multiple dynamic motions of the p53 might make it a perfect spot for compensating the deleterious effects of mutations. Furthermore, the results impose a link between deleterious mutations and their related compensatory mutations in the spectrum of slow modes of motion which suggests the allosteric interaction between deleterious mutations and their compensatory mutations.

3.3.1. Cross Correlation Comperassions

The crystal structures of wild type p53 (PDB ID: 1TSR), p53 with deleterious mutant V157F (PDB ID: 4KVP), and p53 with compensatory mutations V157F/N235K/N239Y (PDB ID: 4LOF) are available structures on Protein Data Bank for compensatory mutation studies. These crystal structures are used in GNM analysis and 2D cross correlation maps are obtained for all three structures. Analysis is made for slowest three modes, slowest ten modes and all modes. In the slowest ten modes analysis, some correlation loss between specific domains of p53 is observed on the deleterious mutant structure. The results for 2D cross correlation maps for slowest ten modes are given in Figure 3.47.

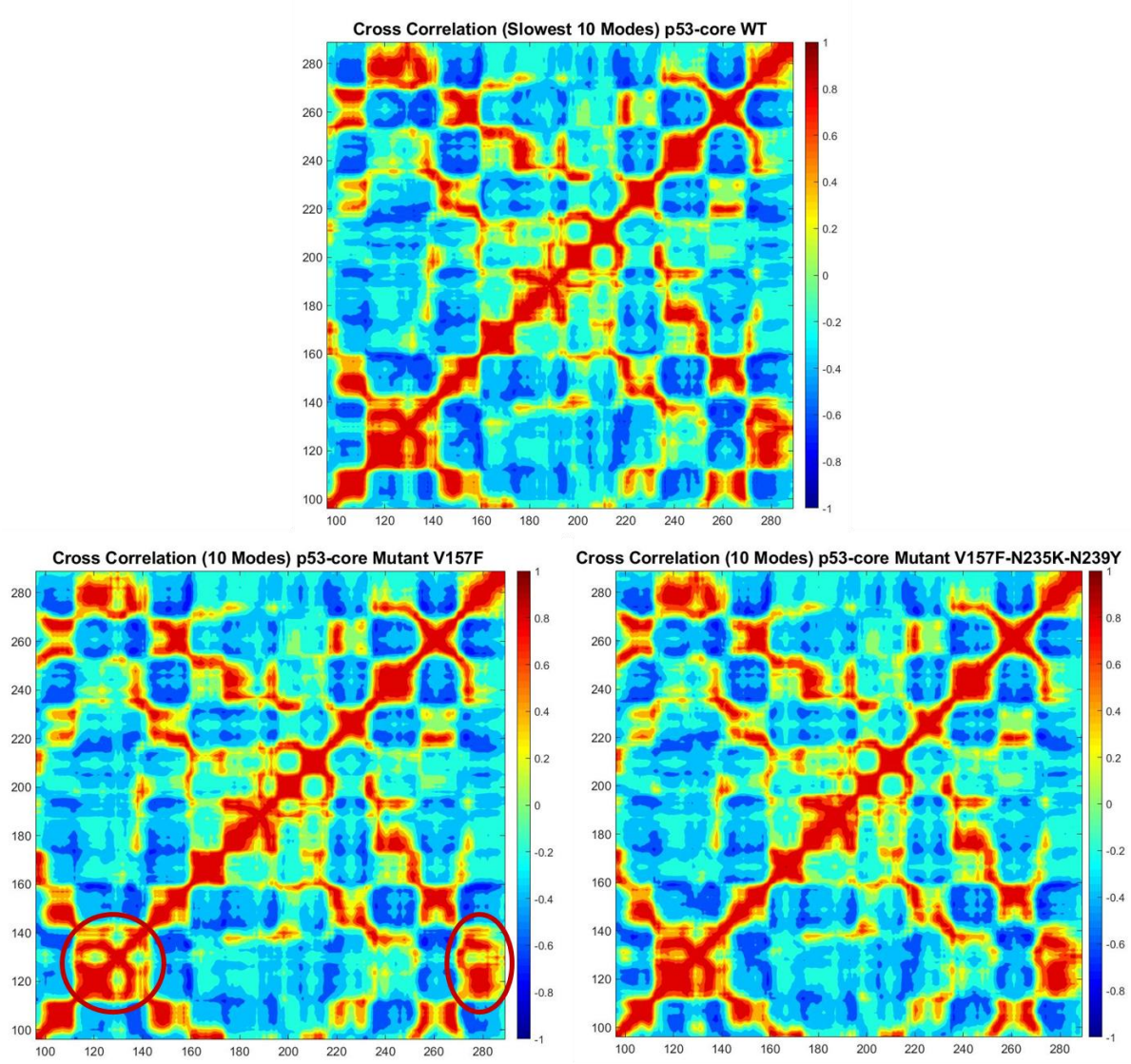


Figure 3.47. 2D cross correlation maps for wild type deleterious mutant and rescued mutant structures of p53.

It is observed that there is a correlation loss between residues in the region 110-140 and 275-285 in the deleterious mutant (V157F- bottom left) structure. This correlation loss may lead to disfunction in the protein thus resulting in deleterious effects. Same region is observed to restore the correlation in the compensatory mutated structure (V157F/N235K/N239Y-bottom right). That correlation loss could be the underlying mechanism of mutation V157F. The mentioned regions are not DNA binding sites of p53, but the loss of correlation in these regions may alter the dynamics of the protein and prevent it from working properly.

4. CONCLUSIONS AND RECOMMENDATIONS

4.1. Conclusions

Proteins are responsible for many vital functions which strongly correlate with their sequence. Mutations that cause loss of function or diseases (deleterious mutations) may either be linked with evolutionary or environmental effects. The effect of a deleterious mutation can allosterically be rescued by another mutation which is called compensatory mutations.

This work is mainly focused on the relationship between evolutionary conserved/reused segments and dynamic domains as well as the dynamic determinants of deleterious mutations and compensatory mutations. Mode perturbation analysis by GNM is also performed to investigate the effect of a mutation on selected residues.

Themes, which are reused segments in protein sequences, are examined in all-beta and all-alpha proteins. Our results suggest that themes show a good correlation with dynamic domains in both all-beta and all-alpha architectures. It is revealed that themes either alone or in combination with other themes may constitute a dynamic domain. Furthermore, shared themes of the proteins are observed to pair with dynamic domains obtained from different modes of motion. Statistical significance analysis which is performed by means of mutual information analysis, indicate a strong correlation between dynamic domains and themes. The correlation between themes and dynamic domains suggests that the functional contribution of a dynamic domain carried through the evolution in forms of themes and revealed itself in different modes of motion in different proteins. Additionally, GNM based perturbation analysis unveiled that the boundaries of the themes give the highest response to perturbations which is expected to occur in the key residues for the global dynamics of the protein.

In order to disclose dynamic traits of deleterious mutations, GNM based mode perturbation analysis is performed on a set of proteins that were analyzed with the deep sequencing method. A perturbation is implemented in the GNM algorithm to mimic the effect of a mutation. Our results indicate a correlation between fluctuation difference caused

by perturbation and mutation sensitivity of residues. Confidence interval regarding the correlation between fluctuation difference caused by perturbation and mutation sensitivity is observed to be more than %90 for each structure except TEM1-β -lactamase. Results for PSD95-PDZ domain shows a highly significant correlation between fluctuation difference caused by perturbation and mutation sensitivity. The same is observed for CcdB monomer structure. Structures PSD95-PDZ domain, CcdB monomer, GAL4, and H-Ras GTPase showed more than %95 confidence for correlation between fluctuation difference caused by perturbation and mutation sensitivity.

Furthermore, it is observed that mutation sensitive residues tend to be the residues that have a high capacity to change eigenvalues upon perturbation. P-value results suggest that confidence interval regarding the correlation between eigenvalue difference caused by perturbation and mutation sensitivity is more than %90 for each structure except GAL4. PSD95-PDZ domain gives the best results for both fluctuation difference analysis and eigenvalue difference analysis. For eigenvalue difference analysis, p- values obtained for PSD95-PDZ domain, CcdB monomer, CcdB dimer, ubiquitin, TEM1-β -lactamase and H-Ras GTPase gives more than %95 confidence for correlation between eigenvalue difference caused by perturbation and mutation sensitivity. So, it is believed that deleterious mutations that affect the global structural content of the protein can be predicted by perturbation methods.

GNM analysis regarding the compensatory mutations revealed that deleterious mutations and their related compensatory mutations are correlated in the spectrum of slow modes of motion. Seventeen deleterious mutation sites of tumor suppressor protein p53 and their related compensatory mutation sites are analyzed with GNM. Our results revealed that these mutation sites are related to hinge residues of slow modes of motion which implies the importance of hinge residues in the concept of compensatory mutations. Additionally, results of GNM analysis regarding the compensatory mutations disclosed the allosteric interaction between deleterious mutations and compensatory mutations. Dynamic traits revealed by GNM analysis regarding the compensatory mutations are believed to be a keystone for further studies about the prediction of compensatory mutations.

Our results revealed that the theme boundaries, deleterious mutations, and compensatory mutations are located at the dynamic interphases of slow modes of motion which implies the importance of coevolution in the subject of deleterious and compensatory mutations.

4.2. Recommendations for Future Studies

Themes dataset used in this thesis consist of only two types of protein architecture, all-beta and all-alpha architectures. In order to extend the scope of this research, other proteins with different architectures should be added to the dataset. Integrity of statistical significance analysis will be also amplified by increasing the number of proteins in the dataset.

Mutation sensitive residues are identified for each protein according to the fitness score data that provided in deep sequencing studies of each mutation site. Instead of selecting twenty residues with highest functional cost of mutation, number of mutation sensitive residues can be obtained with specific threshold value for each structure. After implementing a specific threshold according to characteristics of each protein, number of mutation sensitive residues will be different for each structure. That issue may be the reason of low confidence scores for some of the proteins. The results of the statistical significance analysis can be improved by this method.

In order to obtain more dynamic information concerning compensatory mutations, GNM based perturbation method will be improved and utilized with dynamic traits revealed in this thesis for prediction of compensatory mutations.

REFERENCES

Adkar, B. V., A. Tripathi, A. Sahoo, K. Bajaj, D. Goswami, P. Chakrabarti, M. K. Swarnkar, R. S. Gokhale, and R. Varadarajan, 2012, "Protein model discrimination using mutational sensitivity derived from deep sequencing", *Structure*, 20(2), pp. 371-381.

Anderson, A. C., 2012, "Winning the arms race by improving drug discovery against mutating targets", *ACS chemical biology*, 7(2), pp. 278-288.

Bahar, I., A. R. Atilgan, and B. Erman, 1997, "Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential", *Folding and Design*, 2(3), pp. 173-181.

Bahassi, E. M., M. H. O'Dea, N. Allali, J. Messens, M. Gellert, and M. Couturier, 1999, "Interactions of ccdb with dna gyrase inactivation of gyra, poisoning of the gyrase-dna complex, and the antidote action of ccda", *Journal of Biological Chemistry*, 274(16), pp. 10936-10944.

Bandaru, P., N. H. Shah, M. Bhattacharyya, J. P. Barton, Y. Kondo, J. C. Cofsky, C. L. Gee, A. K. Chakraborty, T. Kortemme, R. Ranganathan, and J. Kuriyan, 2017, "Deconstruction of the Ras switching cycle through saturation mutagenesis", *Elife*, 6, e27810.

Barešić, A., L. E. Hopcroft, H. H. Rogers, J. M. Hurst, and A. C. Martin, 2010, "Compensated pathogenic deviations: analysis of structural effects", *Journal of molecular biology*, 396(1), pp. 19-30.

Baroni, T. E., T. Wang, H. Qian, L. R. Dearth, L. N. Truong, J. Zeng, A. E. Denes, A. W. Chen, and R. K. Brachmann, 2004, "A global suppressor motif for p53 cancer mutants", *Proceedings of the National Academy of Sciences*, 101(14), pp. 4930-4935.

Behjati, S., and P. S. Tarpey, 2013, "What is next generation sequencing?", *Archives of Disease in Childhood-Education and Practice*, 98(6), pp. 236-238.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, P. E. Bourne, 2000, "The Protein Data Bank", *Nucleic Acids Research*, 28, pp. 235-242.

Bernard, P., and M. Couturier, 1992, "Cell killing by the F plasmid CcdB protein involves poisoning of DNA-topoisomerase II complexes", *Journal of molecular biology*, 226(3), pp. 735-745.

Bernard, P., K. E. Kézdy, L. Van Melderen, J. Steyaert, L. Wyns, M. L. Pato, P. N. Higgins, and M. Couturier, 1993, "The F plasmid CcdB protein induces efficient ATP-dependent DNA cleavage by gyrase", *Journal of molecular biology*, 234(3), pp. 534-541.

Brachmann, R. K., K. X. Yu, Y. Eby, N. P. Pavletich, and J. D. Boeke, 1998, "Genetic selection of intragenic suppressor mutations that reverse the effect of common p53 cancer mutations", *Embo Journal*, 17(7), pp. 1847-1859.

Bullock, A. N., and A. R. Fersht, 2001, "Rescuing the function of mutant p53", *Nature Reviews Cancer*, 1(1), p. 68.

Chen, F., W. Wang, and W. S. El-Deiry, 2010, "Current strategies to target p53 in cancer", *Biochemical pharmacology*, 80(5), pp. 724-730.

Cover, T. M., and J. A. Thomas, 1991, *Elements of Information Theory*, John Wiley & Sons, Inc., New York.

Cox, A. D., and C. J. Der, 2010, "Ras history: The saga continues", *Small GTPases*, 1(1), pp. 2-27.

Cui, H., A. Hayashi, H. S. Sun, M. P. Belmares, C. Cobey, T. Phan, J. Schweizer, M. W. Salter, Y. T. Wang, A. Tasker, D. Garman, J. Rabinowitz, P. S. Lu, and M. Tymianski, 2007, "PDZ protein interactions underlying NMDA receptor-mediated excitotoxicity and neuroprotection by PSD-95 inhibitors", *Journal of Neuroscience*, 27(37), pp. 9901-9915.

Danziger, S. A., R. Baronio, L. Ho, L. Hall, K. Salmon, G. W. Hatfield, P. Kaiser, and R. H. Lathrop, 2009, "Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning", *PLoS computational biology*, 5(9), e1000498.

Danziger, S. A., J. Zeng, Y. Wang, R. K. Brachmann, and R. H. Lathrop, 2007, "Choosing where to look next in a mutation sequence space: Active Learning of informative p53 cancer rescue mutants", *Bioinformatics*, 23(13), pp. i104-i114.

Deardorff, J. A., and A. B. Sachs, 1997, "Differential effects of aromatic and charged residue substitutions in the RNA binding domains of the yeast poly (A)-binding protein", *Journal of molecular biology*, 269(1), pp. 67-81.

Duffy, J. B., 2002, "GAL4 system in Drosophila: a fly geneticist's Swiss army knife", *Genesis*, 34(1-2), pp. 1-15.

Emekli, U., D. Schneidman-Duhovny, H. J. Wolfson, R. Nussinov, and T. Haliloglu, 2008, "HingeProt: automated prediction of hinges in protein structures", *Proteins: Structure, Function, and Bioinformatics*, 70(4), pp. 1219-1227.

Firnberg, E., J. W. Labonte, J. J. Gray, and M. Ostermeier, 2014, "A comprehensive, high-resolution map of a gene's fitness landscape", *Molecular biology and evolution*, 31(6), pp. 1581-1592.

Gao, M., H. Zhou, and J. Skolnick, 2015, "Insights into disease-associated mutations in the human proteome through protein structural analysis", *Structure*, 23(7), pp. 362-1369.

Giniger, E., S. M. Varnum, and M. Ptashne, 1985, "Specific DNA binding of GAL4, a positive regulatory protein of yeast", *Cell*, 40(4), pp. 767-774.

Goldman, D., and K. Domschke, 2014, "Making sense of deep sequencing", *International Journal of Neuropsychopharmacology*, 17(10), pp. 1717-1725.

Guarente, L., R. R. Yocum, and P. Gifford, 1982, “A GAL10-CYC1 hybrid yeast promoter identifies the GAL4 regulatory region as an upstream site”, *Proceedings of the National Academy of Sciences*, 79(23), pp. 7410-7414.

Haliloglu, T., I. Bahar, and B. Erman, 1997, “Gaussian dynamics of folded proteins”, *Physical review letters*, 79(16), p. 3090.

Harris, B. Z., and W. A. Lim, 2001, “Mechanism and role of PDZ domains in signaling complex assembly”, *Journal of cell science*, 114(18), pp. 3219-3231.

Inga, A., and M. A. Resnick, 2001, “Novel human p53 mutations that are toxic to yeast can enhance transactivation of specific promoters and reactivate tumor p53 mutants”, *Oncogene*, 20(26), p. 3409.

Kimura, M., 1985, “The role of compensatory neutral mutations in molecular evolution,” Journal *of Genetics*, 64(1), p. 7.

Kitzman, J. O., L. M. Starita, R. S. Lo, S. Fields, and J. Shendure, 2015, “Massively parallel single-amino-acid mutagenesis”, *Nature methods*, 12(3), p. 203.

Kühn, U., and E. Wahle, 2004, “Structure and function of poly (A) binding proteins”, Biochimica *et Biophysica Acta (BBA)-Gene Structure and Expression*, 1678(2-3), pp. 67-84.

Lee, H. J., and J. J. Zheng, 2010, “PDZ domains and their binding partners: structure, specificity, and modification”, *Cell communication and Signaling*, 8(1), p. 8.

Madl, T., L. Van Melderen, N. Mine, M. Respondek, M. Oberer, W. Keller, L. Khatai, and K. Zangger, 2006, “Structural basis for nucleic acid and toxin recognition of the bacterial antitoxin CcdA”, *Journal of molecular biology*, 364(2), pp. 170-185.

Maisnier-Patin, S., and D. I. Andersson, 2004, “Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution”, *Research in microbiology*, 155(5), pp. 360-369.

Mavor, D., K. Barlow, S. Thompson, B.A. Barad, A.R. Bonny, C.L. Cario, G. Gaskins, Z. Liu, L. Deming, S.D. Axen, and E. Caceres, 2016, "Determination of ubiquitin fitness landscapes under different chemical stresses in a classroom setting", *Elife*, 5, e15802.

McLaughlin Jr, R. N., F. J. Poelwijk, A. Raman, W. S. Gosal, and R. Ranganathan, 2012, "The spatial architecture of protein function and adaptation", *Nature*, 491(7422), p. 138.

Melamed, D., D. L. Young, C. E. Gamble, C. R. Miller, and S. Fields, 2013, "Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly (A)-binding protein", *Rna*, 19(11), pp. 1537-1551.

Merz Jr, K. M., D. Ringe, and C. H. Reynolds, 2010, *Drug design: structure-and ligand-based approaches*, Cambridge University Press, New York.

Nepomnyachiy, S., N. Ben-Tal, and R. Kolodny, 2014, "Global view of the protein universe", *Proceedings of the National Academy of Sciences*, 111(32), pp. 11691-11696.

Nepomnyachiy, S., N. Ben-Tal, and R. Kolodny, 2017, "Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths", *Proceedings of the National Academy of Sciences*, 114(44), pp. 11703-11708.

Nesse, R. M., C. T. Bergstrom, P. T. Ellison, J. S. Flier, P. Gluckman, D. R. Govindaraju, D. Niethammer, G. S. Omenn, R. L. Perlman, M. D. Schwartz, M. G. Thomas, S. C. Stearns, and D. Valle, 2010, "Making evolutionary biology a basic science for medicine", *Proceedings of the National Academy of Sciences*, 107(suppl 1), pp. 1800-1807.

Omenn G S., 2010, "Evolution in health and medicine Sackler colloquium: evolution and public health", *Proceedings of the National Academy of Sciences of the United States of America*, 107, pp. 1702–1709.

Poon, A., B. H. Davis, and L. Chao, 2005, "The coupon collector and the suppressor mutation: estimating the number of compensatory mutations by maximum likelihood", *Genetics*, 170(3), pp. 1323-1332.

Romano, S., J. Bailey, V. Nguyen, and K. Verspoor, 2014, “Standardized mutual information for clustering comparisons: one step further in adjustment for chance”, *International Conference on Machine Learning*, pp. 1143-1151.

Salverda, M. L., J. A. G. De Visser, and M. Barlow, 2010, “Natural evolution of TEM-1 β-lactamase: experimental reconstruction and clinical relevance”, *FEMS microbiology reviews*, 34(6), pp. 1015-1036.

Studer, R. A., B. H. Dessailly, and C. A. Orengo, 2013, “Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes”, *Biochemical Journal*, 449(3), pp. 581-594.

Toto, A., S. W. Pedersen, O. A. Karlsson, G. E. Moran, E. Andersson, and C. N. Chi, Stromgaard, K., Gianni, S., and Jemth, P., 2016, “Ligand binding to the PDZ domains of postsynaptic density protein 95”, *Protein Engineering, Design and Selection*, 29(5), pp. 169-175.

Varshavsky, A., 2001, “Ubiquitin”, *Encyclopedia of Genetics*, Elsevier Science Inc., pp. 2091-2093.

Vinh, N. X., J. Epps, and J. Bailey, 2010, “Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance”, *Journal of Machine Learning Research*, 11, 2837-2854.

Williams, C. L., 2003, “The polybasic region of Ras and Rho family small GTPases: a regulator of protein interactions and membrane association and a site of nuclear localization signal sequences”, *Cellular signalling*, 15(12), pp. 1071-1080.

Yue, P., Z. Li, and J. Moult, 2005, “Loss of protein structure stability as a major causative factor in monogenic disease”, *Journal of molecular biology*, 353(2), pp. 459-473.

APPENDIX A: ADDITIONAL FIGURES AND TABLES ABOUT THEMES AND DYNAMIC DOMAINS

Table A1. The themes detected in the 2XYI propeller.

Theme Name	Position on 2XYI
14940	14-53
14945	16-120
14952	47-120
14946	55-120
14942	126-170
14953	126-181
14944	126-213
14812	176-213
14954	176-220
14813	176-240
14937	227-265
14938	227-287
14939	227-309
15218	230-273
15220	230-297
15221	230-330
14949	243-297
14815	244-309 OR 289-353 OR 299-353 OR 343-406
14816	271-309
15219	281-315
14955	283-341

Table A1. The themes detected in the 2XYI propeller. cont.

Theme Name	Position on 2XYI
14861	314-353 OR 368-407
14950	314-408
14956	315-381
15222	336-407
14951	356-415
14941	376-411

Table A2. The themes detected in the 3EMH propeller. Themes that are shared with 2XYI propeller shaded with grey and their position on each structure is given.

Theme Name	Position on 2XYI	Position on 3EMH
14812	176-213	38-77 OR 75-115 OR 81-120 OR 90-125 OR 124-167 OR 127-161 OR 127-162 OR 165-200 OR 166-203
14813	176-231 OR 176-240 OR 227-287	43-87 OR 82-132 OR 90-134 OR 127-178 OR 132-177 OR 166-224 OR 208-267 OR 217-257 OR 217-267
14815	244-309 OR 289-353 OR 299-353 OR 343-406	48-120 OR 93-157 OR 137-204 OR 174-243 OR 192-241 OR 228-294 OR 266-331
14816	271-309	123-157 OR 127-160 OR 169-209 OR 211-245
14861	314-353 OR 368-407 OR 368-408 OR 369-408 OR 371-407	165-199 OR 205-242 OR 292-331
14955	283-341	48-132
14862	-	108-202
14863	-	108-167
14997	-	124-190
14986	-	127-246
14993	-	137-172
14880	-	164-225
14988	-	166-215
14881	-	180-225
14883	-	187-245 OR 229-287
14994	-	192-257
14957	-	205-287 OR 217-288

Table A2. The themes detected in the 3EMH propeller. Themes that are shared with 2XYI propeller shaded with grey and their position on each structure is given. cont.

Theme Name	Position on 2XYI	Position on 3EMH
14859	-	209-245
14989	-	217-257
14958	-	217-288
14884	-	229-330
14995	-	48-132
14987	-	82-132

Table A3. Shared and non-shared themes of 2OF3, 1B3U and 4ADY and their respective sequence positions in the structure.

Theme	Position on 2OF3	Position on 1B3U	Position on 4ADY
c_180_39	734-762,766,768-796,801,804-807,811-834	282-310,323-353,362-389	567-575,577-594,597-604,606-628,633-662
c_180_17	769-797,804-807,811-835	281-310,323-352	604-628,633-665
c_180_4	650-676,679-681,686-693,698-716,721,732-762,767-796,803-807,811-826,830-836	165-193,202-232,241-271,280-310,323-342,346-353	459-488,491-492,495-508,544-556,561-575,577-594,597-602,604-628,633-654,658-664
c_180_29	735-807,811-836	483-587	606-687,690-708
c_180_9	650-676,679-681,686-693,698-716,721,732-761,766-796,801,804-807,811-836	167-193,202-232,241-271,280-310,323-353,362-389	463-488,493-507,509-524,527-538,540-556,561-575,577-594,597-602,604-628,633-662
c_180_19	770-797,804-807,811-835	521-548,557-586	605-628,633-664
c_180_20	770-796,801,804-807,811-835	521-548,557-586	605-628,633-664
c_180_36	698-715,720-721,732,734-762,766,768-795,800-801,804-807,811-833	252-271,280,282-310,323-353,362-388	541-556,561-565,567-575,577-594,597-603,605-628,633-661
c_180_37	735-762,767-795,800-801,804-807,811-834	283-310,323-353,362-389	567-575,577-594,597-604,606-628,633-662
c_180_8	698-716,731-746,749-762,766,768-796,801,804-807,811-835	244-270,279-294,297-310,323-353,362-390	529-538,540-556,561-575,577-580,583-594,597-602,604-628,633-664

Table A3. Shared and non-shared themes of 2OF3, 1B3U and 4ADY and their respective sequence positions in the structure. cont.

c_180_15		166-193,202-232,241-271,280-310,323-353,362-390	459-488,493-507,509-524,527-538,540-556,561-575,577-594,597-602,604-628,633-663
c_180_10		258-271,280-309	617-628,633-664
c_180_31		363-583	496-507,509-537,565-575,577-688,691-707
c_180_18		281-308,311-312,323-353	562-575,577-594,597-601,603-628,633-648
c_180_34		282-317,319-374,376-550	419-433,441-489,491-508,510-537,539-558,561-575,577-667
c_180_6		245-312,314-452	498-536,564-575,577-687,690-713
c_180_7		206-312+314-583	
c_180_22		439-585	
c_180_5		207-312,314-583	
c_180_41		14-23,25-132,171-277,279-317,319-357	
c_180_38			624-628,633-635,637-657
c_180_13			464-488,493-507
c_180_14			605-628,633-648

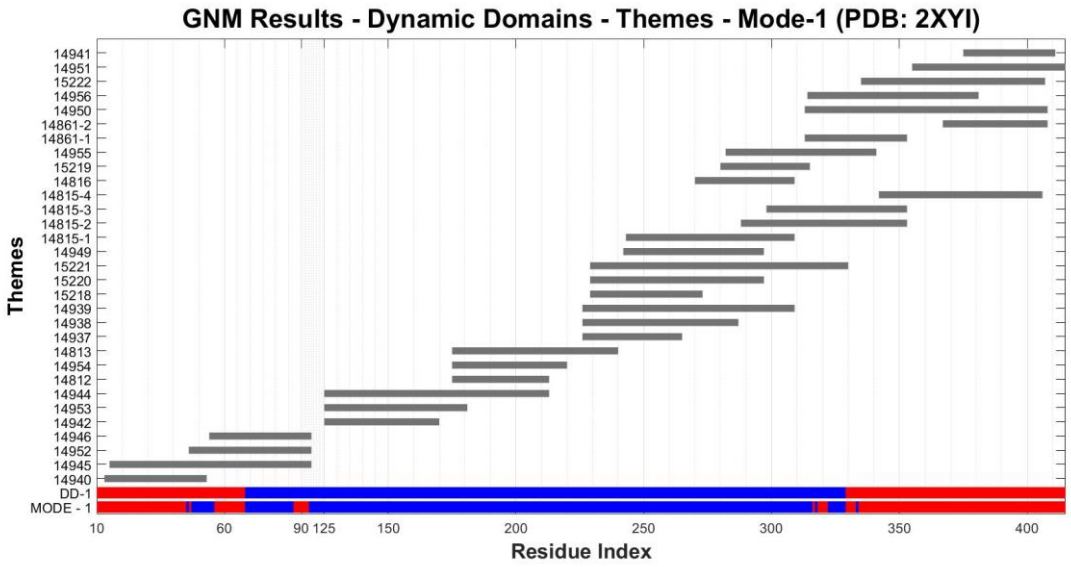


Figure A1. Slowest GNM mode, dynamic domains and themes (PDB ID: 2XYI).

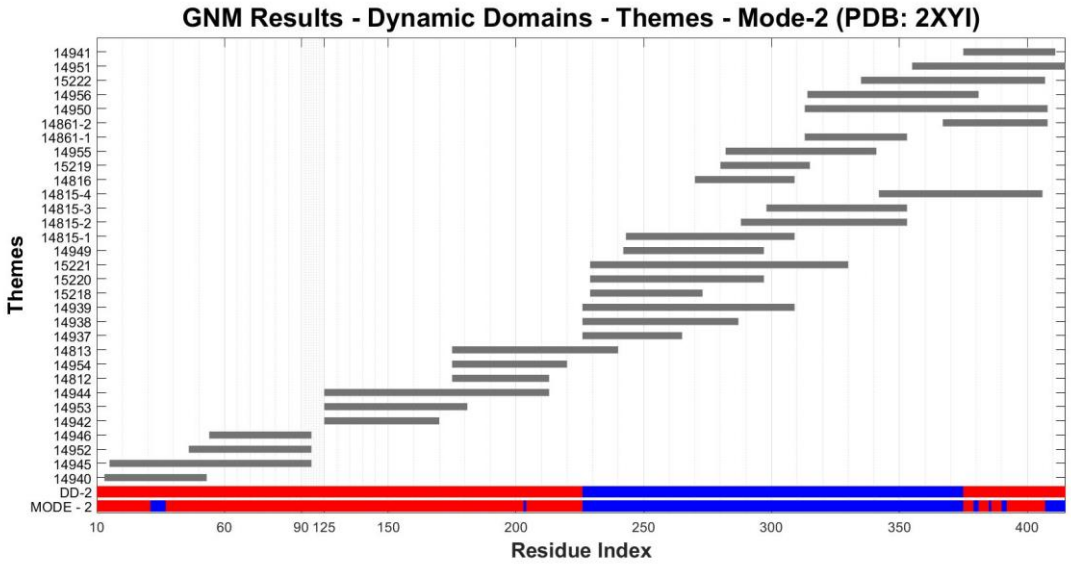


Figure A2. Second slowest GNM mode, dynamic domains and themes (PDB ID: 2XYI).

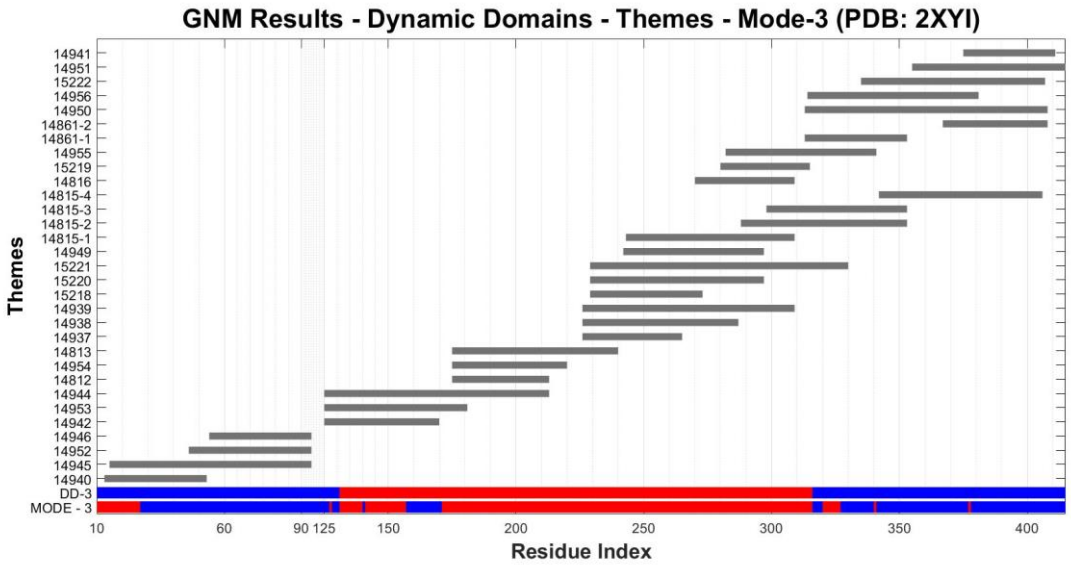


Figure A3. Third slowest GNM mode, dynamic domains and themes (PDB ID: 2XYI).

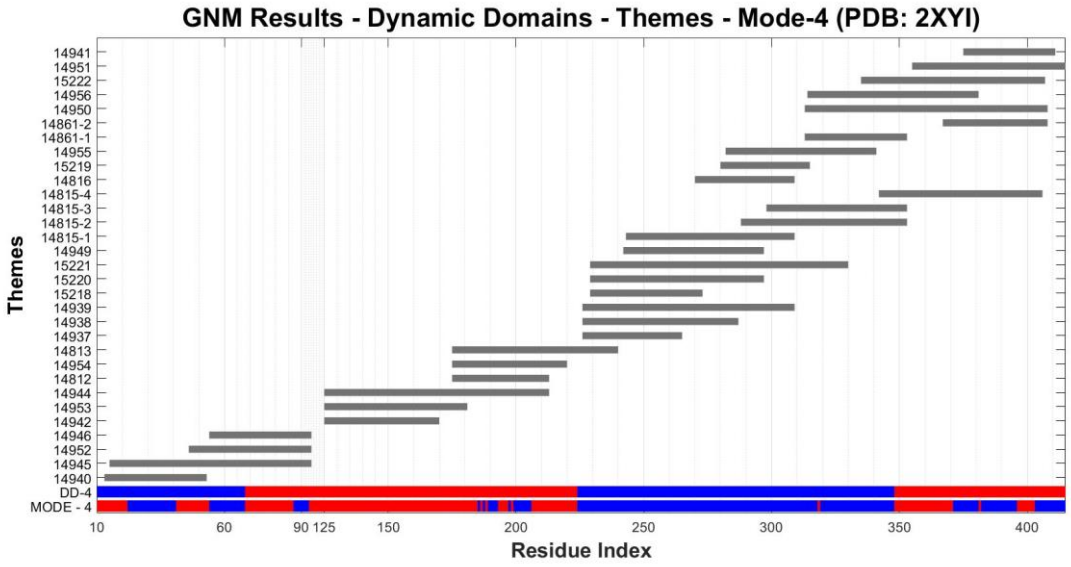


Figure A4. Fourth slowest GNM mode, dynamic domains and themes (PDB ID: 2XYI).

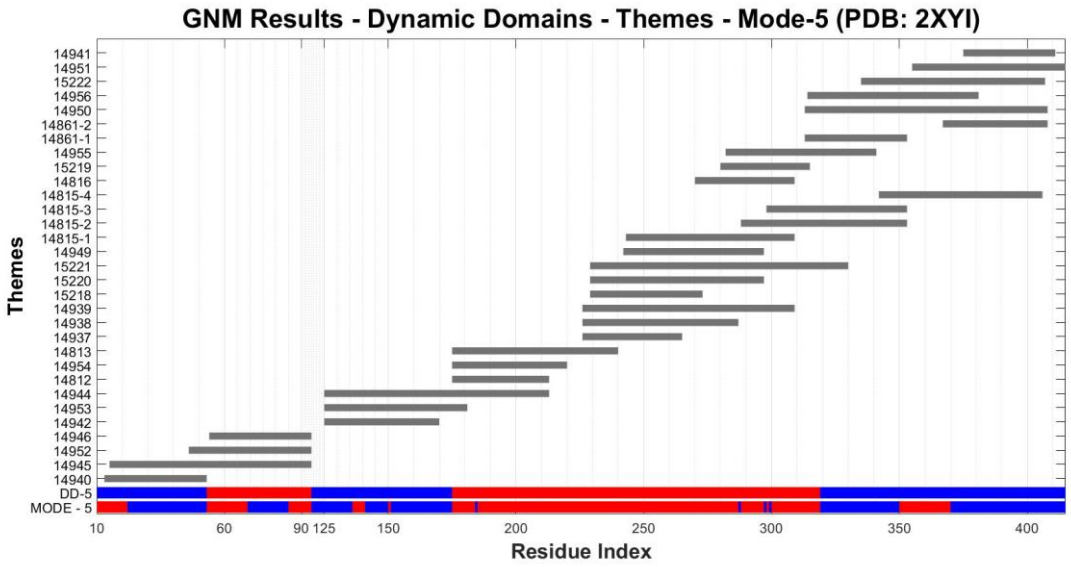


Figure A5. Fifth slowest GNM mode, dynamic domains and themes (PDB ID: 2XYI).

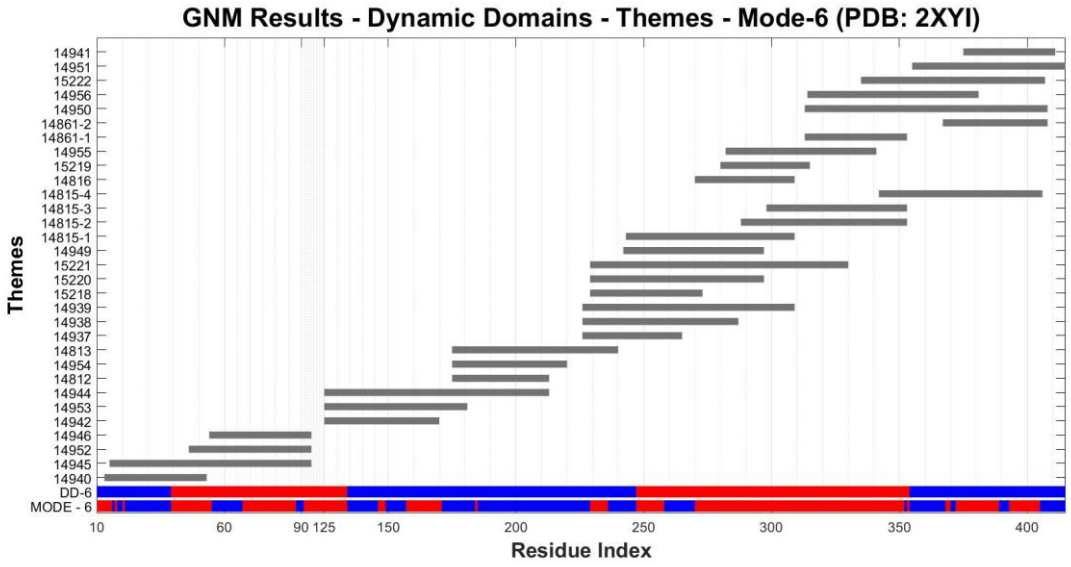


Figure A6. Sixth slowest GNM mode, dynamic domains and themes (PDB ID: 2XYI).

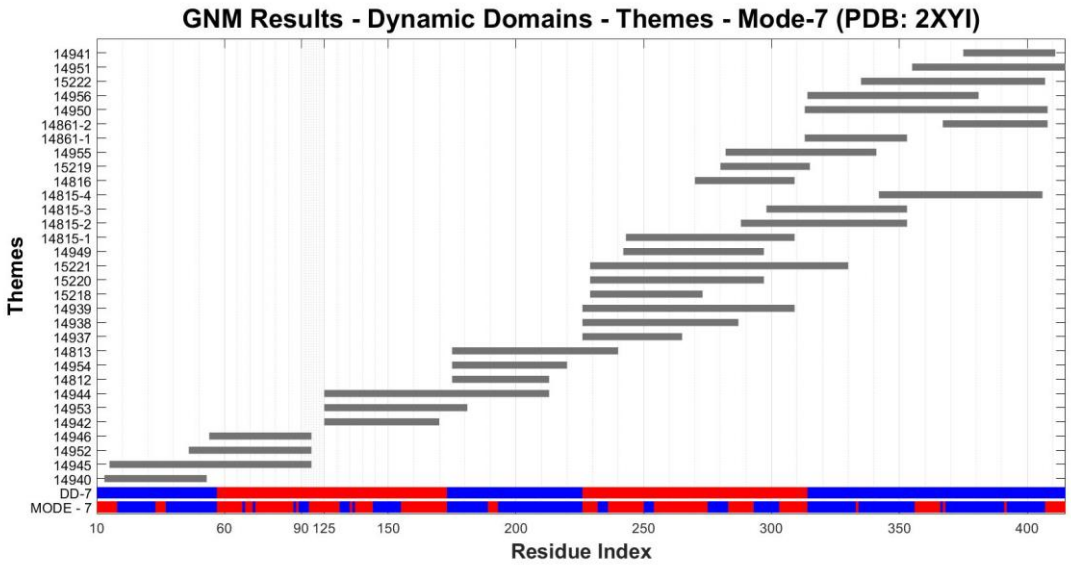


Figure A7. Seventh slowest GNM mode, dynamic domains and themes (PDB ID: 2XYI).

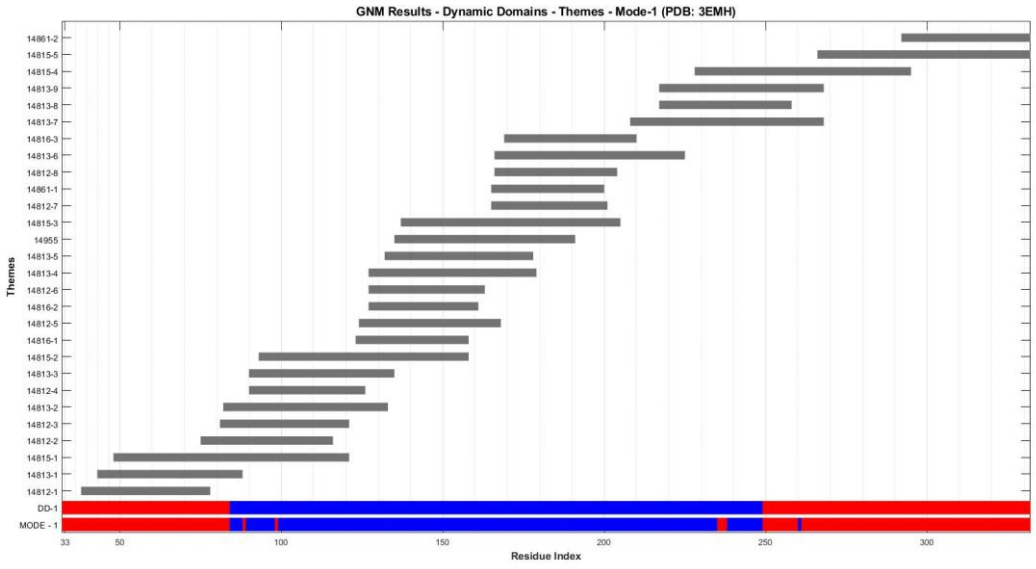


Figure A8. Slowest GNM mode, dynamic domains and themes (PDB ID: 3EMH).

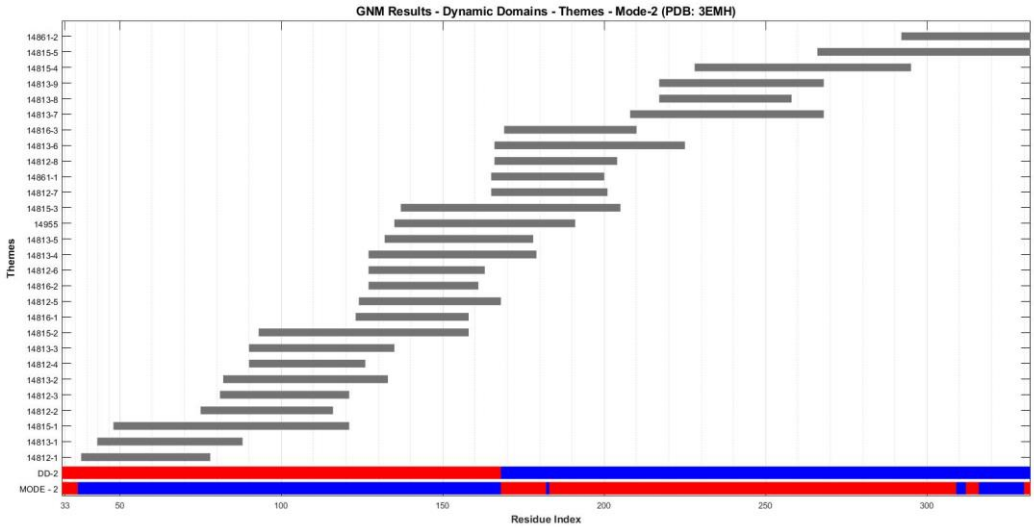


Figure A9. Second slowest GNM mode, dynamic domains and themes (PDB ID: 3EMH).

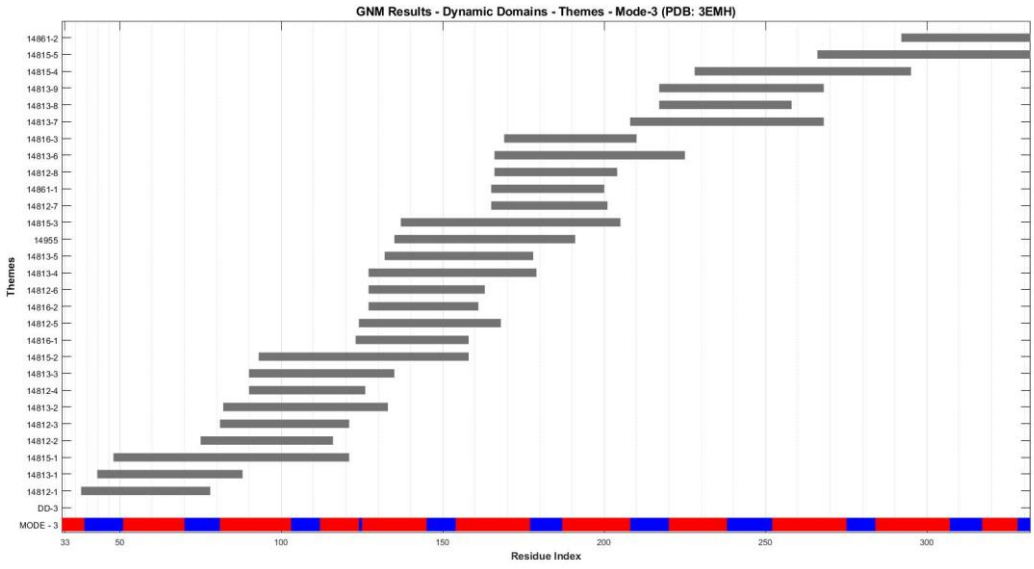


Figure A10. Third slowest GNM mode and themes (PDB ID: 3EMH).

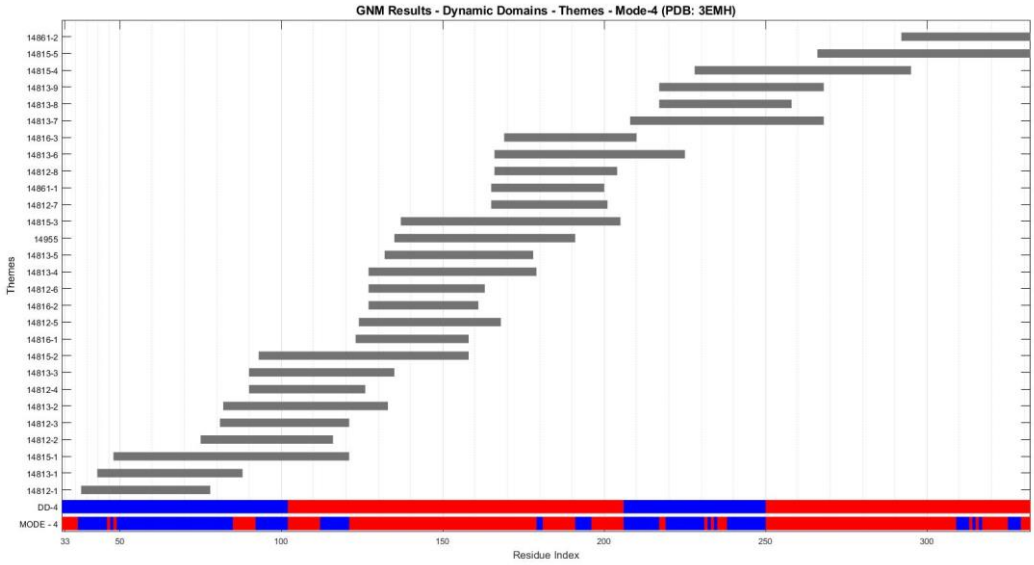


Figure A11. Fourth slowest GNM mode, dynamic domains and themes (PDB ID: 3EMH).

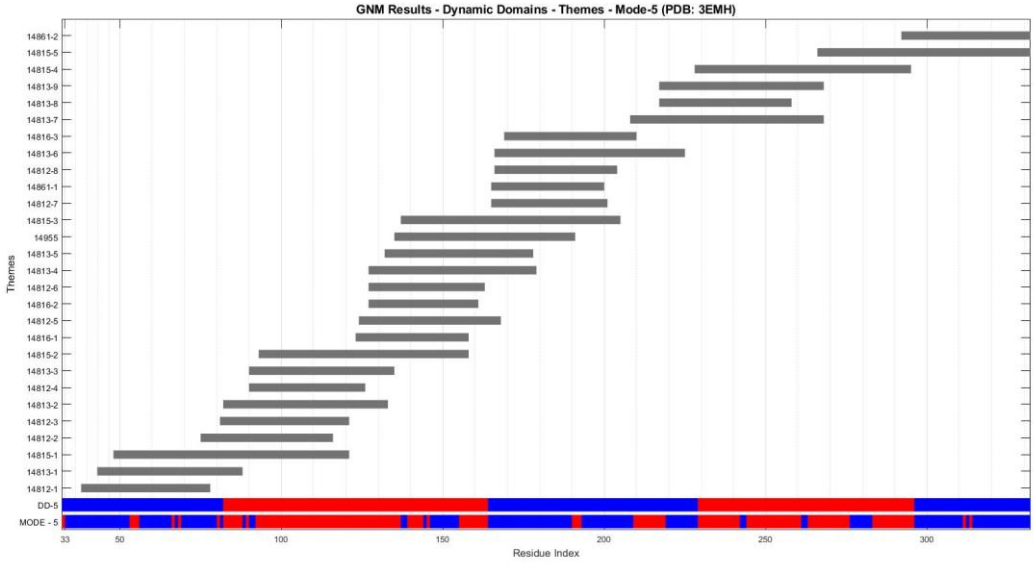


Figure A12. Fifth slowest GNM mode, dynamic domains and themes (PDB ID: 3EMH).

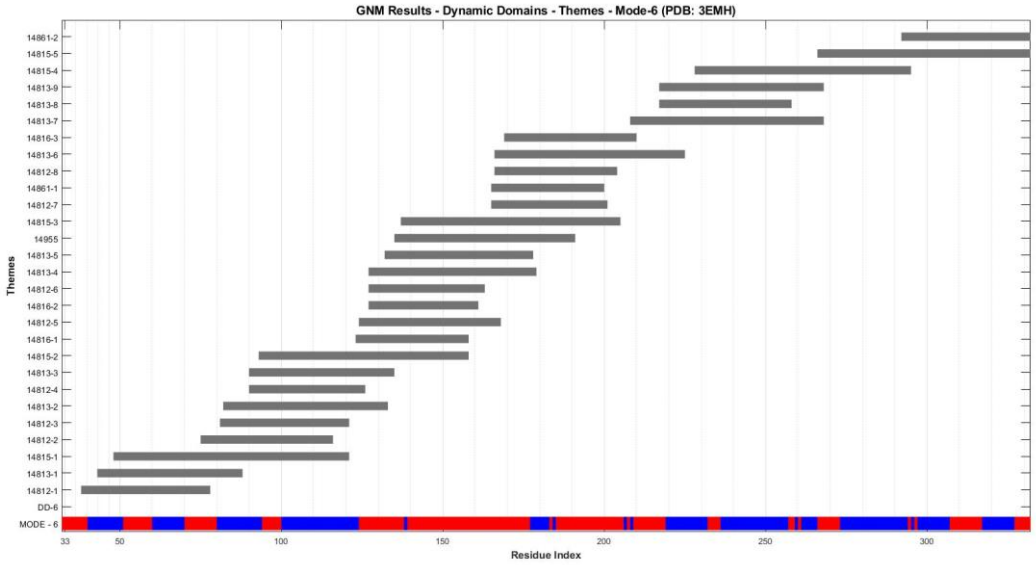


Figure A13. Sixth slowest GNM mode and themes (PDB ID: 3EMH).

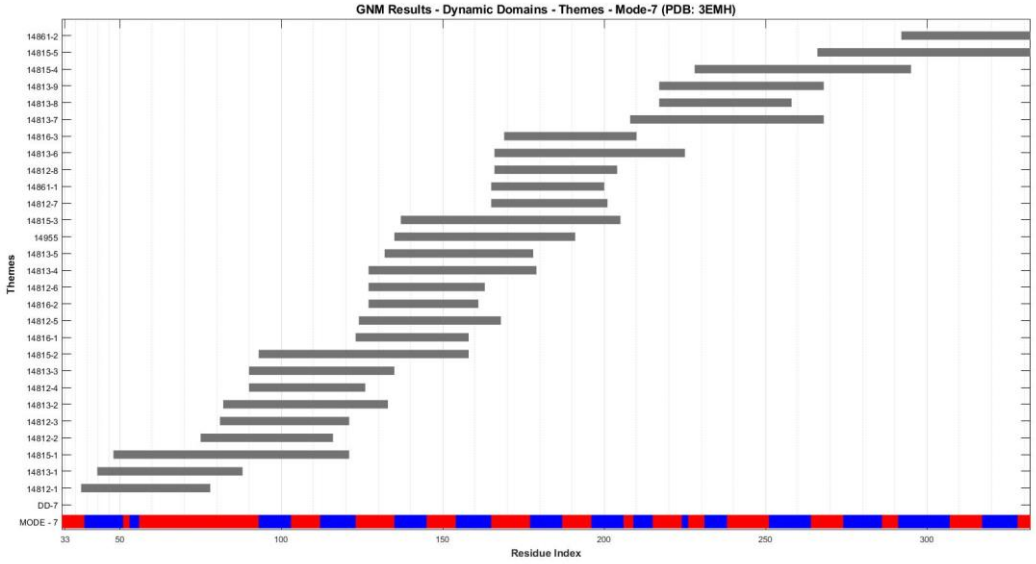


Figure A14. Seventh slowest GNM mode and themes (PDB ID: 3EMH).

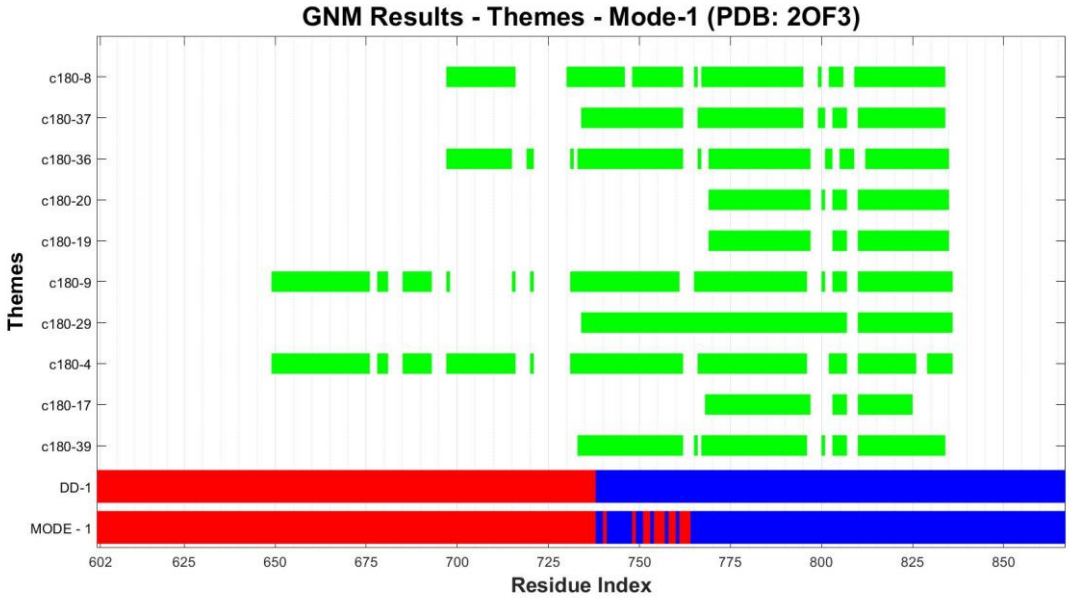


Figure A15. Slowest GNM mode, dynamic domains and themes (PDB ID: 2OF3).

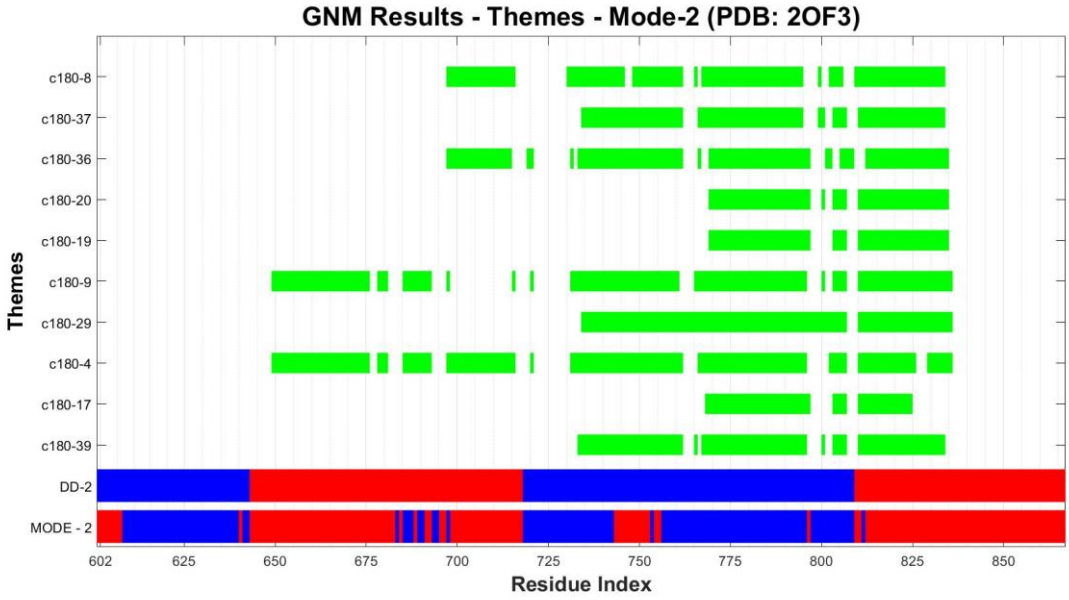


Figure A16. Second slowest GNM mode, dynamic domains and themes (PDB ID: 2OF3).

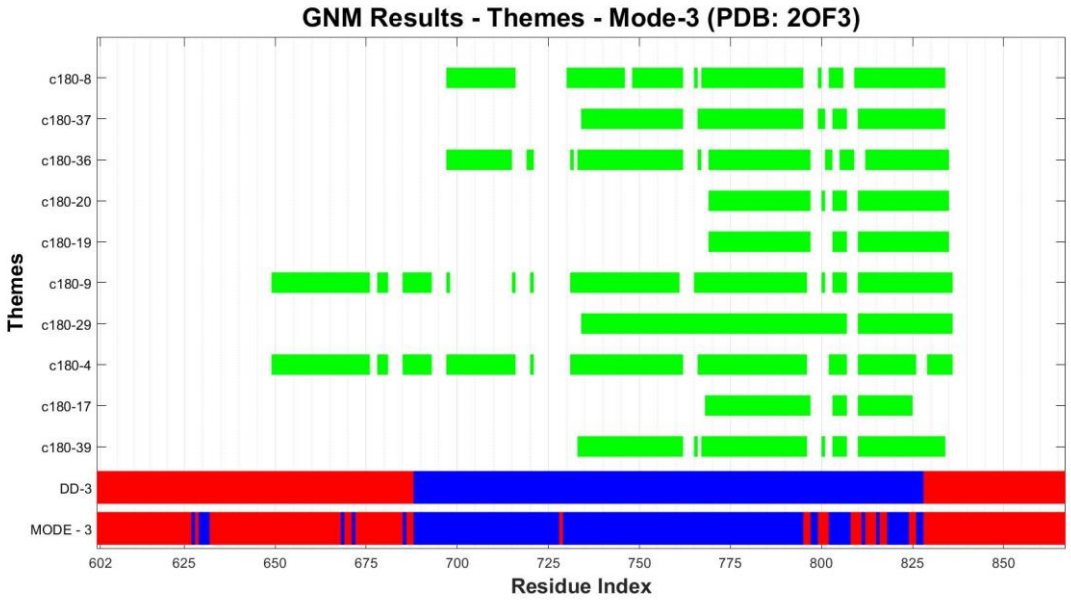


Figure A17. Third slowest GNM mode, dynamic domains and themes (PDB ID: 2OF3).

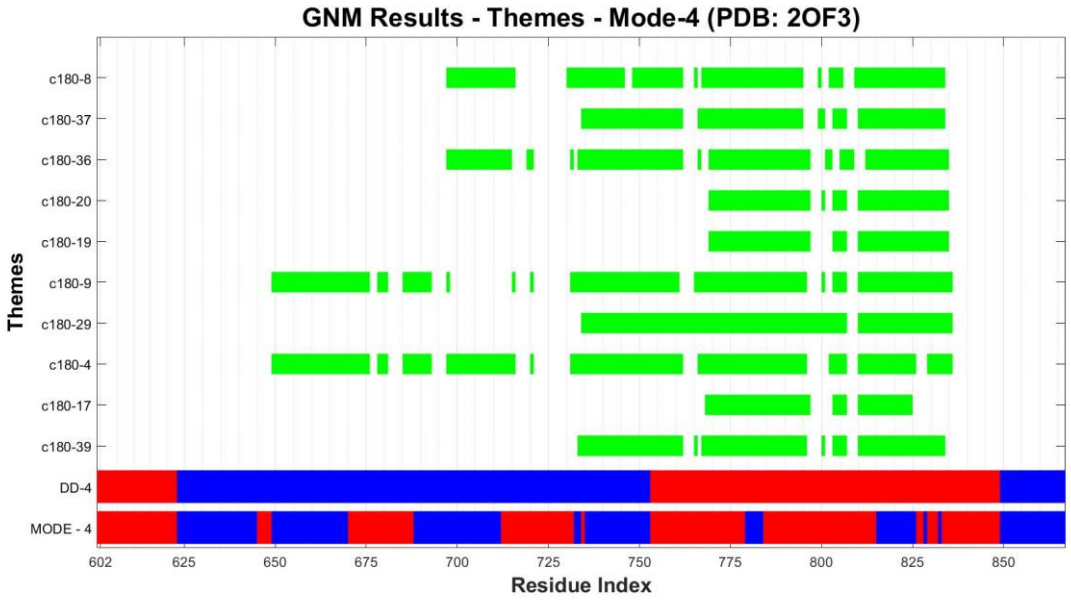


Figure A18. Fourth slowest GNM mode, dynamic domains and themes (PDB ID: 2OF3).

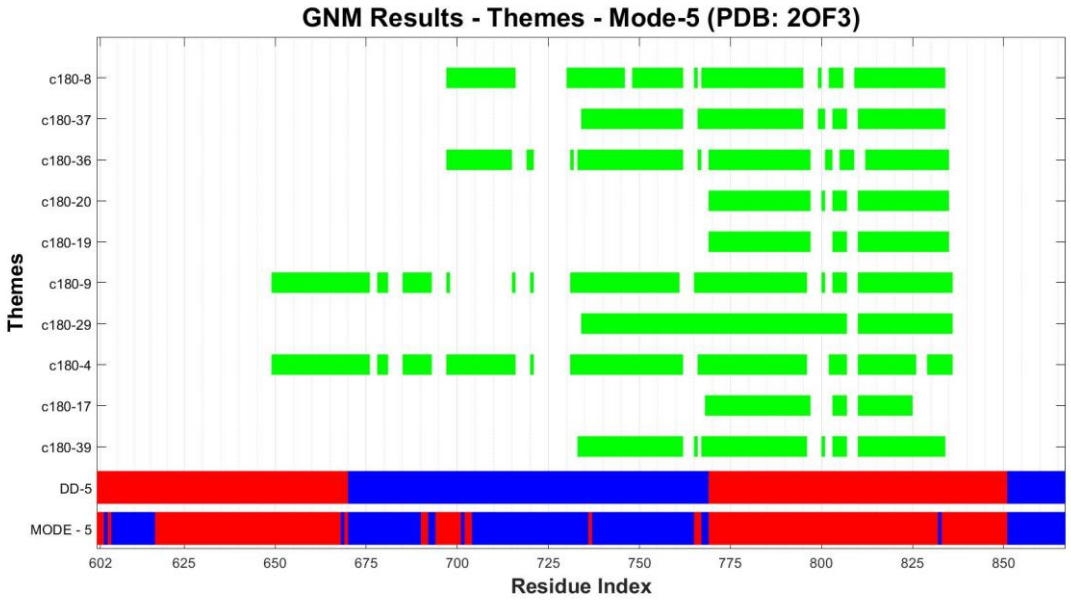


Figure A19. Fifth slowest GNM mode, dynamic domains and themes (PDB ID: 2OF3).

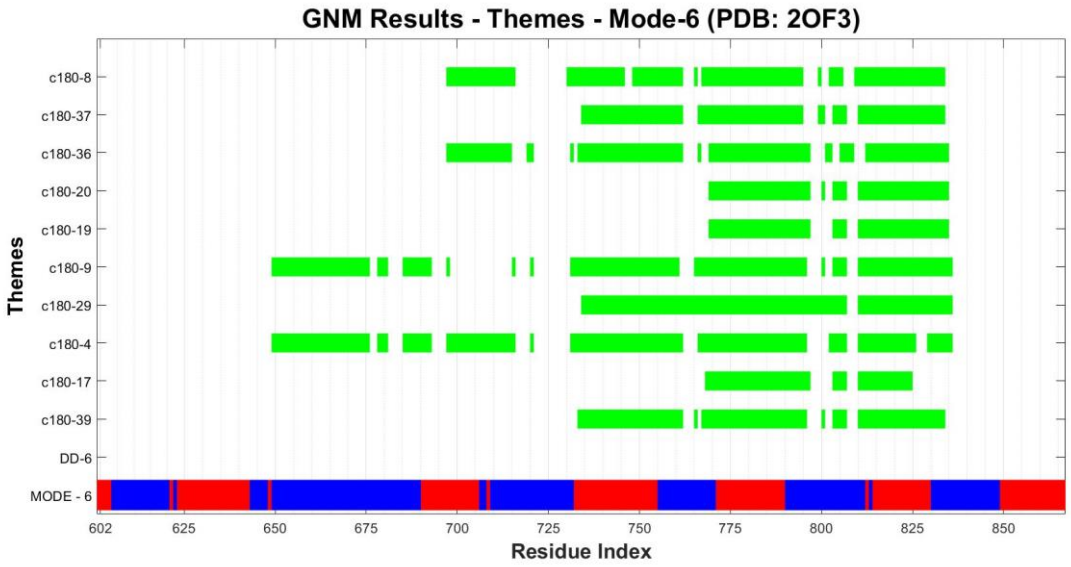


Figure A20. Sixth slowest GNM mode, dynamic domains and themes (PDB ID: 2OF3).

Figure A21. Seventh slowest GNM mode, dynamic domains and themes (PDB ID: 2OF3).

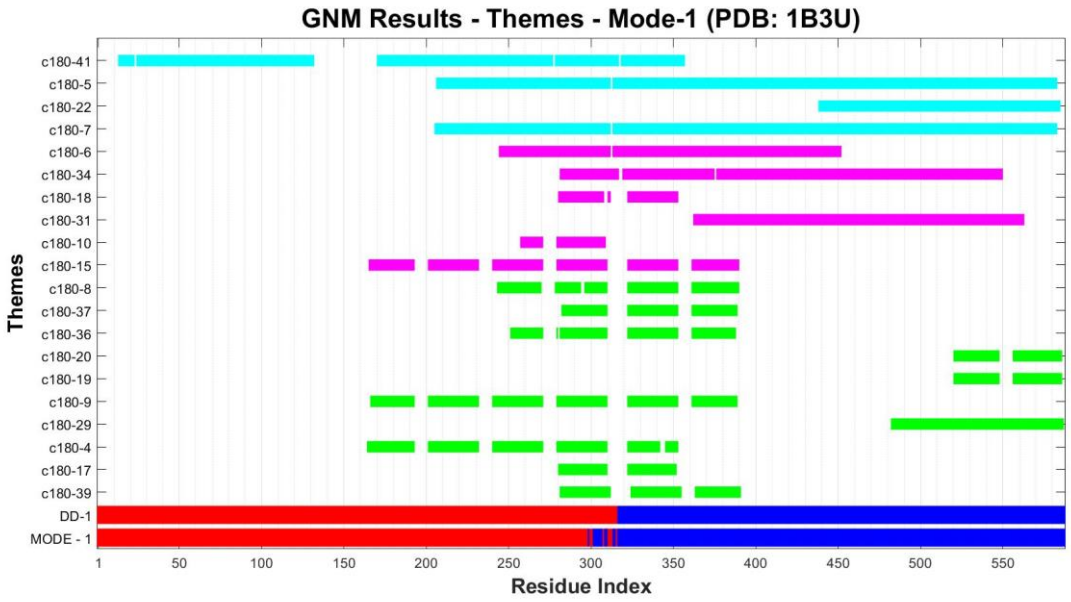


Figure A22. Slowest GNM mode, dynamic domains and themes (PDB ID: 1B3U).

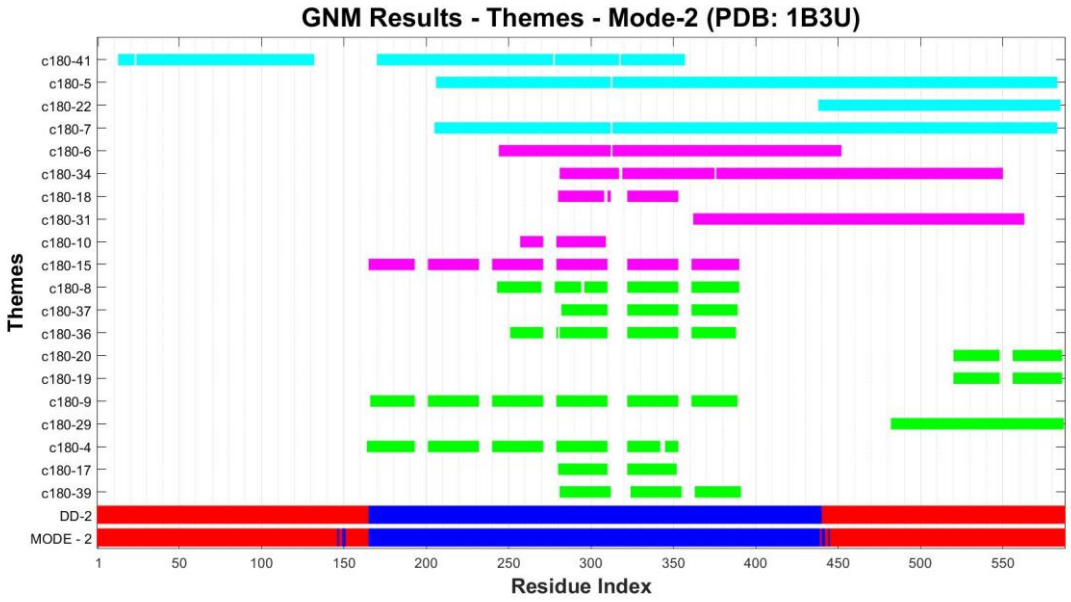


Figure A23. Second slowest GNM mode, dynamic domains and themes (PDB ID: 1B3U).

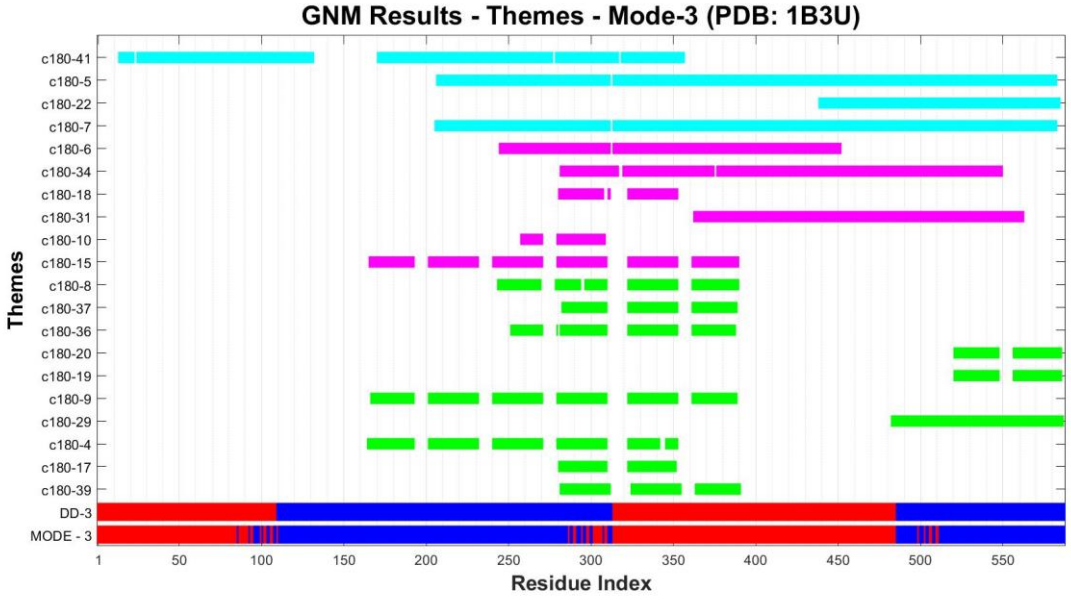


Figure A24. Third slowest GNM mode, dynamic domains and themes (PDB ID: 1B3U).

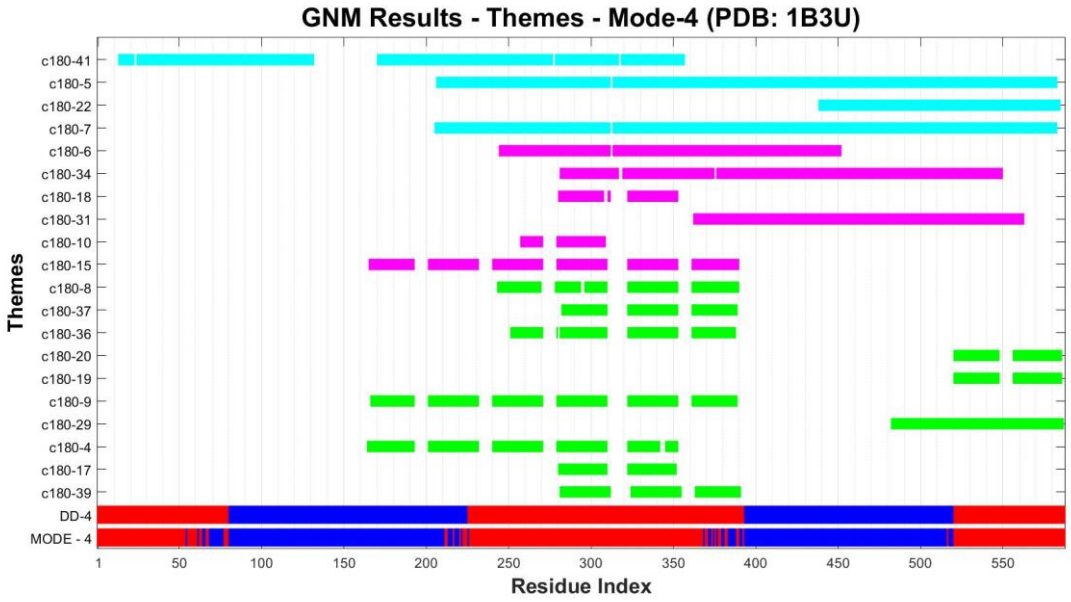


Figure A25. Fourth slowest GNM mode, dynamic domains and themes (PDB ID: 1B3U).

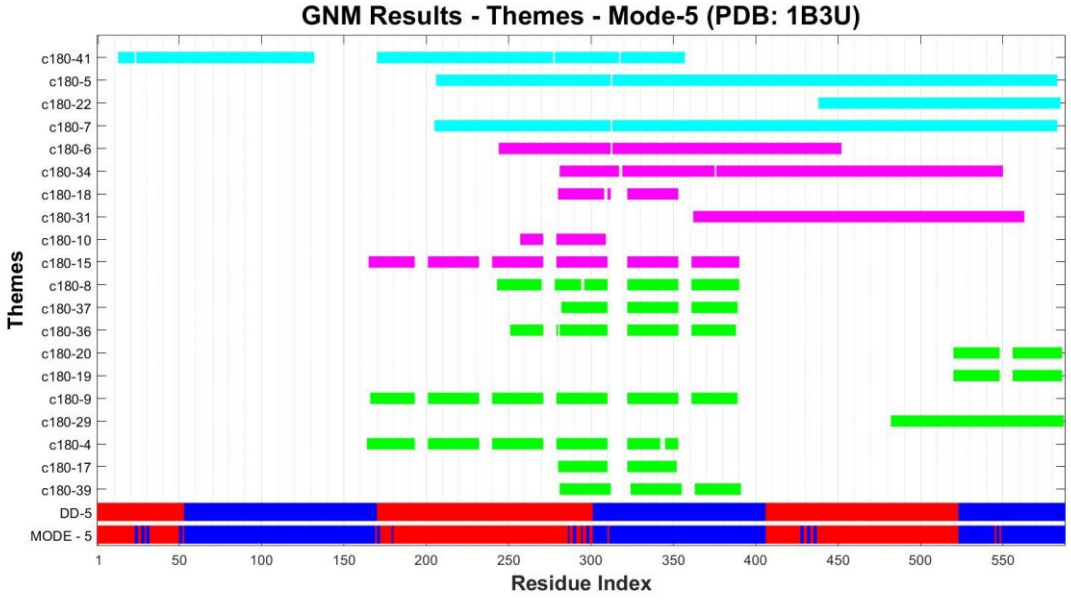


Figure A26. Fifth slowest GNM mode, dynamic domains and themes (PDB ID: 1B3U).

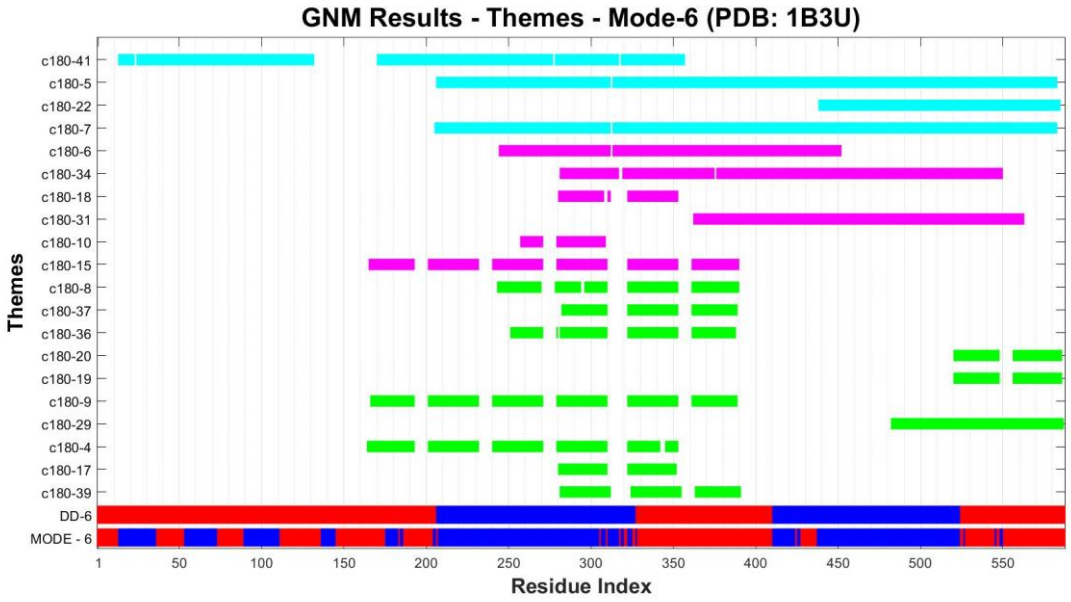


Figure A27. Sixth slowest GNM mode, dynamic domains and themes (PDB ID: 1B3U).

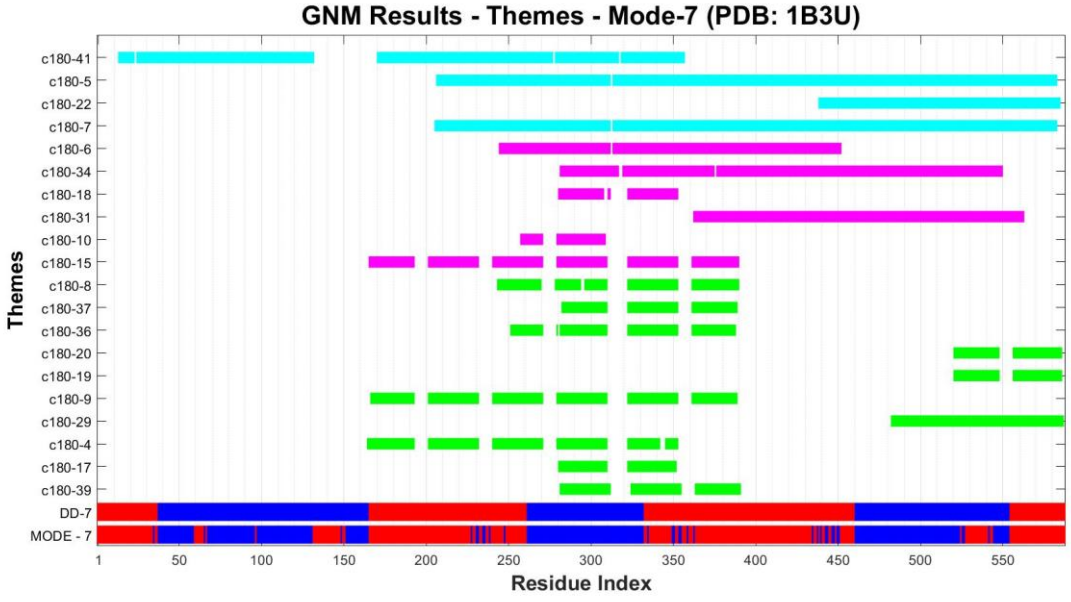


Figure A28. Seventh slowest GNM mode, dynamic domains and themes (PDB ID: 1B3U).

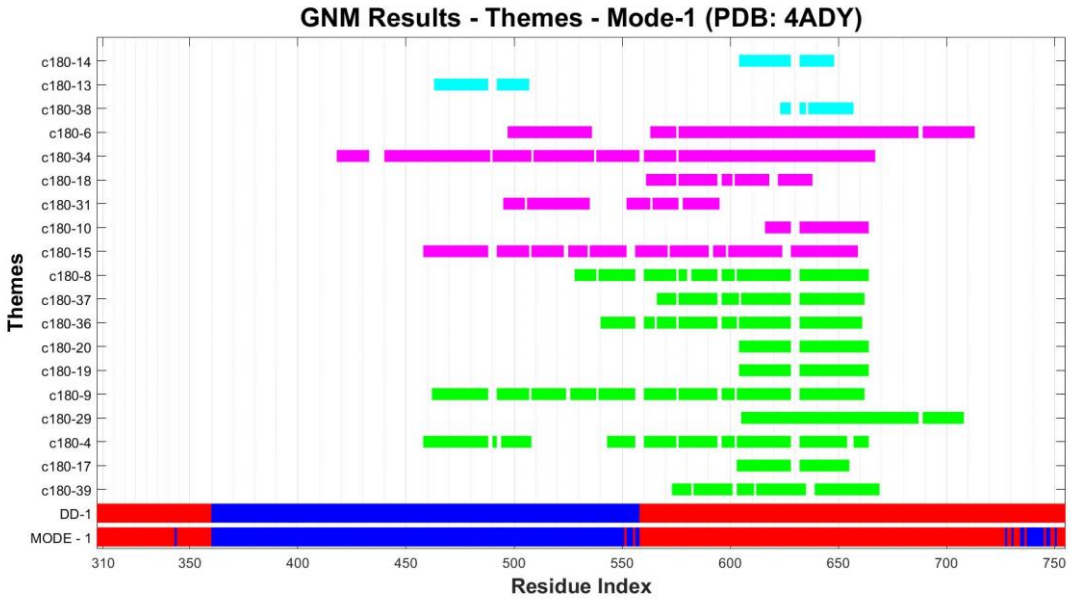


Figure A29. Slowest GNM mode, dynamic domains and themes (PDB ID: 4ADY).

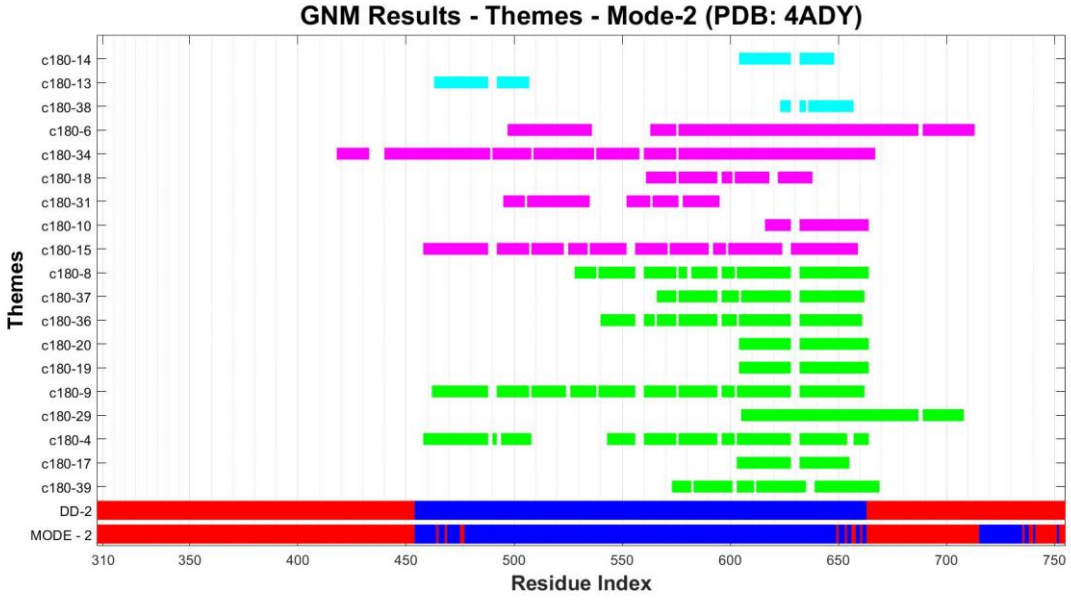


Figure A30. Second slowest GNM mode, dynamic domains and themes (PDB ID: 4ADY).

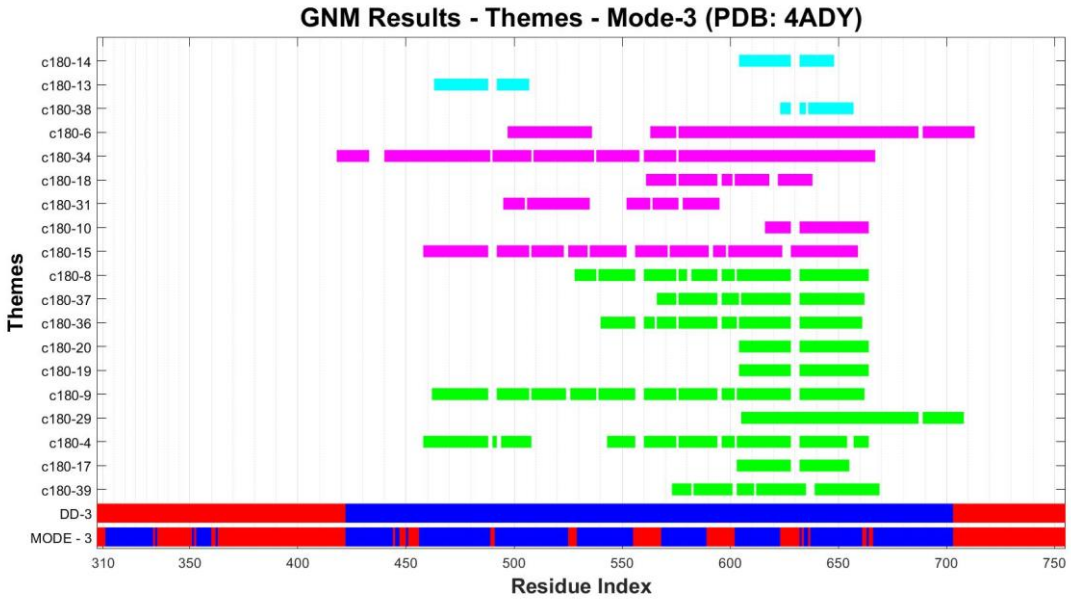


Figure A31. Third slowest GNM mode, dynamic domains and themes (PDB ID: 4ADY).

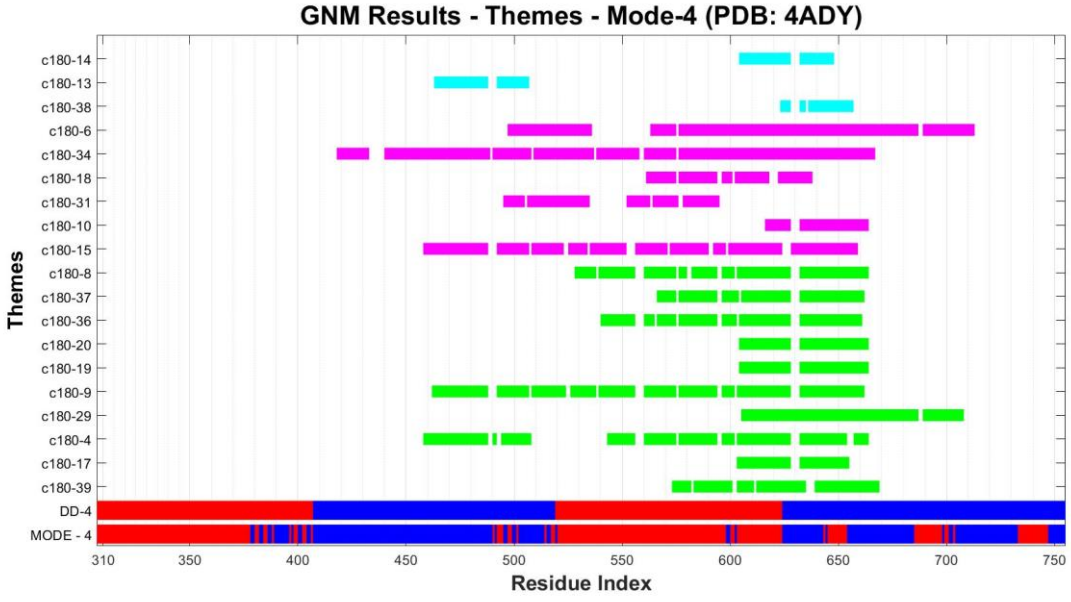


Figure A32. Fourth slowest GNM mode, dynamic domains and themes (PDB ID: 4ADY).

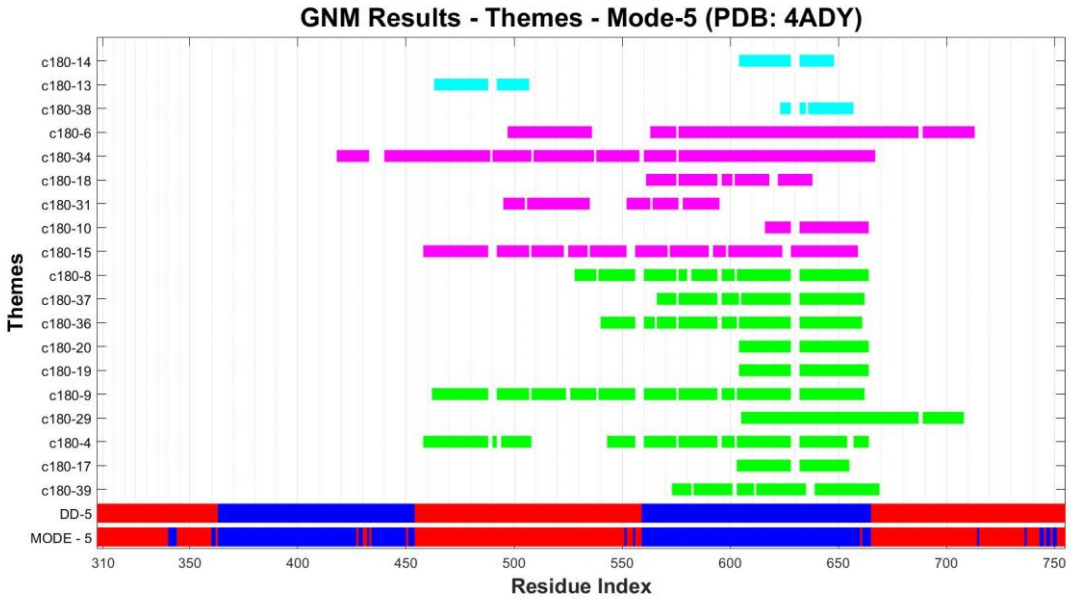


Figure A33. Fifth slowest GNM mode, dynamic domains and themes (PDB ID: 4ADY).

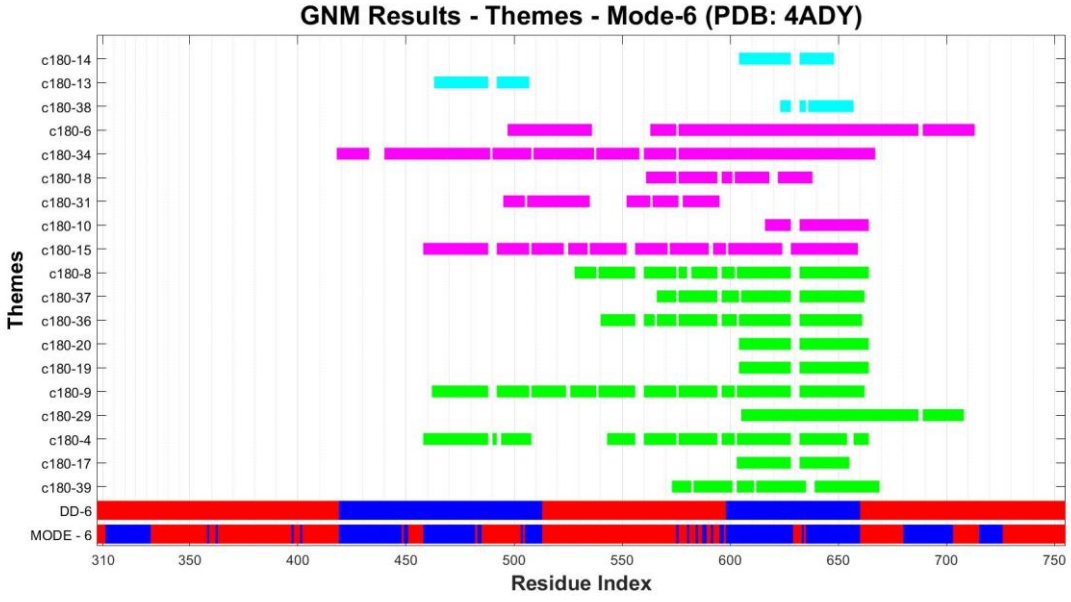


Figure A34. Sixth slowest GNM mode, dynamic domains and themes (PDB ID: 4ADY).

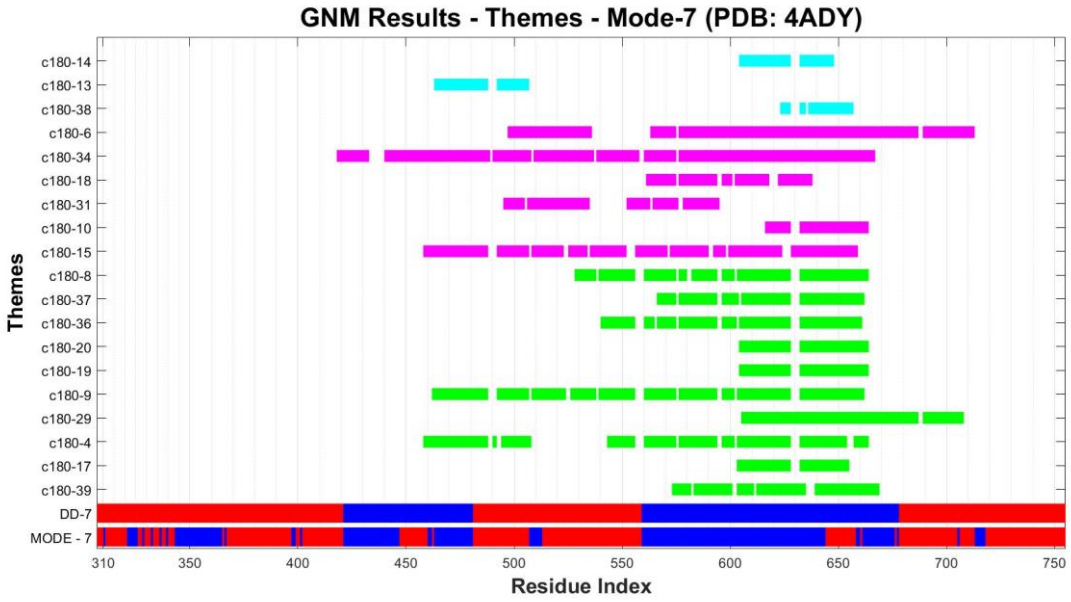


Figure A35. Seventh slowest GNM mode, dynamic domains and themes (PDB ID: 4ADY).

Table A4. All possible theme combinations, filtered with 3 residue overlap and 8 residue gap restriction.

C-1	**Theme**	**14940**	**14946**	**14942**	**14954**	**14937**	**14816**	**14861-1**
	Residue	14-53	55-120	126-170	176-220	227-265	271-309	314-353
C-2	**Theme**	**14940**	**14946**	**14942**	**14954**	**14937**	**14816**	**14950**
	Residue	14-53	55-120	126-170	176-220	227-265	271-309	314-408
C-3	**Theme**	**14940**	**14946**	**14942**	**14954**	**14938**	**14815-2**	**14951**
	Residue	14-53	55-120	126-170	176-220	227-287	289-353	356-415
C-4	**Theme**	**14940**	**14946**	**14942**	**14954**	**14939**	**14861-1**	**14951**
	Residue	14-53	55-120	126-170	176-220	227-309	314-353	356-415
C-5	**Theme**	**14940**	**14946**	**14942**	**14954**	**14939**	**14950**	
	Residue	14-53	55-120	126-170	176-220	227-309	314-408	
C-6	**Theme**	**14940**	**14946**	**14942**	**14813-1**	**14949**	**14815-3**	**14951**
	Residue	14-53	55-120	126-170	176-240	243-297	299-353	356-415
C-7	**Theme**	**14940**	**14946**	**14942**	**14813-1**	**14815-1**	**14861-1**	**14951**
	Residue	14-53	55-120	126-170	176-240	244-309	314-353	
C-8	**Theme**	**14940**	**14946**	**14942**	**14813-1**	**14815-1**	**14950**	
	Residue	14-53	55-120	126-170	176-240	244-309	314-408	
C-9	**Theme**	**14945**	**14942**	**14954**	**14937**	**14816**	**14861-1**	**14951**
	Residue	16-120	126-170	176-220	227-265	271-309	314-353	356-415
C-10	**Theme**	**14945**	**14942**	**14954**	**14937**	**14816**	**14950**	
	Residue	16-120	126-170	176-220	227-265	271-309	314-408	
C-11	**Theme**	**14945**	**14942**	**14954**	**14938**	**14815-2**	**14951**	
	Residue	16-120	126-170	176-220	227-287	289-353	356-415	
C-12	**Theme**	**14945**	**14942**	**14954**	**14939**	**14861-1**	**14951**	
	Residue	16-120	126-170	176-220	227-309	314-353	356-415	

Table A4. All possible theme combinations, filtered with 3 residue overlap and 8 residue gap restriction. cont.

C-13	**Theme**	**14945**	**14942**	**14954**	**14939**	**14950**	
	Residue	16-120	126-170	176-220	227-309	314-408	
C-14	**Theme**	**14945**	**14942**	**14813-1**	**14949**	**14815-3**	**14951**
	Residue	16-120	126-170	176-240	243-297	299-353	356-415
C-15	**Theme**	**14945**	**14942**	**14813-1**	**14815-1**	**14861-1**	**14951**
	Residue	16-120	126-170	176-240	244-309	314-353	356-415
C-16	**Theme**	**14945**	**14942**	**14813-1**	**14815-1**	**14950**	
	Residue	16-120	126-170	176-240	244-309	314-408	

Table A5. AMI values for the correlation between the dynamic domains of each of the seven slowest GNM modes and theme combinations, filtered with 5 residues overlap and 10 residues gap limit.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5	Mode-6	Mode-7
Minimum	0.4625	0.5181	0.6231	0.5863	0.7117	0.6436	0.6900
Maximum	0.6200	0.7000	0.7619	0.7427	0.8934	0.7879	0.9126
Average	0.5116	0.6124	0.6846	0.6759	0.7920	0.6998	0.7970

Table A6. SMI values for the correlation between the dynamic domains of each of the seven slowest GNM modes and theme combinations, filtered with 5 residues overlap and 10 residues gap limit.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5	Mode-6	Mode-7
Minimum	89.896	105.283	129.736	118.365	128.158	111.925	133.278
Maximum	123.551	149.575	183.211	150.557	163.512	148.284	168.774
Average	101.571	123.233	150.718	132.373	142.470	125.637	145.788

Table A7. AMI values for the correlation between the dynamic domains of each of the seven slowest GNM modes and theme combinations, filtered with 5 residues overlap and 15 residues gap limit.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5	Mode-6	Mode-7
Minimum	0.4311	0.5181	0.6082	0.5863	0.6673	0.6436	0.6722
Maximum	0.6200	0.7371	0.7984	0.7969	0.8978	0.7879	0.9126
Average	0.5073	0.6306	0.6789	0.6935	0.7812	0.7111	0.7853

Table A8. SMI values for the correlation between the dynamic domains of each of the seven slowest GNM modes and theme combinations, filtered with 5 residues overlap and 15 residues gap limit.

	Mode-1	Mode-2	Mode-3	Mode-4	Mode-5	Mode-6	Mode-7
Minimum	77.154	103.934	121.997	111.441	115.658	109.057	119.301
Maximum	123.551	164.490	204.081	163.116	164.005	151.004	170.511
Average	98.744	124.644	146.834	133.311	138.537	125.301	141.081

APPENDIX B: ADDITIONAL FIGURES ABOUT PERTURBATION ANALYSIS

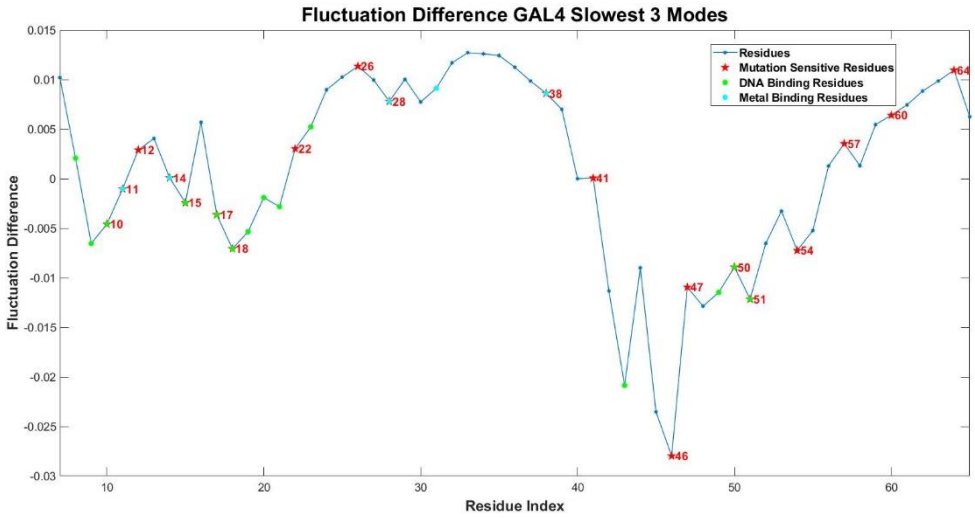


Figure B1. GAL4 perturbation analysis - Fluctuation difference vs residue index in slowest three GNM modes. (PDB ID: 1D66)

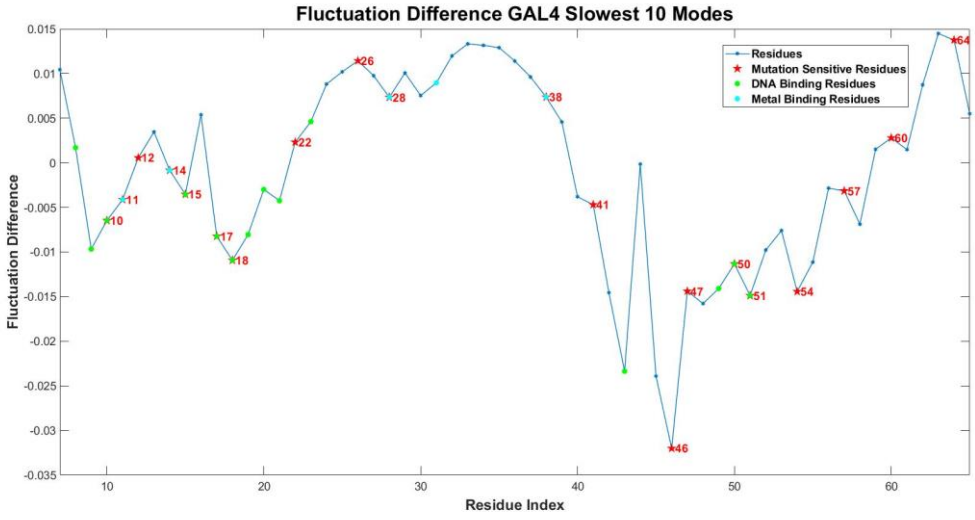


Figure B2. GAL4 perturbation analysis - Fluctuation difference vs residue index in slowest ten GNM modes. (PDB ID: 1D66)

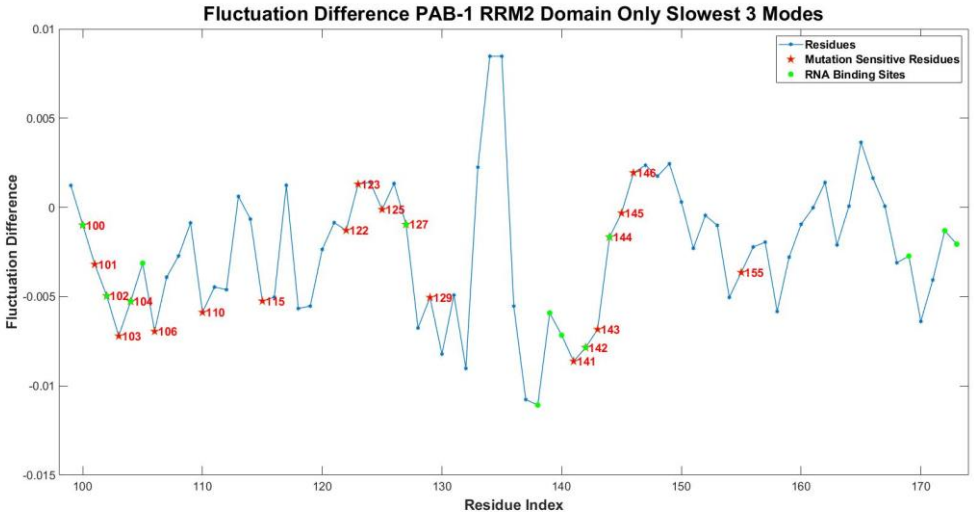


Figure B3. PAB1 monomer RRM2 domain perturbation analysis - Fluctuation difference vs residue index in slowest three GNM modes. (PDB ID: 1CVJ)

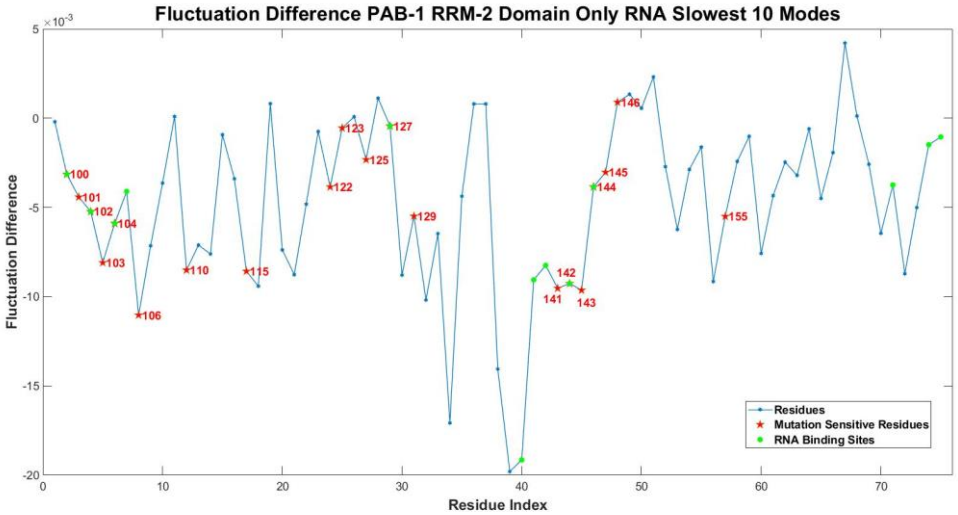


Figure B4. PAB1 monomer RRM2 domain perturbation analysis - Fluctuation difference vs residue index in slowest ten GNM modes. (PDB ID: 1CVJ)

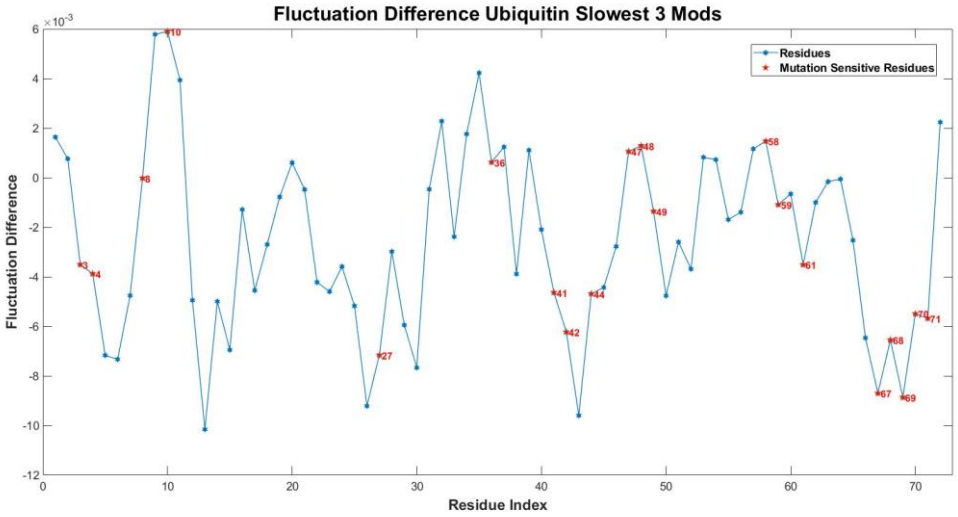


Figure B5. Ubiquitin perturbation analysis - Fluctuation difference vs residue index in slowest three GNM modes. (PDB ID: 1UBQ)

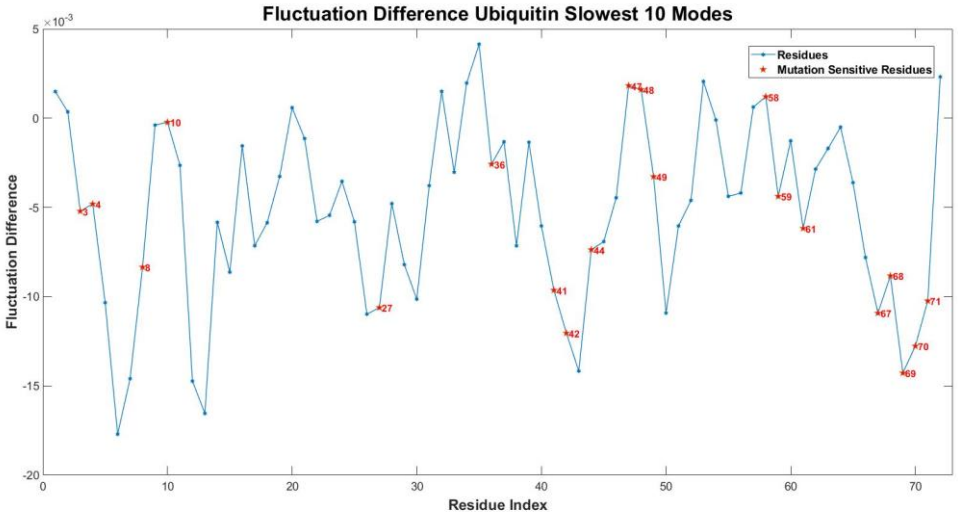


Figure B6. Ubiquitin perturbation analysis - Fluctuation difference vs residue index in slowest ten GNM modes. (PDB ID: 1UBQ)

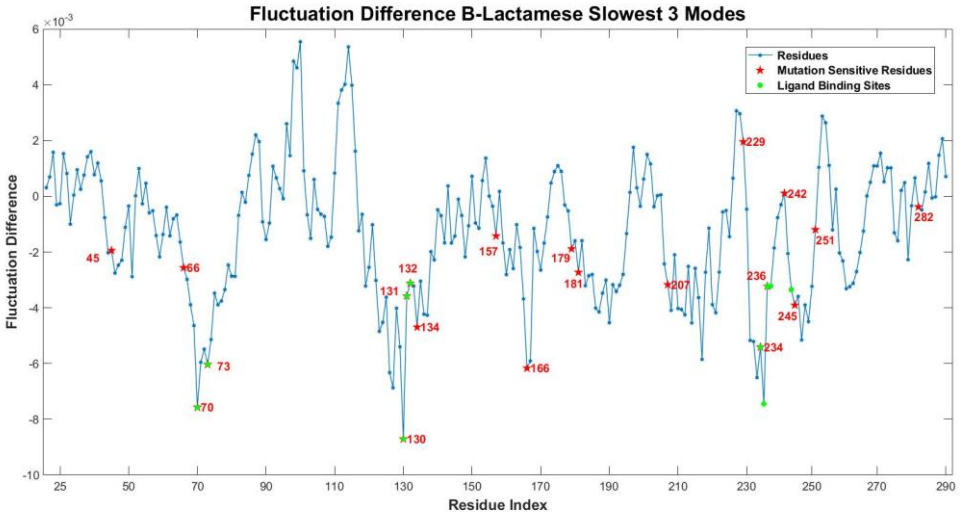


Figure B7. β-Lactamase perturbation analysis - Fluctuation difference vs residue index in slowest three GNM modes. (PDB ID: 1XPB)

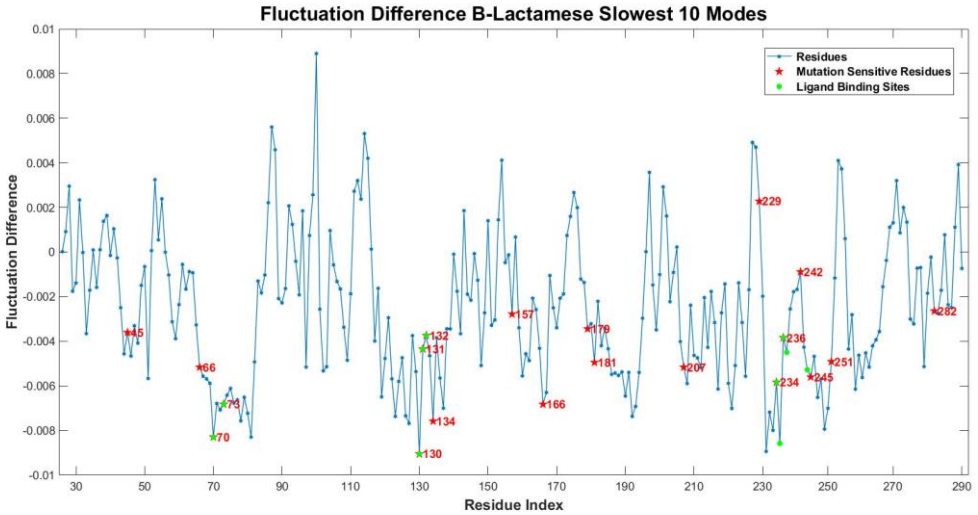


Figure B8. β-Lactamase perturbation analysis - Fluctuation difference vs residue index in slowest ten GNM modes. (PDB ID: 1XPB)

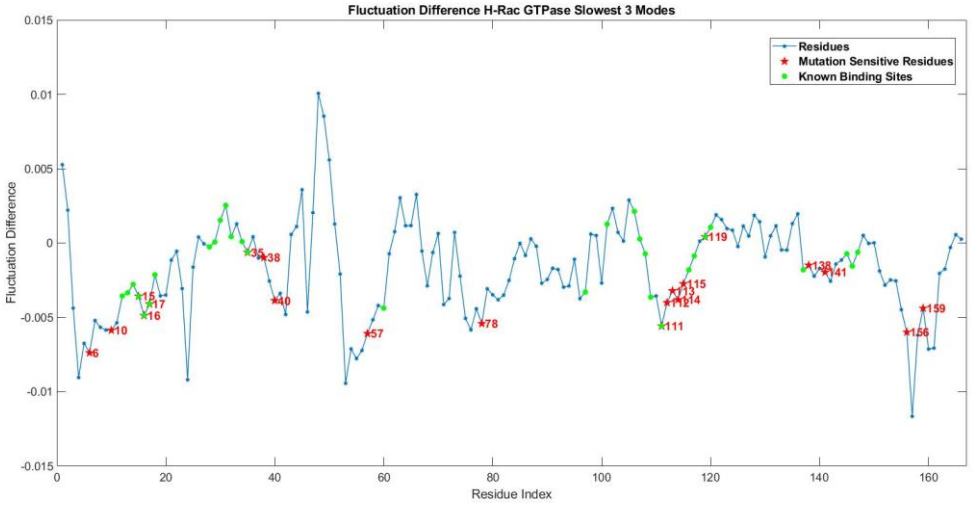


Figure B9. H-Ras GTPase perturbation analysis - Fluctuation difference vs residue index in slowest three GNM modes. (PDB ID: 3K8Y)

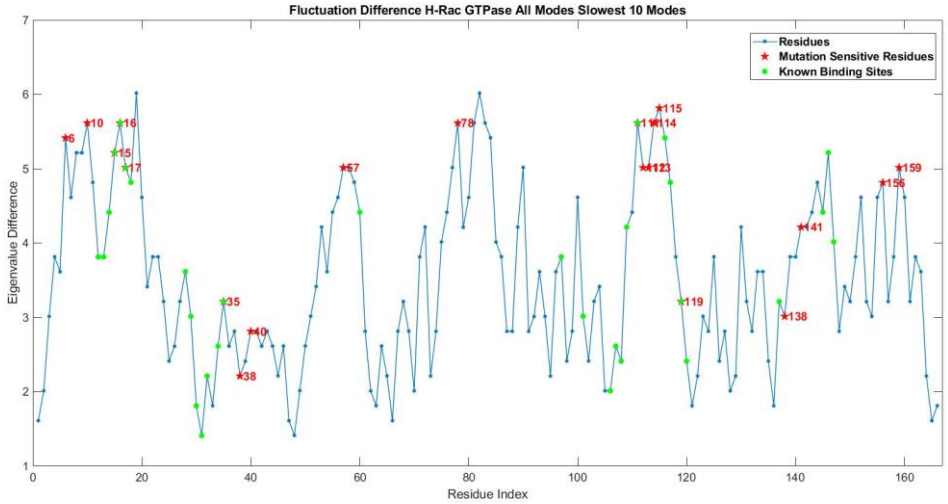


Figure B10. H-Ras GTPase perturbation analysis - Fluctuation difference vs residue index in slowest ten GNM modes. (PDB ID: 3K8Y)