COMPUTATIONAL INSIGHTS INTO THE MUTATIONS OF AN OMEGA-TRANSAMINASE USING EVOLUTIONARY FOOTPRINTS

by
Hande Abeş

Submitted to Graduate School of Natural and Applied Sciences
in Partial Fulfillment of the Requirements
for the Degree of Master of Science in
Chemical Engineering

Yeditepe University
2023

COMPUTATIONAL INSIGHTS INTO THE MUTATIONS OF AN OMEGA-TRANSAMINASE USING EVOLUTIONARY FOOTPRINTS

APPROVED BY:

Prof. Dr. Nihan Çelebi ……
(Thesis Supervisor)
(Yeditepe University)

Prof. Dr. Türkan Haliloğlu ……
(Thesis Co-Supervisor)
(Boğaziçi University)

Prof. Dr. Tuğba Davran Candan ……
(Yeditepe University)

Prof. Dr. Fethiye Aylin Sungur ……
(Istanbul Technical University)

Assist. Prof. Dr. Burcu Dedeoğlu ……
(Gebze Technical University)

DATE OF APPROVAL: ..../..../2023

I hereby declare that this thesis is my own work and that all information in this thesis has been obtained and presented in accordance with academic rules and ethical conduct. I have fully cited and referenced all material and results as required by these rules and conduct, and this thesis study does not contain any plagiarism. If any material used in the thesis requires copyright, the necessary permissions have been obtained. No material from this thesis has been used for the award of another degree.

I accept all kinds of legal liability that may arise in case contrary to these situations.

Name, Last name Hande Abeş

Signature …..…….……..

ACKNOWLEDGEMENTS

My supervisor Prof. Nihan Çelebi is a mary sue as a role model and she leaded to shape my idea of who I want to be and what I want to do in my life as a young researcher. I will remain always grateful to her for all contributions to my academic development, broadening my perspective by treating me to think beyond the limits, teaching how the point of view could be changed if required to reach the truth of the matter. Studying under her guidance for six years was a wondrous experience. I am indebted to her for providing the chance and encouragement to attend international conferences, and to conduct my own project starting from my undergraduate level. She provided a limitless support to make a number of unimaginable opportunity a real during this whole excellent time.

It was a great honor to study under the supervision of Prof. Türkan Haliloğlu as my co-supervisor in this thesis, and also I appreciate her acceptance our offer to conduct this project together even though she had not known me before starting to study, as well as her support and understanding. Acquiring knowledge in a different field of computational modeling with her extensive point of view and experience was a golden opportunity.

All department members of Yeditepe University - Chemical Engineering have my gratitude for their support and having a good faith belief on me. They encourage me every time to take forward steps in academic world. I have special thanks for Assist. Prof. Cem Levent Altan for his mentorship at critical decision points and his advices not only about academic issues but also to lighten the road that I walked by sharing his profound experiences.

I would like to thank my room-mate Dr. Yeşim Çamlısoy, for her self-sacrifice and teaching me all the details of assistant work as well as for introducing me into computational chemistry world in my undergrad degree, together with Dr. Sezen Alsancak.

I would like to thank Özge Duman, who is a PhD student studying with Prof. Türkan Haliloğlu at Boğaziçi University. Her contribution was incontrovertibly one of the critical points of this project.

I am grateful to Barış Emek for his endless support at every time that I abandon myself to despair. His helps turned every bottleneck points into easily conquerable steps.

At last but not least, I would like to thank my family for their love and support.

ABSTRACT

COMPUTATIONAL INSIGHTS INTO THE MUTATIONS OF AN OMEGA-TRANSAMINASE USING EVOLUTIONARY FOOTPRINTS

Transaminases have become highly remarkable enzymes for manufacturing enantiopure chiral amines. These chiral amines hold great significance as intermediates in the production of pharmaceuticals and various fine chemicals. Nonetheless, the restricted activity of transaminases frequently presents a challenge in their application for biocatalysis. Consequently, developing novel enzymes with enhanced activity and the ability to function effectively in industrial process conditions has become a crucial matter. Although a variety of studies focus on the development of new variants by mutations in the active site of the enzymes, it remains unclear how beneficial the mutations far from the active site confer improved catalytic properties. Getting an insight into this issue would aid the design of new enzymes with higher catalytic activities. Moreover, for natural evolution perspective, it is also shrouded in mystery how mutations far from the active site affect the active sites' functions and also stability. In this study, the objective is to develop an understanding on the effects of mutations realized during the natural evolution of an omega-transaminase, using computational tools. To achieve the stated purpose, a modern-day 4-aminobutyrate transaminase, and its three ancestral variants having improved catalytic activity and substrate promiscuity were chosen to study. All mutations that occurred during their evolution of this family were outside the active sites of the enzymes. Thus, effects of the mutations that occurred at the dynamic domain interfaces were put under the scope so that mutations could be classified and the important ones for the catalytic activity in addition to substrate selectivity could be identified. Functional collective motions of the variants were evaluated by using computational methods, Gaussian Network Model (GNM) and Anisotropic Network Model (ANM). It was revealed that the mutations at cross-correlation difference regions and close to hinge points have significant influence on the loop dynamics near the active sites resulting in the activity and selectivity changes. Molecular dynamic simulations results led to verification of the conservation of active site interactions in all variants, as well as understanding of the structural changes in the global behavior of the variants determining their thermotolerances.

ÖZET

EVRİMSEL AYAK İZLERİ KULLANARAK BİR OMEGA-TRANSAMİNAZ MUTASYONLARINA İLİŞKİN HESAPLAMALI ANLAYIŞLAR

Transaminazlar, enantiopür kiral aminlerin üretimi için oldukça dikkate değer enzimler haline gelmiştir. Kiral aminler, farmasötiklerin ve çeşitli ince kimyasalların üretiminde ara ürünler olarak büyük önem taşır. Bununla birlikte, transaminazların sınırlı aktivitesi, sıklıkla biyokataliz uygulamalarında zorluk teşkil etmektedir. Bu nedenle geliştirilmiş aktiviteye ve endüstriyel proses koşullarında etkin bir şekilde işlev görme yeteneğine sahip yeni enzimler geliştirmek çok önemli bir konu haline gelmiştir. Çeşitli çalışmalar, enzimlerin aktif bölgesindeki mutasyonlar yoluyla yeni varyantların geliştirilmesine odaklansa da, aktif bölgeden uzaktaki mutasyonların gelişmiş katalitik özellikler sağlanmasına katkısı yeterince anlaşılamamıştır. Bu etki hakkında fikir edinmek, daha yüksek katalitik aktivitelere sahip yeni enzimlerin tasarımına yardımcı olacaktır. Ayrıca, doğal evrim açısından, aktif bölgeden uzaktaki mutasyonların, aktif bölgelerin fonksiyonlarını ve stabilitesini nasıl etkilediği de hala bir sır perdesi ardındadır. Bu çalışmada, hesapsal yaklaşımlarla bir omega-transaminazın doğal evrimi sırasında gerçekleşen mutasyonların etkileri hakkında bir anlayış geliştirmek amaçlanmaktadır. Belirtilen amaca ulaşmak için, modern bir 4-aminobutirat transaminaz ve gelişmiş katalitik aktiviteyle birlikte substrat seçiciliğinde daha geniş bir yelpazeye sahip olan dört atasal varyant çalışmalarda kullanılmak üzere seçildi. Bu ailenin evrimi sırasında meydana gelen tüm mutasyonlar, enzimlerin aktif bölgelerinin dışında yer almaktadır. Bu sebeple, dinamik domain arayüzlerinde meydana gelen mutasyonların etkileri, mutasyonların sınıflandırılması ve substrat seçiciliği yanında katalitik aktivite için önemli olanların tespit edilmesi amacıyla mercek altına alındı. Varyantların fonksiyonel kollektif hareketleri Gaussian Ağ Modeli (GNM) ve Anizotropik Ağ Modeli (ANM) hesaplama yöntemleri kullanılarak değerlendirildi. Çapraz korelasyon fark bölgelerindeki ve menteşe noktalarına yakın mutasyonların, aktivite ve seçicilik değişikliklerine neden olan aktif bölgelere yakın döngülerin dinamikleri üzerinde önemli bir etkiye sahip olduğu ortaya çıktı. Moleküler dinamik simülasyon sonuçları, tüm varyantlarda aktif bölge içerisindeki etkileşimlerinin korunduğunu doğruladı ve ayrıca varyantların termo-toleranslarını belirleyen global davranışındaki yapısal değişikliklerin anlaşılmasını sağladı.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iv

ABSTRACT v

ÖZET vi

LIST OF FIGURES x

LIST OF TABLES xv

LIST OF SYMBOLS/ABBREVIATIONS xvi

1. INTRODUCTION 1

2. THEORETICAL BACKGROUND 4

2.1. TRANSAMINASES 5

2.1.1. ω - Transaminases 5

2.1.2. Industrial Usage of Transaminases 6

2.1.3. A 4-Aminobutyrate Transaminase, KES23360 7

2.2. COMPUTATIONAL METHODS DEVELOPED FOR THE IDENTIFICATION OF THE DYNAMIC MOTIONS IN PROTEINS 10

2.2.1. Gaussian Network Model (GNM) 11

2.2.2. Anisotropic Network Model (ANM) 15

2.2.3. Molecular Dynamic Simulations (MD) 16

2.2.3.1. Force Fields 18

2.2.3.2. The Algorithm Behind MD Calculations (Classical Mechanics) 19

2.2.3.3. Statistical Mechanics 21

2.2.3.4. Timescales 22

2.2.3.5. Solvent Effect 23

2.2.4. Langevin Dynamics 23

3. METHODOLGY 26

3.1. DETERMINATION OF DIFFERENCE VECTOR MAGNITUDES 26

3.2. ANM-LD SIMULATION PROTOCOL 27

3.3. USAGE OF GNM IN CALCULATIONS 27

3.4. ASSIGNMENT OF ACTIVE SITE RESIDUES AND THEIR FUNCTIONS 28

3.5. MOLECULAR DYNAMIC SIMULATIONS 29
3.5.1. PLP Parametrization 30
3.5.2. Preparation of Enzyme Systems 30
3.5.2.1. Assessment of Protonation States 30
3.5.3. Creation of PLP-bound Transaminase Unit 31
3.5.4. MD Simulation Protocol 35
3.5.5. Minimization 35
3.5.6. Heating 36
3.5.7. Equilibration 36
3.5.8. Production 36
4. RESULTS AND DISCUSSION 37
4.1. DETERMINATION OF THE DRIVING ANM MODES IN NATURAL EVOLUTION OF THE VARIANTS BY ANM-LD SIMULATIONS 38
4.2. DETERMINATION OF THE GNM MODES CORRESPONDING WITH THE SELECTED ANM MODES 42
4.3. HINGE POINT DETERMINATION 48
4.3.1. Mutations Close to Hinges 50
4.4. CROSS CORRELATION DIFFERENCE POINTS AND CLOSE MUTATIONS 55
4.5. EVALUATION OF THE DYNAMIC MOTION OF CATALYTIC RESIDUE ARG267, A LOCKKEEPER FOR SUBSTRATE SELECTIVITY 69
4.5.1. Interactions of ARG267 Obtained by MD Simulations 70
4.5.2. Evaluation of the Changes in Loop Dynamics 74
4.6. VERIFICATION OF THE CONSERVED ACTIVE SITES IN EACH GENERATION 76
4.7. THE REASON BEHIND LOW THERMOTOLERANCES OF THE TWO END VARIANTS OF EVOLUTION 78
5. CONCLUSION 80
6. FUTURE WORK 82
REREFERENCES 83
APPENDIX A 88

APPENDIX B 89
APPENDIX C 94
APPENDIX D 114

LIST OF FIGURES

Figure 1.1. Workflow of computational study 2

Figure 2.1. Transamination reaction mechanism 6

Figure 2.2. Phylogenetic tree of KES23360 8

Figure 3.1. Active site residues 29

Figure 3.2. PLP bonded to LYS288 unit with adjacent residues 32

Figure 3.3. Frozen atoms of PLP bound LYS unit in geometry optimization 33

Figure 3.4. LYP 34

Figure 3.5. NLP 34

Figure 3.6. MD simulation steps 35

Figure 4.1. Difference vector-ANM overlaps for N48-N15 39

Figure 4.2. Difference vector-ANM overlaps for N15-N6 40

Figure 4.3. Difference vector-ANM overlaps for N6-KES23360 41

Figure 4.4. Difference vector-ANM overlaps for N48-KES23360 41

Figure 4.5. ANM3 overlap with GNM1 for N48-N15 42

Figure 4.6. ANM1&ANM2 overlaps with GNM2 for N48-N15 43

Figure 4.7. ANM1&ANM2 overlaps with GNM2 for N15-N6 45

Figure 4.8. ANM3 overlap with GNM4 for N15-N6 46

Figure 4.9. ANM1&ANM3 overlaps with GNM modes for N6-KES23360 47

Figure 4.10. ANM1&ANM2 overlaps with GNM2 for N48-KES23360 48

Figure 4.11. GNM2 mode shapes and hinges for N48, N15, N6 and KES23360 50

Figure 4.12. Mutations and hinges of N48 on 3D structure 51

Figure 4.13. Mutated residues close to hinges of N48..... 51

Figure 4.14. Mutations and hinges of N15 on 3D structure 52

Figure 4.15. Mutated residues close to hinges of N15..... 52

Figure 4.16. Mutations and hinges of N6 on 3D structure 53

Figure 4.17. Mutated residues close to hinges of N6 structure..... 53

Figure 4.18. Mutations and hinges of KES23360 on 3D structure..... 54

Figure 4.19. Mutated residues close to hinges of KES23360 54

Figure 4.20. Cross-map difference of N48-N15 for GNM2 56

Figure 4.21. Relative placement of hinges and cross correlation differences in N48-N15 . 57

Figure 4.22. Cross correlation difference points and mutations in N48 58

Figure 4.23. Cross-map difference of N15-N6 for GNM2 59

Figure 4.24. Relative placement of hinges and cross correlation differences in N15-N6 ... 60

Figure 4.25. Cross correlation difference points and mutations in N15 for GNM2 60

Figure 4.26. Cross-map difference of N15-N6 for GNM4 62

Figure 4.27. Cross correlation difference points and mutations in N15 for GNM4 63

Figure 4.28. Cross-map difference of N6-KES23360 for GNM2 64

Figure 4.29. Relative placement of hinges and closer cross correlation differences in N6-KES23360..... 64

Figure 4.30. Cross correlation difference points and mutations in N6-KES23360 for GNM2 65

Figure 4.31. Cross-map difference of N6-KES23360 for GNM4 66

Figure 4.32. Cross correlation difference points and mutations in N6-KES23360 for GNM4 67

Figure 4.33. Cross-map difference of N48-KES23360 for GNM2 68

Figure 4.34. GNM2 mode shape, cross correlation difference points, close mutations and active site representation for N48-KES23360 system 69

Figure 4.35. H-bond established between ARG267 and TYR59 in apo MD simulation of N48 71

Figure 4.36. H-bond established between ARG267 and TYR59 in apo MD simulation of KES23360 71

Figure 4.37. H-bond established between ARG267 and TYR59 in active site containing free PLP MD simulation of N48 72

Figure 4.38. H-bond established between ARG267 and TYR59 in active site containing free PLP MD simulation of KES23360 72

Figure 4.39. H-bond established between ARG267 and TYR59 in active site containing covalently bound PLP MD simulation of N48 73

Figure 4.40. H-bond established between ARG267 and TYR59 in active site containing covalently bound PLP MD simulation of KES23360 73

Figure 4.41. Cross-map Difference of N48-N15 for the averages of the first five slowest GNM modes 75

Figure 4.42. Close mutations to catalytic arginine in N48 76

Figure 4.43. MD snapshots of the 19th nanoseconds of apo-enzyme structures 79

Figure B.1. Difference vector representation of N48-N15 89

Figure B.2. Difference vector representation of N15-N6 90

Figure B.3. Difference vector representation of N6-KES23360 90

Figure B.4. N48-N15 Sequence Alignment 91

Figure B.5. N15-N6 Sequence Alignment 92

Figure B.6. N6-KES23360 Sequence Alignment 93

Figure C.1. ANM1 overlaps with GNM1 & GNM3 for N48-N15 94

Figure C.2. ANM1 overlaps with GNM4 & GNM5 for N48-N15 95

Figure C.3. ANM2 overlaps with GNM1 & GNM3 for N48-N15 96

Figure C.4. ANM2 overlaps with GNM4 & GNM5 for N48-N15 97

Figure C.5. ANM3 overlaps with GNM2 & GNM3 for N48-N15 98

Figure C.6. ANM3 overlaps with GNM4 & GNM5 for N48-N15 99

Figure C.7. ANM1 overlaps with GNM1 & GNM3 for N15-N6 100

Figure C.8. ANM1 overlaps with GNM4 & GNM5 for N15-N6 101

Figure C.9. ANM2 overlaps with GNM1 & GNM3 for N15-N6 102

Figure C.10. ANM2 overlaps with GNM4 & GNM5 for N15-N6 103

Figure C.11. ANM3 overlaps with GNM1 & GNM2 for N15-N6 104

Figure C.12. ANM3 overlaps with GNM3 & GNM5 for N15-N6 105

Figure C.13. ANM1 overlaps with GNM1 & GNM3 for N6-KES23360 106

Figure C.14. ANM1 overlaps with GNM4 & GNM5 for N6-KES23360 107

Figure C.15. ANM3 overlaps with GNM1 & GNM2 for N6-KES23360 108

Figure C.16. ANM3 overlaps with GNM3 & GNM5 for N6-KES23360 109

Figure C.17. ANM1 overlaps with GNM1 & GNM3 for N48-KES23360 110

Figure C.18. ANM1 overlaps with GNM4 & GNM5 for N48-KES23360 111

Figure C.19. ANM2 overlaps with GNM1 & GNM3 for N48-KES23360 112

Figure C.20. ANM2 overlaps with GNM4 & GNM5 for N48-KES23360 113

Figure D.1. Cross-map Difference of N15-N6 for the averages of the first five slowest GNM modes 114

Figure D.2. Cross-map Difference of N6-KES23360 for the averages of the first five slowest GNM modes 115

LIST OF TABLES

Table 2.1. Specific activity data of four transaminase variants for 10 substrates..... 10

Table 2.2. Timescales of some protein simulation types..... 23

Table 3.1. List of PDB ID's of the studied transaminase variants..... 26

Table 4.1. Hinge residues determined by Hingeprot for the second slowest GNM mode .. 49

Table 4.2. Mutation points close to hinges of GNM2 55

Table 4.3. MD Results of N48 and KES23360 indicating the interactions of the active site residues 77

Table A.1. Proposed active site residues and their functions 88

LIST OF SYMBOLS/ABBREVIATIONS

∇	Gradient
A	Elastic free energy
ΔA	Elastic free energy change
a	Acceleration
C	Single parameter in the Hookean potential
Cα	Alpha-Carbon atom
$CC_{avg}(i,j)$	Average cross correlation of the ith and jth residues
E	Potential enegy
F	Force
$\boldsymbol{H}$	Hessian matrix
$\boldsymbol{H_{i,j}}$	Hessian matrix element of raw i, column j
K	Configurational factor constant
K_b	Bond strength force constant
K_θ	Bond angle force constant
k	Mode number
k_B	Boltzmann constant
l	Bond length
M	Number of simulation steps
m	Mass
$\boldsymbol{R}$	Fluctuation
$\Delta\boldsymbol{R_{ij}}$	Junction fluctuations of the ith and jth alpha-Carbons
$[\Delta\boldsymbol{R}]$	Column vector formed by the fluctuations
s_{ij}	Instantaneous distances between residues i and j
s_{ij}^0	Equilibrium distances between residues i and j
T	Temperature
^{T}	Transpose operator
t	Time
tr	Trace operator
$\boldsymbol{U}$	Orthogonal matrix
$\boldsymbol{U(i,k)}$	Eigenvector of residue i of the kth mode

$\boldsymbol{U(j,k)}$	Eigenvector of residue j of the kth mode
$\boldsymbol{u_k}$	Eigenvector
q	Partial charge
V	Harmonic potential
v	Velocity
$W(\Delta R_{ij})$	Gaussian-distributed fluctuations
$w(k)$	Eigenvalue of the kth mode
Z_N	Configurational factor for a protein with N residues
$\boldsymbol{\Gamma}$	Kirchhoff matrix
$\boldsymbol{\Gamma^{-1}}$	Inverse of the Kirchhoff matrix
γ	Phase angle
γ_f	Force constant
ε	Electroctatic interactions
θ	Bond angle
$\boldsymbol{\Lambda}$	Diagonal matrix of the eigenvalues of the Kirchhoff matrix
λ	Eigenvalue
π	Pi
3D	Three-dimensional
ANM	Anisotropic network model
AMBER	Assisted model building with energy refinement
ALA, A	Alanine
ASN, N	Asparagine
ASR	Ancestral Reconstruction Technique
ASP, D	Aspartic acid
ARG, R	Arginine
CHARMM	Chemistry at Harvard molecular mechanics
c.c.	Cross correlation
CYS, C	Cysteine
DV	Difference vector
e.e.	Enantiomeric excess

ENM	Elastic network model
GAFF	Generalized Amber force field
GLU, E	Glutamate
GLY, G	Glycine
GNM	Gaussian network model
GROMOS	Groningen molecular simulation
HIS, H	Histidine
ILE, I	Isoleucine
LD	Langevin dynamics
LEU, L	Leucine
LYP	Covalently bonded Lysine
LYS, K	Lysine
MD	Molecular dynamics
MET, M	Methionine
mVT	Grand canonical Ensemble
NLP	Covalently bonded PLP
NMA	Normal mode analysis
NPT	Isobaric-isothermal ensemble
NVE	Micro canonical ensemble
NVT	Canonical Ensemble
OPLS	Optimized potentials for liquid simulations
PDB	Protein data bank
PHE, F	Phenylalanine
PMP	Pyridoxamine phosphate
PLP	Pyridoxal-5′-phosphate
PRO, P	Proline
SER, S	Serine
TA	Transaminase
THR, T	Threonine
TRP, W	Tryptophan
TYR, Y	Tyrosine
VAL, V	Valine

1. INTRODUCTION

Transaminases (TAs) have been extensively focused on for their applications in synthesizing amino acids and chiral amines. These enzymes are crucial in the metabolism of amino acids and can be found widely in microorganisms and eukaryotic cells. Transaminases belong to the group of enzymes that depend on pyridoxal 5'-phosphate (PLP) as a cofactor. They facilitate the reversible transfer of an amino group from an amine donor to a ketone or aldehyde substrate. TAs can be categorized into two main types: α-transaminases (α-TAs) and ω-transaminases (ω-TAs), based on the relative position of the amine group being transferred in relation to the carboxyl group of the substrate. α-TAs specifically require the presence of a carboxylic group in the α-position to the carbonyl functionality of the substrate. On the other hand, ω-TAs encompass all other types of TAs where at least one of the two substances involved is not an α-amino acid or α-keto acid. In fact, they have the potential to accept virtually any ketone or amine, making them highly versatile and of great interest to the pharmaceutical industry. ω-TAs offer several advantages over their α-counterparts. They exhibit higher equilibrium constants, providing more favorable reaction conditions. Additionally, ω-TAs possess broader substrate specificity, high enantioselectivity, and do not require cofactor regeneration. These characteristics make them highly useful in various applications and preferable over hydrolase and dehydrogenase enzymes in certain scenarios [1]. While TAs offer several advantages such as broad substrate specificity and high stability, their industrial usage has been restricted due to the low equilibrium constants and limited operability conditions associated with the TA reactions [2]. Although a number of improvements have been made in the activity of specific processes and reactions, studies are still ongoing for the regulation of the activity of transaminases [3].

In the last decades, enzyme engineering studies usually focus on the development of new variants by mutations in the active site of the enzymes or usage of directed evolution to improve enzyme activities. However, understanding the mechanisms by which mutations occurring far from the active site can enhance catalytic properties in enzymes remains puzzle. The impact of mutations that are distal to the active site on its function and stability is also a mystery in natural evolution. Thus, the current challenge is to figure out how nature alters the function and biophysical properties of the enzymes by amino acid substitutions in both allosteric sites and active sites during evolution.

In this study, main aim is to get insights into the mutations which are outside the active site of an ω-transaminase occurred in natural evolution by a computational approach. The ω-transaminase put under the scope in this work is a modern-day 4-aminobutyrate transaminase, KES23360. Wilding et al. [4] examined this transaminase as well as its several ancestral variants experimentally and they discovered that all realized mutations in evolution are outside the active site regions of these variants, but they show different catalytic activity, substrate selectivity and thermo-tolerance. Modern-day transaminase, KES23360 shows promiscuous activity towards 12-aminododecanoic acid, however this substrate specificity is broadened for the second and third generation variants. Moreover, up to 20 fold activity improvements were obtained for the ancestral variants with respect to KES23360. Thermostability of most ancestral variants are also respectively higher than the last generation. Apart from these, there is other significant experimental results for the most ancient variant indicating that it has almost no reasonable catalytic activity for any tested substrates and its melting point is same with the modern-day extant. In the light of these experimental results, in this study, it was aimed to categorize the mutations as activity determinant, selectivity determinant and thermotolerance determinant mutations relating with the dynamic mode transition regions.

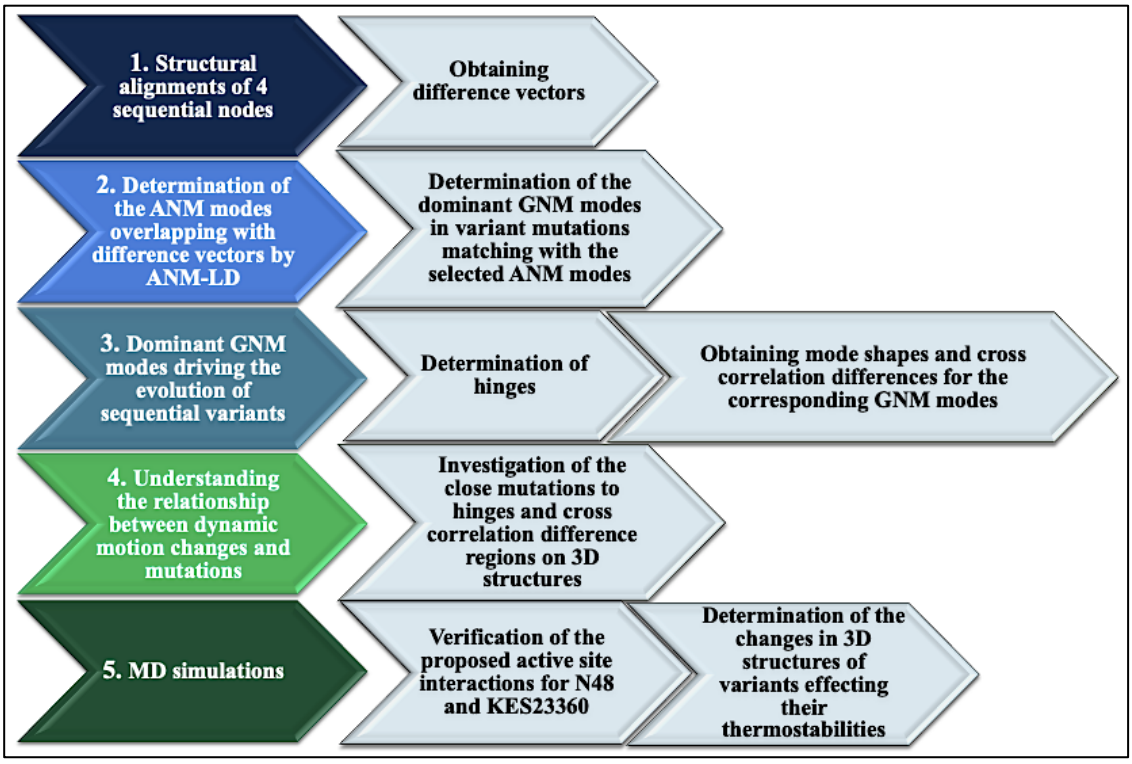

Figure 1.1. Workflow of computational study

All steps followed to achieve this aim are summarized in Figure 1.1. Necessary theoretical information about the computational methods used in the workflow steps, as well as about the transaminases are supplied in the next chapter, Chapter 2. The methodology followed to work with ANM and GNM for the determination of dominant dynamic mode domains and their relation with mutations in addition to MD simulation protocols are explained in Chapter 3. The results obtained after calculations and simulations leading to the determination of the dynamic motion changes in 3D structures of variants and their relation with the mutations are discussed in Chapter 4. All concluding remarks achieved by this computational study are given in Chapter 5 and, the studies may be conducted in the future by the light of this work are stated in the last chapter, Chapter 6.

2. THEORETICAL BACKGROUND

Enyzmes are the building blocks of every living creature. They are biological catalysts accelerating the chemical reactions that could take very long times without the help of the enzymes. Through billions of years of evolution, enzymes have emerged that can facilitate chemical transformations that would typically occur too slowly to be detected. Enzymes improve the catalytic efficiency for the specific reactions and substrates. Most enzymes require cofactors or coenzymes for their function and environmental factors such as pH, temperature, and solvent are highly deterministic. Despite this, they can function within a wide range of these conditions which makes them ideal components for the complex metabolic pathways [5]. Through billions of years of evolution, enzymes acquire new functions and their operating conditions have been modified. Precambrian thioredoxin enzymes (Trx) originated around 1.4 to 4 billion years ago, in example, exhibit remarkable stability, being up to 32°C more stable than their modern variants. Moreover, the oldest Trx enzymes display significantly greater activity than the ones found today at a pH level of 5. While the oxidoreductase property of Trx enzymes has remained constant, they have undergone adaptations over the course of 4 billion years to accommodate the changes in temperature and ocean acidity that characterize the evolution of Earth's global environment from ancient times to the present [6].

There are two models that describe how substrates bind to the active site of an enzyme. The first model, which is well-known as “lock and key mechanism”, suggests that the shape and chemical properties of the substrate perfectly match those of the active site on the enzyme. In this model, when the substrate enters the active site, it fits precisely, and the two components bind together, forming the enzyme-substrate complex. The second model, “induced fit model”, proposes that the enzyme and the substrate initially do not have an exact fitting in shape or alignment. Instead, the enzyme alters its shape since it has a flexible structure, and it let the substrate to induce its alignment within the active site, leading to a better fit between the two. These substrate binding to an enzyme is typically stabilized by local molecular interactions with the amino acid residues on the polypeptide chain [7].

Enzymes can be classified into six main categories: oxidoreductases, transferases, hydrolases, lyases, isomerases, and ligases. The general type of reactions taken place in

active sites specify each category, but they have many sub-categories depending on a wide range of specific reactions [7].

2.1. TRANSAMINASES

Transaminases are enzymes that need pyridoxal-5′-phosphate (PLP) as their cofactor. A number of enzymes like transaminases, are inactive until a cofactor binds them and activates the enzyme. Cofactors might be either metal ions or organic compounds like PLP that attach to the enzyme by covalent or non-covalent bonds [7]. Transaminases facilitate the transfer of an amino group from a donor amine to the carbonyl carbon atom of an α-keto acid, ketones, or an aldehyde. In recent years, amine transaminases have gained attention as effective biocatalysts for producing optically pure amines, which are precious intermediates or products in the pharmaceutical and agricultural industries. However, to utilize them efficiently as biocatalysts for stereoselective amination, significant protein engineering is often necessary due to their limited substrate range [8]. There are various biocatalytic approaches for generating enantiopure chiral amines through kinetic resolution utilizing hydrolases, monoamine oxidases, and ω-transaminases, a recent advancement involves the engineering of two amino acid dehydrogenases into amine dehydrogenases. However, currently, ω-transaminases remain the sole natural enzymes capable of directly synthesizing enantiopure chiral amines through the asymmetric amination of prochiral ketones [9].

2.1.1. ω - Transaminases

ω-Transaminases are enzymes that facilitate the transfer of an amino group from a primary amino donor to a carbonyl acceptor, utilizing pyridoxal 5-phosphate (PLP) as a cofactor. The reaction can be divided into two distinct half-reactions. Initially, the amino group is transferred to PLP, resulting in the formation of pyridoxamine phosphate (PMP). Subsequently, the amino group from PMP is transferred onto the carbonyl group of the acceptor. If the reaction mechanism is enlarged upon, the resting state of the enzyme involves the aldehyde group of PLP forming a Schiff base with the side chain amino group of the lysine residue within the active site (known as the internal aldimine). The amino group of the amino donor needs to be positioned toward this aldimine in order to enable a

transaldimination reaction between the aldimine and the amino donor, leading to the formation of a Schiff base between PLP and the amino donor (external aldimine). This external aldimine is further converted to a ketimine through a 1,3-prototropic shift. The released lysine residue serves as a catalytic base during the hydrolysis of the ketimine, resulting in the formation of a ketone, which then exits the active site. The generated PMP subsequently reacts with a newly bound amine acceptor (ketone) and undergoes reductive amination, effectively reversing the aforementioned reaction. This process leads to the formation of a new amine product while regenerating the internal aldimine for another cycle of the reaction [9]. Schematic representation of the detailed reaction mechanism is shown in Figure 2.1.

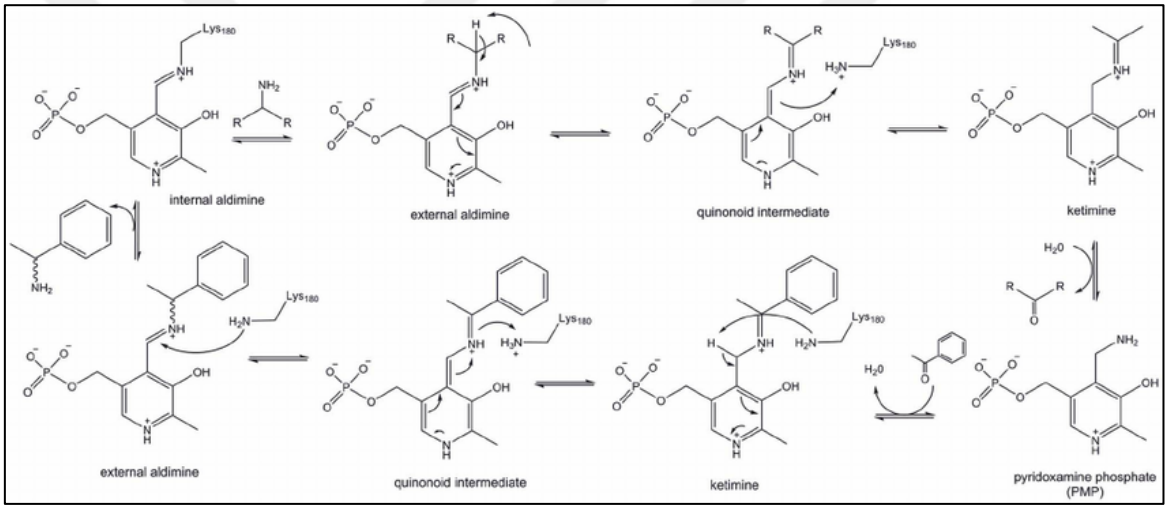

Figure 2.1. Transamination reaction mechanism [9]

2.1.2. Industrial Usage of Transaminases

In the last decades, industrial synthesis benefit from the advantages of enzymatic catalysis. A breakthrough was reported in 2010 by Savile and Janey et al. which is an efficient enzyme catalyzed asymmetric enamine hydrogenation process to replace a recently implemented rhodium-catalyzed synthesis mechanism for the large-scale manufacture of a antidiabetic compound, Sitagliptin. The previous method of Sitagliptin manufacturing involved usage of a rhodium-based chiral catalyst to asymmetrically hydrogenate an enamine at high pressure of 250 psi. However, this process had some drawbacks: It had insufficient selectivity in producing the desired stereoisomers and resulted in a product contaminated with rhodium,

which required additional purification steps. These purification steps negatively impact the overall yield, as they were necessary to enhance both the enantiomeric excess (e.e.) and chemical purity. However, a new approach has been developed that significantly improves the efficiency of Sitagliptin manufacturing. This involves utilizing a transaminase scaffold and employing various protein engineering technologies to create a catalyst and process that address the previous shortcomings [10]. Even after this study, research for the development of more suitable transaminases to catalyze the Sitagliptin manufacturing with more economic and sustainable pathways continues until the date with the developments in enzyme engineering, because most ω-aminotransaminases are *R*-selective and suffer from weak stability and low activity toward bulky prochiral ketones. Three years ago, a selective ω-aminotransferase (ω-ATA) with a preference for the *R*-stereoisomer was isolated from a bacteria, Arthrobacter cumminsii, through cloning. Through a semi-rational protein design approach, the two of the most promising variants of the enzyme were identified, demonstrating enhanced activity, improved affinity for the substrate, and increased stability in an aqueous solution containing 20% dimethyl sulfoxide. When these variants were used as biocatalysts, a remarkable conversion rate of 92% was achieved in the asymmetric amination of 50 g/L prositagliptin ketone. Moreover, this process yielded an exceedingly high enantiomeric excess of greater than 99%. In a larger-scale experiment involving kilograms of the substrate, approximately 40 kg of (*R*)-product with enantiomeric excess exceeding 99% was produced within 30 hours. This outcome was achieved by employing 50 kg of the substrate [11].

2.1.3. A 4-Aminobutyrate Transaminase, KES23360

KES23360 is an enzyme known as 4-aminobutyrate transaminase, which has activity towards 12-aminododecanoic acid. This particular compound is not naturally occurring and holds significant industrial importance as a key building block for nylon-12. It is noteworthy that only a limited number of proteins have been identified with the ability to catalyze this compound, making this enzymatic activity quite uncommon [12]. The utilization of 12-aminododecanoic acid as a target for biocatalysis is particularly attractive due to the associated risks involved in traditional chemical methods of producing polyamides, including nylon-12. These conventional routes have previously resulted in global shortages of this material. Therefore, the enzyme's capability to convert 12-aminododecanoic acid

offers a safer and more sustainable alternative for the production of nylon-12 and related materials. Wilding et al. obtained the phylogenetic tree of KES23360 which is shown in the Figure 2.2. In Figure 2.2, the modern-day transaminase is written in blue and the most ancient variant, N48, is given at the top of the tree. All ancient variants examined in this study are indicated in red frames and these variants were obtained structurally by Ancestral Reconstruction Technique (ASR) by Wilding et al.

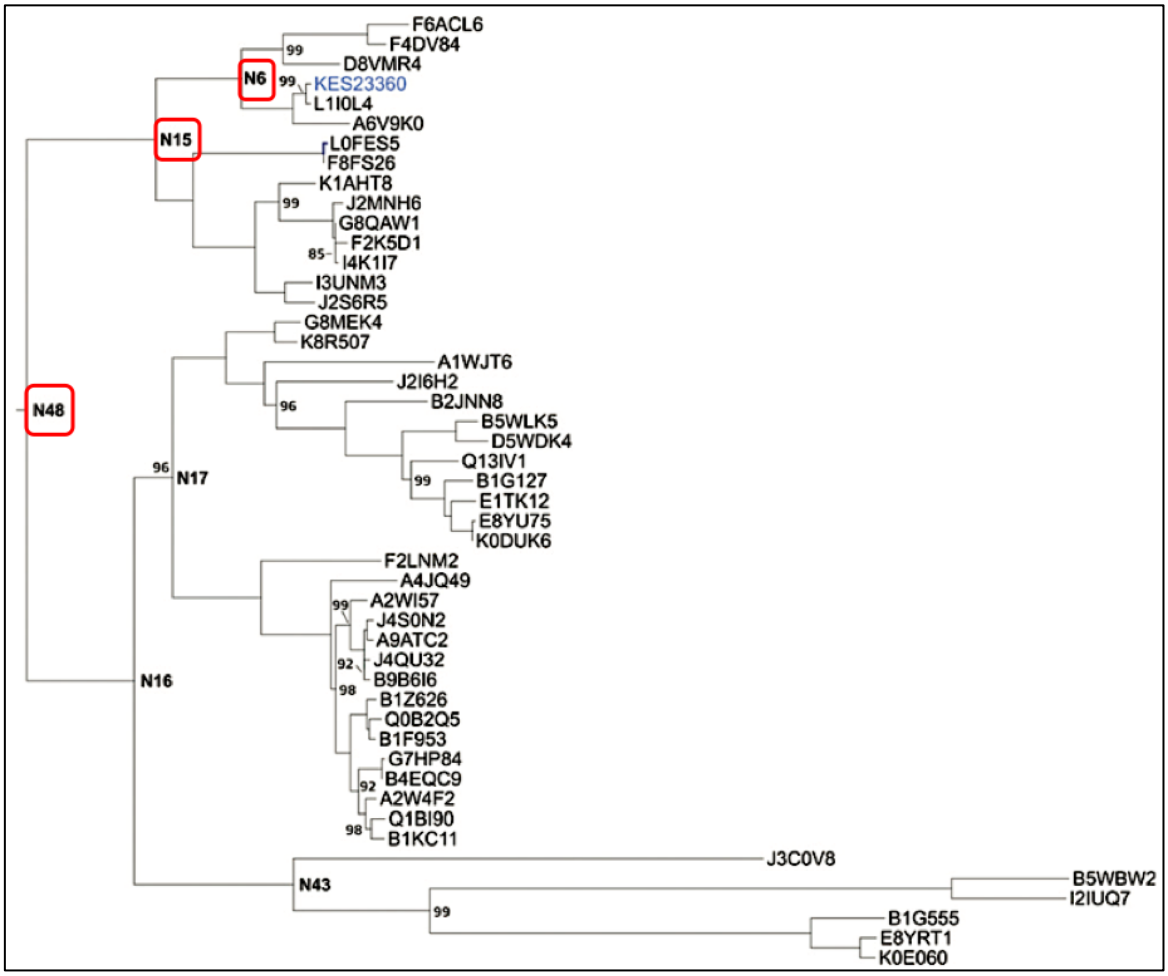

Figure 2.2. Phylogenetic tree of KES23360 [4]

Ancestral sequence reconstruction is a technique used to infer the sequence of ancient proteins based on the sequences of their modern descendants. By analyzing the genetic information of extant organisms, researchers are able to make educated guesses about the amino acid sequences of proteins that existed in the past. ASR is particularly powerful when combined with experimental characterization of these proteins, as it allows for the investigation of problems that would otherwise be difficult to study. One important

application of ASR is in the identification of key amino acid residues that are crucial for protein function. By comparing the sequences of related proteins from different species, researchers can identify the residues which have been conserved throughout evolution, suggesting their functional importance. This information can provide insights into protein structure and function, helping to guide experimental studies and understand the molecular basis of biological processes. This technique also has applications in studying the longevity of mutations. By reconstructing ancestral protein sequences, researchers can investigate how mutations have influenced protein evolution. This information can shed light on the functional consequences of specific mutations and provide a better understanding of the factors that drive protein evolution. Additionally, ASR can be used to explore the contribution of gene duplications to the functional diversification of enzymes. Gene duplications can create new copies of genes, which can then evolve independently and acquire new functions. By reconstructing ancestral sequences, researchers can trace the evolutionary history of enzymes and gain insights into how gene duplication events have led to the functional diversification of enzymes in different organisms [13].

After the structures of the variants were obtained, variants were treated by 40 different substrates including ω-amino acids, α-amino acids, and α,ω-diamines and substrate scope were determined. In this study it was shown that KES23360 has the activity of 95.1 $nmol.min^{-1}.mg^{-1}$ as the highest rate, and it has high activity for four additional ω-amino acids but not for the other kind of substrates. On the other hand, this substrate specificity is broadened for the second and third generation variants. Additionally, up to 20 fold activity improvements were obtained for the ancestral variants with respect to the modern-day transaminase. When the specific activity data given for N48 are considered, it is clearly seen that N48 has almost no reasonable activity for any substrate tasted. Specific activity data for the 10 different substrates of the variants that were examined in this study are given in Table 2.1. Errors are stated near each activity data values.

Table 2.1. Specific activity data of four transaminase variants for 10 substrates [4]

Substrate	Specific Activity (nmol.min^{-1}.mg^{-1})			
	N48	**N15**	**N6**	**KES23360**
4-Aminobutyrate	1.4 ± 0.004	21.9 ± 0.4	60.9 ± 2.0	48.1 ± 0.7
4-NH_2-(S)-2-hydroxybutyrate (AHBA)	2.2 ± 0.01	14.9 ± 0.5	27.0 ± 1.0	95.1 ± 0.7
Cadaverine	0.3 ± 0.002	21.5 ± 0.6	48.6 ± 0.5	4.7 ± 0.3
1,8-Octamethylenediamine	0.3 ± 0.01	13.9 ± 0.2	24.2 ± 0.5	6.8 ± 0.3
1,10-Decamethylenediamine	0.6 ± 0.006	13.7 ± 0.2	47.3 ± 0.3	7.2 ± 0.2
L-Lysine	N.D.	N.D.	N.D.	2.2 ± 0.1
D-Lysine	0.1 ± 0.001	6.2 ± 0.08	10.0 ± 0.09	3.9 ± 0.03
L-Arginine	N.D.	1.0 ± 0.02	1.7 ± 0.04	0.7 ± 0.05
2-Aminoindan	0.1 ± 0.003	4.7 ± 0.06	9.7 ± 0.06	0.7 ± 0.1
6-Aminohexanol	0.3 ± 0.01	6.4 ± 0.2	13.8 ± 0.1	4.3 ± 0.2

In case of thermotolerance, most ancient variants show higher thermostability (N15: 61°C and N6: 62°C) the modern one. However, the most ancient transaminase N48 has almost the same melting point (56.5 °C) with the KES23360 (57°C) [4].

2.2. COMPUTATIONAL METHODS DEVELOPED FOR THE IDENTIFICATION OF THE DYNAMIC MOTIONS IN PROTEINS

Biomolecular structures exhibit various degrees of freedom under equilibrium conditions, which encompass both subtle atomic fluctuations and large-scale motions of domains, subunits, or entire molecules. These motions are not random but rather governed by intra- and intermolecular interactions, which are, in turn, influenced by the three-dimensional (3D) structures. Consequently, these dynamics can be considered as structurally encoded and they primarily arise from the native contacts present in the structure which refer to the non-

covalent interactions between the residues that are in close proximity to each other in the folded state of proteins. These contacts are crucial to stabilize the structure and determine the functional properties. The topology or geometry of these native contacts plays a fundamental role in the specification of the structure-encoded dynamics of proteins. In recent years, there has been a significant increase in the utilization of Elastic Network Models (ENMs) and Normal Mode Analysis (NMA) to explore the structure-encoded dynamics of biomolecules. ENMs provide a simplified representation of the proteins, where alpha-carbons of residues are considered as junction points while the interactions between the residues are represented as the linear springs, with the same analogy of the elasticity theory of random polymer networks. NMA, on the other hand, involves the computation of the collective motions that a biomolecule can undergo while maintaining its overall structure which is called as normal mode. NMA is a widely used method for studying the long-time dynamics and elasticity of biomolecules. It is based on the premise of complex semiempirical potentials that describe the covalent and noncovalent interactions between atoms in biomolecular systems.

By employing ENMs and NMA, researchers can study the low-frequency normal modes of a biomolecular structure. These modes correspond to the most cooperative and significant motions accessible to the molecule, such as domain motions, hinge bending, or global conformational changes. The advantage of using ENMs and NMA lies in their ability to capture these large-scale motions based solely on the native contact topology of the structure, without considering specific atomic interactions. The insights gained from studying structure-encoded dynamics using ENMs and NMA can aid in understanding the functional mechanisms of proteins, such as enzyme catalysis, protein-protein interactions, and allosteric regulation. Overall, the application of elastic network models and normal mode analysis has contributed to our understanding of the structure-encoded dynamics of biomolecular systems, shedding light on the relationship between structure and function at a fundamental level [14].

2.2.1. Gaussian Network Model (GNM)

GNM is a simplified 3D elastic network model used to predict the equilibrium fluctuations of proteins in their folded states. According to GNM, Cα atoms, the junction points of the

network model, undergo Gaussian-distributed fluctuations which means that they fluctuate under the potentials exerted by their neighboring atoms. In GNM, the interactions within the proteins are modeled by a single-parameter harmonic potential. This parameter characterizes the strength of the interactions between closely located pairs of Cα atoms. The model takes into account both nonbonded interactions and chain connectivity. Nonbonded interactions are typically intermolecular forces and electrostatic interactions between different atoms in the protein. Chain connectivity refers to the constraints imposed by the first neighboring Cα atoms along the protein backbone. By using single parameter for the harmonic potential, based on the work of Tirion [15], the GNM can accurately predict the equilibrium fluctuations of the Cα atoms in various proteins, in close agreement with experimental observations. The simplicity and reliability of the GNM, with its single-parameter Gaussian model, is proven successfully in the prediction of the results for the folded proteins from their crystal structures by its ability to accurately reproduce temperature factors (B-factors) and cross-correlations [16,17].

In this model by evaluating the junction fluctuations $\Delta \boldsymbol{R_{ij}}$ of the ith and jth Cα atoms in the folded protein as Gaussian-distributed fluctuations, $W(\Delta \boldsymbol{R_{ij}})$:

$$W(\Delta \boldsymbol{R_{ij}}) = (C/\pi)^{3/2} \exp\left(C \Delta \boldsymbol{R_{ij}}^2\right) \tag{2.1}$$

Where C is the single parameter in the Hookean potential obtained in the study of Tirion[15]. In the elasticity theory of polymer networks, the distribution function $W(\Delta \boldsymbol{R_{ij}})$ represents the probability distribution of the fluctuation $\Delta \boldsymbol{R_{ij}}$ in the end-to-end vector connecting junctions i and j of the polymer network. Then by substituting this distribution function into Equation 2.2:

$$\Delta A = -k_B T ln W(\Delta \boldsymbol{R_{ij}}) \tag{2.2}$$

where ΔA is the elastic free energy change associated with the fluctuation $\Delta \boldsymbol{R_{ij}}$; the harmonic potential ($k_B T C \Delta \boldsymbol{R_{ij}}^2$) and Hookean force constant ($2k_B TC$) are obtained. The Hookean force constant characterizes the stiffness of the elastic network and how much force is required to induce a given displacement in the end-to-end vector connecting junctions i and j of the polymer network. Then in analogy to the theory of random Gaussian networks, Equation 2.3 provides a mathematical expression for the configurational factor of a protein

of N residues based on the fluctuations of Cα atoms, incorporating the Kirchhoff matrix, **Γ**, and reflecting the statistical behavior of different fluctuation configurations.

$$Z_N = Kexp(-[\Delta \boldsymbol{R}^{\boldsymbol{T}}]\boldsymbol{\Gamma}[\Delta \boldsymbol{R}]) \tag{2.3}$$

In Equation 2.3: Z_N is the configurational factor for a protein with N residues, K is a constant, $[\Delta \boldsymbol{R}]$ is an N-dimensional column vector formed by the fluctuations {ΔR1, ΔR2, ..., ΔRN} of the Cα atoms, The superscript ^{T} denotes the transpose in matrix operation, and **Γ** is a symmetric matrix known as the Kirchhoff matrix. The Kirchhoff matrix includes the connectivity information of the protein network which represents the interactions and relationships between the Cα atoms in the protein structure. Symmetry of this matrix ensures that the fluctuations of one Cα atom affect its neighboring Cα atoms consistently. The exponential term, $exp(-[\Delta \boldsymbol{R}^{\boldsymbol{T}}]\boldsymbol{\Gamma}[\Delta \boldsymbol{R}])$, quantifies the statistical weight or probability of a specific configuration of Cα atom fluctuations. It takes into account the energy associated with these fluctuations and allows for the exploration of different conformations in the protein. The elements of the Kirchhoff matrix are determined according to Equation 2.4:

$$\Gamma_{ij} = \begin{cases} -C \;\; if \; i \neq j \text{ and } R_{ij} \leq r_c \\ 0 \;\; if \; i \neq j \text{ and } R_{ij} > r_c \\ -\sum_{i \neq j} \Gamma_{ij} \;\; if \; i = j \end{cases} \tag{2.4}$$

where r_c is the is the cutoff distance identifying the range of the separation of nonbonded contacts [17–19]. Using the given expression for the inverse of the Kirchhoff matrix in Equation 2.5, cross correlations between the fluctuations of the Cα atoms can be obtained by the off-diagonal elements of the inverse of the Kirchhoff matrix, $\boldsymbol{\Gamma^{-1}}$

$$\boldsymbol{\Gamma^{-1}} = \boldsymbol{U}(\Lambda^{-1})\boldsymbol{U^T} \tag{2.5}$$

where $\boldsymbol{U}$ is an orthogonal matrix whose columns are composed of the eigenvectors of the Kirchhoff matrix, and **Λ** is the diagonal matrix of the eigenvalues of **Γ** [16].

In the analysis of protein dynamics using GNM, the square of each eigenvector provides information about the mean square fluctuations, also known as auto-correlations, of residues from their equilibrium positions along the principal coordinates. To study the correlations between residues, the correlation matrix of the structure is computed. This involves calculating the correlation between each pair of residues for a given mode. The correlation

between two residues reflects their tendency to move in the same or opposite directions. In the correlation matrix, positive correlations are denoted by "+1" and indicate that the residues tend to move in the same direction. This suggests a coordinated motion or concerted behavior between the residues. On the other hand, negative correlations are represented by "–1" and indicate that the residues move in opposite directions. This implies an anti-correlated motion or a tendency for the residues to move in opposite directions. By examining the correlations between residues and identifying positive and negative correlations, we can gain insights into the collective dynamics and cooperative motions of proteins. This information is valuable for understanding protein function, allosteric communication, and conformational changes that contribute to protein folding, ligand binding, and enzymatic activity [20].

For the calculation of the normalized average cross correlations of Cα atoms for a protein with N residues, Equations 2.6 and 2.7 are used.

$$CC_{avg}(i,j) = \sum_{k=1:N-1} \boldsymbol{U(i,k)} * \boldsymbol{U^T(j,k)}/w(k) \tag{2.6}$$

$$normalized_{CC_{avg}}(i,j) = \frac{CC_{avg}(i,j)}{\sqrt{CC_{avg}(i,i) * CC_{avg}(j,j)}} \tag{2.7}$$

Here $CC_{avg}(i,j)$ is the average cross correlation of the ith and jth residues, $\boldsymbol{U(i,k)}$ is the eigenvector of residue i of the kth mode, $\boldsymbol{U(j,k)}$ is the eigenvector of residue j of the kth mode, and $w(k)$ is the eigenvalue of the kth mode.

Proteins are known to exhibit high flexibility, thus understanding and predicting their molecular flexibility is crucial for comprehending protein function and elucidating functional mechanisms. Protein movements are often classified into two main types: hinge movements and shear movements. Hinge movements involve the rotation of protein parts, typically domains, around a specific region known as a “hinge”. Hinge motions play a key role in facilitating large-scale conformational changes, such as domain rearrangements, in proteins. On the other hand, shear movements involve a sliding motion of protein parts relative to each other. This movement occurs along an interface plane between protein segments. Shear motions are often more restricted compared to hinge motions, involving small conformational changes across the movement interface. Both hinge and shear movements contribute to the overall flexibility of proteins and play essential roles in their biological functions [21].

2.2.2. Anisotropic Network Model (ANM)

ANM (Anisotropic Network Model) is a simplified tool used for the analysis of vibrational motions in molecular systems. It is based on the principles of NMA and utilizes the ENM methodology. The overall potential energy of the system is determined by the sum of harmonic potentials between interacting nodes of the network model. These harmonic potentials includes the interactions and connectivity between the residues. The interactions taken into account in ANM are limited to those within a specified cutoff distance. The cutoff distance is a predetermined parameter in the model and defines the range in which interactions are considered. This allows for a simplified representation of the system, focusing on the most relevant interactions. Importantly, ANM incorporates the information about the orientation of each interaction with respect to the global coordinate system. This information is encoded within the Force constant matrix, which characterizes the strength of the interactions between nodes. By considering the anisotropic motions, ANM provides a computationally efficient approach to analyze vibrational motions in proteins leading to get insights into the dynamics of macromolecules [22,23].

The Force constant matrix is expressed by a Hessian matrix with the elements ($\boldsymbol{H_{i,j}}$) of three by three submatrices encoding the anisotropic motions of the nodes with respect to their orientations. Elements of the submatrices are composed of the second derivatives of the harmonic potential, V. Harmonic potential for the interaction of the i^{th} and j^{th} residues represented as linear spring in ENM can be obtained as given in Equation 2.8.

$$V = \left(\frac{1}{2}\right)\gamma_f(s_{ij} - s_{ij}^0)^2 \tag{2.8}$$

In equation 2.8, γ_f is the force constant which is considered as constant in ANM with the same analogy in GNM, s_{ij} is the instantaneos distances between residues i and j, s_{ij}^0 is the equilibrium distances between residues i and j.

$$\boldsymbol{H_{i,j}} = \begin{bmatrix} \frac{\partial^2 V}{\partial x_i x_j} & \frac{\partial^2 V}{\partial x_i y_j} & \frac{\partial^2 V}{\partial x_i z_j} \\ \frac{\partial^2 V}{\partial y_i x_j} & \frac{\partial^2 V}{\partial y_i y_j} & \frac{\partial^2 V}{\partial y_i z_j} \\ \frac{\partial^2 V}{\partial z_i x_j} & \frac{\partial^2 V}{\partial z_i y_j} & \frac{\partial^2 V}{\partial z_i z_j} \end{bmatrix} \tag{2.9}$$

Fluctuations of residues i and j from their equilibrium positions for a protein with N residues can be correlated in 3N-6 modes contributed by the movement along mode k. This decomposition is given in terms of the eigenvalues (λ_k) and the corresponding eigenvector ($\boldsymbol{u_k}$) of the Hessian matrix [$\boldsymbol{H}$] as in Equation 2.11.

$$< \Delta \boldsymbol{R_i}. \Delta \boldsymbol{R_j} >= \frac{3k_B T}{\gamma_f} tr[\boldsymbol{H^{-1}}] \tag{2.10}$$

$$< \Delta \boldsymbol{R_i}. \Delta \boldsymbol{R_j} >= \frac{3k_B T}{\gamma_f} \sum_{k=1}^{N} tr\left[{\lambda_k}^{-1} \boldsymbol{u_k} \boldsymbol{u_k^T}\right]_{ij} \tag{2.11}$$

In Equation 2.11; $< \Delta \boldsymbol{R_i}. \Delta \boldsymbol{R_j} >$ represents the correlation between the fluctuations of residues i and j, k_B is the Boltzmann's constant, T is the temperature and tr is trace operation defined as the sum of the elements on the main diagonal of a square matrix.

When the fluctuation vectors are known, it is possible to explore the alternative conformational space that could potentially be sampled by individual or combinations of modes. This provides insights into the conformational dynamics and flexibility of the system [23,24].

2.2.3. Molecular Dynamic Simulations (MD)

The laws governing motion in the microscopic world, as described by quantum mechanics, differ significantly from the laws of macroscopic dynamics. Quantum mechanical phenomena are characterized by deterministic laws rather than probabilistic behavior. In the aspect of macroscopic dynamics, the behavior of subatomic particles is described by wave functions that represent the probability of finding a particle in a particular state. Chemical bonds, for example, arise from the interaction of shifting clouds of electrons, which exhibit both particle and wave-like behavior. These electrons occupy specific energy levels around atoms and can transition between these levels, participating in chemical reactions and the formation of bonds [25].

The traditional concept of ligand binding which is often referred as the "lock and key" theory, posited that a receptor remained motionless and static while accommodating a ligand.

However, this idea has been abandoned in favor of newer binding models that consider conformational changes and the random movements of receptors and ligands. It is now recognized that both receptors and ligands exhibit dynamic behavior and undergo fluctuations and jiggling, which are essential for successful binding interactions [26–29]. To explore and understand these dynamic processes, molecular dynamics simulations are employed. These simulations incorporate factors such as conformational changes, as well as the effects of temperature, pressure, and solvation. By simulating the behavior of a system over time, it becomes possible to predict its future states based on its current state.

The calculation of energy is the key step in molecular modeling, as it provides crucial information about the stability, behavior, and properties of a molecular system. There are various methods available to perform these calculations, each with its own strengths and limitations. Quantum mechanics offers the most fundamental approach to calculate the energy of a molecular system, solving the Schrödinger equation. By solving this equation, it is possible to obtain the wave function, from which various properties and energies can be derived. However, quantum mechanical calculations are computationally demanding, especially for large systems, due to the complexity of the calculations involved. As a result, these methods are typically reserved for smaller systems or specific regions of interest within larger systems.

In molecular dynamics, an alternative approach is used to explore the energy surface of a system. Instead of explicitly considering the electronic motions, molecular mechanics force field models are employed. These models approximate the intramolecular and intermolecular interactions within the system and calculate the energy as a function of the positions of the atoms ignoring the electronic motions within the system. By solving Newton's laws of motion, the system's trajectory is determined, allowing for the exploration of the energy landscape. Molecular mechanics force field models consider various contributions to the energy of the system. These include the stretching of bonds, bending of angles, rotations about single bonds, and steric and electrostatic interactions between pairs of non-bonded atoms. By summing up these contributions, an approximation of the total energy of the system is obtained.

While molecular mechanics force field models provide a computationally efficient means to explore large systems, it's important to note that they rely on simplified representations of the molecular interactions. They do not explicitly account for electronic effects. Therefore,

while they are useful for studying larger systems, they may have limitations when it comes to accurately describing certain aspects of molecular behavior [30].

2.2.3.1. Force Fields

In MD simulations, force fields are essential mathematical models that describe the potential energy of a molecular system based on the positions and interactions of its atoms. Several widely used force fields have been developed for MD simulations are AMBER, CHARMM, GROMOS, and OPLS/AMBER. These force fields have been extensively utilized in both fundamental research and applied studies. The AMBER force field is a popular choice known for its accuracy in modeling biomolecular systems. It includes parameters for a wide range of biomolecules, such as proteins, nucleic acids, and carbohydrates. The CHARMM force field, developed by the Chemistry at HARvard Molecular Mechanics (CHARMM) group, is also widely used and has been refined over the years to accurately represent various biomolecular systems. GROMOS (Groningen Molecular Simulation) force fields are predominantly used for simulations of small organic molecules, while the OPLS/AMBER force field combines features from both the AMBER and OPLS (Optimized Potentials for Liquid Simulations) force fields [31].

The simplest force field type used in MD simulations is the Generalized Amber Force Field (GAFF), which provides a simplified representation of the potential energy as expressed in Equation 2.12. GAFF is often used in cases where a more computationally efficient model is required, although it may sacrifice some accuracy compared to more detailed force fields.

$$E = \sum_{bonds} K_b\left(l_i - l_{i,0}\right)^2 + \sum_{angles} K_\theta\left(\theta_i - \theta_{i,0}\right)^2 + \sum_{torsions} \frac{V_n}{2}\left(1 + cos\left(n\omega - \gamma\right)\right) + \sum_{i=1}^{N}\sum_{j>i}^{N}\left(\left[\left(\frac{A_{ij}}{{R_{ij}}^{12}}\right) - \left(\frac{B_{ij}}{{R_{ij}}^{6}}\right)\right] + \frac{q_i q_j}{\varepsilon R_{ij}}\right) \tag{2.12}$$

The first term of Equation 2.12 accounts for the stretching of bonds between atoms. It is a harmonic potential that represents the energy associated with the displacement of the bond length, denoted as l_i, from its ideal value, $l_{i,0}$. The force constant $K_b$determines the strength of the bond and is specific to each pair of bound atoms. The second term represents the

bending of angles between atoms. It also uses a harmonic potential to describe the energy associated with deviations in the bond angle, denoted as θ_i, from its ideal value, $\theta_{i,0}$. Similar to the bond term, the force constant K_θ is specific to the types of atoms involved in the angle. The third term accounts for torsional or dihedral angle interactions. It represents the energy associated with rotations around a middle bond. The potential function used for torsional interactions is periodic and often expressed as a cosine function. The parameters in this term include the symmetry coefficient, n and the phase angle, γ that determine the shape and magnitude of the torsional potential. The last term of Equation 2.12 incorporates non-bonded interactions, such as solvent effects. It includes both van der Waals interactions and electrostatic interactions. The van der Waals interactions are represented by the terms involving A_{ij} and B_{ij}, which describe the attractive and repulsive components, respectively, of the Lennard-Jones potential. The electrostatic interactions are represented by the terms involving the partial charges, q_i and q_j and the distance between atoms, R_{ij}. The parameters ε and R_{ij} determine the strength and range of the electrostatic interactions [31].

Force field methods in molecular mechanics are empirical approaches that rely on the derivation and adjustment of parameters based on experimental and theoretical data. The choice of functional form and the parameters within it are not inherently superior, but rather they are selected to reproduce observed behavior. Additionally, force fields assume transferability, allowing parameters developed for a limited set of reference systems to be applied to a wider range of molecules. These features enable force field methods to be versatile tools for studying molecular systems of varying sizes and complexities [30].

2.2.3.2. The Algorithm Behind MD Calculations (Classical Mechanics)

Molecular dynamics algorithm follows the general steps described below:

1. Time discretization: The continuous time is divided into small time steps, typically on the order of femtoseconds (10^{-15} s) increments. This allows the simulation to capture the dynamics of atoms at a molecular level.
2. Force calculation: At each time step, the forces acting on each atom are computed using a molecular mechanics force field. These forces are derived from the potential energy function associated with the force field and are calculated based on the current positions of the atoms.

3. Update positions and velocities: With the forces determined, the positions and velocities of the atoms are updated to simulate their motion over the next time step. This update is carried out using Newton's laws of motion. Newton's second law states that the force acting on an object is equal to the mass of the object, m_i multiplied by its acceleration, a_i as it is given in Equation 2.13. By numerically integrating the equations of motion, the new positions and velocities of the atoms are calculated.
4. Repeat steps 2 and 3: The process is repeated for subsequent time steps, iterating through the calculation of forces, updating positions and velocities, and advancing the simulation in time. The simulation continues until the desired time or system-specific criteria are reached.

$$F_i = m_i \times a_i \tag{2.13}$$

$$F_i = -m_i \frac{dv_i}{dt} = m_i \frac{d^2x_i}{dt^2} \tag{2.14}$$

In Equation 2.14, v_i is the velocity of the atom i and x_i is the position of the atom i. Both are functions of time. The force acting on an atom is directly correlated with its acceleration and, correspondingly, to the potential energy. This relationship is given by Equation 2.15.

$$F_i = -\nabla_i E \tag{2.15}$$

For N atoms, there are 3N position coordinates and 3N velocity coordinates, thus the algebraic solution of the equations is impossible. Numerical solution is;

$$x_{i+1} = x_i + \delta_t v_i \tag{2.16}$$

$$v_{i+1} = v_i + \delta_t \frac{F(x_i)}{m} \tag{2.17}$$

Time symmetric integration methods are used in the numeric solutions such as Velocity Leapfrog Verlet [32]:

$$x_{i+1} = x_i + \delta_t v_{i+1/2} \tag{2.18}$$

$$v_{i+1/2} = v_{i-1/2} + \delta_t \frac{F(x_i)}{m} \tag{2.19}$$

2.2.3.3. Statistical Mechanics

MD simulations provide detailed information about the positions and velocities of atoms at the microscopic level. However, to obtain macroscopic observables such as pressure, energy, and heat capacities from this microscopic data, statistical mechanics principles are required. Statistical mechanics is employed to bridge the gap between the microscopic behavior of atoms and the macroscopic properties of the system. By applying statistical mechanics, it becomes possible to relate the ensemble average of microscopic quantities to observable macroscopic quantities [31].

a) Ensembles

In statistical mechanics, an ensemble refers to a set of all possible systems that exhibit various microscopic states while sharing the same macroscopic or thermodynamic state. Average values in statistical mechanics are defined as ensemble averages, which take into account the contributions from all the different microscopic states within the ensemble.

$$\langle A \rangle_{ensemble} = \iint dp^N dr^N A(p^N, r^N)\, \rho(p^N, r^N) \tag{2.20}$$

$$\langle A \rangle_{time} = \lim_{\tau \to \infty} \frac{1}{\tau} \int_{t=0}^{\tau} A(p^N(t), r^N(t)) dt \approx \frac{1}{M} \sum_{t=1}^{M} A(p^N, r^N) \tag{2.21}$$

Where t is the simulation time, M is the number of time steps in the simulation.

According to the Ergodic hypothesis, when a system is allowed to evolve over an extended period of time, it will eventually traverse through all possible states. Consequently, the ensemble averages of the system will be equivalent to the time averages. In the context of molecular dynamics simulations, the thermodynamic averages, which are representative of macroscopic properties, are obtained by taking time averages of the system as it evolves over time [31]. Therefore a Molecular Dynamics simulation must be sufficiently long so that enough representative conformations have been sampled.

$$\langle A \rangle_{ensemble} = \langle A \rangle_{time} \tag{2.22}$$

i) NVE – Micro canonical ensemble: In the NVE ensemble, the system is isolated and does not exchange particles (N), volume (V), or energy (E) with its surroundings. This corresponds to an adiabatic process with no heat transfer. The equilibrium state of the system

in the NVE ensemble occurs at maximum entropy.

ii) NVT – Canonical Ensemble: The NVT ensemble, also known as the canonical ensemble, conserves the number of particles (N), volume (V), and temperature (T). It is often referred to as constant temperature molecular dynamics (CTMD). A thermostat is used to maintain a fixed temperature, allowing for energy exchange between endothermic and exothermic processes. The equilibrium state in the NVT ensemble corresponds to the minimum Helmholtz free energy.

iii) NPT – Isobaric-Isothermal Ensemble: In the NPT ensemble, the system conserves the number of particles (N), pressure (P), and temperature (T). To maintain a constant pressure, a "barostat" is employed in addition to a thermostat. The NPT ensemble closely resembles laboratory conditions, with the system resembling a flask open to the ambient temperature and pressure. The equilibrium state of the NPT ensemble corresponds to the minimum Gibbs free energy.

iv) mVT – Grand canonical Ensemble: The mVT ensemble, also known as the grand canonical ensemble, is characterized by a fixed chemical potential (m), a fixed volume (V), and a fixed temperature (T). This ensemble allows for the exchange of particles with a particle reservoir while maintaining a constant chemical potential, volume, and temperature [31].

The majority of molecular dynamics simulations are typically conducted using the NVE ensemble (microcanonical ensemble), where the number of particles (N), volume (V), and energy (E) are held constant. However, more recent approaches aim to replicate experimental conditions more accurately by simulating systems at constant values of N (number of particles), T (temperature), and P (pressure).

2.2.3.4. Timescales

Biological systems demonstrate a broad spectrum of time scales associated with specific processes. MD simulations are capable of capturing various protein dynamics events within certain timescales within certain timescales, including but not limited to, the following ranges given in Table 2.2:

Table 2.2. Timescales of some protein simulation types [33]

Protein Dynamics Event	MD Simulation Timescale
Vibrational motions	femtoseconds (10^{-15} s) – picoseconds (10^{-12} s)
Rotational motions	picoseconds (10^{-12} s) – nanoseconds (10^{-9} s)
Loop dynamics	picoseconds (10^{-12} s) – milliseconds (10^{-3} s)
Ligand binding/unbinding	nanoseconds (10^{-9} s) – seconds
Protein folding/unfolding	microseconds (10^{-6} s) – seconds
Aggregation	Seconds and beyond

2.2.3.5. Solvent Effect

Effect of the solvent (water is the typical solvent for the protein simulations in MD calculations) must be taken into account in simulations. In order to account it, two ways are present, explicit solvent method and implicit solvent method. In explicit solvation, a solvent box is built around the system with solvent molecules. It brings high computational expense but also accuracy. On the other hand, implicit solvation is a mathematical model to approximate the average effects of the solvent by taking into account a solvation parameter depending on the solvent type. It is much faster but less accurate.

In summary, molecular dynamics simulations involve integration of the Newton's equations of motion to capture the time-dependent behavior of a macromolecular system. The simulation generates a sequence of conformations over time, named as a "trajectory" or a "path," which describes the positions and movements of each atom as governed by Newton's laws of motion.

2.2.4. Langevin Dynamics

In order to speed up the integration of the equations of motion in MD simulations, various techniques have been developed. One approach is to separate the faster components of the dynamic propagator from the slower components, or even eliminate the high-frequency motions altogether. The stability and resonance limitations of the integrators primarily depend on the fast modes and their coupling. Consequently, several integrators have been created to address both stability and resonance issues, such as multiple-timestep integrators

combined with stochastic dynamics. Stochastic methods, such as Langevin dynamics (LD) show promise in this regard. These approaches have coarse-grained characteristics that enable the study of large systems over long time scales, where proving ergodicity is also more feasible. Numerous algorithms have been developed based on these principles, including the widely used SHAKE algorithm, normal mode multiple time stepping Langevin dynamics integrator (NML), driven molecular dynamics (DMD), amplified-collective-motion (ACM) method, and steered/interactive MD (SMD/IMD).

In the normal mode multiple time stepping Langevin dynamics integrator (NML) approach, the kinetics and thermodynamics of the molecule are approximated by separating the dynamics into slow, functionally relevant modes (essential subspace) and fast modes (near-constraint coordinates). This separation is based on a normal mode decomposition of the dynamical space. The reduced model focuses on the low-frequency dynamical modes. To mitigate resonance effects, Langevin dynamics is employed instead of Newtonian dynamics. The low-frequency modes are propagated using a Langevin integrator, while the high-frequency modes are damped near their energy minimum and advanced using a computationally less expensive accelerated Brownian dynamics (BD) propagator to maintain the fast vibrations around their equilibrium values [24,34].

The fundamental equation of motion in stochastic dynamics is the Langevin equation, given by the Equation 2.23:

$$M_i \frac{d^2 r_i}{dt^2} = -\frac{\partial V}{\partial r_i} - M_i \gamma \frac{dr_i}{dt} + R_i(t) \tag{2.23}$$

This equation describes the motion of an atom i in the system. The left-hand side represents the mass (M_i) multiplied by the acceleration ($d^2 r_i / dt^2$). The first term on the right-hand side, $-\partial V / \partial r_i$, accounts for the gradient of the potential function V, which is commonly used in Newtonian (classical) molecular dynamics simulations and describes the molecular interactions. The second term, $-M_i \gamma \, dr_i / dt$, represents the damping or frictional force with the damping coefficient γ, which reflects the effect of friction due to the implicitly simulated solvent. It acts as a dissipative force, removing energy from the system. The last term, $R_i(t)$, represents the random force that arises from stochastic collisions between solvent molecules and the solute. The stochastic force introduced by $R_i(t)$ adds energy to the system. This

equation describes the motion of an atom i. It can be expressed in a more concise form as in Equation 2.24:

$$M_i\ddot{r}_i = -\nabla_i V - \zeta_i \dot{r}_i + R_i \tag{2.24}$$

The term $\nabla_i V$ represents the gradient of the potential function V. The other two terms, ζ_i and R, are unique to stochastic dynamics. The term, ζ_i, represents the friction or damping coefficient and the term, R_i, represents the random force as same as in Equation 2.23. When ζ_i is equal to zero, the LD simulation is equivalent to conventional MD, where only Newtonian dynamics is considered. However, when ζ_i is greater than zero, the random force represented by R_i can assist in barrier-crossing motions, potentially enhancing the exploration of conformational space compared to conventional MD. The relationship between ζ_i and R_i is crucial for maintaining the temperature of the simulated system [34,35].

3. METHODOLGY

This study is composed of the usage of computational methods to get insights into the mutations realized in a transaminase family during their natural evolution. Crystal structure of each variant were obtained from Protein Data Bank(PDB)[36] to use in computations. Protein Data Bank identifiers (PDB ID) of four variants examined in this study are given in Table 3.1 with their node names on phylogenetic tree. These node identifications were used for the naming of variants throughout the study.

Table 3.1. List of PDB ID's of the studied transaminase variants

PDB ID	NODE NAME
5KR6	N48
5KQW	N15
5KQU	N6
5KQT	KES23360

3.1. DETERMINATION OF DIFFERENCE VECTOR MAGNITUDES

As the computational approach followed in this study, at first, each sequential variant were aligned structurally using "Alignment" tool of PyMOL [37]. Alignment parameters were used same with the default values, number of cycles were set to 5.0 and cutoff was set to 2.0. Root mean square deviation (RMSD) values were obtained for each aligned variant couple and difference vector magnitudes were computed by taking the difference of the alpha-carbons' 3D coordinates of each residue and calculating the vector magnitudes of these differences. Graphical representations of the obtained difference vectors for the compared variant couples are given in Figures B.1 – B.3 in Appendix B. Sequence alignment results were also obtained and they are presented in Figures B.4 – B.6.

3.2. ANM-LD SIMULATION PROTOCOL

Crystal structures of each sequential variant couple, i.e., N48-N15, were given as initial and target structures to ANM-LD and it constructed the Hessian matrix for the current initial structure in each iteration cycle. This was achieved by defining neighbor interactions between α-carbon atoms using a distance threshold radius (R_{cut} = 13 Å). The pseudo-inverse of the Hessian matrix revealed 3N-6 intrinsic modes of motion. Difference vectors between the initial and target structures and ANM mode vectors were compared, and the mode having the highest overlap value were determined by the tool. To perturb the initial structure, the best overlapping ANM mode was applied with a deformation factor (DF = 0.2 Å). The resulting conformation was then subjected to energy minimization for 500 steps. The perturbations generated in the system were relaxed through a 100-step LD simulation, with each step lasting 0.2 fs, at a temperature of T = 310 K and a damping constant $\gamma = 5\ ps^{-1}$. The duration of 20 fs corresponded to the relaxation time of the perturbed structure by LD in each cycle. In general, iterative steps of ANM-LD are performed until RMSD between the new generated structure and the target conformation reaches a plateau, resulting in the generation of a transition pathway for the compared structures. However, in this study the aim was to determine the dominant ANM modes driving the transition of sequential variants during natural evolution rather than the generation of transition pathways. Therefore, the best overlapping ANM modes were searched for 10 cycles. At the end of these simulations, ANM modes having the best overlap with the difference vectors were determined as the driver modes of the evolution.

3.3. USAGE OF GNM IN CALCULATIONS

Eigenvectors and eigenvalues of the first 10 slowest GNM modes and mode shapes were obtained using the DynOmics web server [39]. GNM mode vectors were compared with the hitting ANM modes determined in the previous step and the GNM modes suitable to model the dynamic motion transitions of individual generations in their evolution were detected. Cross-correlation value data of each residue of each variant and hinge residues of first two slowest GNM modes in addition to the animated dynamic motions modeled for the first two slowest ANM modes were obtained by HingeProt [20]. For the selected GNM modes driving

the evolutionary sub-steps, cross correlation differences were computed by creating the cross-correlation (c.c.) matrices of each variants and they were represented on cross-map difference graphs. Residues having different cross-correlations in different variants were identified on the 3D structures of the variants together with the close mutation points. Normalized averages of the cross correlations of first five slowest GNM modes were also computed for each variant and their differences were obtained. For the normalized c.c. calculations, Equations 2.6 and 2.7 were used. For the matrix including calculation steps MATLAB[40] was used.

3.4. ASSIGNMENT OF ACTIVE SITE RESIDUES AND THEIR FUNCTIONS

For neither the modern-day transaminase, KES23360, nor the other variants, all residues having a function in transamination reaction mechanism were not identified. Thus, assignment of these residues were necessary in this study. For this purpose, a computational work based on MD simulations performed for one of the members of PLP dependent enzymes, γ-aminobutyric acid aminotransferase (GABA-AT) was examined. Molecular dynamics simulations were conducted by Sungur et al.[41] to investigate the functions of protein residues and the cofactor PLP in GABA-AT. The study explored three distinct states: the apoenzyme, the holoenzyme, and the inactive state. Additionally, the researchers examined various protonation states for PLP and two crucial residues in the active site, Asp298 and His190. In this paper, residues taking roles in the transamination reaction mechanism and protonation states of the catalytic residues as well as of PLP are proposed in details and crystal structure of the studied structures are available in PDB. By comparing the structures of GABA-AT with our transaminases, active sites of variants were identified residue by residue as shown in Figure 3.1. Cyan stick represents PLP and green lines represent proposed active site residues in Figure 3.1. Functions assigned to each residue matching with the GABA-AT system are given in Table A.1.

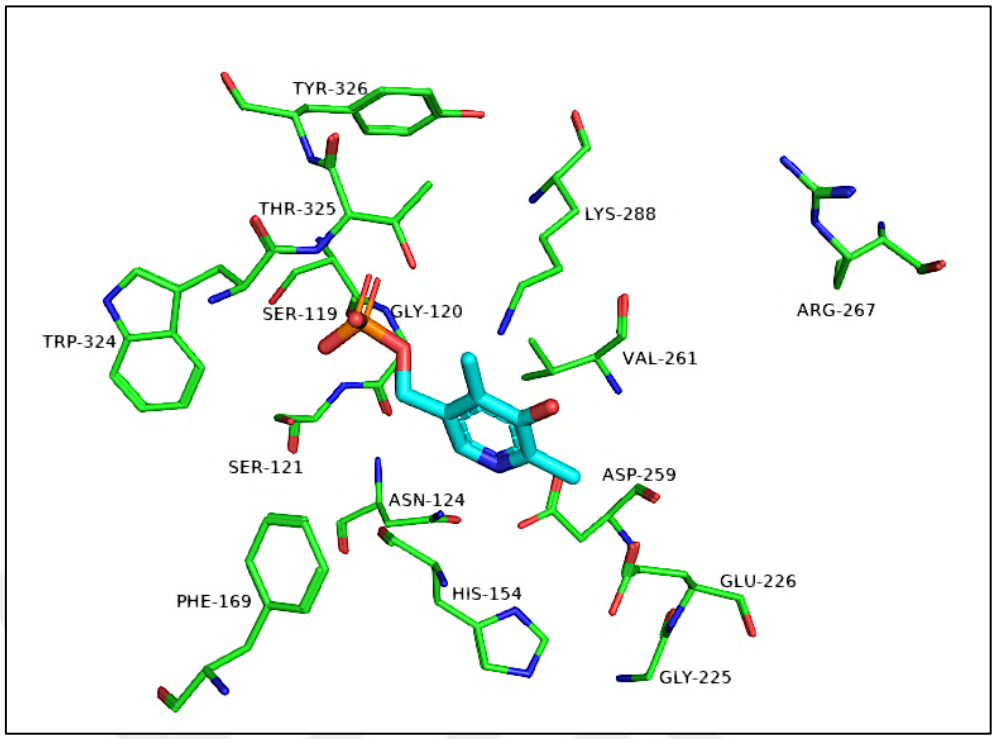

Figure 3.1. Active site residues

3.5. MOLECULAR DYNAMIC SIMULATIONS

Molecular Dynamic (MD) simulations were performed for N48 and KES23360 for their apo-protein (only the enzyme structure without any ligand or cofactor), free-PLP containing in their active site, and holo-enzyme (PLP is covalently bonded to LYS288) forms for 20 ns. Apo-enzyme MD simulations were also performed for N15 and N6 with the same time-scale. AMBER12[42] was used to perform MD simulations. Topology and coordinate files were generated using “tleap” which is a modified program of “LEaP” algorithm to create new systems in AMBER12 module. Amino acid residues has been modeled using the ff99SB force field parameters. Periodic boundary conditions were imposed on the systems. Na^+ ions were added to neutralize the complexes. TIP3P force field, which is a water model developed in a specific force field in Amber was used for explicit solvation and a water box around the systems with the edge length of 10 Å were added.

3.5.1. PLP Parametrization

Geometry optimization of PLP was performed by Gaussian16 [43]. Hartree-Fock (HF) method was used for geometry optimization with the basis set of 6-31G(d). Single point energy calculation was performed using the output of the geometry optimization with HF method, 6-31G(d) basis set. Extra keywords of SF=tight, POP=MK iop(6/33=2), iop(6/42=6), was used. “Antechamber” module was used to extract partial charges obtained as Gaussian output for the further simulations.

3.5.2. Preparation of Enzyme Systems

Firstly, crystal structure of each variant to be used in the simulations were prepared as apo-proteins; hydrogens of the protein residues and additional ions within the protein complexes were removed. Then, cysteine amino acids of the protein were examined in order to determine whether they form disulfide bridges. The distances between sulfur atoms were used for this consideration. The limit for disulfide bridge formation was 2-2.5 Å. There was not any disulfide bridge formation within the studied variants. Histidine amino acids were also examined. If a histidine residue had the hydrogen bond formation possibility from delta (δ) positions instead of epsilon (ε), then the name of this residue was changed from HIS to HID. HIS154 is protonated from delta position instead of the default position of epsilon in addition to HIS203, HIS322 and HIS423.

3.5.2.1. Assessment of Protonation States

To set up a simulation system, it is necessary to assign protonation states. The protonation states of certain amino acids (Arginine, Lysine, Tyrosine, Cysteine, Histidine, Glutamine, Aspartate) depend on the local micro-environment and its pH. In a highly polar microenvironment, the charged form is more stable, whereas in a less polar microenvironment, the neutral form is favored. Under physiological pH conditions, TYR, LYS, CYS, and ARG typically maintain their standard protonation states, while GLU, ASP, and HIS may adopt non-standard forms. The protonation pattern plays a critical role in the catalytic function and structural stability of proteins. Before initiating a classic molecular dynamics (MD) simulation, the protonation states must be determined and assigned. To

ensure realistic MD simulations, it is crucial to accurately assign the appropriate protonation states [44,45].

Protonation states of the catalytic residues, ASP259 and HIS154 were assigned before the MD simulations as follows: ASP259 was negatively charged and HIS154 was kept in its neutral form, and protonated from the delta position instead of epsilon in order to form the required ionic diad for the transamination reaction. For the protonation states of the other titratable residues, Molecular Operating Environment (MOE)[46] was used to detect pKa values for each variant and no further regulations were made for the default protonation states of AMBER according to the obtained results from MOE.

3.5.3. Creation of PLP-bound Transaminase Unit

In the resting state of the transaminase, PLP is covalently bound to Lysine288. It is necessary to create bonded-case PLP and LYS units' modified force field parameter files (frcmod) and library files (lib) so that MD simulations could be run. The steps followed to create these units are listed below. For this purpose, at first, PLP(covalently bound to LYS288), LYS288, and its adjacent residues, ALA287 and GLY289 were extracted as a new pdb file. The lastly created pdb file including PLP, LYS and its adjacent residues were loaded into "Xleap", which is a version of LEaP program of Amber module with a graphical user interface, allowing visualization of the molecular structures and their modifications. The necessary atoms of adjacent residues (starting from the Lysine288 ends, the atoms until α-Carbons, including α-Carbons) were kept to consider their peptide bonds with LYS and the rest of them were erased. Deleted atoms are selected as pink in Figure 3.2.

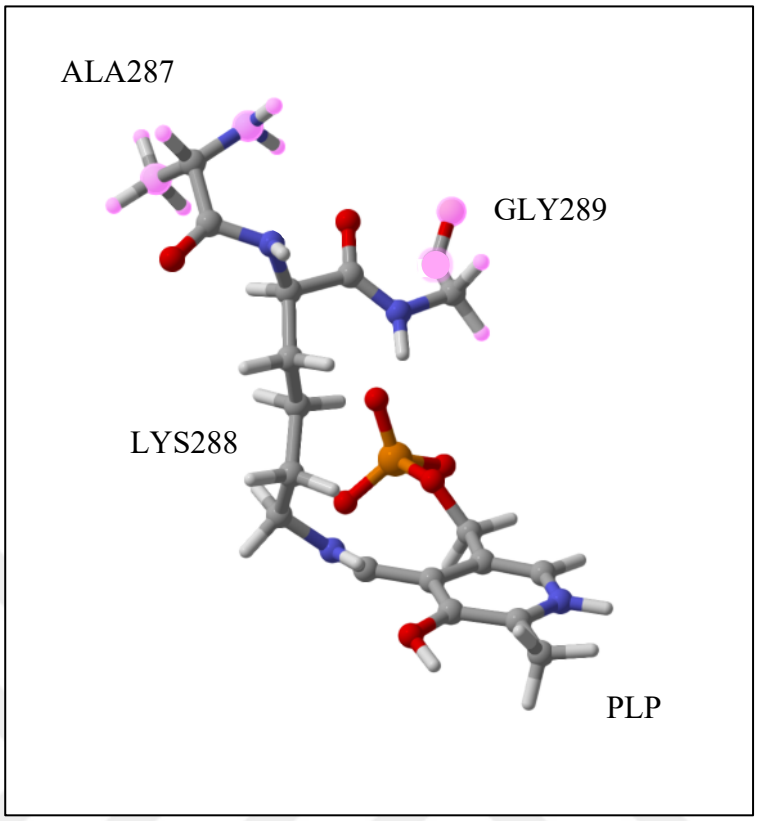

Figure 3.2. PLP bonded to LYS288 unit with adjacent residues

The resulting structure was used for geometry optimization. Some atoms were needed to be frozen during the geometry optimization in order to conserve the general structure and to prevent extra intramolecular bonds. All atoms were frozen except for CH_3 and OH groups of pyridine ring of PLP, oxygens of Phosphate group, oxygens of the adjacent residues and all the hydrogens present in the structure. Frozen atoms are shown in Figure 3.3 as pink.

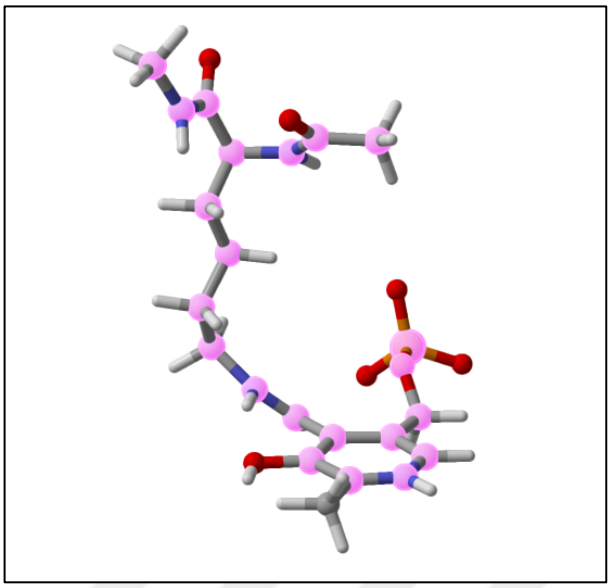

Figure 3.3. Frozen atoms of PLP bound LYS unit in geometry optimization

Geometry optimization and single point energy calculation were performed with the same method used in PLP parametrization. Prep input (prepin) file was generated using output file of the single point energy calculation. In generated prepin file, extra atoms mimicking peptide bonds with LYS288 were deleted and the atom types listed in prepin file were changed to proper Amber atom types for all residue atoms. After the removal of the extra atoms mimicking peptide bonds, total charge of the system was -0.045769. Therefore positive charge of +0.045769 was added onto the Phosphorus atom in order to get a neutral system. Total charge was checked again and parameter file of the resulting structure was generated in order to add missing parameters in MD setup.

Since the resulting unit includes both LYS and PLP, it was necessary to split them into two to generate two different library files for covalently bound LYS residue and covalently bound PLP. At first, “LYP” (LYS covalently bound to PLP) residue was created as follows. All atoms coming from PLP was removed in Xleap. Resulting LYP structure is given in Figure 3.4. Head and tail atoms of LYP were defined as N is the head and C is the tail. LYP was set as a residue type protein and the resulting system was saved as a library file.

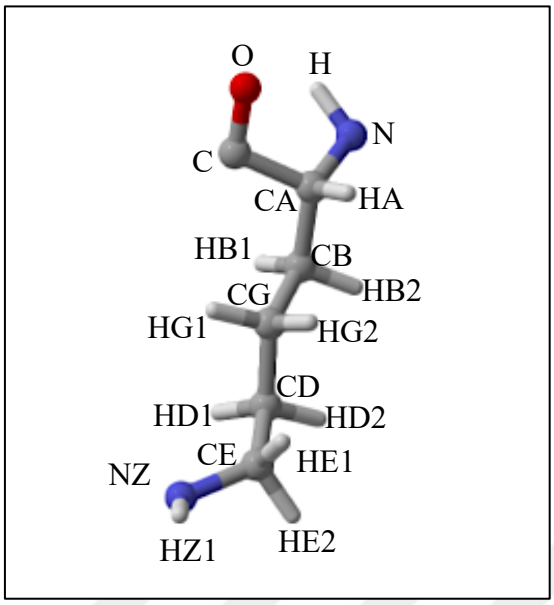

Figure 3.4. LYP

Secondly, library file of covalently bonded PLP (NLP) was generated only keeping the atoms of bound PLP and deleting the rest. Resulting NLP structure is given in Figure 3.5. Final structure was saved as a library file.

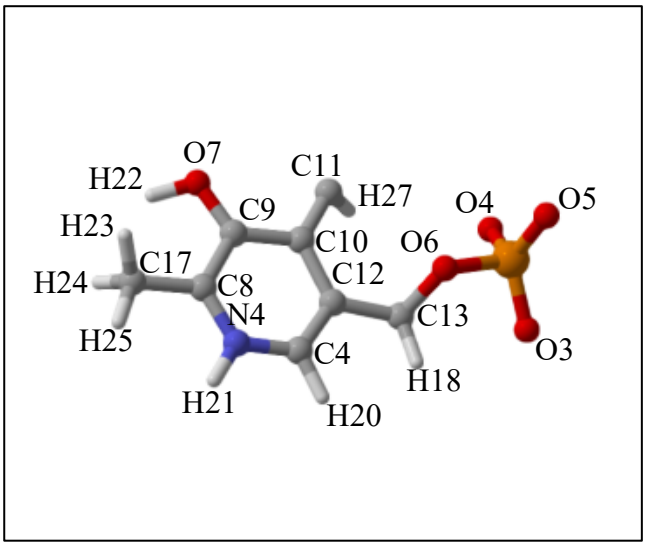

Figure 3.5. NLP

Resulting library files were loaded to tleap module while preparing MD setups of the complexes of PLP bonded to enzyme. In these MD setups, NZ atom of Lysine residue was bonded to the Carbon atom of PLP numbered as 11 by a special line.

3.5.4. MD Simulation Protocol

After the completion of MD setups, the protocol followed in MD simulation runs are given in Figure 3.6.

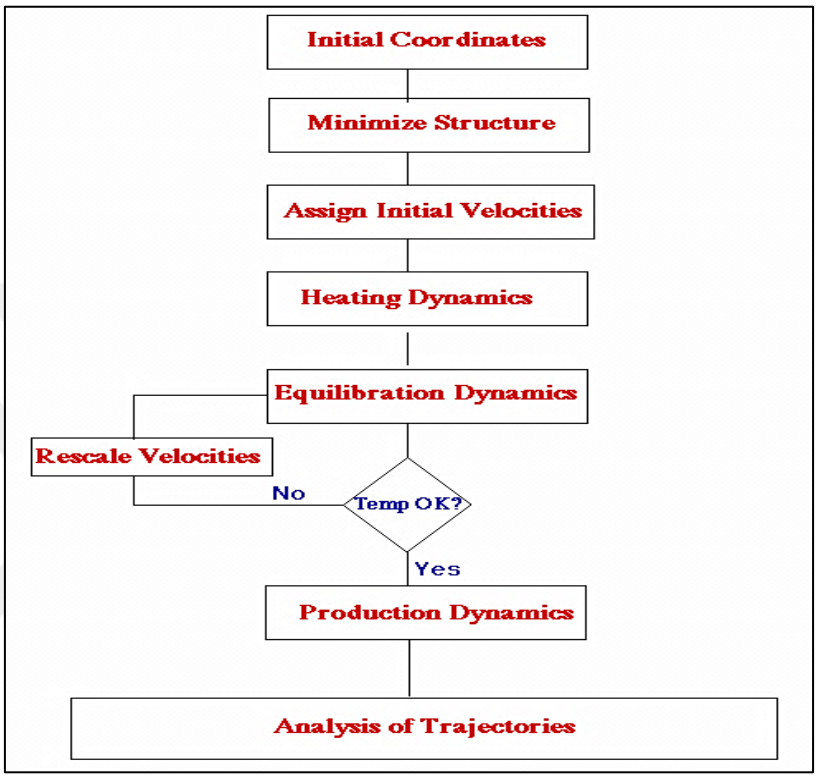

Figure 3.6. MD simulation steps

3.5.5. Minimization

First simulation type in MD was minimization of the systems. Minimizations were performed in 2 steps. Each step was run for 20,000 time steps. In the minimization of the enzyme-cofactor complexes, a constrain having the value of 150 was put onto whole protein to freeze protein residues and only the water molecules was released, for the first minimization. In the second one, no constrain was used, all residues in the complex was released. The non-bonded cutoff distance was specified as 8 Å. Systems were specified as constant volume system by setting “ntb” to 1. “Igb” which is a parameter for the generalized Born or Poisson-Boltzmann implicit solvent models was taken in its default value of zero. SHAKE was not performed in minimization steps. (In SHAKE algorithm, the system of non-

linear constraint equations is solved using the Gauss–Seidel method which approximates the solution of the linear system of equations using the Newton–Raphson method.)

3.5.6. Heating

After minimization step, minimized systems were heated from 0 K to 300 K in 6 steps (50 K heating for each). In the heating steps for cofactor-enzyme complexes, PLP was restrained by putting a constraint of 30 onto it. Langevin thermostat was used.

3.5.7. Equilibration

Restart files generated after the last heating step was used for equilibration runs. Equilibrium was performed in 2 steps. First one was performed in NVT (Canonical Ensemble), whereas the second one is in NPT (Isobaric-Isothermal Ensemble). Therefore a thermostat was used for NVT, and both thermostat and barostat was used for NPT. The same thermostat was used with the heating steps, as ntt=3, Langevin thermostat. For barostat, ntb=2, pres0=1.0, ntp=1, taup=2.0 parameters were used. ntb=2 means constant pressure periodic boundary conditions were used. Pres0=1.0 refers to reference pressure (in units of bars) at which the system is maintained at its default value, 1.0 (when NTP > 0). Ntp=1 means simulation were performed with isotropic position scaling. Taup parameter was used for pressure relaxation.

3.5.8. Production

The last step of the MD simulations was production run. This step included only one sub-step with the simulation time of 20 ns. No constraint was applied to any residue in production runs. Restart files generated after the last equilibration step was used for the production runs.

4. RESULTS AND DISCUSSION

From a critical perspective, one could consider the results obtained from MD simulations, which is based on full-atomic representation and well-established force fields, as the most accurate models for protein models. Bahar and co-workers[47] proved the accuracy of the coarse-grained GNM and ANM approaches in their study to assess the dynamics of α-amylase inhibitors with their MD simulations serve as a benchmark. By quantitatively comparing the slow mode curves of GNM and ANM with those obtained from MD simulations, it was found that the network models demonstrate a well agreement with MD simulations and it also means an improvement achieved by considering anisotropic effects within the calculations. With the support of these results, GNM and ANM models were used to get insights into the dynamic motions and basic functions of the examined transaminases lowering the high computational cost and time to complete the MD simulations with full-atomic representations for long scales.

In particular, while GNM and ANM exhibit similar performance to simulations or experiments when considering all modes, they diverge in the slow mode regime. ANM emerges as a more realistic model for estimating the slowest modes, indicating its greater suitability in capturing these modes that are likely to play a significant role in molecular function [47]. While the GNM assumes isotropic fluctuations in 1D and considers each residue as an independent site, ANM incorporates the directional preferences and 3D character of motions. In ANM, the molecule is still represented as a collection of N sites, one for each residue. However, instead of considering N-1 independent modes as in the GNM, ANM considers 3N-6 modes, which corresponds to the number of possible independent directions of motion in three-dimensional space. By considering the directions of motion, it becomes possible to understand the mechanisms and functional implications of these motions [48]. For example, the identification of concerted motions along specific directions can provide insights into allosteric communication, ligand binding, and other important biological processes. By elucidating both the magnitudes and directions of fluctuation vectors, ANM provides a more comprehensive understanding of the dynamics and function of biomolecular systems. Thus, at first, the dominant ANM modes seem acting on the evolution of variants were determined and global motions of the systems were brought into light. Then, corresponding GNM modes overlapping with these selected ANM modes

were specified and further used to determine hinge residues and correlation differences points of each variant. In order to achieve the overall goal, mutations close to hinges and cross correlation differences were determined to understand their effect on functional changes of the variants. Additionally, MD simulation results were evaluated to get insights into the active site interactions and to investigate the influence of PLP presence. Apart from this, MD simulations of apo-enzyme form of each variant provided a significant point of view on thermo-tolerance differences.

4.1. DETERMINATION OF THE DRIVING ANM MODES IN NATURAL EVOLUTION OF THE VARIANTS BY ANM-LD SIMULATIONS

In order to select the dominant ANM modes through the conversion of one variant to its next generation, ANML-LD Simulations[38] were used. Difference vectors of the structurally aligned variant couples (N48-KES23360, N48-N15, N15-N6, N6-KES23360) were compared with ANM mode vectors by ANM-LD and the modes having the highest overlap values were selected. At the end of the first cycle of ANM-LD simulations, highest mode overlaps were 0.34 for the transition from the most ancient variant to the modern-day transaminase, 0.30 for the transition from N48 to N15, 0.23 for the transition from N15 to N6, and 0.42 for the transition from N6 to KES23360. Overlap values for the examined variant transitions increased in next cycles and they were not lower than 0.5 in any simulations. Overlaps of the ANM mode vectors and difference vectors are shown in Figures 4.1 – 4.4. In the following figures, purple lines represent the difference vectors of the alignments and gray lines indicate ANM mode vectors. Corresponding ANM modes are attached right-hand-side of the graphs given in the following figures. Overlap of the minimum points of the selected ANM mode vectors with difference vectors indicate the accuracy of the selected modes.

ANM1, ANM2, and ANM3 are the selected ANM modes overlapping with the difference vectors of the alignment of N48 (the most ancient) and N15 (second generation) as shown in Figure 4.1. First three ANM modes are selected rather than a single mode since none of the individual ANM modes matches completely with the difference vectors for these sequential variants.

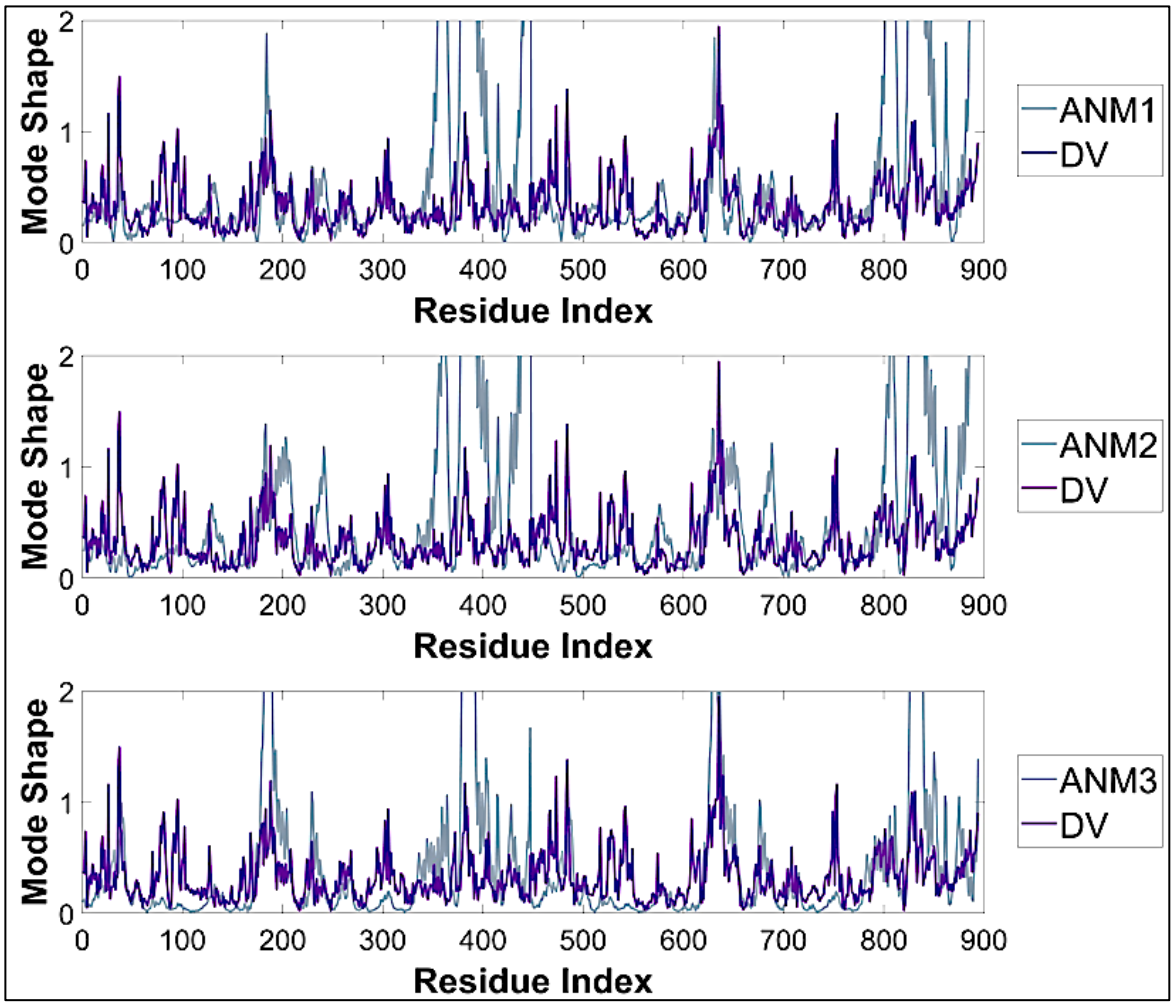

Figure 4.1. Difference vector-ANM overlaps for N48-N15

ANM1, ANM2, and ANM3 are the ANM modes coinciding with the difference vectors of the alignment of N15 (second generation) and N6 (third generation) as shown in Figure 4.2. It is again not possible to discriminate one ANM mode having a perfect match with the difference vectors.

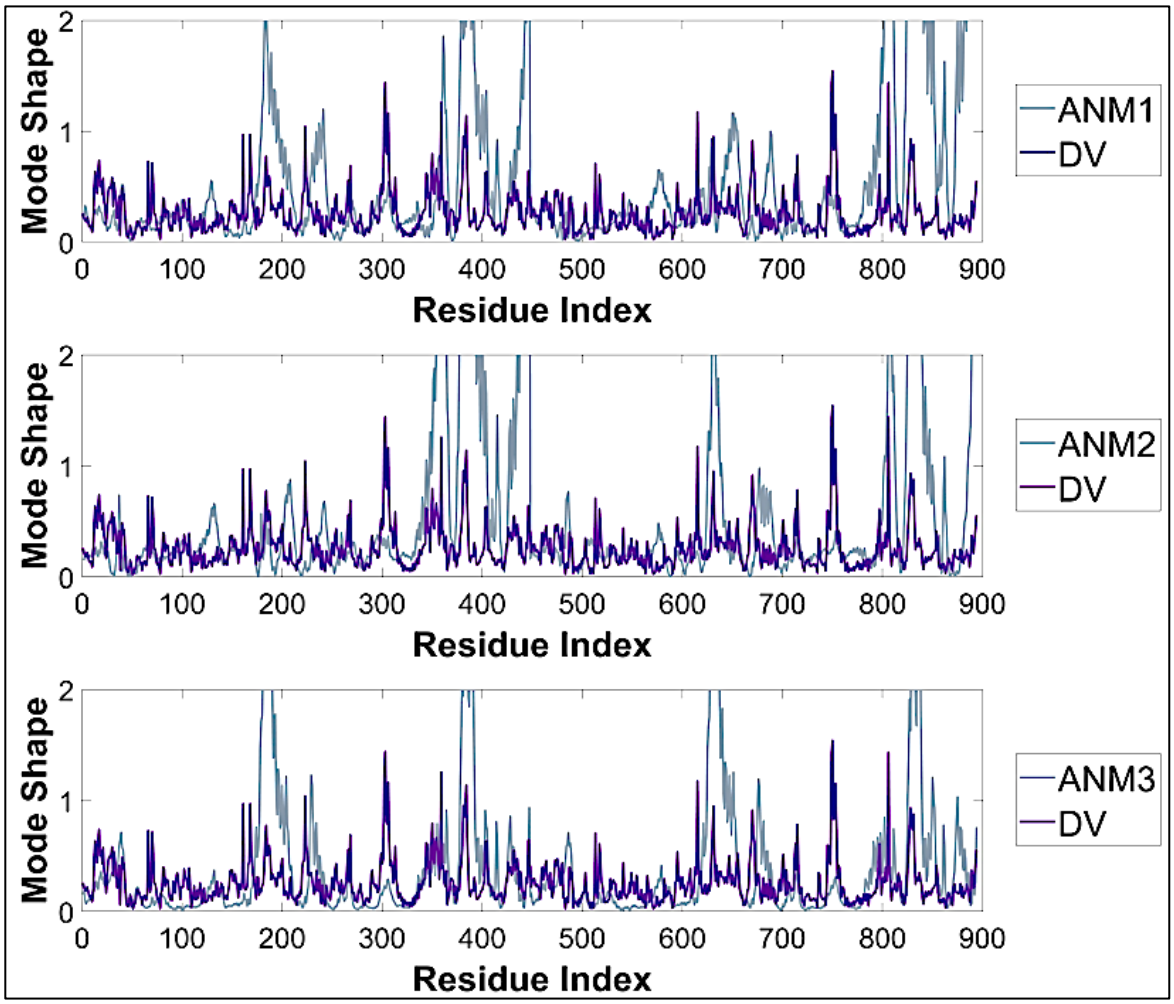

Figure 4.2. Difference vector-ANM overlaps for N15-N6

For the evolution from N6 to KES23360, ANM1 is the better overlapping mode with the difference vectors in the case of chain A (residues:1-455), whereas ANM3 shows better alignment with the difference vectors of chain B (residues: 455-910). Therefore both ANM1 and ANM3 were selected as matching modes.

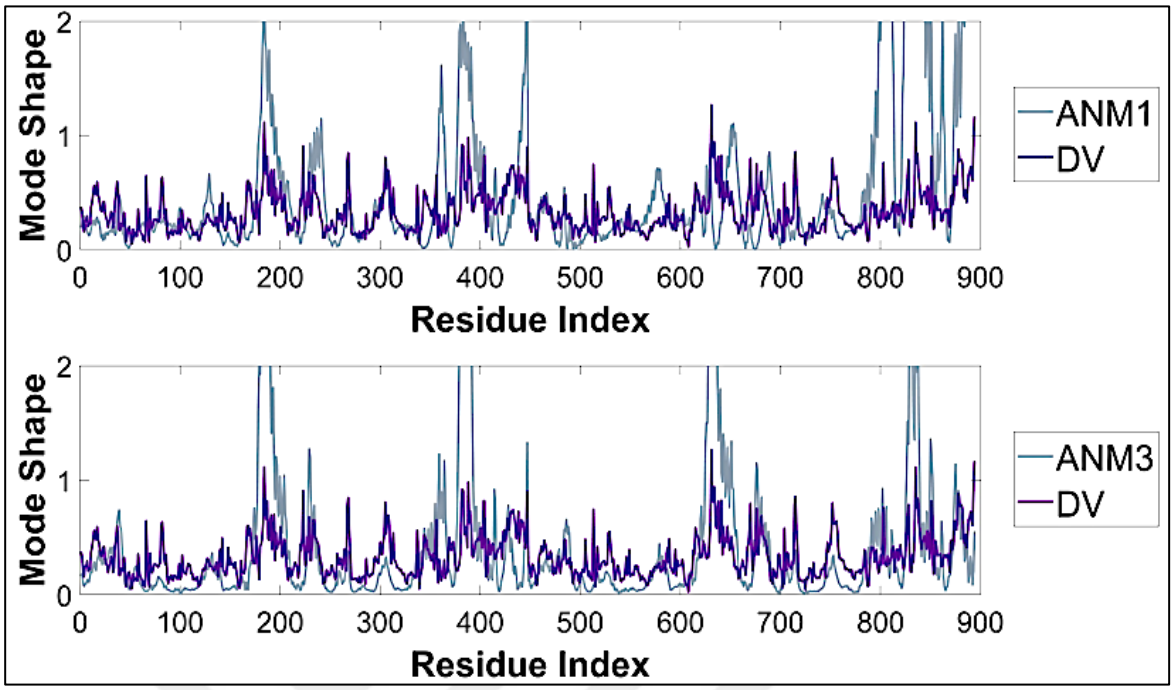

Figure 4.3. Difference vector-ANM overlaps for N6-KES23360

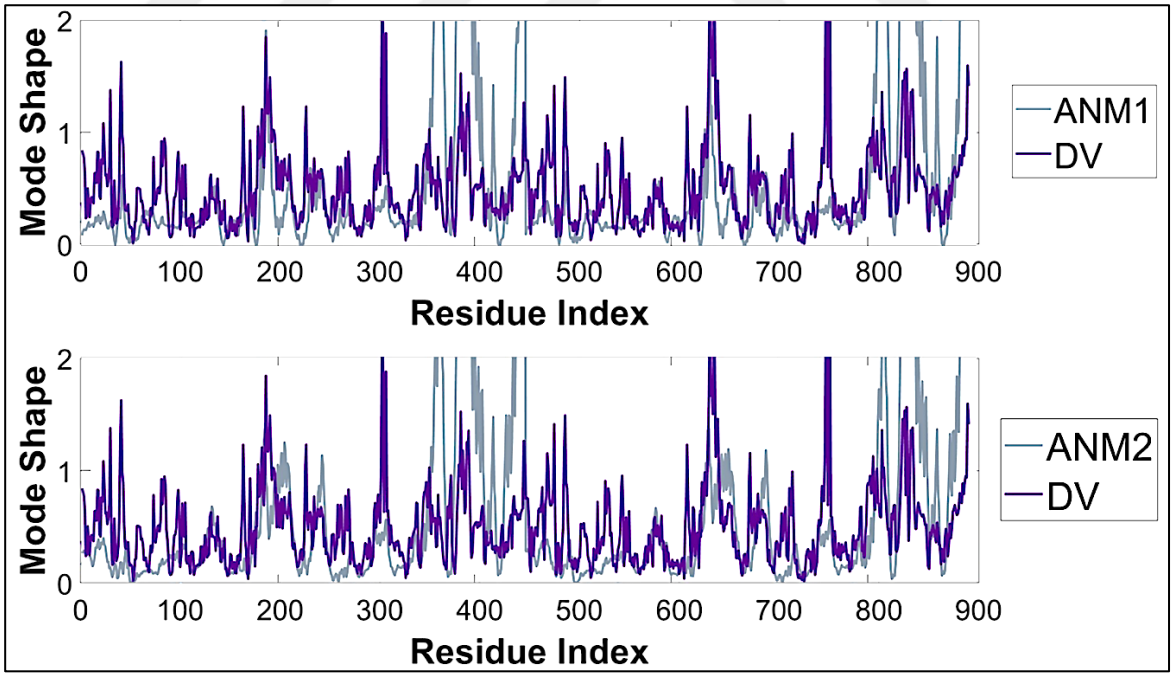

Figure 4.4. Difference vector-ANM overlaps for N48-KES23360

Although KES23360 and N48 are not two sequential variants, difference vectors and corresponding ANM modes were investigated in order to determine the dominant ANM

modes through the evolution from the most ancient to the most modern transaminase. In this case, ANM1 and ANM2 are the best fitting modes to difference vectors of the structure alignments of N48 and KES23360. Additionally, both modes have almost the same vectors.

4.2. DETERMINATION OF THE GNM MODES CORRESPONDING WITH THE SELECTED ANM MODES

Dominant ANM modes' vectors that were selected in the previous section were compared with the slowest GNM modes' vectors in order to determine the driving GNM modes in evolution of the variants. Determination of the dominant GNM modes is important to be able to designate the hinge residues and difference points in dynamic motions of the variants during evolution and to associate them to the mutated residues. GNM modes having better matches with the dominant ANM modes are given in Figures 4.5 – 4.10. All graphs belonging to the comparison of the mode vectors of the selected ANM modes by ANM-LD and first five slowest GNM modes are given in Appendix C.

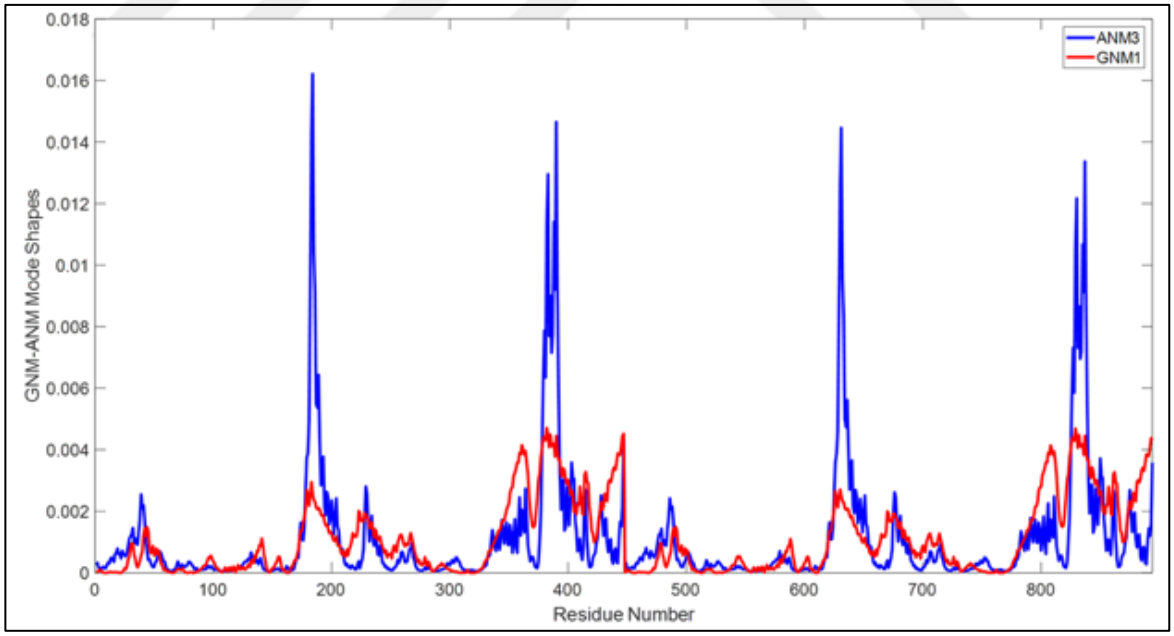

Figure 4.5. ANM3 overlap with GNM1 for N48-N15

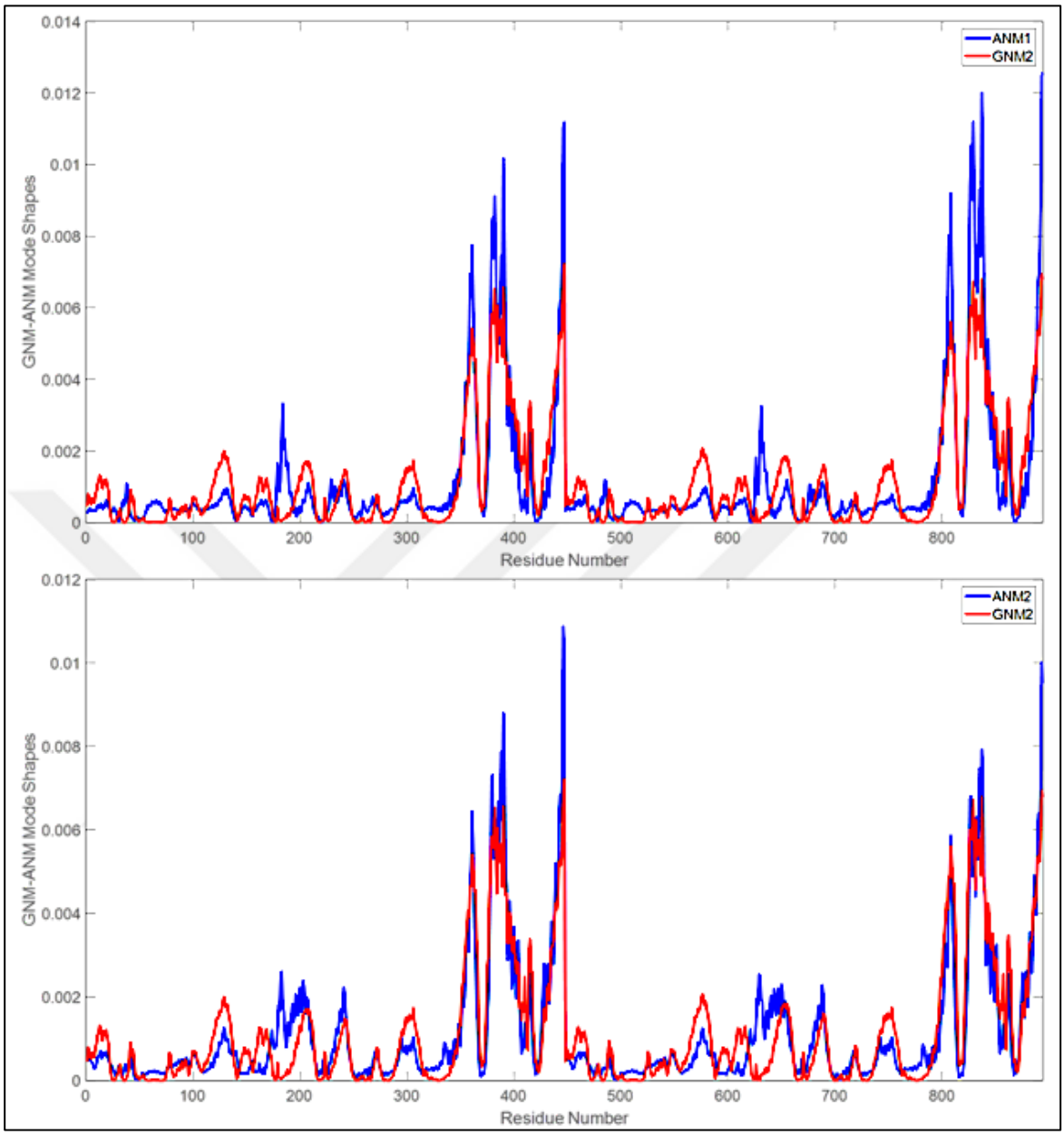

Figure 4.6. ANM1&ANM2 overlaps with GNM2 for N48-N15

In the case of the evolution from N48 to N15; ANM1, ANM2 and ANM3 are the dominant ANM modes matching with the alignment difference vectors. GNM2 is the corresponding GNM mode with both ANM1 and ANM2 as indicated in Figure 4.6. ANM3 is also one of the selected ANM modes driving the evolution from N48 to N15, however when ANM3 vectors are compared with the GNM mode vectors of the slowest 10 modes, no considerable overlap is detected for ANM3, except for GNM1 as seen in Figure 4.5. GNM predicts N-1 modes, while ANM predict 3N-6 modes, where N is the number of residues of the protein.

This means that each mode in the GNM roughly coincides with three modes in ANM. To visualize and compare the dispersion of modes for GNM and ANM, the GNM results generally plotted at every third mode index, starting from GNM2, rather than GNM1 [47]. For this reason, although it shows an overlap with ANM3 to a certain extent, GNM1 was not taken into consideration as one of the dominant modes for GNM case.

ANM1, ANM2 and ANM3 are the ANM modes driving the evolution from N15 to N6 and when the ANM vectors of these 3 modes are superimposed with the slowest 10 GNM modes, GNM2 is the best matching mode with ANM1 and ANM2 which is shown in Figure 4.7. In Figure 4.8, for ANM3, GNM4 is the best overlapping mode. As a result, GNM2 and GNM4 were chosen as the dominant modes in evolution of second to third generation transaminases.

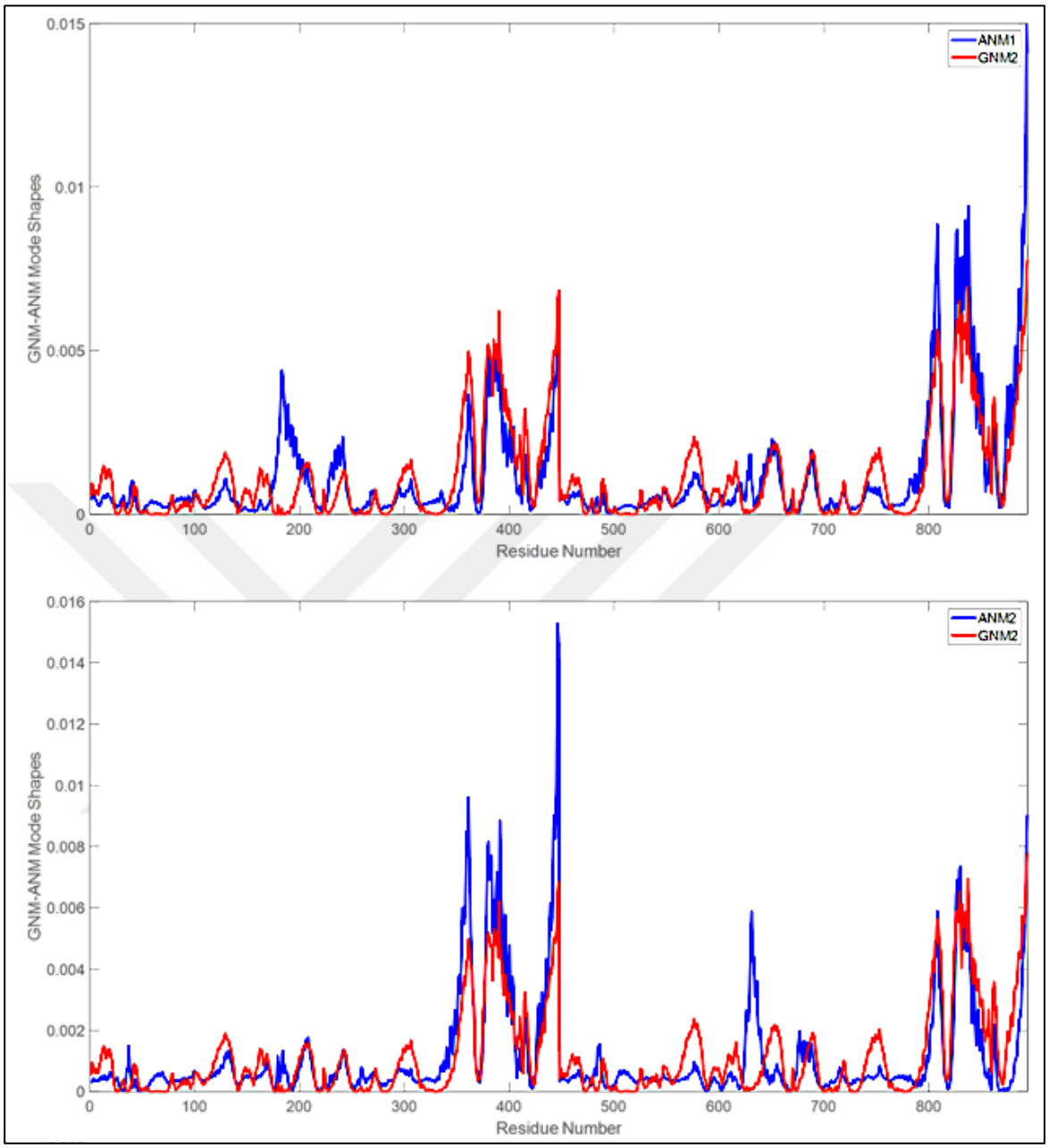

Figure 4.7. ANM1&ANM2 overlaps with GNM2 for N15-N6

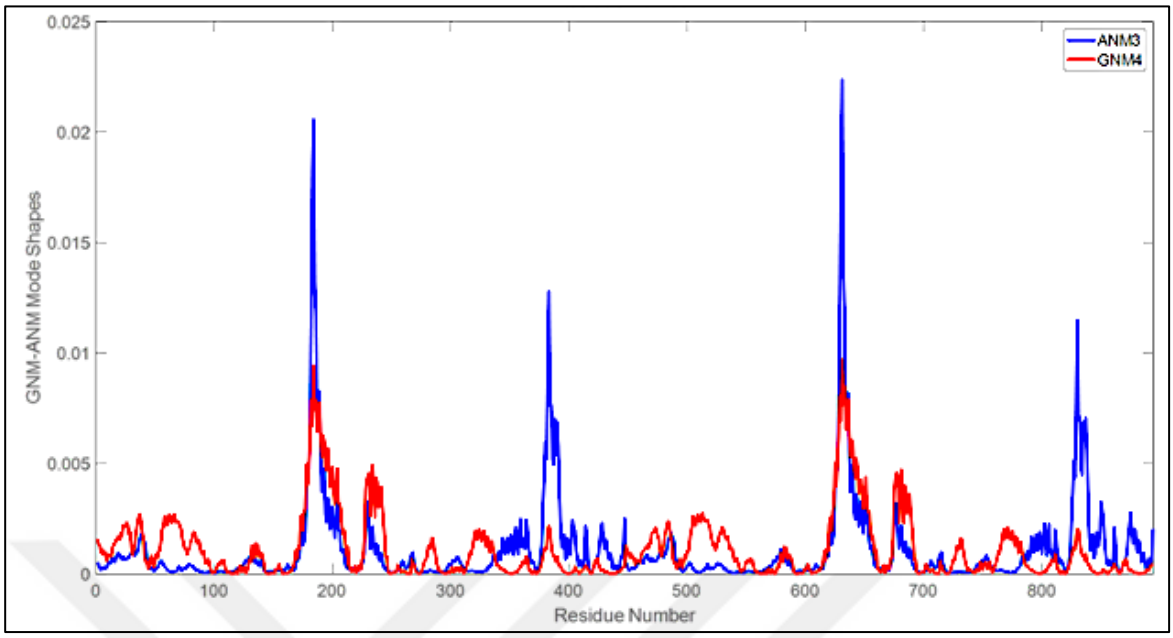

Figure 4.8. ANM3 overlap with GNM4 for N15-N6

Figure 4.9 was used to determine corresponding GNM modes with the dominant ANM modes for the variant couple, N6 and KES23360. According to the vector representations on the graphs given in this figure, GNM2 has the best alignment on ANM1. On the other hand, GNM4 vectors have better overlap with ANM3 than the other GNM modes.

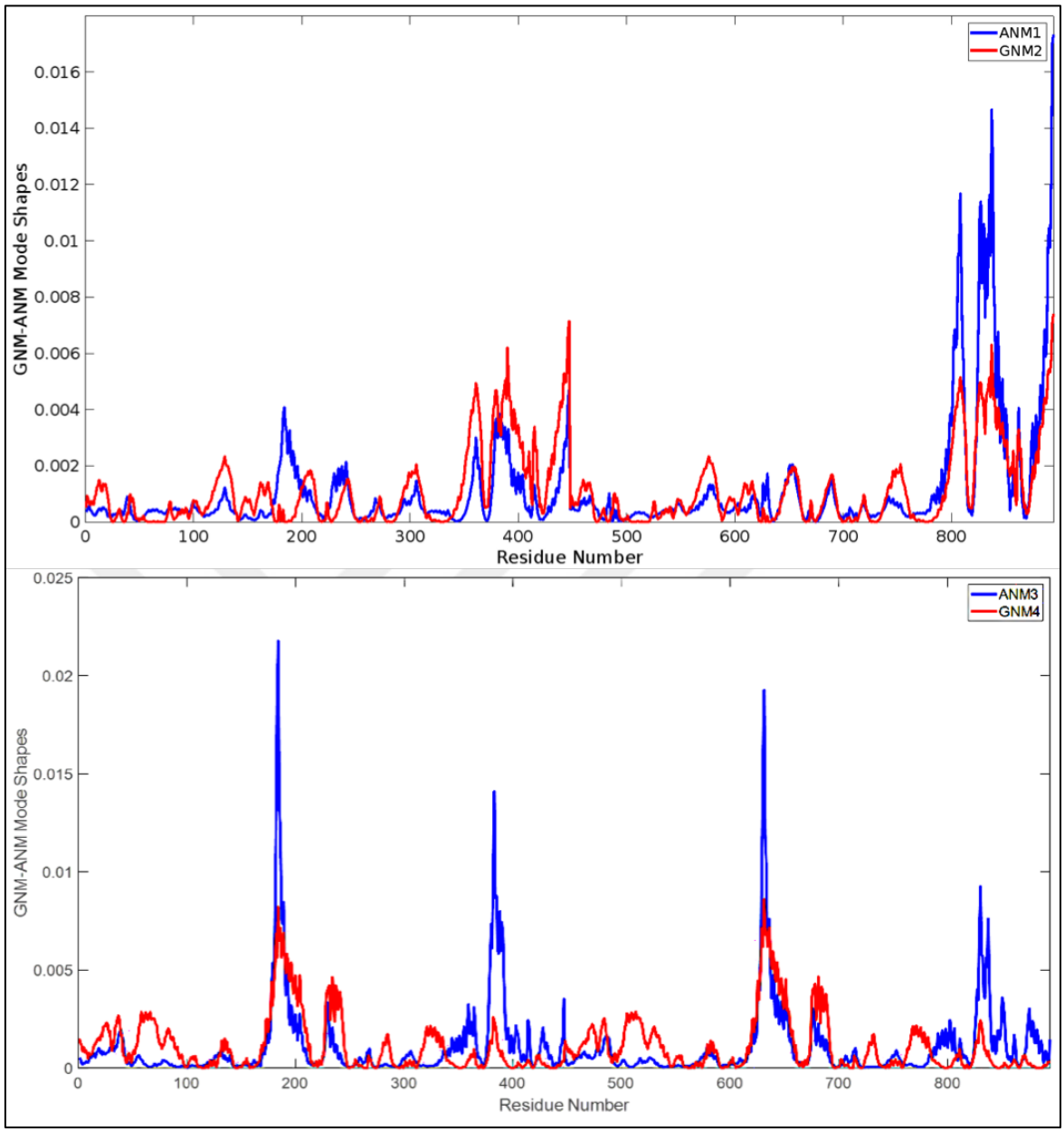

Figure 4.9. ANM1&ANM3 overlaps with GNM modes for N6-KES23360

When the dominant mode driving the global motion transitions during the overall natural evolution from the most ancient variant to the most modern one was tried to detect, both ANM1 and ANM2 were suitable modes for the alignment difference vectors. Both of the selected ANM modes' vectors coincide with GNM2 as represented in Figure 4.10, thus GNM2 was chosen as the dominant mode for the overall evolution in this transaminase family.

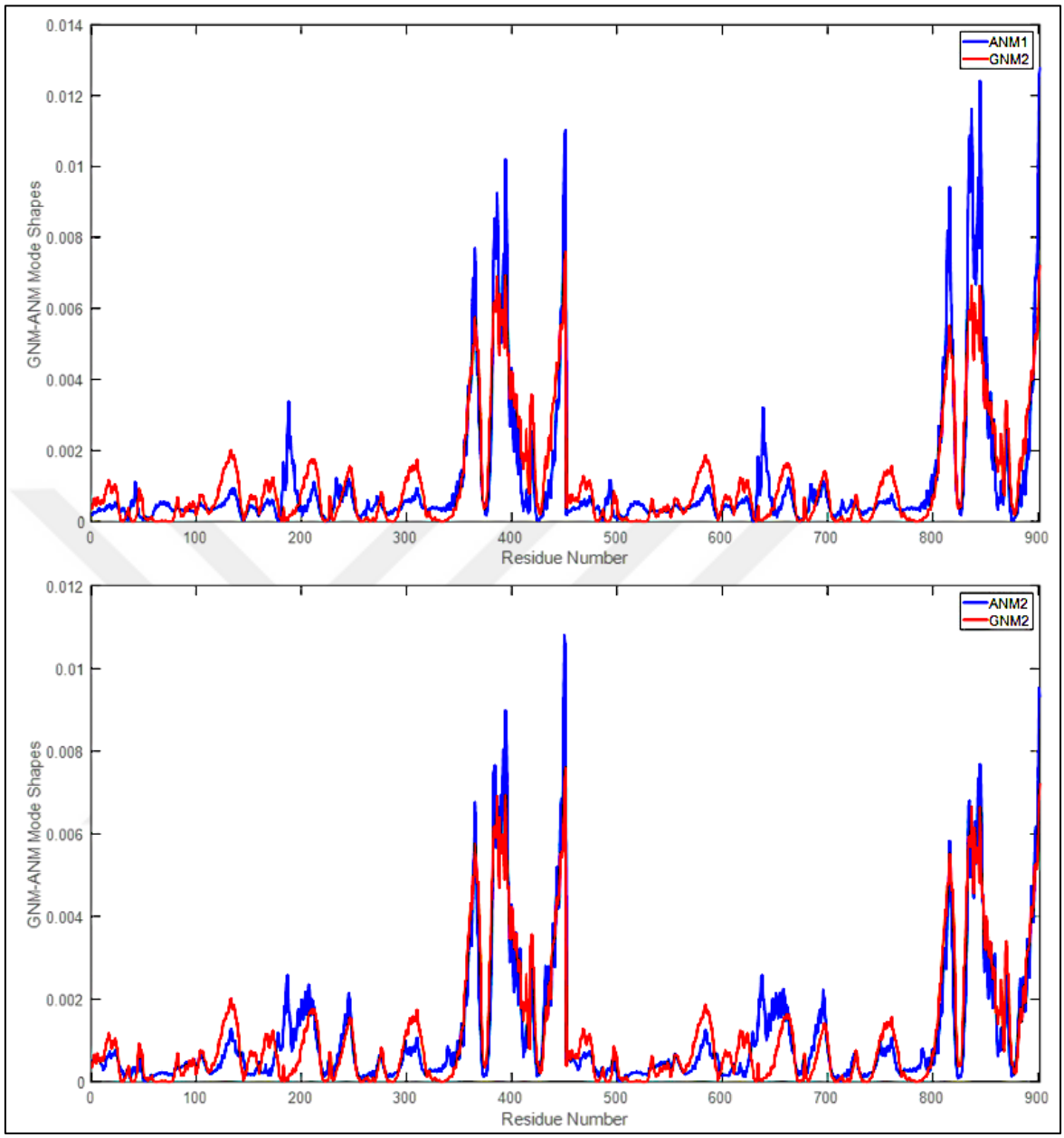

Figure 4.10. ANM1&ANM2 overlaps with GNM2 for N48-KES23360

4.3. HINGE POINT DETERMINATION

In the context of collective dynamics in proteins, hinges play a crucial role in facilitating the coupling of motions between different domains or regions within a complex. In the global modes predicted by the Gaussian Network Model (GNM), hinge residues tend to exhibit the lowest mobilities or fluctuations which means that these residues have a higher degree of stability compared to other residues in the molecule. The lower mobility of hinge residues

indicates that they are less prone to undergo large-scale conformational changes or fluctuations, making them essential for maintaining the overall structure and stability of the molecule [49]. Hingeprot[20] web server was used to determine the hinges of transaminases. Hingeprot designates the hinge residues for the first two slowest GNM modes. Table 4.1 includes the hinge residues of each variant determined by Hingeprot. Only the hinges of GNM2 were considered since GNM2 is one of the selected GNM modes corresponding to dominant ANM modes for each examined variant couple whereas GNM1 was out of our scope for the evaluation of these variants.

Table 4.1. Hinge residues determined by Hingeprot for the second slowest GNM mode

Transaminase Variant	**Hinge Residues for GNM2**
N48	37A 67A 342A 458A 37B 67B 342B
N15	37A 67A 342A 458A 37B 67B 342B
N6	37A 67A 342A 458A 37D 67D 342D
KES23360	37A 67A 342A 458A 37B 67B 342B

Regarding the residues listed in Table 4.1, hinges are in a complete agreement in both chains of these homodimeric transaminases and do not differ from one variant to another as expected. Mutations or perturbations in hinge residues can have significant effects on the dynamics and function of biomolecules since they restrict the motions of adjacent domains, thereby contributing to the overall stability of the complexes. GNM2 mode shapes and hinge residues on 3D structure of each variant are shown in Figures 4.11. Blue colored residues represent negatively correlated residues and red residues are positively correlated residues. Green spheres indicate hinge points for GNM2. In Figure 4.11, it is clearly seen that hinges are placed between the correlated domains.

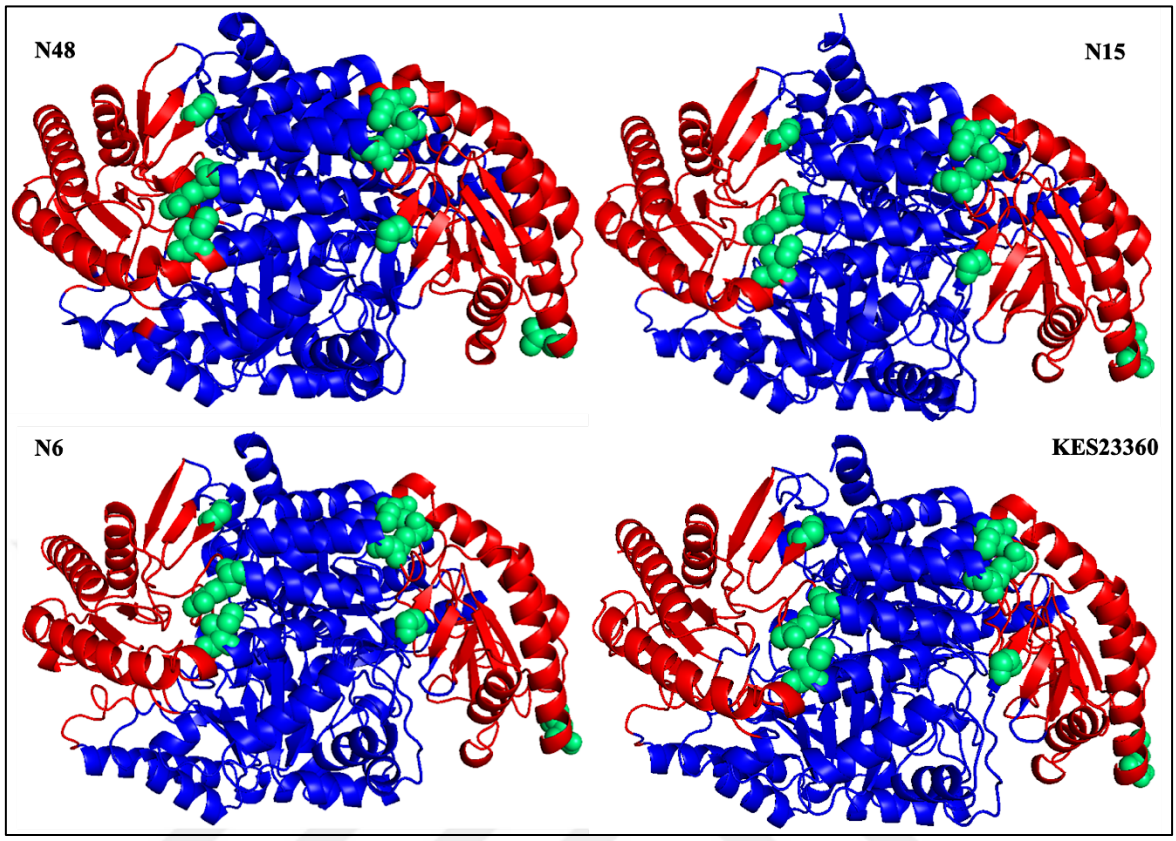

Figure 4.11. GNM2 mode shapes and hinges for N48, N15, N6 and KES23360 (green spheres: hinge residues, red: positive correlation regions, blue: negative correlation regions)

4.3.1. Mutations Close to Hinges

Since hinges are crucial points for the function and stability of enzymes, whether there are some mutations nearby the hinges in 3D structures is investigated. In Figures 4.12 and 4.13, mutated residues passing from N48 to N15 are shown in red color and hinges are represented as dark blue meshes. In Figures 4.14 and 4.15, red regions are the mutations in evolution from N48 to N15 whereas dark blues are the mutations in evolution from N15 to N6.

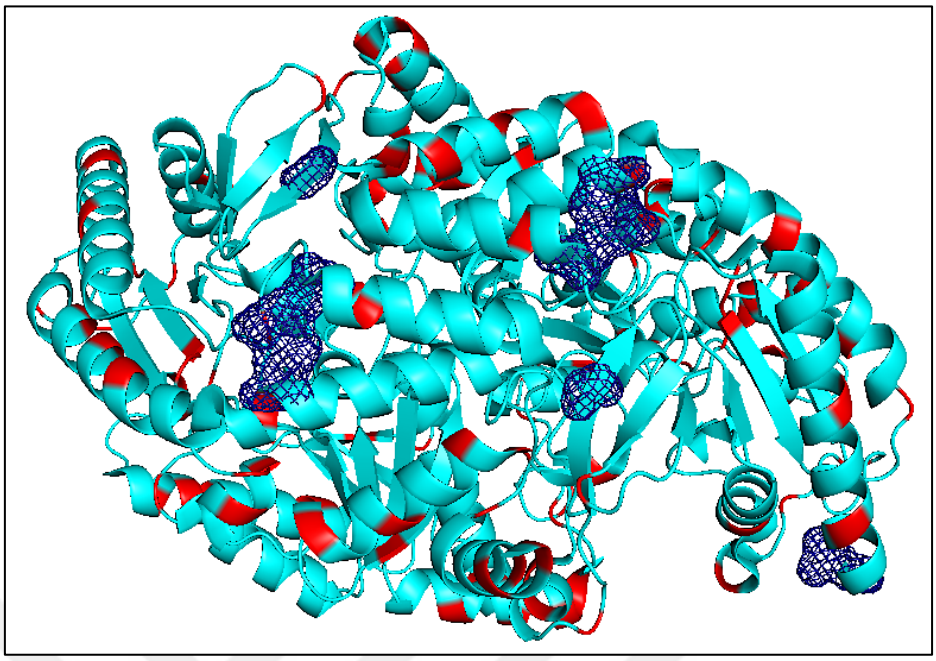

Figure 4.12. Mutations and hinges of N48 on 3D structure (meshes: hinge residues, red: mutations in evolution of N48 to N15)

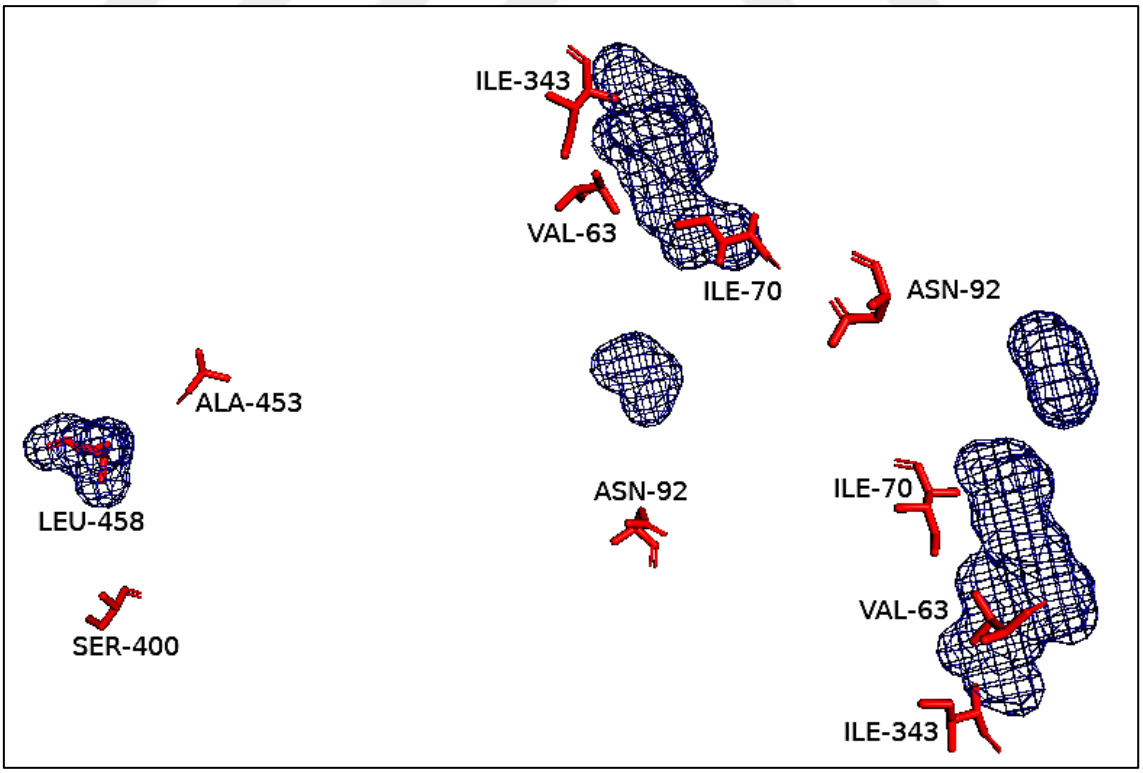

Figure 4.13. Mutated residues close to hinges of N48

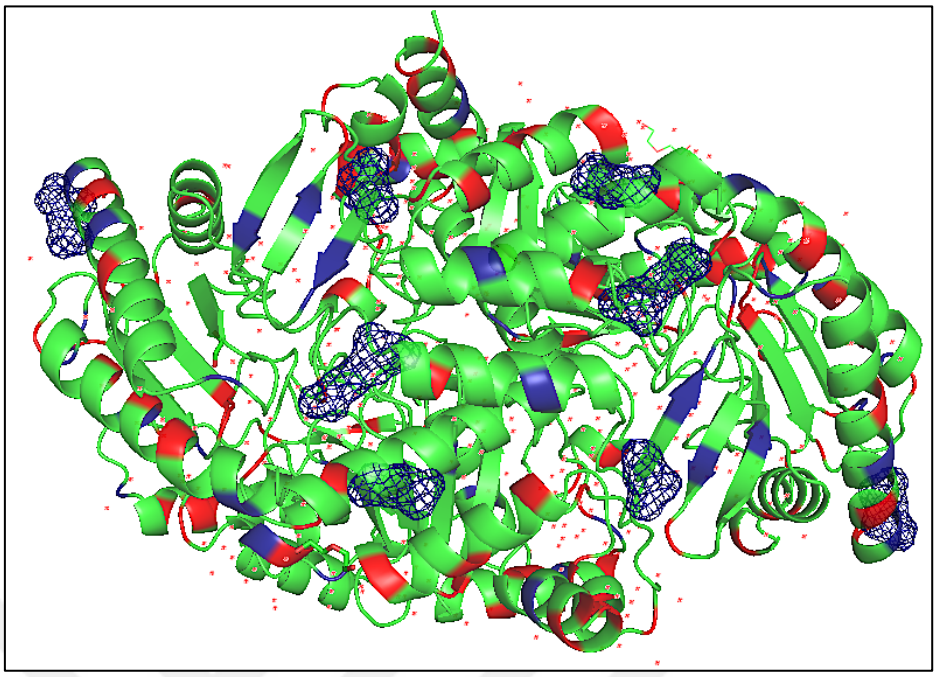

Figure 4.14. Mutations and hinges of N15 on 3D structure (meshes: hinge residues, red: mutations in evolution of N48 to N15, dark blue: mutations in evolution of N15 to N6)

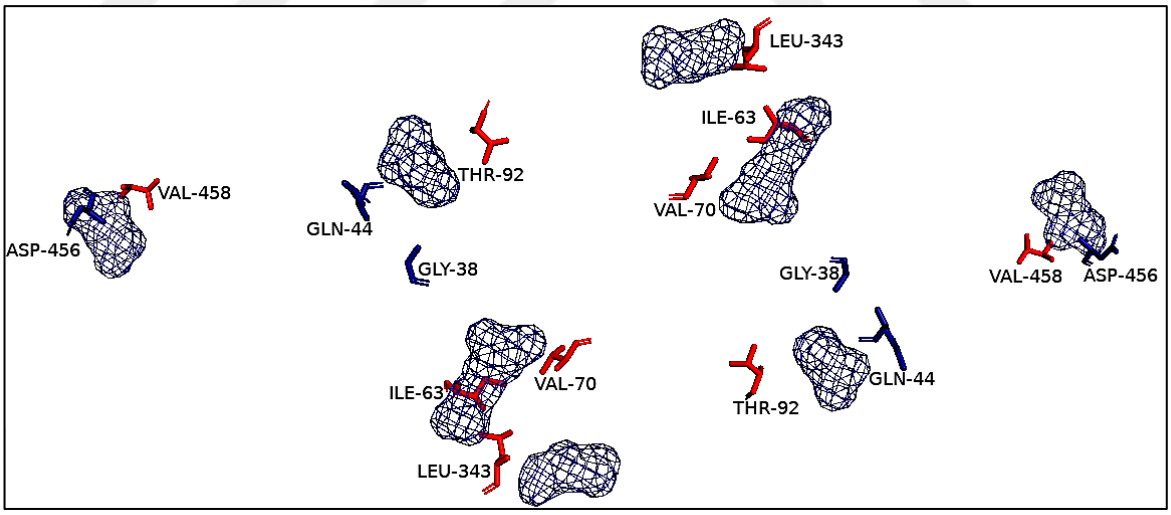

Figure 4.15. Mutated residues close to hinges of N15 (meshes: hinge residues, red: mutations in evolution of N48 to N15, dark blue: mutations in evolution of N15 to N6)

In Figures 4.16 and 4.17, there is an additional color used for the mutations in addition to red and dark blue which is magenta, represents the double-mutated points. Red and dark blue are used as in the same analogy with the previous figures, red regions are the mutations in

evolution from previous generation to N6 and dark blues are the mutations in evolution from N6 to the modern-day transaminase.

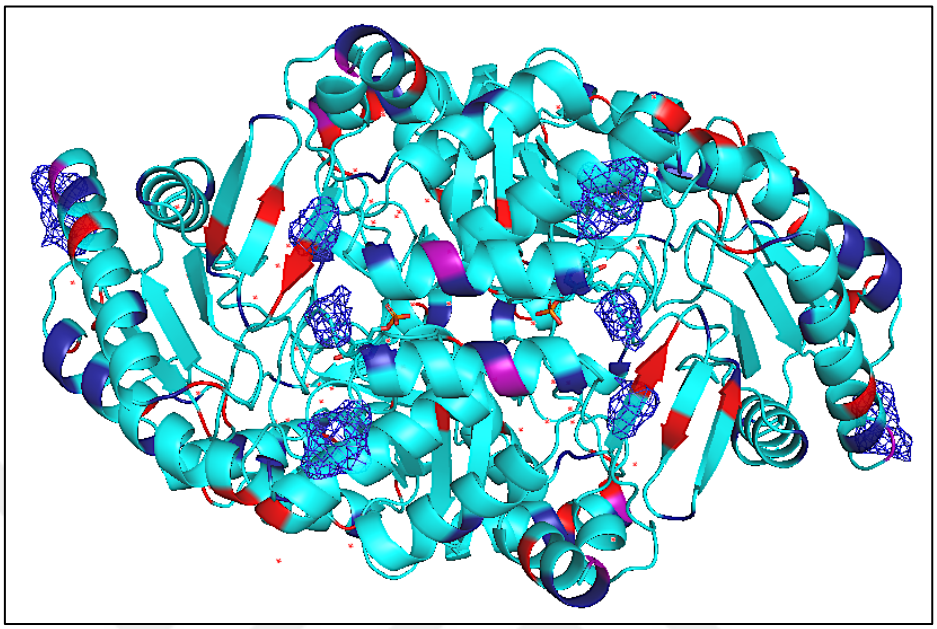

Figure 4.16. Mutations and hinges of N6 on 3D structure (meshes: hinge residues, red: mutations in evolution of N15 to N6, dark blue: mutations in evolution of N6 to KES23360, magenta: double mutation point)

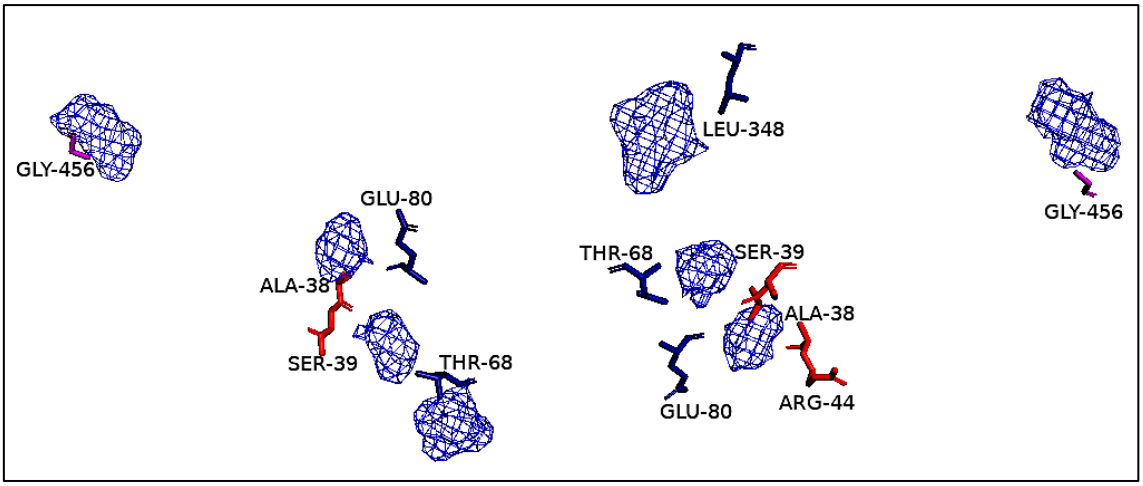

Figure 4.17. Mutated residues close to hinges of N6 structure (meshes: hinge residues, red: mutations in evolution of N15 to N6, dark blue: mutations in evolution of N6 to KES23360, magenta: double mutation point)

In Figures 4.18 and 4.19, only red color is used to identify mutation points same as the N48 case because there is no further mutation in KES23360. All mutations shown in KES23360 as red are the residues mutated from N6 to modern-day transaminase.

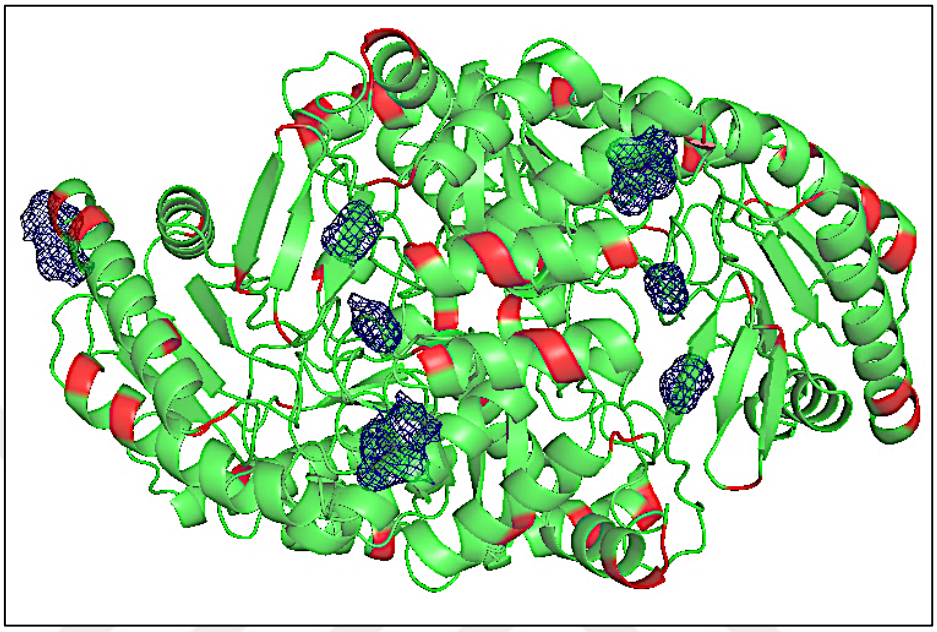

Figure 4.18. Mutations and hinges of KES23360 on 3D structure (meshes: hinge residues, red regions: mutation points)

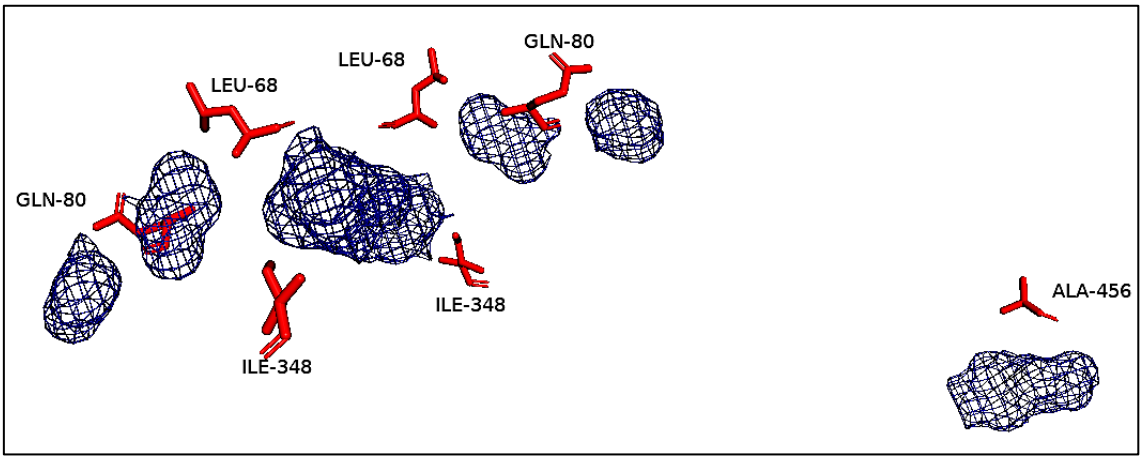

Figure 4.19. Mutated residues close to hinges of KES23360

Mutations occurred close to hinge points during the natural evolution of this transaminase family are summarized in Table 4.2. 456th residue mutated two times by natural evolution and it is close to hinge point at the end of A-chains, 458. When the residue 456 and hinge

458 are evaluated regarding the 3D structures, they are on the outer surface of the enzyme and located at the chain end. Considering the mobility of the chain ends distant from the inner surfaces or chain interfaces as well as the solvent exposure of the outer surfaces of enzymes, frequent mutation of 456th residue and can be related with the changes in the local environmental conditions affecting to that the enzyme has to accommodate itself by altering the surface physio-chemistry.

Table 4.2. Mutation points close to hinges of GNM2

Transaminase Variant	Mutation Points Close to Hinges
N48	63, 70, 92, 343, 400, 453, 458
N15	38, 44, 63, 70, 92, 343, 456, 458
N6	38, 39, 44, 68, 80, 348, 456
KES23360	68, 80, 348, 456

4.4. CROSS CORRELATION DIFFERENCE POINTS AND CLOSE MUTATIONS

After the determination of the corresponding GNM modes acting on the evolution of variants, cross correlation (c.c.) difference points for the sequential variants and also for N48-KES23360 were investigated. The residues having differences in their cross correlations from one variant to next generation are significant, because mutations at or near these residues can be related with the global motion and basic function transitions gained after the evolution. Cross correlation values of the residues can take the value of -1 or +1. +1 means the residue couples having positive correlation value show the tendency to move in the same direction, on the other hand -1 indicates that these residues have negative correlation. Cross-map differences show the residues that display different correlations for the compared variants. In these maps, cross correlation difference (c.c. difference $\neq 0$) points are represented as red lines and remaining residues having the same correlation for both variants (c.c. difference = 0) are the blue regions. Cross-map differences are evaluated only for the selected GNM modes in section 4.2.

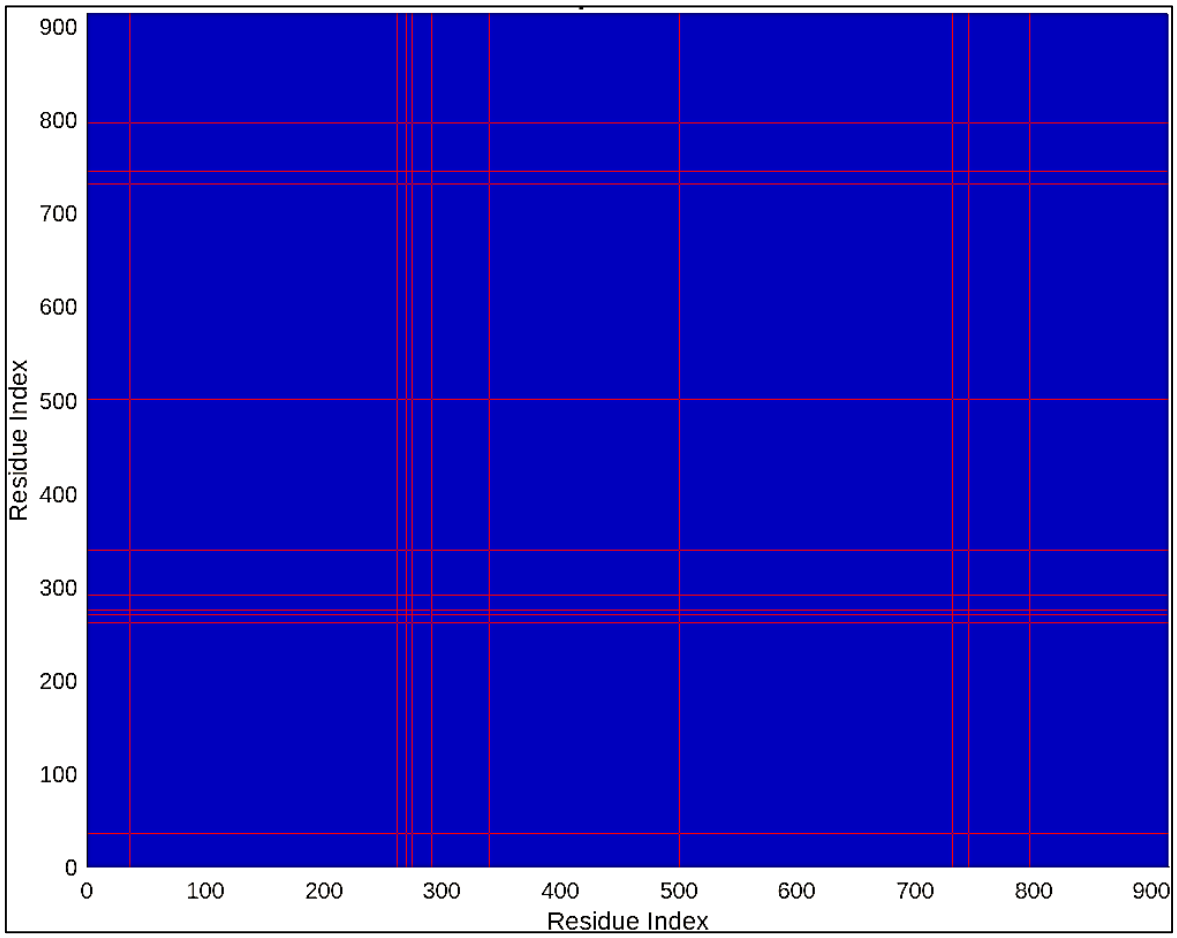

Figure 4.20. Cross-map difference of N48-N15 for GNM2

According to Figure 4.20, residues having different cross correlations in N48 and N15 are in A-chain: 37, 263, 271, 276, 292, 341; in B-chain: 45, 276, 289, 341. Here, residues 37 and 341 attract the attention as the residues in hinge regions. Additionally, residues 289 and 292 are close to hinge point 67 in 3D structure. Placement of the cross correlation difference points with respect to the hinges can be seen in Figure 4.21 in which meshes indicate the hinge points, red and blue small spheres represent cross correlation difference points.

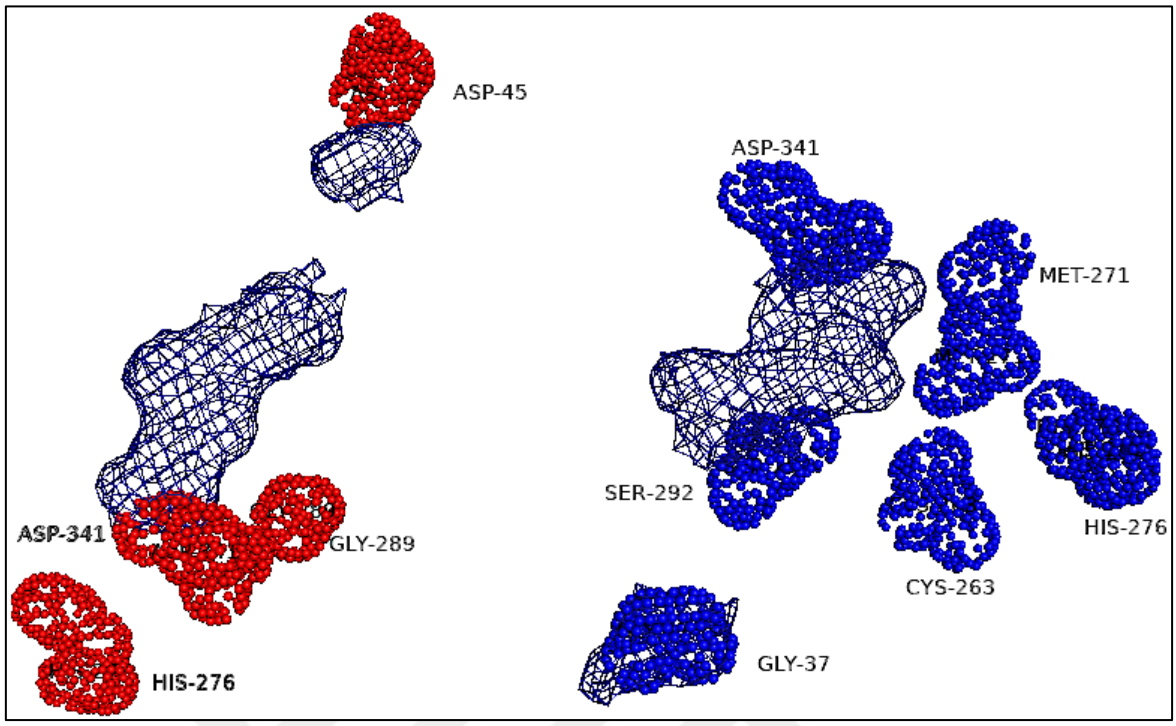

Figure 4.21. Relative placement of hinges and cross correlation differences in N48-N15

When these cross correlation difference points located on the 3D structure are tried to link with the mutations occurred in the evolution of N48 to N15, it is clear that cross correlation difference points are very close to mutated resides. 38th residue mutated in N48 and 37th residue shows cross correlation difference in A-chain. Nearby mutations for 263A, 271A and 276A-B on 3D structure: 105, 268, 272, 275. Moreover, there is a mutation at 264 while passing from N15 to N6. Mutations nearby 341A are 343 and 70. For the 45th residue in B-chain, there are 2 close mutations in 3D structure which are residues 4 and 46. These emphasized mutations are shown as cyan surfaces and the locations of the cross correlation difference points are represented as red and blue small spheres in Figure 4.22.

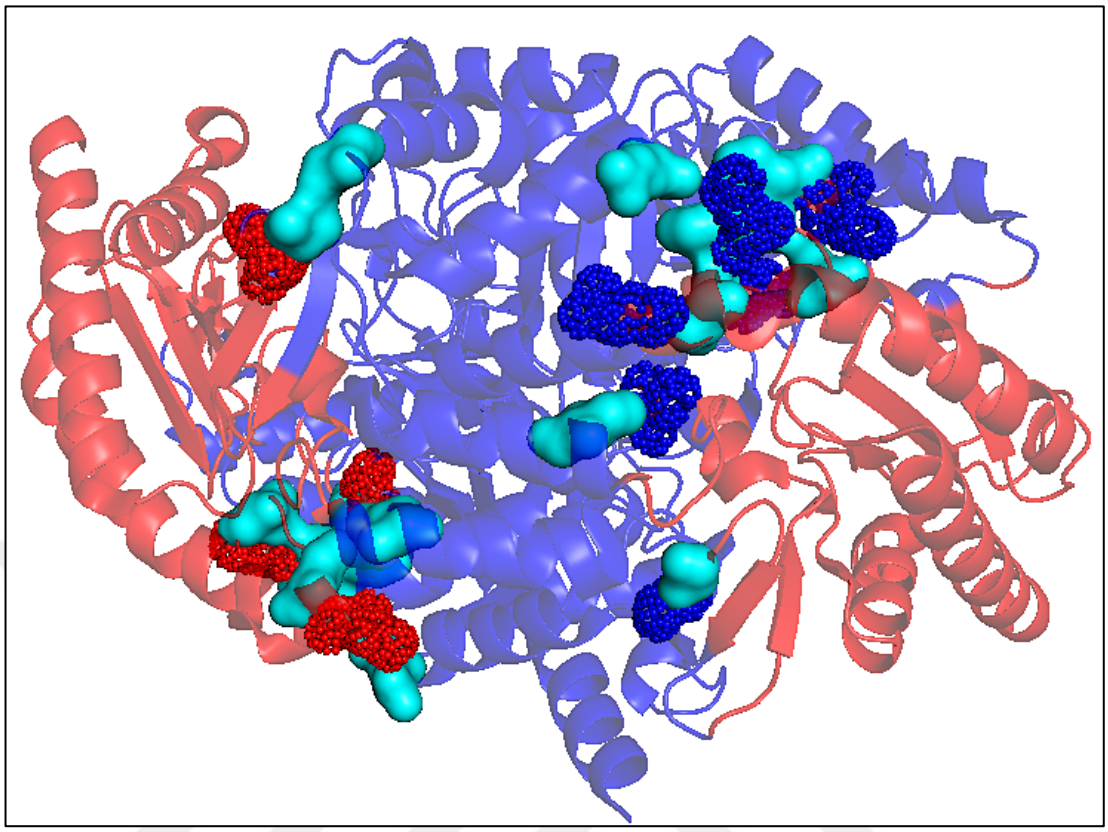

Figure 4.22. Cross correlation difference points and mutations in N48 (cyan surfaces: mutation points, red and blue spherical regions: c.c. difference points)

In Figure 4.23, cross-map difference of N15 and N6 for the second slowest GNM mode is given. Cross correlation difference points are 37, 67, 187, 188, 191, 192, 194, 195, 196, 197, 200, 229, 238 for chain A and 37, 45, 67, 187, 188, 191, 192, 194, 195, 196, 229, 292, 342 for chain B. As it is recalled in N48-case, 37 and 342 are hinge residues and these regions are again cross correlation difference points in the passage from N15 to N6 and 292 is close to hinge point 67. Another significant results is, one of the c.c. changing point, residue 67, is also a hinge. Positions of hinges with respect to the c.c. difference points are given in Figure 4.24 and three hinges are c.c. difference points as shown in the figure.

For the c.c. difference series starting from 187 to 200, 196th and 199th residues are mutation points within this range. Nearby mutations for 238A are 236A and 239A. In B-chain, there is a mutation in 44B neighboring the 45B, c.c. difference point. All mutations close to c.c. changing regions are represented in Figure 4.25 as cyan surfaces. Although there is no close

mutation for 229, having a c.c. difference is important for this residue since it is nearby the active sites of the both chains.

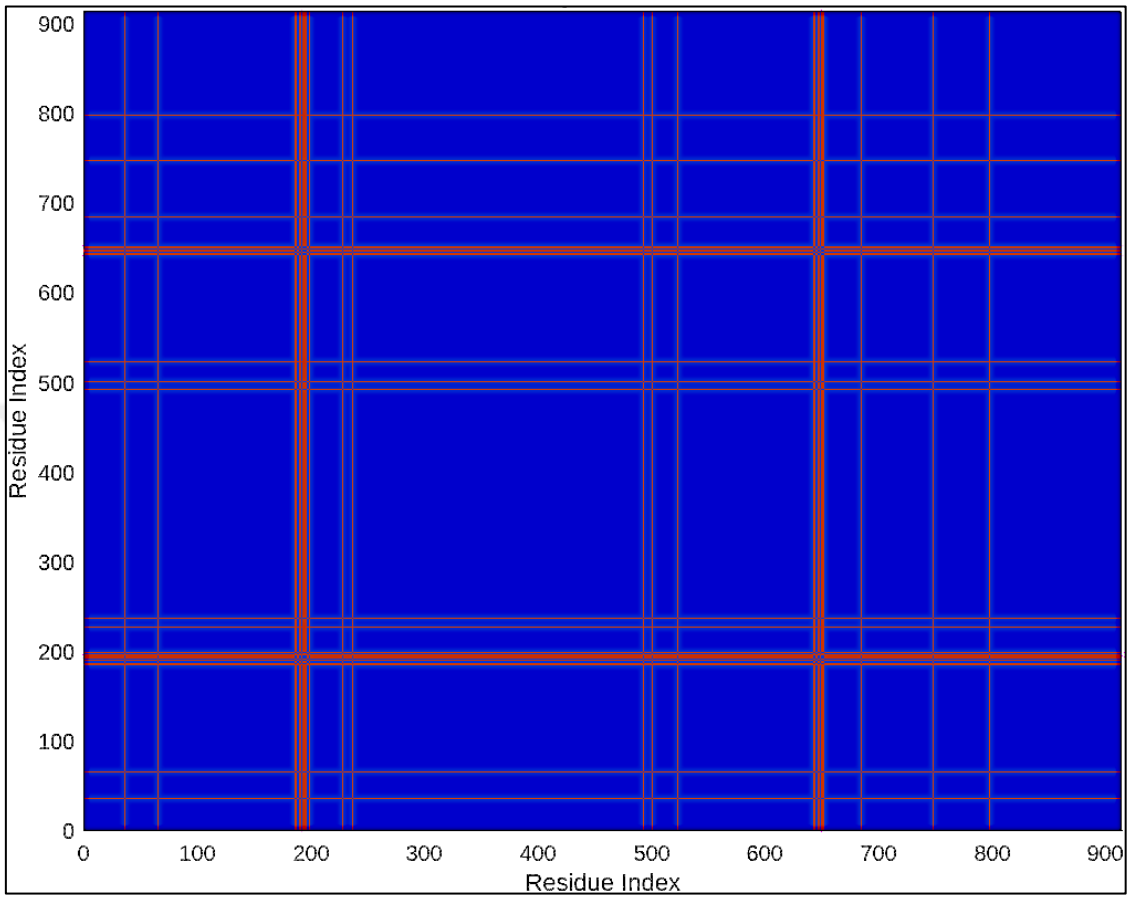

Figure 4.23. Cross-map difference of N15-N6 for GNM2

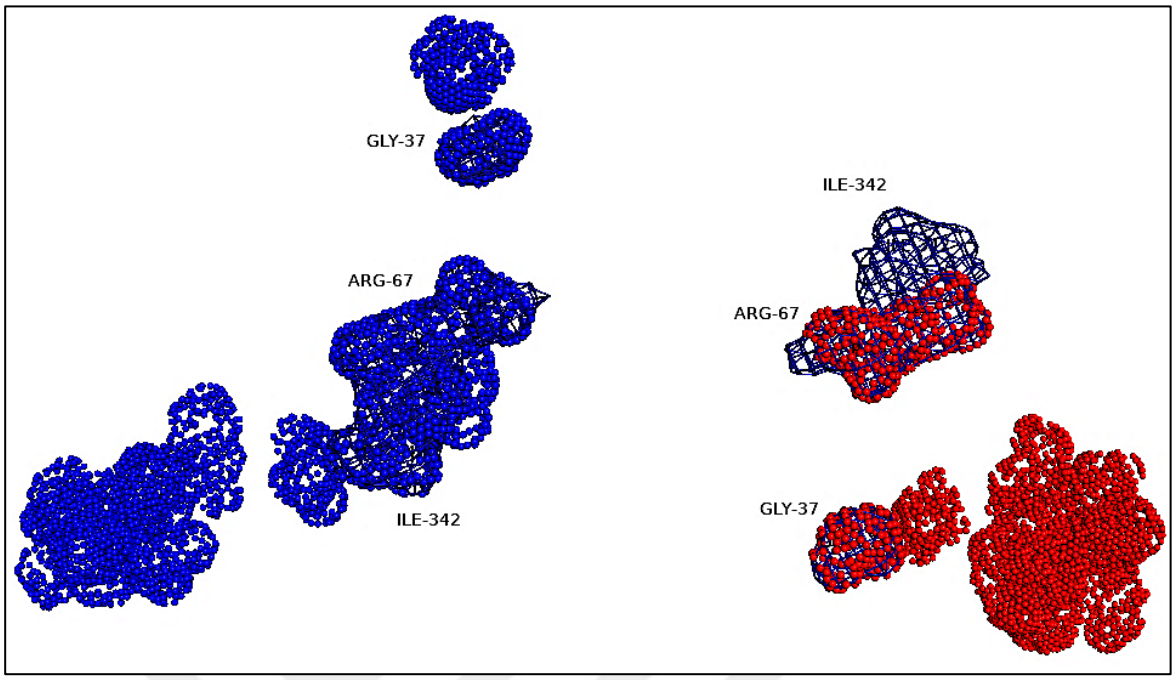

Figure 4.24. Relative placement of hinges and cross correlation differences in N15-N6

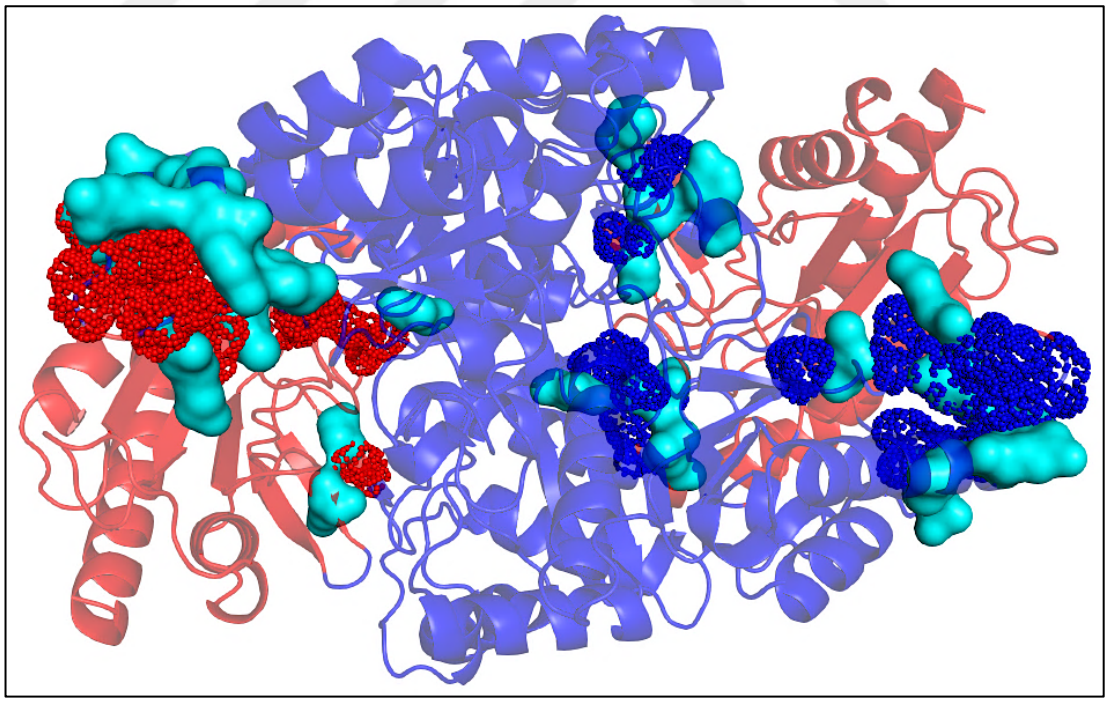

Figure 4.25. Cross correlation difference points and mutations in N15 for GNM2 (cyan surfaces: mutation points, red and blue spherical regions: c.c. difference points)

Mode shift in variants N15 and N6 can be seen comparing the Figures 4.25 and 4.30. The region indicated as red spheres becomes blue and blue spherical regions becomes red in N6. This means that cross correlation changes between these sequential variants drive a mode shift in two ends of the mode domains.

As it was discussed in section 4.2, GNM4 is one of the driving GNM modes together with GNM2 in the transform of N15 to N6. Therefore cross-map difference of these variants for GNM4 given in Figure 4.26 was also evaluated for the mutation classifications. Cross correlation difference points in chain-A are 125, 259, 304, 305, 308, 320, 360, 362, 378, 407, 448, and 449, whereas chain-B has the difference points of 306, 361, 363, 379, 408, 448, 449, and 450. Here, 125A is nearby the active site of the chain A where residues 121 and 124 are catalytic residues, and 259A is a catalytic residue. If mutation points are related with the red lines in cross-map difference, nearby mutations for the 320^{th} residue in A-chain are 316A, 317A, 319A. For the c.c. differences seen in both chains at 360, 361, 362, 363; 360^{th} residue is mutated. Additionally, 379^{th} residue is a mutation point where 378 and 379 are the red lines observed in Figure 4.26. At last, nearby mutation for the series of 447, 448, 449, 450 is 449^{th} residue.

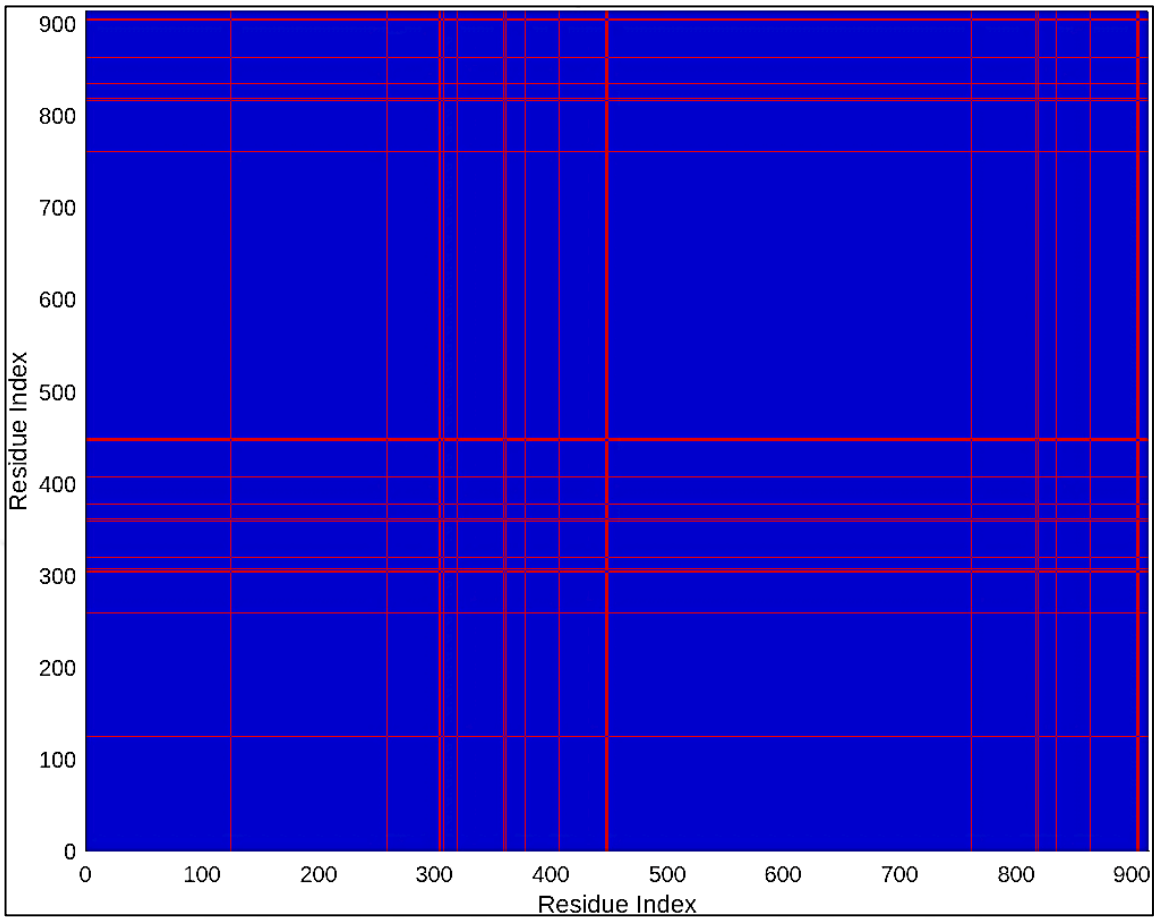

Figure 4.26. Cross-map difference of N15-N6 for GNM4

Determined mutations are represented in Figure 4.27 as cyan surfaces near the red and blue small spheres representing the c.c. difference regions.

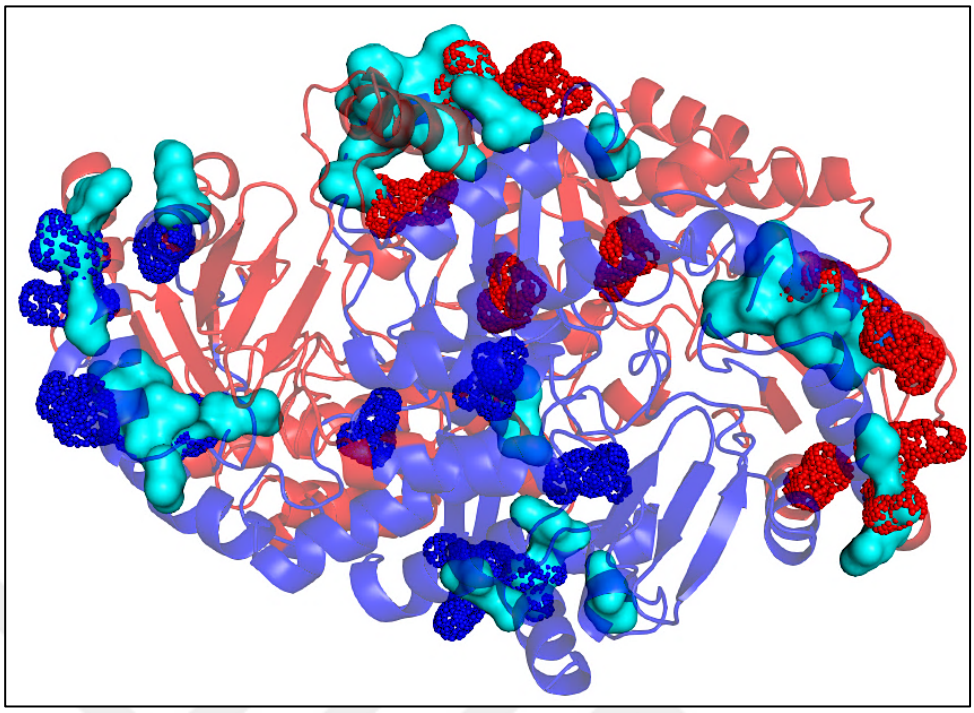

Figure 4.27. Cross correlation difference points and mutations in N15 for GNM4 (cyan surfaces: mutation points, red and blue spherical regions: c.c. difference points)

In the evolution of N6 to KES23360, it was determined that GNM2 and GNM4 are the dominant GNM modes mimicking the transitions. Thus, cross correlation change points in these two modes were investigated to relate with the mutations. Figure 4.28 shows the cross-map difference in GNM2 for N6-KES23360. Chain-A c.c. difference points are 194, 196, 199, 236, 237, 291, and chain B difference points are 195, 237, 292, 342 according to Figure 4.28. For these two variants, only one residues which is 342 is a hinge as different from the other sequential generation cases examined. From 194 to 200, there is a mutation at 203 passing from N6 to KES23360, however as it is figured out in GNM2 cross correlation evaluation of the sequential variants N15 and N6, 196 and 199 were mutated resides, too. In N6 and KES23360, 291A is close to hinge 67B and 292B is nearby the hinge 67A as the previous generations. In the context of last two variants in family, 237 and 236 are near the catalytic residue GLU226. Relative placements of the c.c. difference points close to hinges are shown in Figure 4.29.

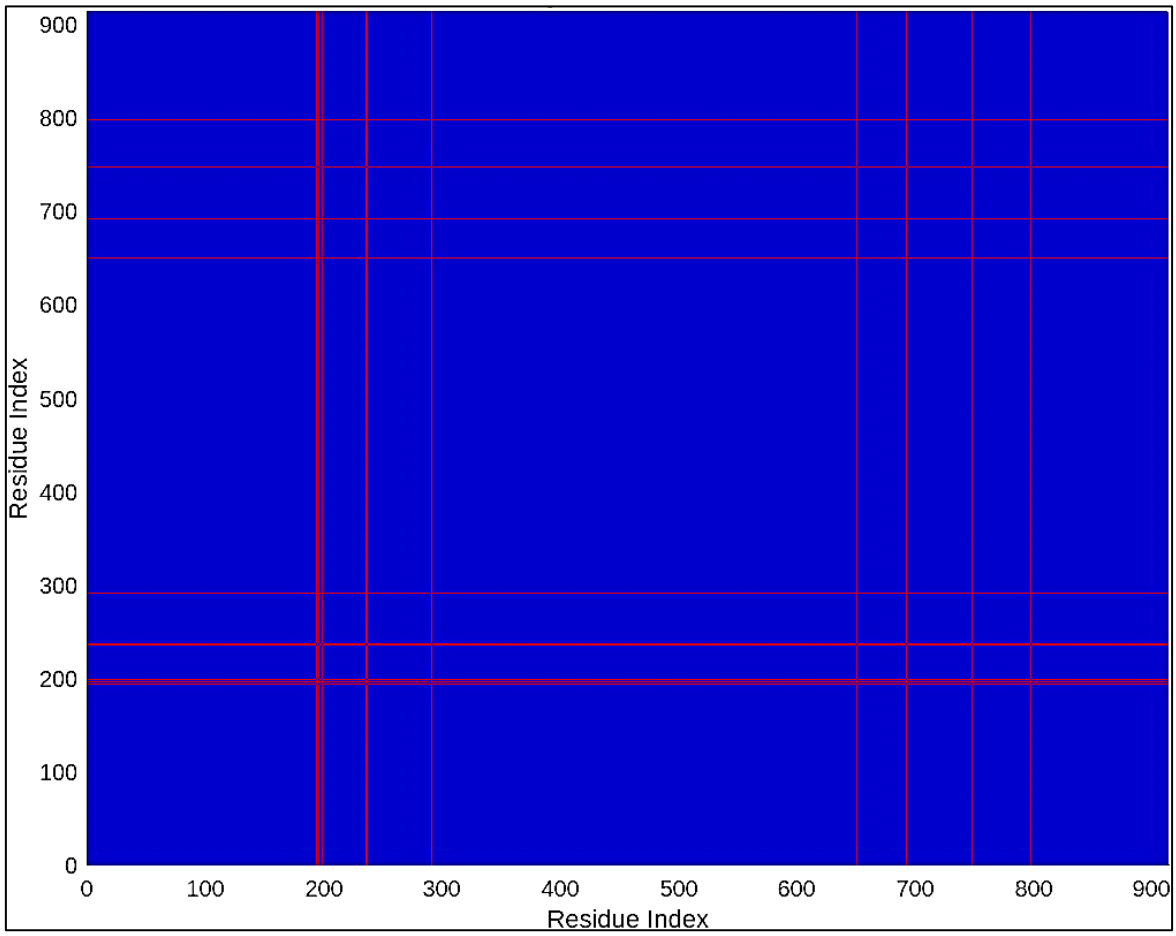

Figure 4.28. Cross-map difference of N6-KES23360 for GNM2

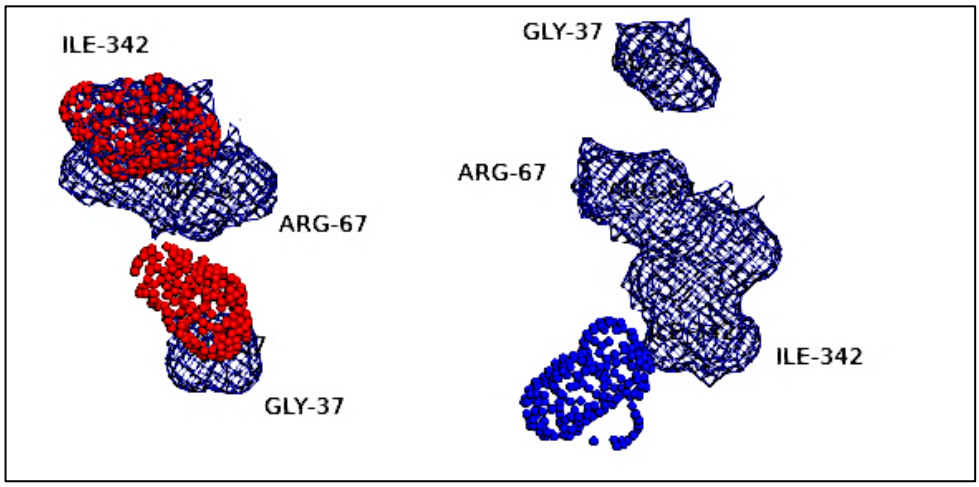

Figure 4.29. Relative placement of hinges and closer cross correlation differences in N6-KES23360

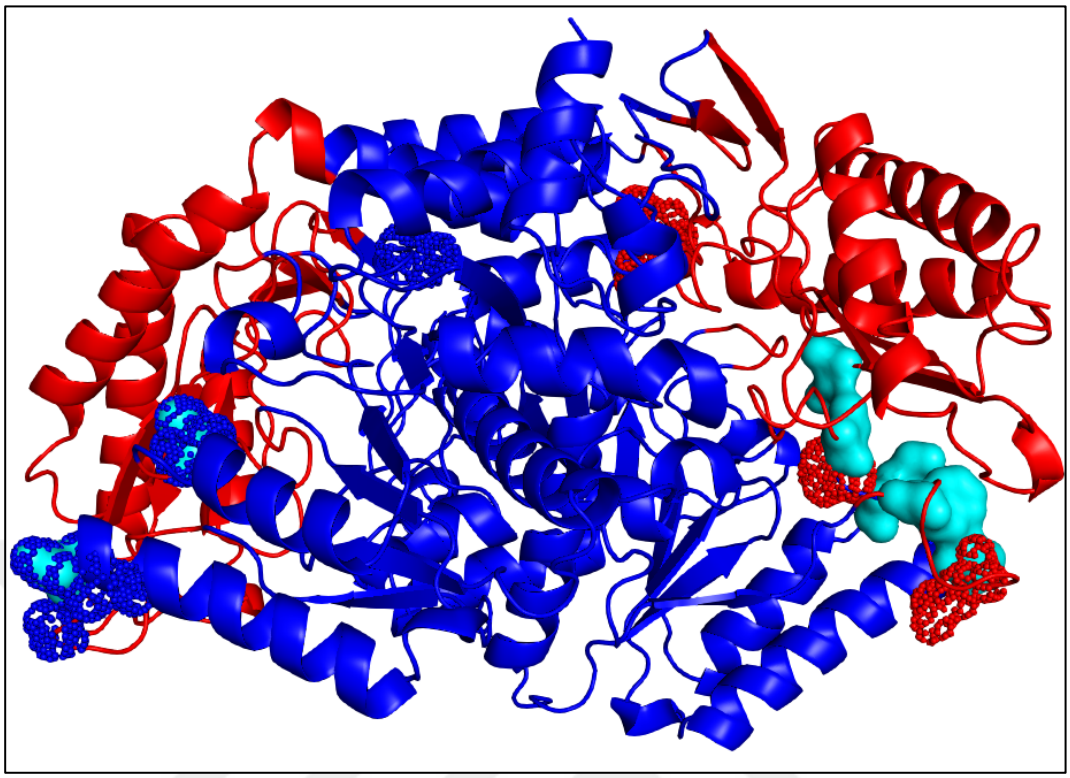

Figure 4.30. Cross correlation difference points and mutations in N6-KES23360 for GNM2 (cyan surfaces: mutation points, red and blue spherical regions: c.c. difference points)

In Figure 4.30, mutations close to c.c. changing regions from N6 to KES23360 are shown as cyan surfaces. If Figure 4.30 is examined in details, it is seen that there is not a significant mode shift in GNM2 in the transition of these two sequential variants which means that global motion represented by GNM2 is almost kept constant in the last generations.

When the representation of GNM4 on the transition from N6 to KES23360 is taken under the scope, cross-map difference was obtained as given in Figure 4.31. Here, chain A has c.c. differences between N6 and KES23360 at the residues 125, 259, 304, 308, 378, 407 and chain B has at 321, 363, 379, 408. As it is emphasized in the case of GNM4 evaluation for N15-N6, 125A is neighbor the active site of the A-chain in which residues 121 and 124 are catalytic residues, and 259A is itself has a catalytic task in the transamination reaction mechanism. In addition, 128A and 159A/B are the mutated residues close to residue 125. For 308A nearby mutations are 311 and 312; for 363B close mutations are 362B and 366B. When the nearby mutation for the residues 378 and 379 are investigated, there is no mutation close to this region while passing from N6 to KES23360, but 379 is a mutation point for the

transition from N15 to N6 and also residue 360. For the residues 407 and 408, 449 and 453 are the close mutations in 3D-structure. All close mutations to c.c. difference regions of N6 and KES23360 for GNM4 are indicated in Figure 4.32.

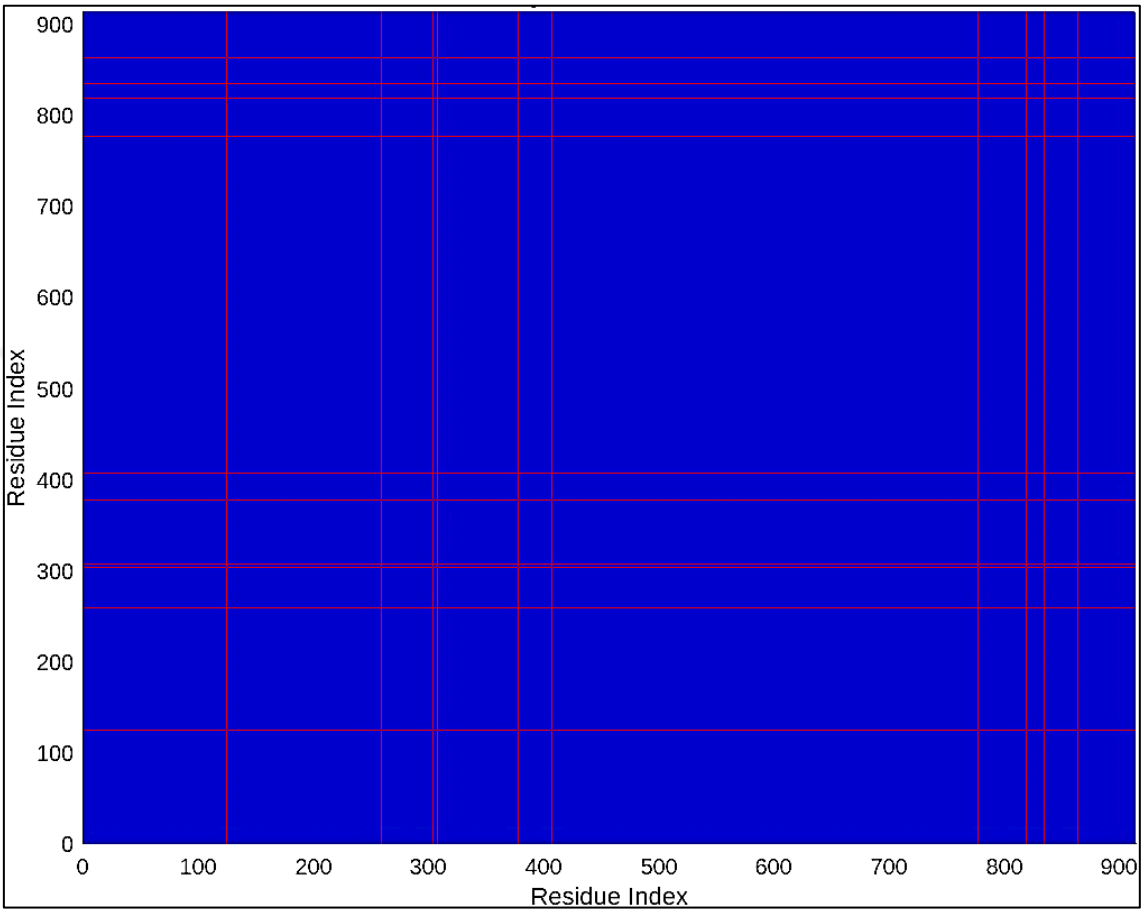

Figure 4.31. Cross-map difference of N6-KES23360 for GNM4

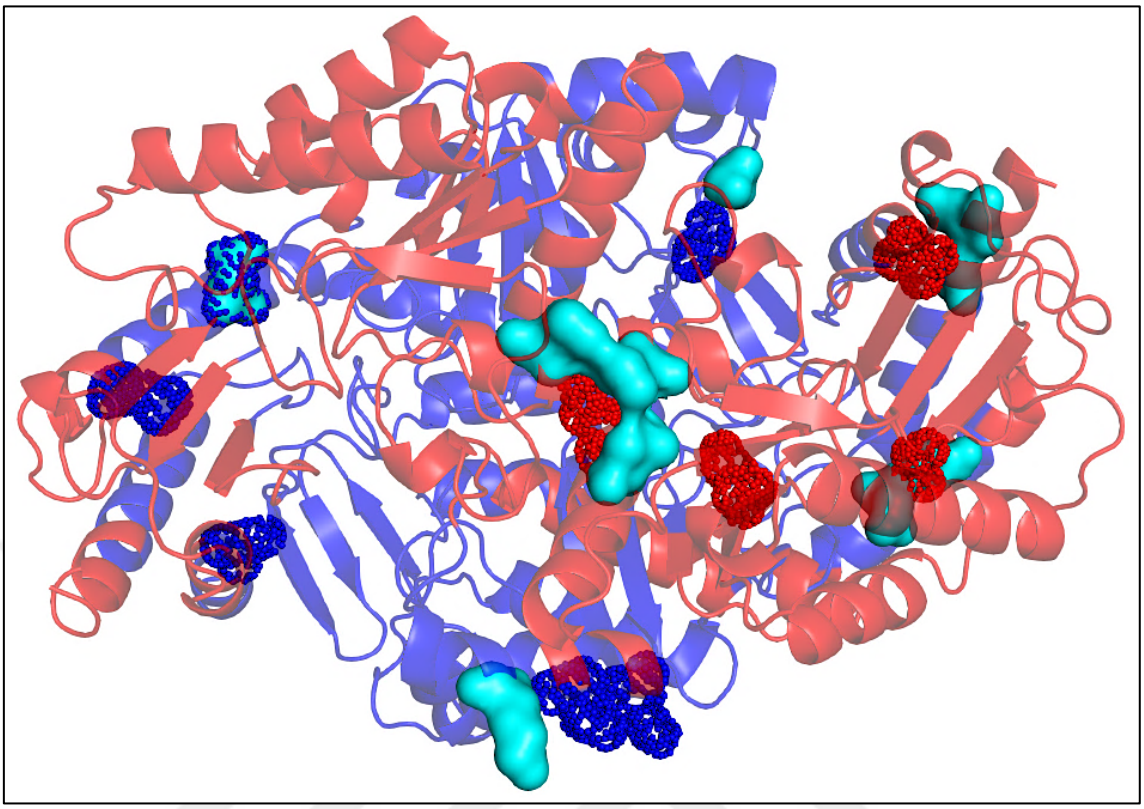

Figure 4.32. Cross correlation difference points and mutations in N6-KES23360 for GNM4 (cyan surfaces: mutation points, red and blue spherical regions: c.c. difference points)

While ending the “Cross Correlation Difference Points And Close Mutations” section, it is also crucial to evaluate the linkage between the most ancient variant and modern-day extant. As it was previously determined, GNM2 is the dominant GNM mode to relate and model the global dynamics of N48 and KES23360. Cross-map difference for GNM2 of these two variants is given in Figure 4.33. Cross-map difference of N48-KES23360 covers both the GNM2 graphs of N48-N15 given in Figure 4.20 and of N15-N6 given in Figure 4.23 and almost same with the Figure 4.23. Moreover, as it is shown in Figure 4.34, close mutations to c.c. difference points are also the combination of the mutations examined for N48-N15 evolution and N15-N6 evolution as GNM2.

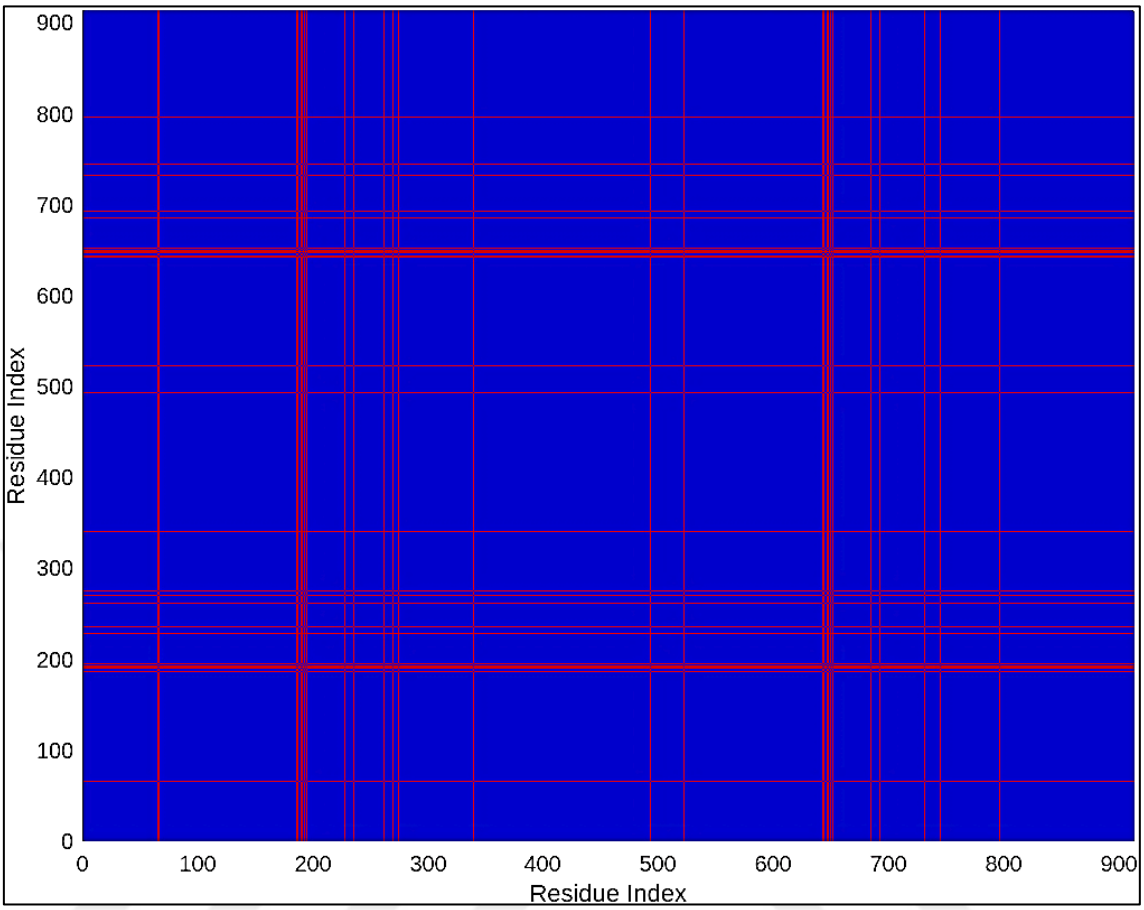

Figure 4.33. Cross-map difference of N48-KES23360 for GNM2

Figure 4.34 actually summarizes the significant results achieved in this section. Cyan surfaces are mutated residues close to c.c. changing regions which are shown as red and blue meshed spheres, and active sites are also indicated in this figure as orange meshes. The green-stick representations is belong to PLP which is the co-factor of these transaminases. It can be clearly seen in Figure 4.34 that cross correlation difference regions are very close to mutations occurred from the most ancient extant to the modern-day variant. Even though none of the mutations are within the active sites, some mode shape shifting regions and also mutations are close to active sites. Moreover, taking into consideration the mutations and cross correlation changes near the hinge points, it can be conclude that these mutations alter the global dynamic motions of the variants through the evolution and it might be crucial for the functionality changes in variants since hinges are responsible for the stability and functionality of enzymes. There is a clearly distinguishable mode shift between N48 and KES23360 from left to right when the locations of the c.c. difference points are regarded.

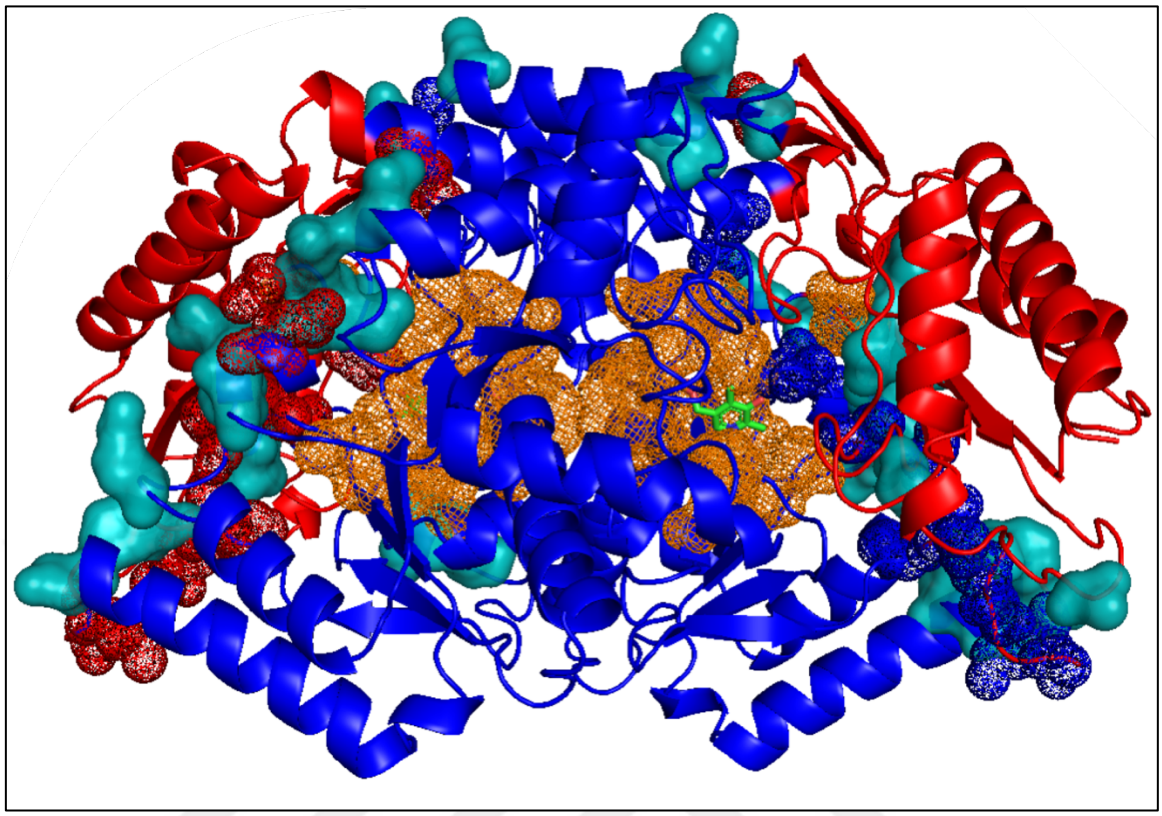

Figure 4.34. GNM2 mode shape, cross correlation difference points, close mutations and active site representation for N48-KES23360 system (orange mesh region: active site, yellow stick: PLP, cyan surfaces: mutation points, red meshes and blue meshes: c.c. difference points)

4.5. EVALUATION OF THE DYNAMIC MOTION OF CATALYTIC RESIDUE ARG267, A LOCKKEEPER FOR SUBSTRATE SELECTIVITY

PLP-dependent transaminases have an active site that consists of a large-binding pocket and a small-binding pocket, which together form the substrate-binding site of the enzymes. This arrangement is achieved through dimerization of the two monomers that compose the enzyme. Typically, the small binding pocket of TAs can only accommodate a methyl group, indicating the significance of engineered TAs that can accept bulkier substituents. One remarkable feature of TAs is their ability to accept both hydrophilic and hydrophobic substrates within the same active site. This ability is attributed to the presence of an "flipping" arginine residue located in a loop near the active site. The guanidinium group of this arginine residue can move within the active site, allowing it to interact either with the

carboxylate group of the substrate or move outward, creating a hydrophobic space for bulkier substrates in that region of the active site. This movement ability makes this loop a determinant of substrate specificity. This loop's flexibility and the position of this catalytic arginine residue play a crucial role in the enzyme's ability to modify substrate specificity through protein engineering. By modifying this loop and the positioning of the arginine residue, researchers can potentially alter the substrate preferences of TAs [50].

This catalytic arginine coincides with ARG267 in our transaminase family. In this study, placement of ARG267 relative to active sites and also the movement of whole loop containing ARG267 were evaluated in order to find out the origin of the substrate specificity modifications in this family. For this purpose, the interactions established by ARG267 with the neighbor residues were investigated for N48 and KES23360 as apo, active site including nonbonded PLP, and covalently bound PLP cases by the help of MD simulations. In addition, movement models of the ARG267 containing loops of all variants were obtained by ANM and normalized average cross correlation differences for these regions were determined for each sequential variants to discover at which step of the natural evolution loop dynamics were modified.

4.5.1. Interactions of ARG267 Obtained by MD Simulations

Hydrogen bonds established by ARG267 in 20 ns MD simulations are represented in the distance-angle graphs given in Figures 4.35 – 4.40. There is no significant change in the established bonds by ARG267 as seen in the following figures, catalytic arginine interacts only with the 59th residue which is tyrosine. In the case of free PLP present in the active sites of N48 and KES23360, ARG267 again interacts only with TYR59 but its conformation undergoes a small change with respect to the apo form of N48 resulting in the change in number of the hydrogen atoms of arginine interacting with tyrosine, and these interactions are more fluctuated with respect to the other simulations. Same conformational change in arginine is seen in covalently bound PLP containing MD simulations of both variants. If the strength of the bonds are evaluated for each case, H-bond with a considerable strength is obtained only in the simulations of the variants including covalently bound PLP. Population of the distance-angle data of N48 given in Figure 4.39 points out a stronger interaction for N48 than in the case of KES23360.

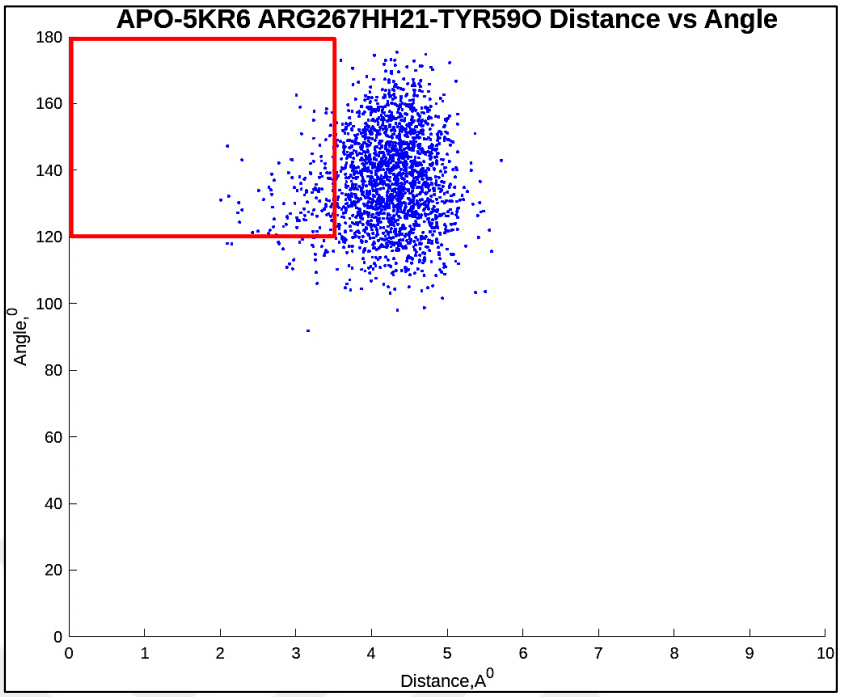

Figure 4.35. H-bond established between ARG267 and TYR59 in apo MD simulation of N48

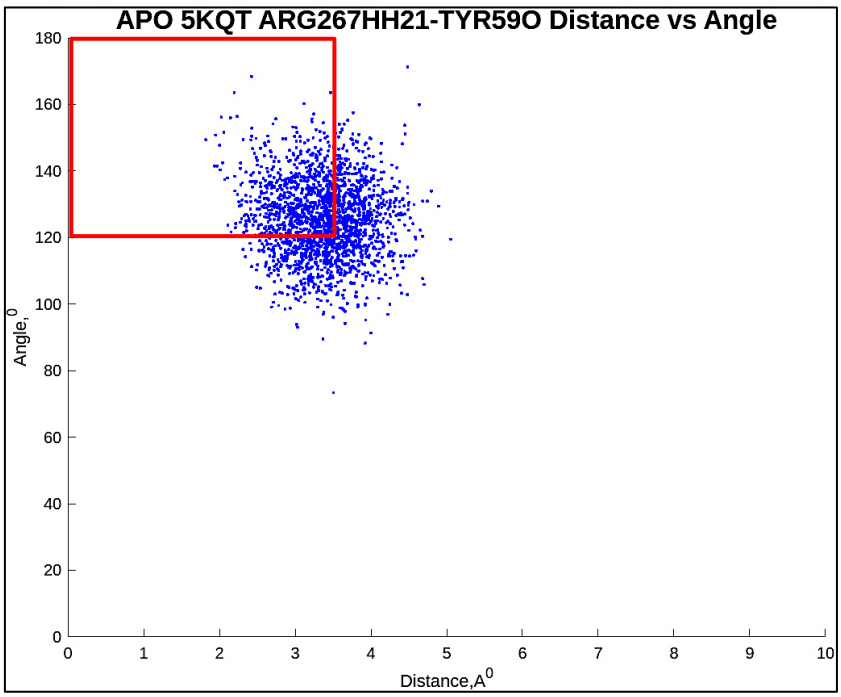

Figure 4.36. H-bond established between ARG267 and TYR59 in apo MD simulation of KES23360

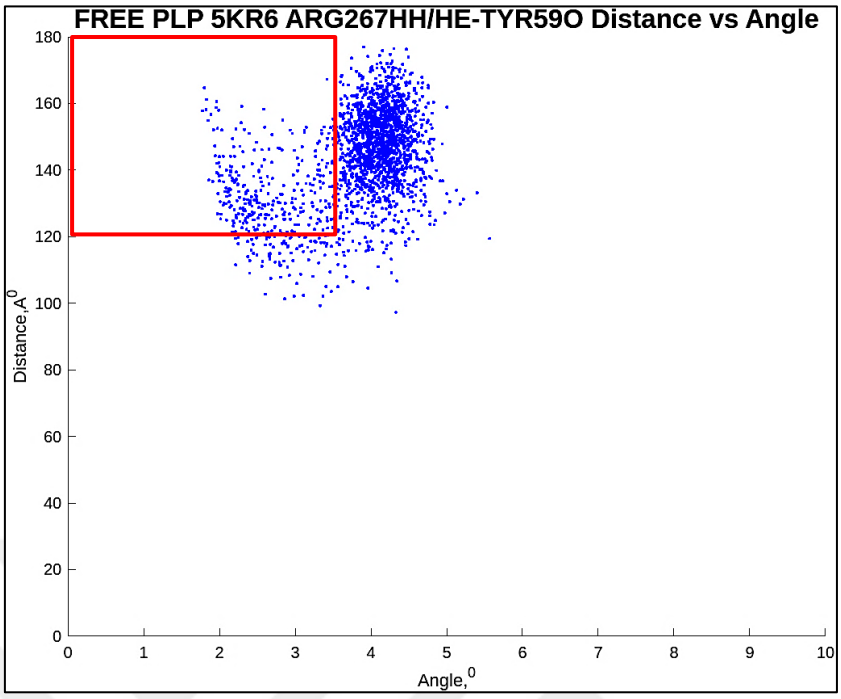

Figure 4.37. H-bond established between ARG267 and TYR59 in active site containing free PLP MD simulation of N48

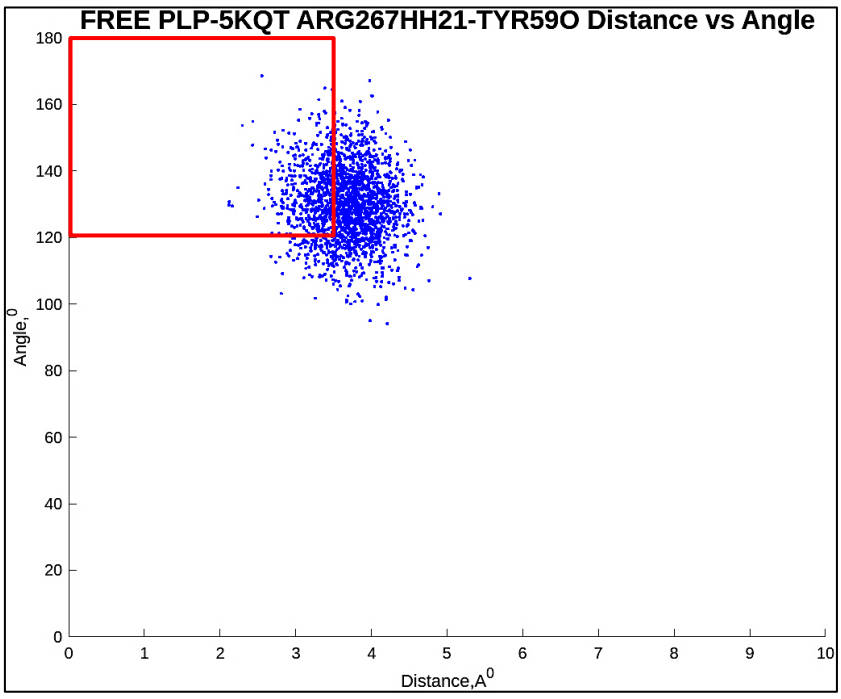

Figure 4.38. H-bond established between ARG267 and TYR59 in active site containing free PLP MD simulation of KES23360

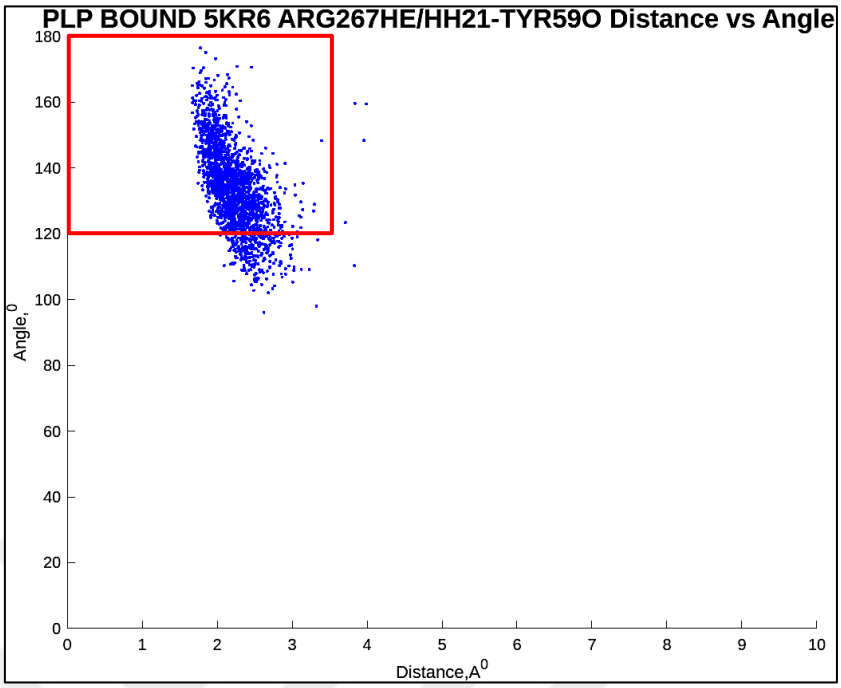

Figure 4.39. H-bond established between ARG267 and TYR59 in active site containing covalently bound PLP MD simulation of N48

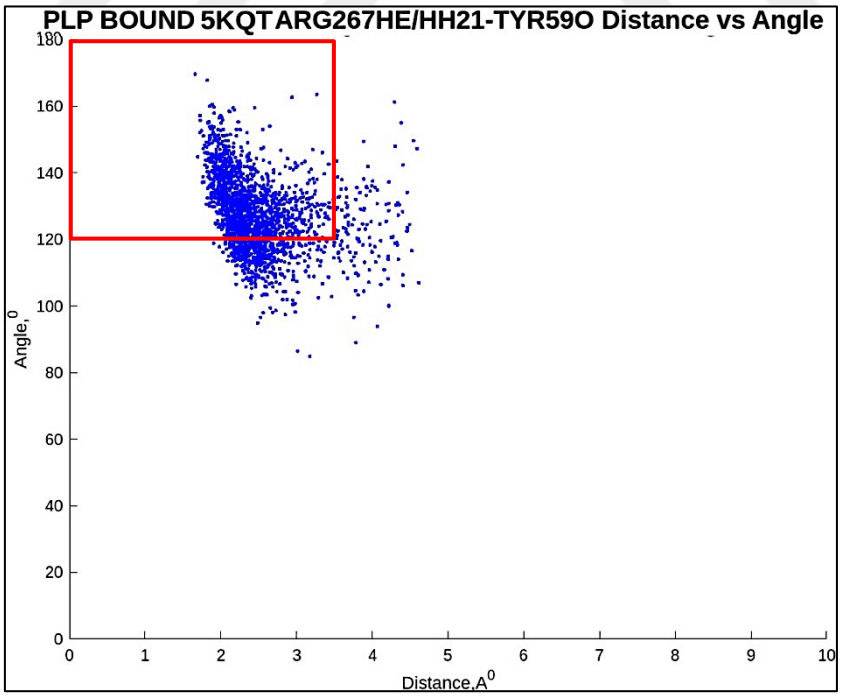

Figure 4.40. H-bond established between ARG267 and TYR59 in active site containing covalently bound PLP MD simulation of KES23360

4.5.2. Evaluation of the Changes in Loop Dynamics

Visualization of fluctuations obtained by ANM calculations for the first two slowest modes are available in Hingeprot sever [20]. Animations of the fluctuations of ARG267 containing loops for GNM2 were examined and it was observed that the loop responsible for the substrate specificity moves in opposite direction in N48 with respect to N15, N6 and KES233360. In N48, this loop moves toward inside the active site, closing the intake of the active site and providing a more hydrophobic property. On the other hand, in other generations' variants, this loop have the same direction, moving outside the active site and generates a hydrophilic entry. This finding actually supports the experimental activity data obtained by Wilding et al.[4] which states that N48 does not show any detectable activity for the tested 40 substrates. Since the tested substrates in their experiments are hydrophilic substrates and N48 has a more hydrophobic active site according to the loop dynamics obtained by ANM, their experimental findings reach significance.

Starting from N48, each sequential variant passages were examined in terms of cross correlation changes of the average first five slowest GNM modes to verify the dynamic motion changes. In the evolution of N48 to N15, it was obtained that the catalytic arginine containing loop's residues have important cross correlation differences. Cross-map difference of N48 and N15 for the cross correlations of the average of the first five slowest GNM modes are given in Figure 4.41. In Figure 4.41, a color-bar with the range of (-0.3 – 0.3) is used since all average cross correlation difference values are within this range. The regions emphasized with red ellipses are the regions of the loop of catalytic arginine (ARG267) in both chains of the homodimers. In these regions, cross correlation difference values are in between 0.24 and 0.30 which are the highest values for the total average values. There are no significant c.c. differences in these indicated regions in the evolution of the next generations and these results accord with the dynamic motion models previously obtained. The figures for the cross-map differences with the average of the first five slowest GNM modes of the variant couples, N15-N6 and N6-KES23360 are given in Appendix D.

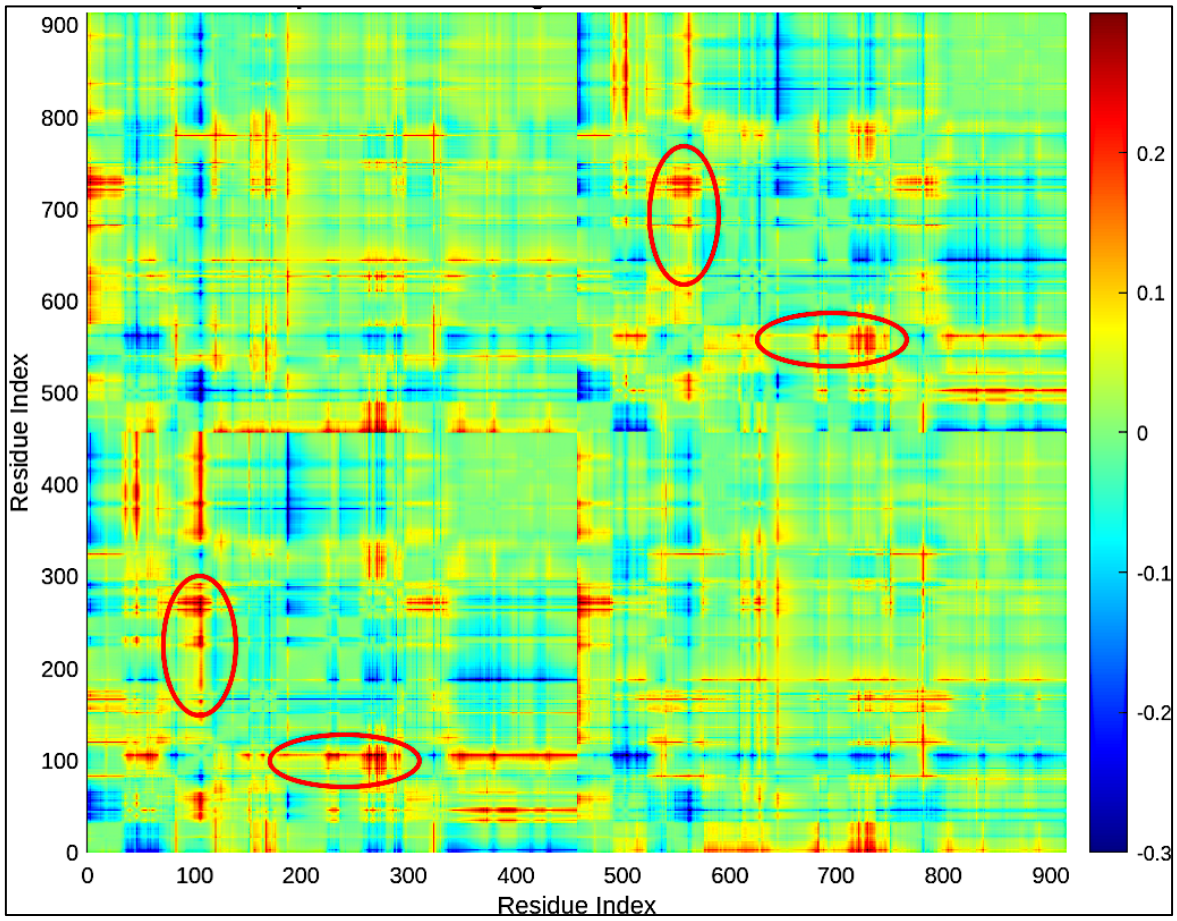

Figure 4.41. Cross-map Difference of N48-N15 for the averages of the first five slowest GNM modes

The idea that there must some mutations affecting the dynamic motion of the catalytic arginine and the loop where it belongs to was investigated after the findings of the direction change in the loop motions in natural evolution. Therefore close mutations to c.c. difference points were analyzed. In Figure 4.42, residues represented as red sticks are the mutated residues in evolution of N48 to N15. These residues are close to catalytic arginine (ARG267, shown with colored stick representation of green for carbon, blue for nitrogen, and red for oxygen atoms) and placed on the same loop responsible for the substrate selectivity. Orange surface indicates the active site and the molecule shown with a stick is PLP-cofactor within the active site. Mutations at the very close points in 3D structure such as 268th, 382nd and 272nd residues may force the catalytic arginine to a conformational change especially in its guanidinium group resulting in different loop dynamics.

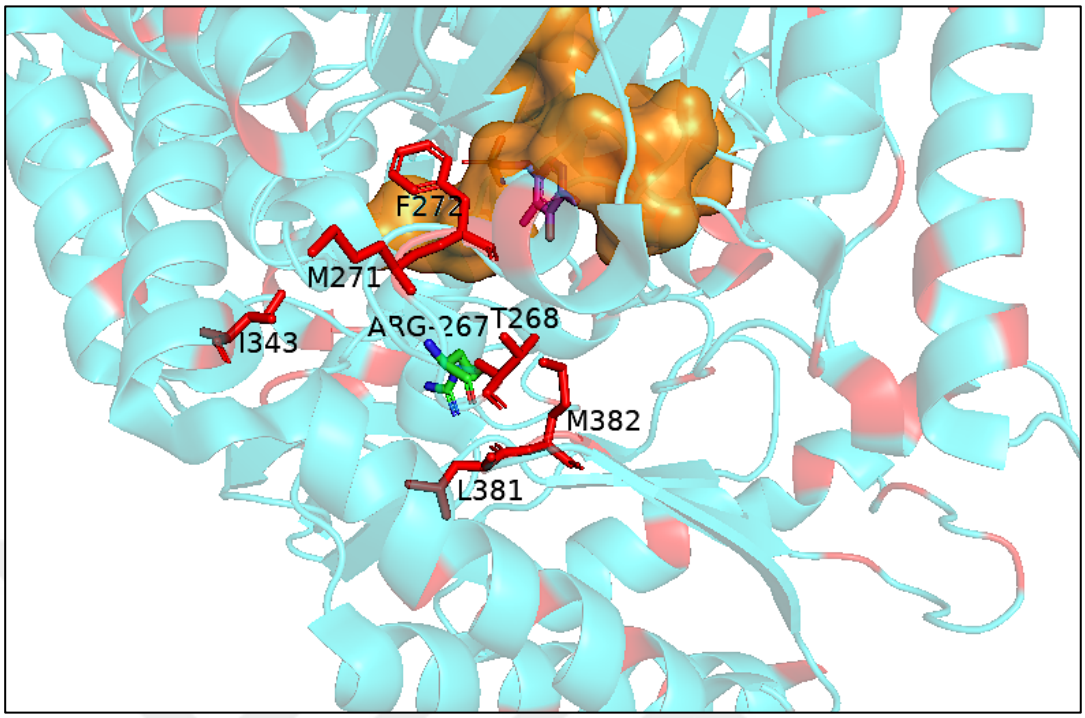

Figure 4.42. Close mutations to catalytic arginine in N48 (orange surface: active site, green stick: catalytic ARG267, red sticks: mutated residues close to catalytic ARG267, red ribbon regions: mutation points)

4.6. VERIFICATION OF THE CONSERVED ACTIVE SITES IN EACH GENERATION

MD simulations with the time-scale of 20 ns were used to compare the interactions occurred within the active sites of the most ancient and the most modern variants. In literature, examples of the usage of 20 ns MD simulations to successfully model the active site interactions are known [5]. Three different systems as apo-protein, PLP covalently bonded to the enzymes and free PLP present in the active sites of the enzymes were used in MD runs in order to determine the effect of covalently-bound PLP in active site present as cofactor. Expected interactions as well as the MD results are listed in Table 4.3. Obtained interactions within the active sites of all systems for both N48 and KES233360 mainly the same verifying the conserved active site during the evolution. One of the active site residues, SER121, was supposed to interact with one of the oxygen atoms of the phosphate group of PLP, however, it does not show this interaction in covalently bound PLP simulations. Only in free-PLP case of KES23360 an interaction is observed with PLP. Other interactions established by SER121

within the active site do not show difference. For the covalently bound PLP to enzymes simulations, other residue interactions meet the expectations as listed in Table 4.3.

Table 4.3. MD Results of N48 and KES23360 indicating the interactions of the active site residues

RESIDUE	FUNCTION	APO	FREE PLP	BOUND PLP
LYS288	covalently bound to PLP	**KES23360:** LEU58, VAL261	**KES23360:** TYR783, LEU58, THR782	**KES23360:** VAL261
		N48: No interaction	**N48:** PLP-O4(weak), THR782	**N48:** VAL261, TYR783
ARG267	interacts with the carbonyl group of the substrates	**Both:** TYR59	**Both:** TYR59	**Both:** TYR59
SER121	H-bond with the Phosphate group of PLP	**KES23360:** SER119, SER156, TYR153, ASN124	**KES23360:** PLP-O4, SER119, ASN124, SER156	**KES23360:** SER119, ASN124, SER156
		N48: SER119, ASN124, HIS154	**N48:** SER119, ASN124, SER156	**N48:** SER119, ASN124, SER156
THR782	H-bond with the Phosphate group of PLP	**KES23360:** TRP781 (weak)	**KES23360:** PLP-O4, LYS288	**KES23360:** PLP-O3 (strong)
		N48: No interaction	**N48:** LYS288	**N48:** PLP-O3 (strong)

Table 4.3. Continued

GLY120	H-bond with the Phosphate group of PLP	**Both:** ASN124 (strong)	**Both:** ASN124, PLP-O4 (strong)	**Both:** ASN124, PLP-O6 (strong)
HIS154 ASP259	help the activation of the 1st half rxn. Of the catalytic mech. By forming an "ionic diad"	**Both:** Strong interaction (100% occupancy)	**Both:** Strong interaction (100% occupancy)	**Both:** Strong interaction (100% occupancy)
ASN124	favors a negative ASP259	**Both:** HIS154 (Strong) THR285 (Strong) ASP259 (Strong) GLY120 (Strong)	**Both:** HIS154 (Strong) THR285 (Strong) ASP259 (Strong) GLY120 (Strong)	**Both:** HIS154 (Strong) THR285 (Strong) ASP259 (Strong) GLY120 (Strong)
GLY225	favors a negative ASP259	**Both:** ASP259 (Strong)	**Both:** ASP259 (Strong)	**Both:** ASP259 (Strong)

Combinations of the PLP presence in active site do not affect the expected interactions of GYL225-ASP259, ASN124 favoring a negative ASP259 in active site, ionic diad formation by HIS154 and ASP259, and ARG 267. On the other hand, interactions of LYS288 having the function of bonding to co-factor, GLY120 and THR782 which establish H-bonds with one of the phosphate group oxygens depend on the presence of PLP in active site as well as its state (free or bonded) as expected.

4.7. THE REASON BEHIND LOW THERMOTOLERANCES OF THE TWO END VARIANTS OF EVOLUTION

Wilding and co-workers [4] determined the melting point of the most ancient transaminase, N48 as almost the same with the modern-day transaminase, KES23360. Taking into

consideration the environmental conditions with that the ancestral enzymes dealt with, this is an unexpected result and other ancestral variants have relatively high melting points compared to modern-day extant. Apo-enzyme MD simulations performed for each variant to discover the reason behind this experimental results. Snapshots obtained at the end of apo-protein simulations were examined and it was seen that homodimeric chains were separated in simulations of N48 and KES23360 whereas 3D structures of chains were kept together until the end of MD simulations of N15 and N6. Snapshots of 19th nanoseconds of the MD simulations of each variant are given in Figure 4.43. Green enzyme is N48, purple structure is N15, N6 is shown as an orange cartoon structure and cyan complex is KES23360 in Figure 4.43.

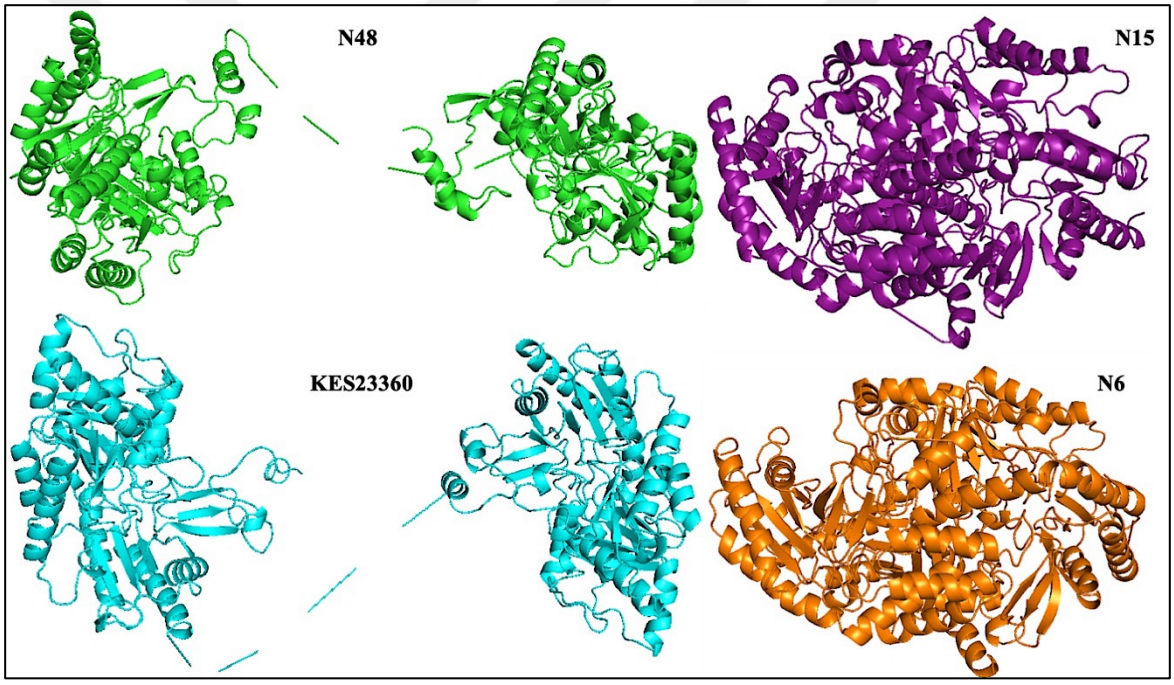

Figure 4.43. MD snapshots of the 19th nanoseconds of apo-enzyme structures

Chain separations observed in N48 and KES23360 explain the lower thermo-tolerance of the most ancient and the most modern transaminases than the other variants obtained by experimental studies.

5. CONCLUSION

Determination of the effective mutations on the activity, selectivity and thermo-tolerance variations of the naturally evolved transaminases was the main objective of this study.

Starting from the most ancient to the most modern, one over third of the whole enzyme is mutated. Additionally, at some points, residues mutated more than one time. Even though all mutations are outside the active site, it was shown in the study of Wilding et al.[4] that variants' activities, substrate selectivities and thermotolerances were different from each other. When the interactions in active sites of the most ancient variant, N48, and the most modern one, KES23360, were compared by the help of MD simulations, all interactions were almost the same and established by the same residues. Since all active site residues, 3D structures of the active sites and interactions are identical, the reason behind the activity and selectivity changes must be related with allostery and modified dynamic motions of the variants during evolution. To this aspect, mutations seen during the evolution of this transaminases were classified by associating with the global motion transitions. Difference vectors obtained after the alignment of variant couples by structurally were compared with the slow modes' ANM vectors of each variants. Slowest ANM modes were matched with the corresponding difference vectors as expected since it is the core knowledge that the slowest ANM and GNM modes indicates the global motions of the whole structures and environmental adaptation proceeds by slowest modes characteristics. When the GNM mode shapes of the dominant slowest modes were examined and cross correlation difference points were identified on the 3D structures of the enzymes, it was clearly seen that mutations coincide with the mode shape transition regions. This means that mutations have been realized in order to modify the global motions of the enzymes by natural evolution resulting the activity and selectivity changes. 28 mutations were determined for the selected GNM modes at the cross correlation difference points. Additionally, there are 14 mutated residues close to hinges which are the key points for the domain separations of dynamic movements. These mutations can be considered as the potential drivers in the change of variant functions during natural evolution. It was determined that GNM2 is the dominant mode for the evolution from N48 to KES23360, and from N48 to N15. GNM2 and GNM4 were the determinant modes driving the evolution of N15 to N6, as well as the transition of N6 to KES23360. When the higher catalytic activity data proposed for N6 compared to N15 and

KES23360 by Wilding et al.[4] are taken into consideration and if it is combined with the computational work performed in this study, it can be concluded that mutations realized nearby the GNM2 and GNM4 modes' transition regions are responsible for the change in activities of the variants. If the narrowed and promiscuous substrate selectivity for KES23360 with respect to N6 is considered, GNM2 and GNM4 can explain the substrate selectivity effecting motions and corresponding mutations realized in this evolution step.

The experimental work conducted by Wilding et al.[4] is denoted that N48 did not show significant activity for the tested substrates covering ω-amino acids, α-amino acids, and α,ω-diamines. Apart from this, its thermotolerance was nearly identical with the most modern variant. These published results are actually contrary to the knowledge gained based on the natural evolution since environmental conditions was tough in the ancient terms and proteins needed to have high thermotolerance to survive, as well as substrate selectivity has not acquired yet. In the second slowest ANM mode simulations of each variant, it was clearly seen that the active site's catalytic arginine residues (ARG267) and the whole loops containing this arginine residues have an inward-movement through the active site in the case of N48 whereas this movement is in the opposite direction for the other variants. The inward loop movement seen in N48 makes the active site hydrophobic by closing the active site's entry to hydrophilic substrates. This result explains why N48 did not show any activity for the tested substrates since all substrates need hydrophilic active sites. For the perspective of unexpected experimental thermotolerance data, MD simulations performed for the apo forms of each dimeric transaminase variant in this study showed that chains are separated from each other in the simulations of N48 and KES23360, however there was no chain separation in N15 and N6 at the end of the MD simulations. Chain separation clarifies the lower thermotolerance of N48 and KES23360 because when the chains are separated lower melting points are measured.

In conclusion, by this computational study, experimental results obtained by Wilding et al.[4] are verified and the reasons behind them are elicited. Network models used to understand the dynamic motions of each family members are helped to get an insight into the effect of mutations on the functional behaviors of the enzymes.

6. FUTURE WORK

It has not clear yet that every mutation categorized in this study is responsible for the assigned task individually and the role of allosteric communications within the variants stays still in the dark. Apart from this, some mutations should be realized in order not to disconcert the physio-chemical properties of the mutated regions by keeping the necessary interactions between the residues as constant in 3D structures and folding states, rather than to bring in an effect on the enzyme functionality. From these point of views, it is important to investigate the allostery to discover the presence of the allosteric communication between the active site residues and the mutated ones as a future work. Neighboring mutations in the close regions in 3D structures should be considered in terms of physio-chemical characteristics, in order to diminish the number of mutations that must be focused on.

Hydrophobicity of N48 should be check by modelling a reaction with a more hydrophobic substrate so that the results of the present study could be verified. In addition, performing longer MD simulations in a microsecond scale allows the observation of loop motions including the motion of catalytic arginine with its connected loop. Apart from this, several protonation state regulations can be considered in MD simulations. For example in apo-MD simulations, lysine residue covalently bonding to PLP was treated as positively charged. It can be replaced by a neutral lysine in apo-MD simulations to examine its effect in the generation of Schiff base and on the active site interactions. Another protonation state regulation can be performed for the catalytic residue HIS154 by converting it into positively charged form as stated in the study of Sungur et al. [41] which was used as a base for the active site determinations of the examined variants.

The overall objective of this project is modification of the modern-day transaminase using evolutionary footprints to get a new transaminase with improved catalytic activity together with the broadened selectivity and high thermotolerance for the industrial usage. Therefore specified mutations responsible for the change in these functions should be introduced into KES23360 and several reactions with the substrates from different amino-containing families should be studied to achieve this goal.

REREFERENCES

1. Kelly SA, Pohle S, Wharry S, Mix S, Allen CCR, Moody TS, Gilmore BF. Application of ω-Transaminases in the pharmaceutical industry. *Chem Rev*. 2018;118(1):349–67.

2. Yun H, Hwang BY, Lee JH, Kim BG. Use of enrichment culture for directed evolution of the Vibrio fluvialis JS17 ω-transaminase, which is resistant to product inhibition by aliphatic ketones. *Appl Environ Microbiol*. 2005;71(8):4220–4.

3. Cao JR, Fan FF, Lv CJ, Wang HP, Li Y, Hu S, Zhao WR, Chen HB, Huang J, Mei LH. Improving the thermostability and activity of transaminase From Aspergillus terreus by charge-charge interaction. *Front Chem*. 2021;9.

4. Wilding M, Peat TS, Kalyaanamoorthy S, Newman J, Scott C, Jermiin LS. Reverse engineering: Transaminase biocatalyst development using ancestral sequence reconstruction. *Green Chemistry*. 2017;19(22):5375–80.

5. Kiss G, Çelebi-Ölçüm N, Moretti R, Baker D, Houk KN. Computational enzyme design. *Angewandte Chemie International Edition*. 2013;52(22):5700–25.

6. Perez-Jimenez R, Inglés-Prieto A, Zhao ZM, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, Garcia-Manyes S, Kappock TJ, Tanokura M, Holmgren A, Sanchez-Ruiz JM, Gaucher EA, Fernandez JM. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. *Nat Struct Mol Biol*. 2011;18(5):592–6.

7. Robinson PK. Enzymes: principles and biotechnological applications. *Essays Biochem*. 2015;59:1–41.

8. Weiß MS, Pavlidis I V., Spurr P, Hanlon SP, Wirz B, Iding H, Bornscheuer UT. Protein-engineering of an amine transaminase for the stereoselective synthesis of a pharmaceutically relevant bicyclic amine. *Org Biomol Chem*. 2016;14(43):10249–54.

9. Łyskowski A, Gruber C, Steinkellner G, Schürmann M, Schwab H, Gruber K, Steiner K. Crystal structure of an (R)-Selective ω-Transaminase from Aspergillus terreus. *PLoS One*. 2014;9(1):e87350.

10. Savile CK, Janey JM, Mundorff EC, Moore JC, Tam S, Jarvis WR, Colbeck JC, Krebber A, Fleitz FJ, Brands J, Devine PN, Huisman GW, Hughes GJ. Biocatalytic asymmetric synthesis of chiral amines from ketones applied to Sitagliptin manufacture. *Science (1979)*. 2010;329(5989):305–9.

11. Cheng F, Chen XL, Li MY, Zhang XJ, Jia DX, Wang YJ, Liu ZQ, Zheng YG. Creation of a robust and *R*-selective ω-amine transaminase for the asymmetric synthesis of sitagliptin intermediate on a kilogram scale. *Enzyme Microb Technol*. 2020;141.

12. Wilding M, Peat TS, Newman J, Scott C. A β-Alanine catabolism pathway containing a highly promiscuous ω-Transaminase in the 12-Aminododecanate-Degrading Pseudomonas sp. Strain AAC. *Appl Environ Microbiol*. 2016;82(13):3846–56.

13. Merkl R, Sterner R. Ancestral protein reconstruction: techniques and applications. *Biol Chem*. 2016;397(1):1–21.

14. Bahar I, Lezon TR, Yang LW, Eyal E. Global dynamics of proteins: Bridging between structure and function. *Annu Rev Biophys*. 2010;39(1):23–42.

15. Tirion MM. Large amplitude elastic motions in proteins from a single-parameter, atomic analysis. *Phys Rev Lett*. 1996;77(9):1905–8.

16. Haliloglu T, Bahar I, Erman B. Gaussian dynamics of folded proteins. *Phys Rev Lett*. 1997;79(16):3090–3.

17. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Fold Des*. 1997;2(3):173–81.

18. Flory PJ. Statistical thermodynamics of random networks. *Proceedings of the Royal Society of London A Mathematical and Physical Sciences*. 1976;351(1666):351–80.

19. Eichinger BE. Elasticity theory. I. Distribution functions for perfect phantom networks. *Macromolecules*. 1972;5(4):496–505.

20. Emekli U, Schneidman-Duhovny D, Wolfson HJ, Nussinov R, Haliloglu T. HingeProt: Automated prediction of hinges in protein structures. *Proteins: Structure, Function and Genetics*. 2008;70(4):1219–27.

21. Lesk AM, Chothia C. Mechanisms of domain closure in proteins. *J Mol Biol*. 1984;174(1):175–91.

22. Doruker P, Atilgan AR, Bahar I. Dynamics of proteins predicted by molecular dynamics simulations and analytical approaches: Application to-Amylase inhibitor. *PROTEINS: Structure, Function, and Genetics*. 2000;40(1):512–24.

23. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J*. 2001;80(1):505–15.

24. Aykaç-Fas AB. Computational and experimental investigations into the mechanism of functional motion in biomolecular systems. Doctoral Thesis, Boğaziçi University, 2016.

25. Durrant JD, McCammon JA. Molecular dynamics simulations and drug discovery. *BMC Biol*. 2011;9(1):71.

26. Teague SJ. Implications of protein flexibility for drug discovery. *Nat Rev Drug Discov*. 2003;2(7):527–41.

27. Ma B, Kumar S, Tsai CJ, Nussinov R. Folding funnels and binding mechanisms. *Protein Engineering, Design and Selection*. 1999;12(9):713–20.

28. Kumar S, Ma B, Tsai CJ, Wolfson H, Nussinov R. Folding funnels and conformational transitions via hinge-bending motions. *Cell Biochem Biophys*. 1999;31(2):141–64.

29. Ma B, Shatsky M, Wolfson HJ, Nussinov R. Multiple diverse ligands binding at a single protein site: A matter of pre-existing populations. *Protein Science*. 2009;11(2):184–97.

30. Leach AR. Ligand-based approaches: Core molecular modeling. *Comprehensive Medicinal Chemistry*. 2007;87–118.

31. B. D. Madej, R. Walker. An introduction to molecular dynamics simulations using AMBER. AMBER Tutorial B0. 2015.

32. Becker OM, D. MacKerell Jr, A, Roux B, Watanabe M. Computational Biochemistry and Biophysics. 1st ed. New York: Marcel Dekker, Inc.; 2001. 44–48 p.

33. Pandya A, Howard MJ, Zloh M, Dalby PA. An evaluation of the potential of NMR spectroscopy and computational modeling methods to Inform biopharmaceutical fFormulations. *Pharmaceutics*. 2018;10(125).

34. Sweet CR, Petrone P, Pande VS, Izaguirre JA. Normal mode partitioning of Langevin dynamics for biomolecules. *Journal of Chemical Physics*. 2008;128(14).

35. Becker OM, D. MacKerell Jr, A, Roux B, Watanabe M. Computational Biochemistry and Biophysics. 1st ed. New York: Marcel Dekker, Inc.; 2001. 42–57 p.

36. Berman HM. The Protein Data Bank. *Nucleic Acids Res*. 2000;28(1):235–42.

37. The PyMOL Molecular Graphics System. Schrödinger, LLC.

38. Yang M, Livnat Levanon N, Acar B, Aykac Fas B, Masrati G, Rose J, Ben-Tal N, Haliloglu T, Zhao Y, Lewinson O. Single-molecule probing of the conformational homogeneity of the ABC transporter BtuCD. *Nat Chem Biol*. 2018;14(7):715–22.

39. Li H, Chang YY, Lee JY, Bahar I, Yang LW. DynOmics: dynamics of structural proteome and beyond. *Nucleic Acids Res*. 2017;45(W1):W374–80.

40. MATLAB. Natick, Massachusetts: The MathWorks Inc.; 2022.

41. Gökcan H, Monard G, Sungur Konuklar FA. Molecular dynamics simulations of apo, holo, and inactivator bound GABA-at reveal the role of active site residues in PLP dependent enzymes. *Proteins: Structure, Function and Bioinformatics*. 2016;84(7):875–91.

42. Case DA, Darden TA, Cheatham TE, Simmerling CL, Wang J, Duke RE, Luo R, Walker RC, Zhang W, Merz KM, Roberts B, Hayik S, Roitberg A, Seabra G, Swalis J, Götz AW, Kolossváry I, Wong KF, Peasani F, Vanice J, Wolf RM, Liu J, Wu X, Brozell SR, Steinbrecher T, Gohlke H, Cai Q, Ye X, Wang J, Hsieh MJ, Cui G, Roe DR, Mathews DH, Seetin MG, Salomon-Ferrer R, Sagui C, Babin V, Luchko T,

Gusarov S, Kovalenko A, Kolman PA. AMBER 12. University of California, San Francisco; 2012.

43. Frisch M, Trucks G, Schlegel H et al. Gaussian 16. Wallingford CT: Gaussian, Inc.; 2016.

44. Alvarado O, García-Meseguer R, Javier Ruiz-Pernía J, Tuñon I, Delgado EJ. A molecular dynamics study on the role of the protonation state in the biosynthesis of R-PAC by AHAS. *Chem Phys Lett*. 2019;716:247–51.

45. Varma S, Chiu SW, Jakobsson E. The Influence of Amino Acid Protonation States on Molecular Dynamics Simulations of the Bacterial Porin OmpF. *Biophys J*. 2006;90(1):112–23.

46. Molecular Operating Environment (MOE). Montreal: Chemical Computing Group ULC; 2023.

47. Doruker P, Atilgan AR, Bahar I. Dynamics of proteins predicted by molecular simulations and analytical approaches: Application to α-amylase inhibitor. *Proteins: Structure, Function and Genetics*. 2000;40(3):512–24.

48. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. *Biophys J*. 2001;80(1):505–15.

49. Zhu Y, Ye F, Zhou Z, Liu W, Liang Z, Hu G. Insights into conformational Dynamics and allostery in DNMT1-H3Ub/USP7 interactions. *Molecules*. 2021;26(17):5153.

50. Rocha JF, Pina AF, Sousa SF, Cerqueira NMFSA. PLP-dependent enzymes as important biocatalysts for the pharmaceutical, chemical and food industries: a structural and mechanistic perspective. *Catal Sci Technol*. 2019;9(18):4864–76.

APPENDIX A: FUNCTIONS OF PROPOSED ACTIVE SITE REDIUES

Table A.1. Proposed active site residues and their functions

RESIDUE (in GABA-AT) [41]	**FUNCTION**	**MATCHED RESIDUE FROM KES23360**
HIS190 ASP298	help the activation of the 1st half rxn. of the catalytic mech. by forming an "ionic diad"	HIS154 ASP259
VAL300 PHE181	-sandwich PLP with PHE -work as a gate-keeper with the assistance of PHE	VAL261 PHE169
LYS329	covalently bonds to PLP	LYS288
GLU265	an intermediate function in catalytic mechanism	GLU226
ARG192	interaction with the carbonyl group of the substrates	ARG267
ASN140 HIS190	favor a negative ASP in active site	ASN124 HIS154
GLY136 SER137 THR353B	H-bond with one of the Oxygens of the Phosphate group of PLP	GLY120 SER121 THR325B

APPENDIX B: DIFFERENCE VECTOR MAGNITUDE GRAPHS OF VARIANT COUPLE ALIGNMENTS

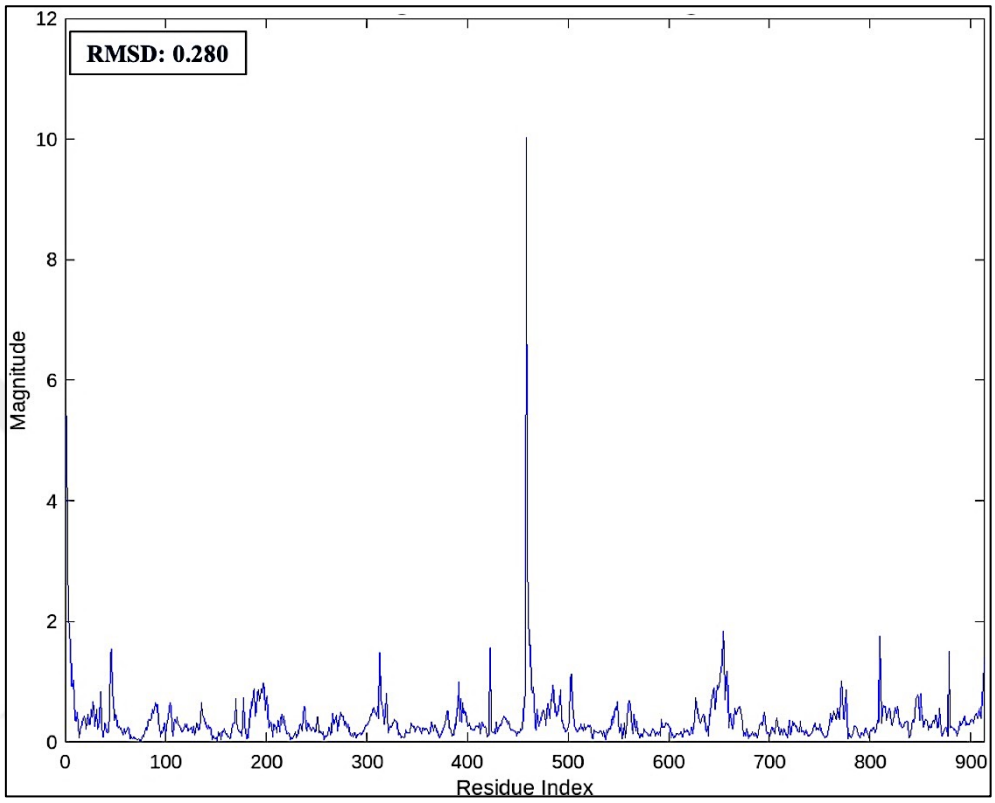

Figure B.1. Difference vector representation of N48-N15

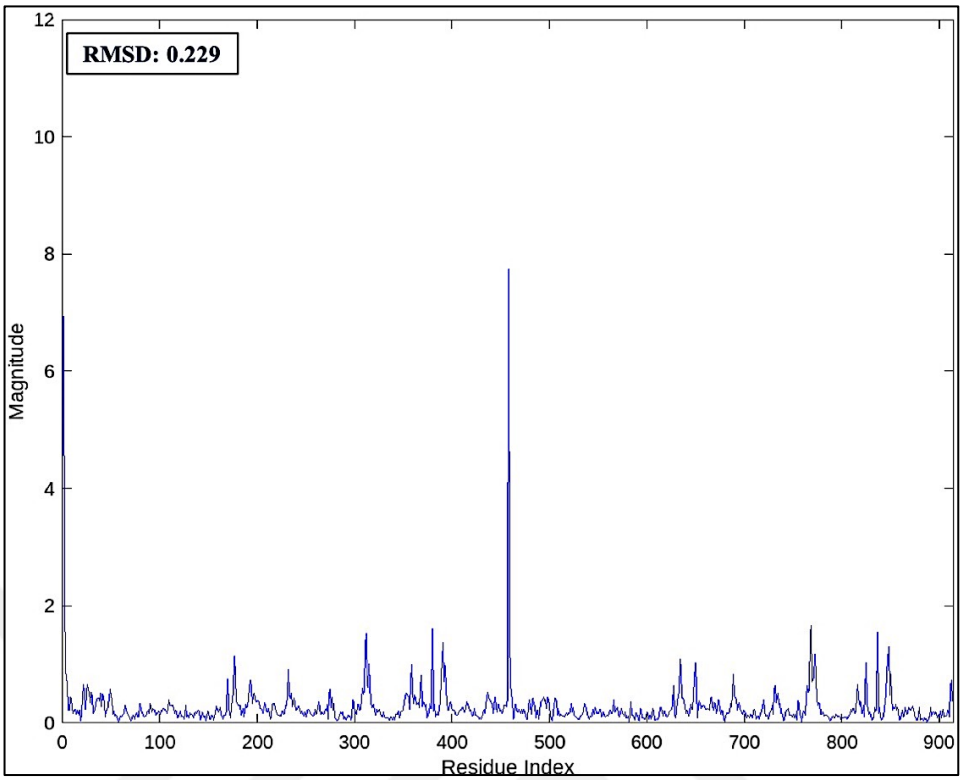

Figure B.2. Difference vector representation of N15-N6

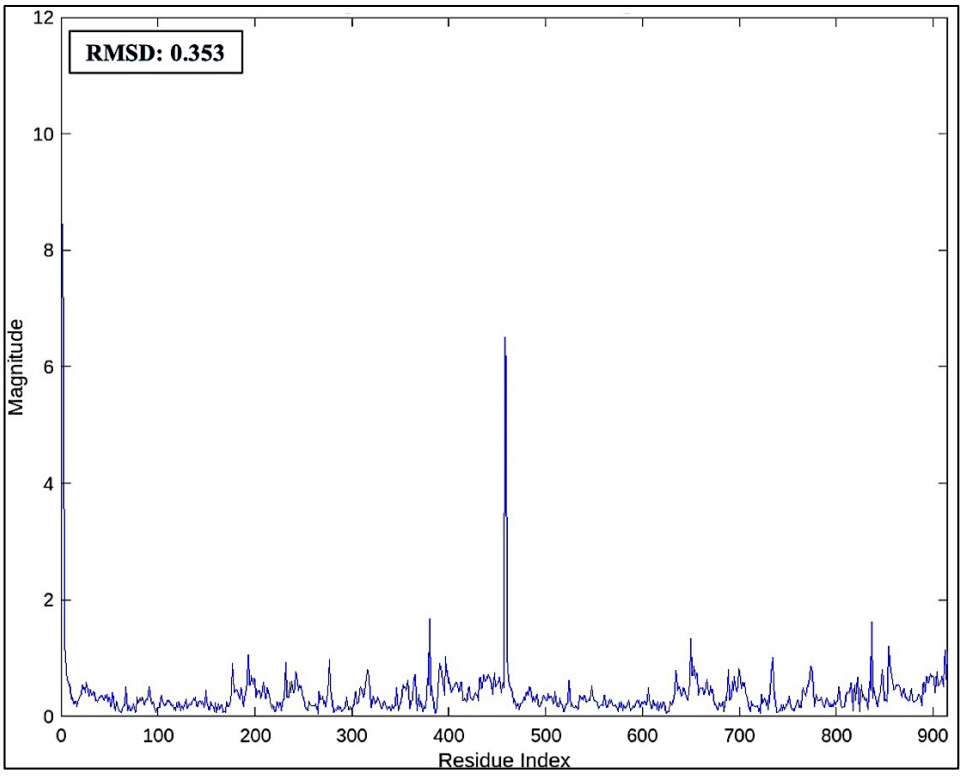

Figure B.3. Difference vector representation of N6-KES23360

```
N48: TDTDDLLEMDRAHFFHPSTHLRDHASGELPGRIITGGKGIRIQDSEGREYIDAFAGLYCV
N15: TDLDQLFEMDRAHFMHPSTHAHDHASGALPGRIITGGKGIRIQDHEGREYIDAFAGLYCV
     **.*.*.*******.*****..*****.****************.****************

N48: NVGYGRTEIADAIYEQAKELAYYHTYVGHSNEPIIELSERIIREWAPAGMSKVYYGMSGS
N15: NIGYGRTEVADAIYEQAKELAYYHTYVGHSTEAIIELSSRIIRDWAPAGMKKVYYGMSGS
     *.******.**********************.*.*****.****.******.*********

N48: DANETQIKLVWYYNNVLGRPQKKKIISRQRGYHGSGIMTGSLTGLPAFHNHFDLPLEPIR
N15: DANETQIKLVWYYNNVLGRPQKKKIISRQRGYHGSGIMTGSLTGLPSFHQHFDLPIERIK
     ***********************************************.**.*****.*.*.

N48: HTTCPHYYRAPAGMSEAEFSRHCADELEKMILAEGPDTVAAFIGEPVMGTGGIVPPPEGY
N15: HTVCPHWYKAPAGMSEAQFVRYCADELEKLILAEGPDTVAAFIGEPVMGTGGIIPPPQGY
     **.***.*.********.*.*.*******.************************.***.**

N48: WEAIQAVLNKYDILLIADEVVCGFGRTGSMFGSHHYGMKPDLITVAKGLTSAYAPLSGVI
N15: WEAIQAVLNKYDILLIADEVVCGFGRLGSKMGSQHYGMKPDLITVAKGLTSAYAPLSGVI
     **************************.**..**.****************************

N48: VGEKVWKVLEQGSDQYGPIGHGWTYSGHPICAAAALANLDIIERENLTGNAADTGAYFQQ
N15: VGEKVWDVIEKGSQEHGPMGHGWTYSGHPICAAAALANLDILERENLTGNAADVGAYLQQ
     ******.*.*.**...**.**********************.***********.***.**

N48: RMRETFGDHPLVGEVRGVGLMAALEFVADKDKRTRFDPSLKVGPRVSAACLEDGMIARAM
N15: RLRETFGGHPLVGEVRGVGMLAALEFMADKDARTPFDPALKVGPKVSAACLEDGMIARAM
     *.*****.***********..*****.****.**.***.*****.***************

N48: PHGDILGFAPPLVITRAEVDEIVDIAKQAVDAVADELTDTDDLLEMDRAHFFHPSTHLRD
N15: PHGDILGFAPPLVITRAEVDEIVGIAKQAVDEVADEVTDLDQLFEMDRAHFMHPSTHAHD
     **********************.*******.****.**.*.*.*******.******..*

N48: HASGELPGRIITGGKGIRIQDSEGREYIDAFAGLYCVNVGYGRTEIADAIYEQAKELAYY
N15: HASGALPGRIITGGKGIRIQDHEGREYIDAFAGLYCVNIGYGRTEVADAIYEQAKELAYY
     ****.****************.****************.*******.***************

N48: HTYVGHSNEPIIELSERIIREWAPAGMSKVYYGMSGSDANETQIKLVWYYNNVLGRPQKK
N15: HTYVGHSTEAIIELSSRIIRDWAPAGMKKVYYGMSGSDANETQIKLVWYYNNVLGRPQKK
     *******.*.*****.****.******.**********************************

N48: KIISRQRGYHGSGIMTGSLTGLPAFHNHFDLPLEPIRHTTCPHYYRAPAGMSEAEFSRHC
N15: KIISRQRGYHGSGIMTGSLTGLPSFHQHFDLPIERIKHTVCPHWYKAPAGMSEAQFVRYC
     **********************.**.*****.*.*.*.**.***.*.********.*.*.*

N48: ADELEKMILAEGPDTVAAFIGEPVMGTGGIVPPPEGYWEAIQAVLNKYDILLIADEVVCG
N15: ADELEKLILAEGPDTVAAFIGEPVMGTGGIIPPPQGYWEAIQAVLNKYDILLIADEVVCG
     ******.************************.***.**************************

N48: FGRTGSMFGSHHYGMKPDLITVAKGLTSAYAPLSGVIVGEKVWKVLEQGSDQYGPIGHGW
N15: FGRLGSKMGSQHYGMKPDLITVAKGLTSAYAPLSGVIVGEKVWDVIEKGSQEHGPMGHGW
     ***.**..**.*********************************.*.*.**...**.****

N48: TYSGHPICAAAALANLDIIERENLTGNAADTGAYFQQRMRETFGDHPLVGEVRGVGLMAA
N15: TYSGHPICAAAALANLDILERENLTGNAADVGAYLQQRLRETFGGHPLVGEVRGVGMLAA
     ******************.***********.***.***.*****.***********..**

N48: LEFVADKDKRTRFDPSLKVGPRVSAACLEDGMIARAMPHGDILGFAPPLVITRAEVDEIV
N15: LEFMADKDARTPFDPALKVGPKVSAACLEDGMIARAMPHGDILGFAPPLVITRAEVDEIV
     ***.****.**.***.*****.******************************************

N48: DIAKQAVDAVADEL
N15: GIAKQAVDEVADEV
     .*******.****.
```

Figure B.4. N48-N15 Sequence Alignment

```
N15: TDLDQLFEMDRAHFMHPSTHAHDHASGALPGRIITGGKGIRIQDHEGREYIDAFAGLYCV
 N6: TDFDQLFEQDRAHFMHPSTHAHDHASGALPGRIITGASGIRIRDHEGRELIDAFAGLYCV
     **.*****.*********** ***   ********..****.***** .**********

N15: NIGYGRTEVADAIYEQAKELAYYHTYVGHSTEAIIELSSRIIRDWAPAGMKKVYYGMSGS
 N6: NIGYGRTEVADAIYKQAKELAYYHTYVGHSTEAIIELSSRIIRDWAPAGMKKVYYGLSGS
     **************.*********************************************.***

N15: DANETQIKLVWYYNNVLGRPQKKKIISRQRGYHGSGIMTGSLTGLPSFHQHFDLPIERIK
 N6: DANETQIKLVRYYNNVLGRPQKKKIISRQRGYHGSGIMTGSLTGLPSFHQHFDLPVEGIK
     **********.****************************************** *****  .**

N15: HTVCPHWYKAPAGMSEAQFVRYCADELEKLILAEGPDTVAAFIGEPVMGTGGIIPPPQGY
 N6: HTVCPHWYKAPAGMDEAAFVRYCADELEKLILAEGPDTVAAFIGEPVMGTGGIIVPPKGY
     ************ *.* .************************************.**.**

N15: WEAIQAVLNKYDILLIADEVVCGFGRLGSKMGSQHYGMKPDLITVAKGLTSAYAPLSGVI
 N6: WEAIQAVLNKYDVLLIADEVVCAFGRLGSKMGSQRYGMRPDLITTAKGLTSAYAPLSAVI
     ************.*********.***********.***.*****.************.**

N15: VGEKVWDVIEKGSQEHGPMGHGWTYSGHPICAAAALANLDILERENLTGNAADVGAYLQQ
 N6: VGEKVWDVIEKASQKEGAMGHGWTYSGHPICAAAALANLDILERENLTANAADVGAYLNQ
     ******* **  * ..*.*********************************.**** ***  *

N15: RLRETFGGHPLVGEVRGVGMLAALEFMADKDARTPFDPALKVGPKVSAACLEDGMIARAM
 N6: RLRETFEGHPLVGEVRGDGMLAALEFMADREARTPFDPALKVGPKVSAACLEDGMIARAM
     ******. *********.**********     ****************************

N15: PHGDILGFAPPLVITRAEVDEIVGIAKQAVDEVADEVTDLDQLFEMDRAHFMHPSTHAHD
 N6: PHGDILGFAPPLVLTRAEADEIVGIAKAAVDEVAGEVTDFDQLFEQDRAHFMHPSTHAHD
     *************.****.********.******.**  .*****.***************

N15: HASGALPGRIITGGKGIRIQDHEGREYIDAFAGLYCVNIGYGRTEVADAIYEQAKELAYY
 N6: HASGALPGRIITGASGIRIRDHEGRELIDAFAGLYCVNIGYGRTEVADAIYKQAKELAYY
     **************..****.******.*************************.*********

N15: HTYVGHSTEAIIELSSRIIRDWAPAGMKKVYYGMSGSDANETQIKLVWYYNNVLGRPQKK
 N6: HTYVGHSTEAIIELSSRIIRDWAPAGMKKVYYGLSGSDANETQIKLVRYYNNVLGRPQKK
     **********************************.*************.*************

N15: KIISRQRGYHGSGIMTGSLTGLPSFHQHFDLPIERIKHTVCPHWYKAPAGMSEAQFVRYC
 N6: KIISRQRGYHGSGIMTGSLTGLPSFHQHFDLPVEGIKHTVCPHWYKAPAGMDEAAFVRYC
     ******************************** . .*************  *.**.*****

N15: ADELEKLILAEGPDTVAAFIGEPVMGTGGIIPPPQGYWEAIQAVLNKYDILLIADEVVCG
 N6: ADELEKLILAEGPDTVAAFIGEPVMGTGGIIVPPKGYWEAIQAVLNKYDVLLIADEVVCA
     **********************************.**.***************.*********.

N15: FGRLGSKMGSQHYGMKPDLITVAKGLTSAYAPLSGVIVGEKVWDVIEKGSQEHGPMGHGW
 N6: FGRLGSKMGSQRYGMRPDLITTAKGLTSAYAPLSAVIVGEKVWDVIEKASQKEGAMGHGW
     ************ ***.*****.************.********* **  *   *.*****

N15: TYSGHPICAAAALANLDILERENLTGNAADVGAYLQQRLRETFGGHPLVGEVRGVGMLAA
 N6: TYSGHPICAAAALANLDILERENLTANAADVGAYLNQRLRETFEGHPLVGEVRGDGMLAA
     *************************.*********.*******. *********.*****

N15: LEFMADKDARTPFDPALKVGPKVSAACLEDGMIARAMPHGDILGFAPPLVITRAEVDEIV
 N6: LEFMADREARTPFDPALKVGPKVSAACLEDGMIARAMPHGDILGFAPPLVLTRAEADEIV
     ******..*********************************************.****.****

N15: GIAKQAVDEVADEV
 N6: GIAKAAVDEVAGEV
     ****.******.**
```

Figure B.5. N15-N6 Sequence Alignment

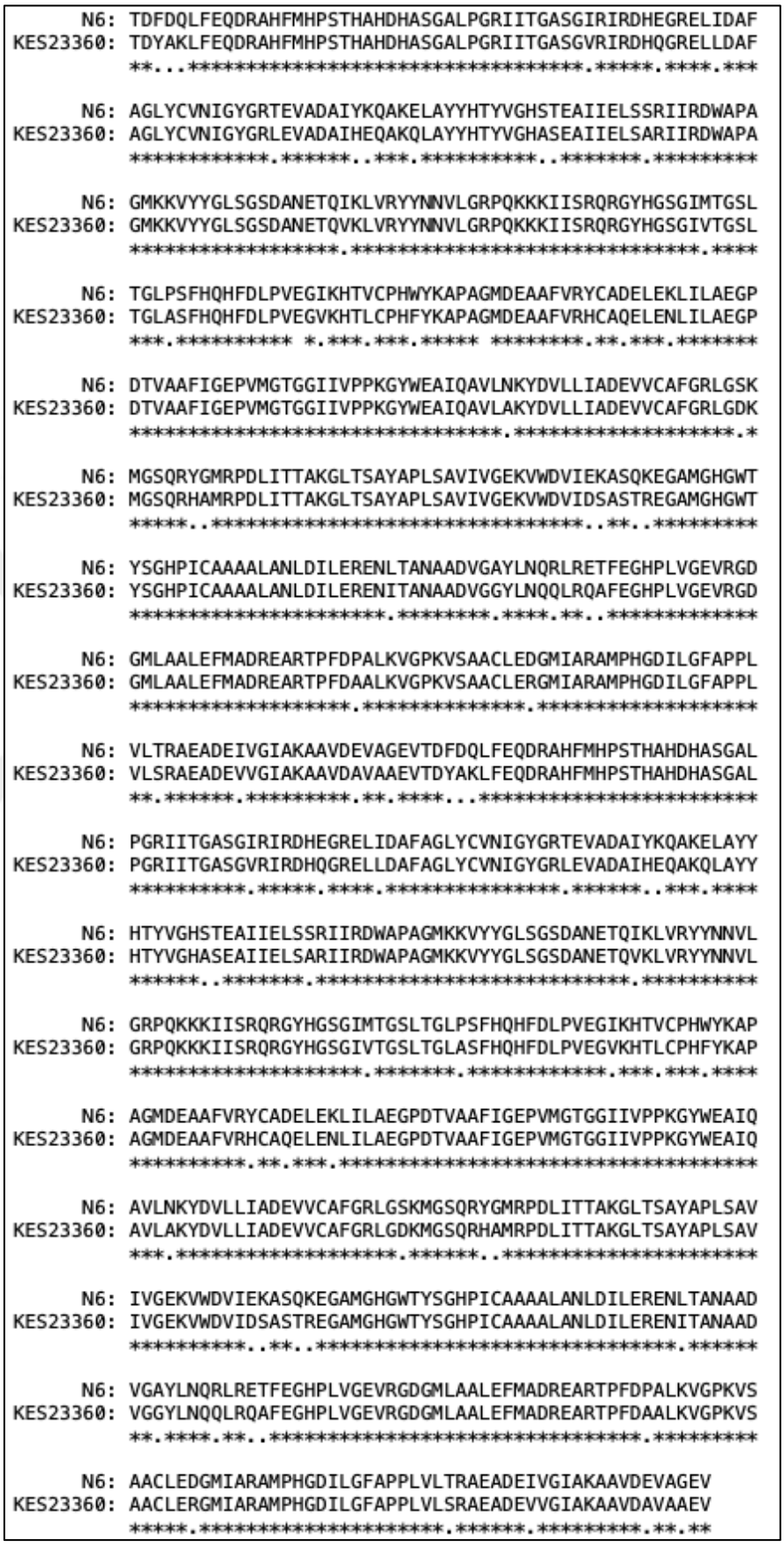

Figure B.6. N6-KES23360 Sequence Alignment

APPENDIX C: MODE VECTOR OVERLAPS OF THE SELECTED ANM MODES AND THE FIRST FIVE SLOWEST GNM MODES

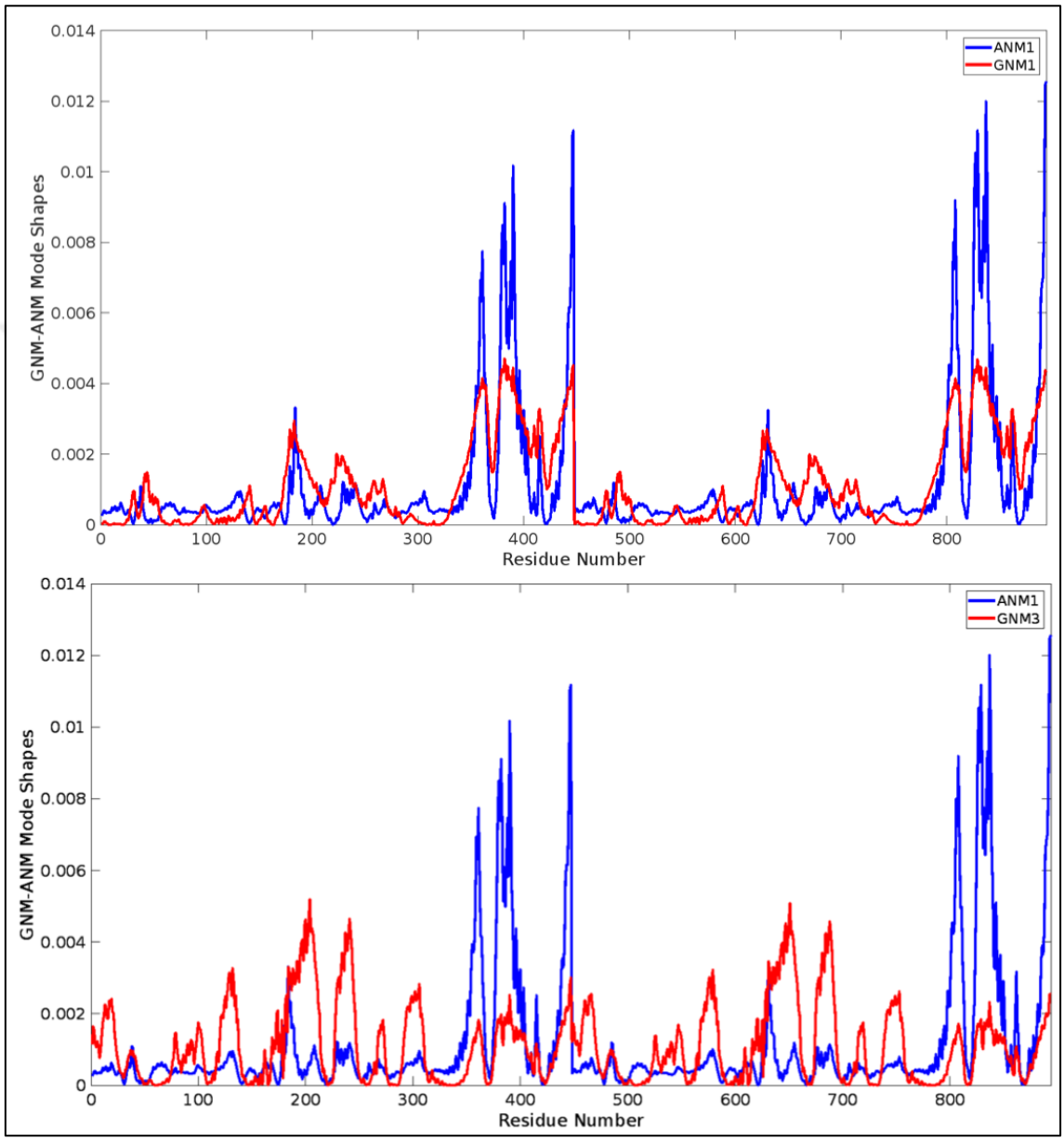

Figure C.1. ANM1 overlaps with GNM1 & GNM3 for N48-N15

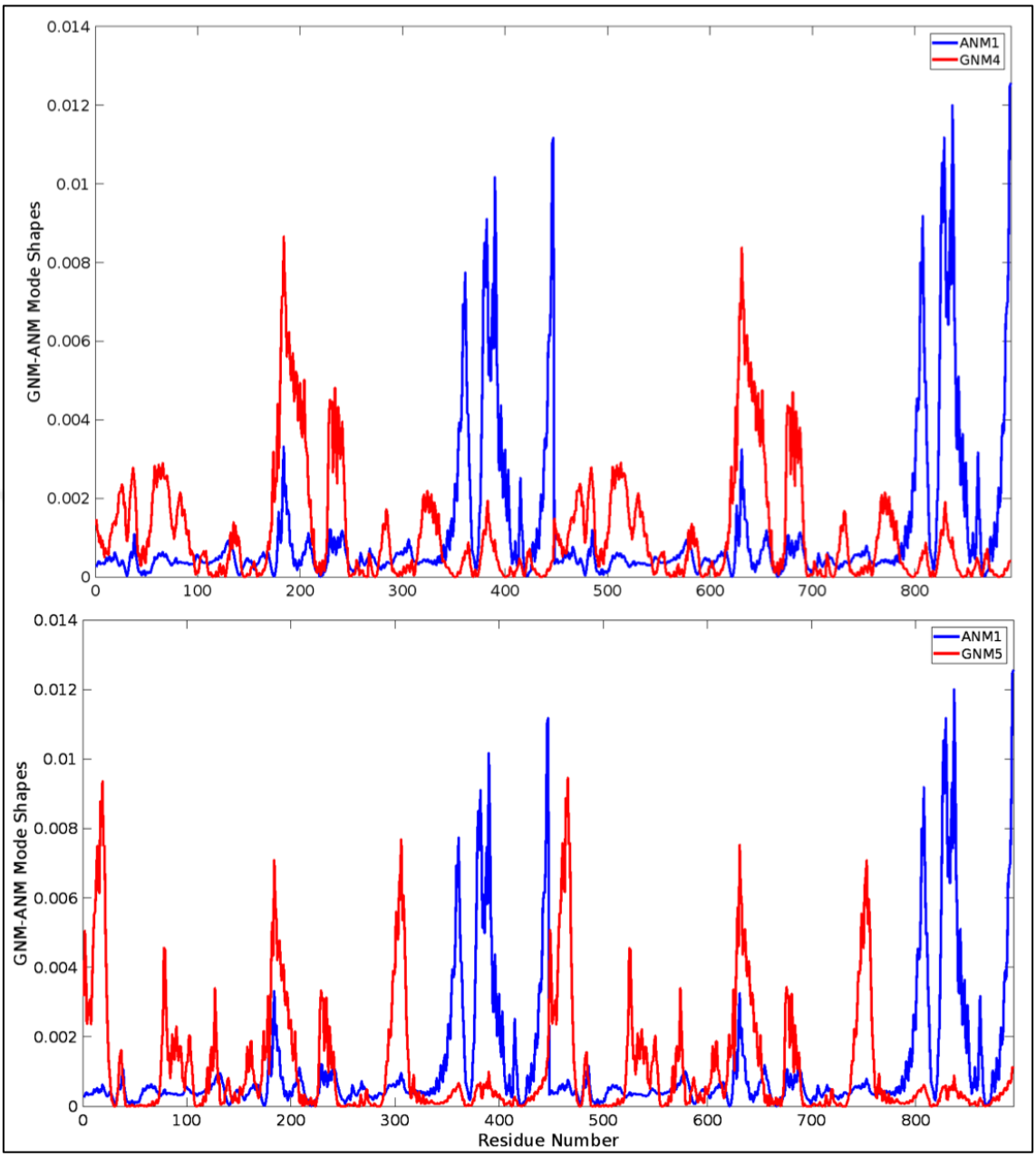

Figure C.2. ANM1 overlaps with GNM4 & GNM5 for N48-N15

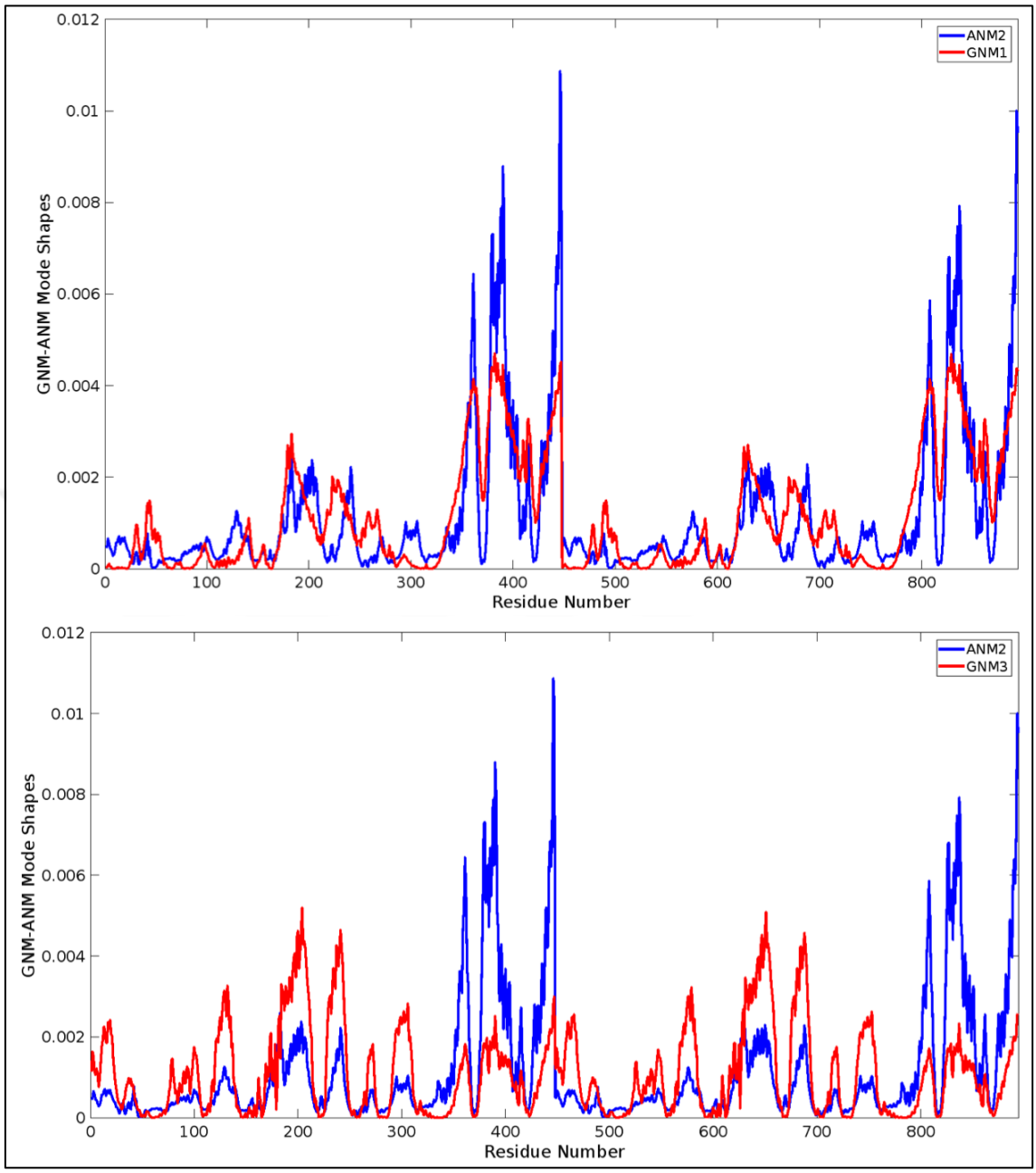

Figure C.3. ANM2 overlaps with GNM1 & GNM3 for N48-N15

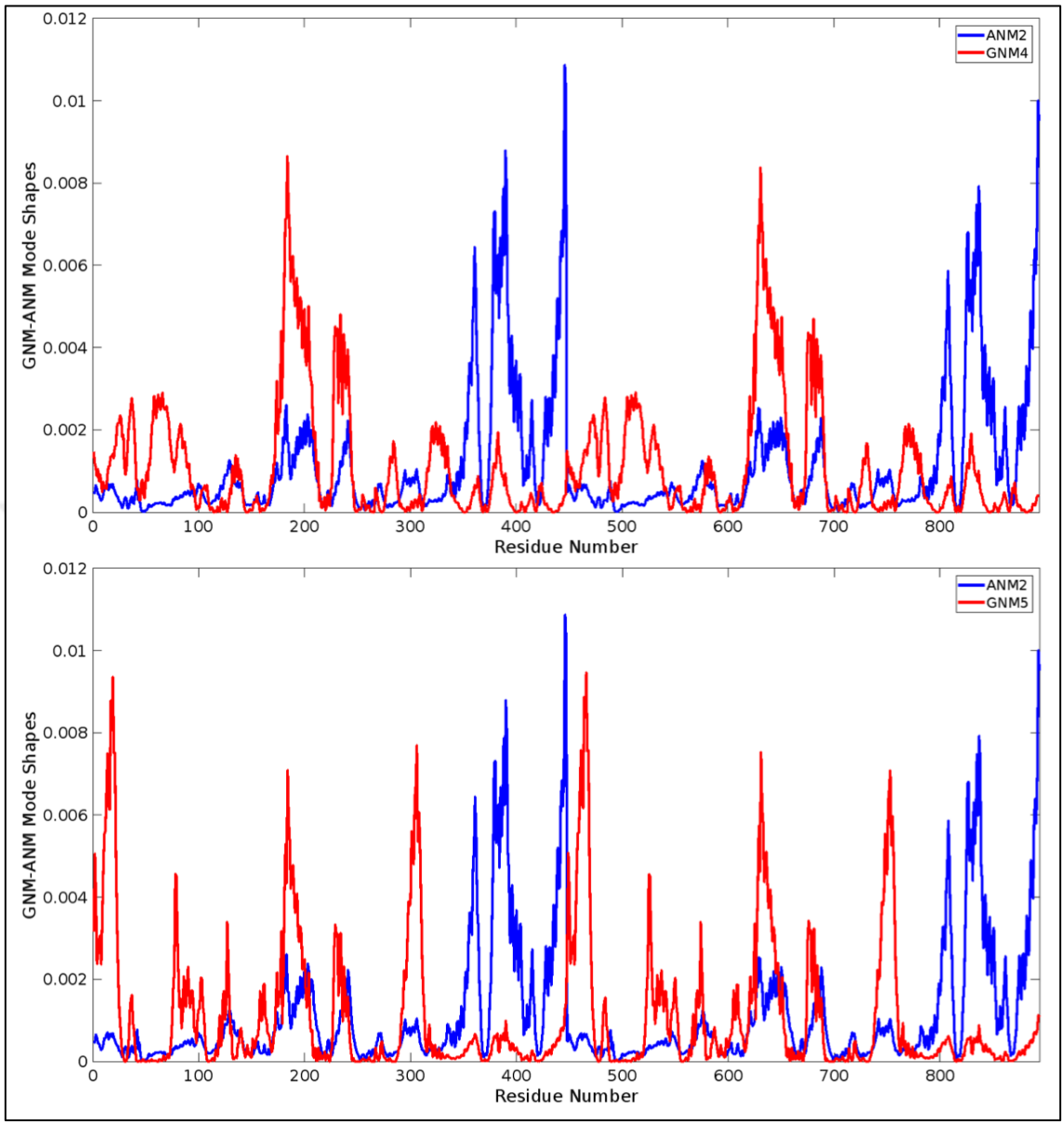

Figure C.4. ANM2 overlaps with GNM4 & GNM5 for N48-N15

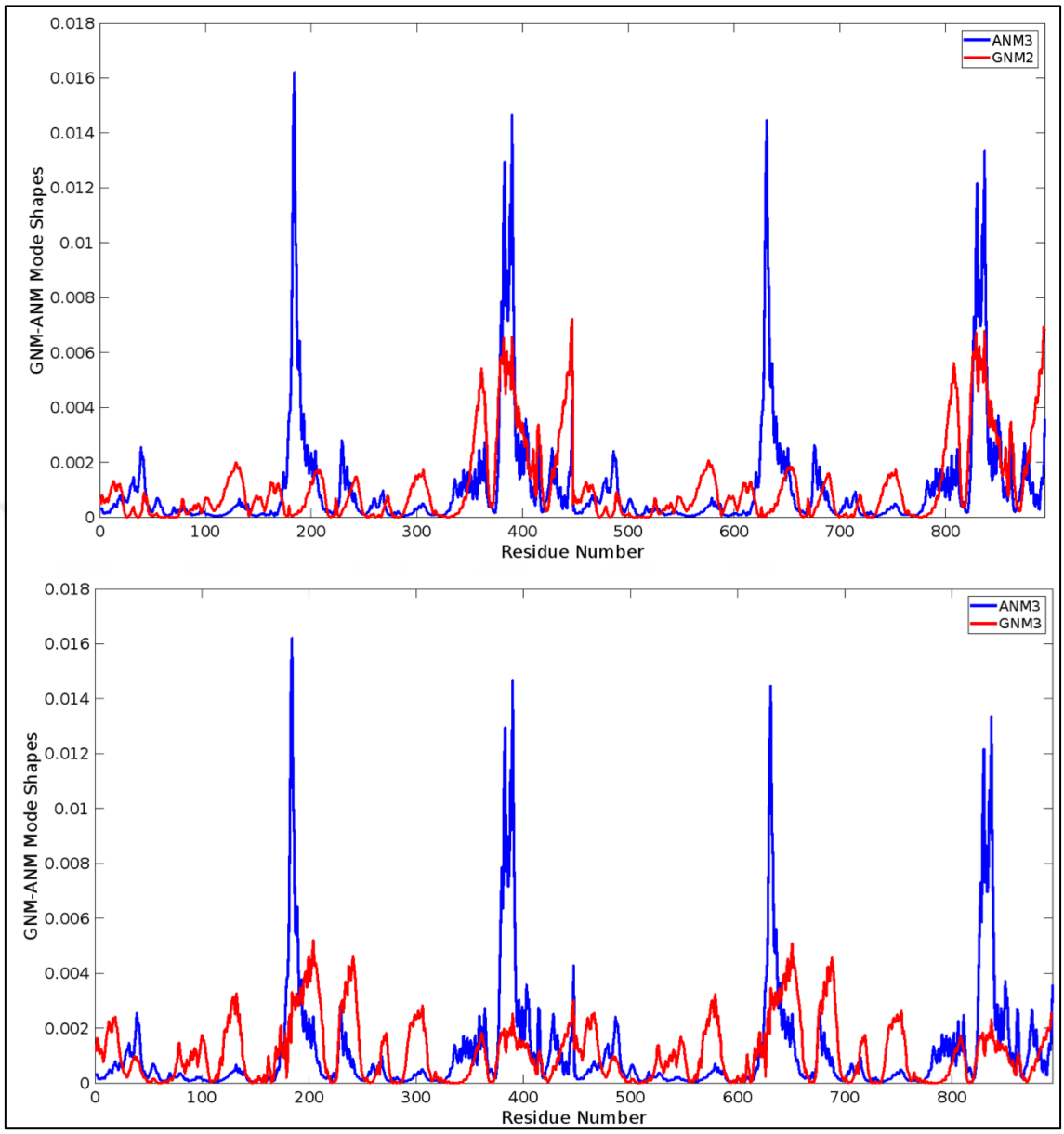

Figure C.5. ANM3 overlaps with GNM2 & GNM3 for N48-N15

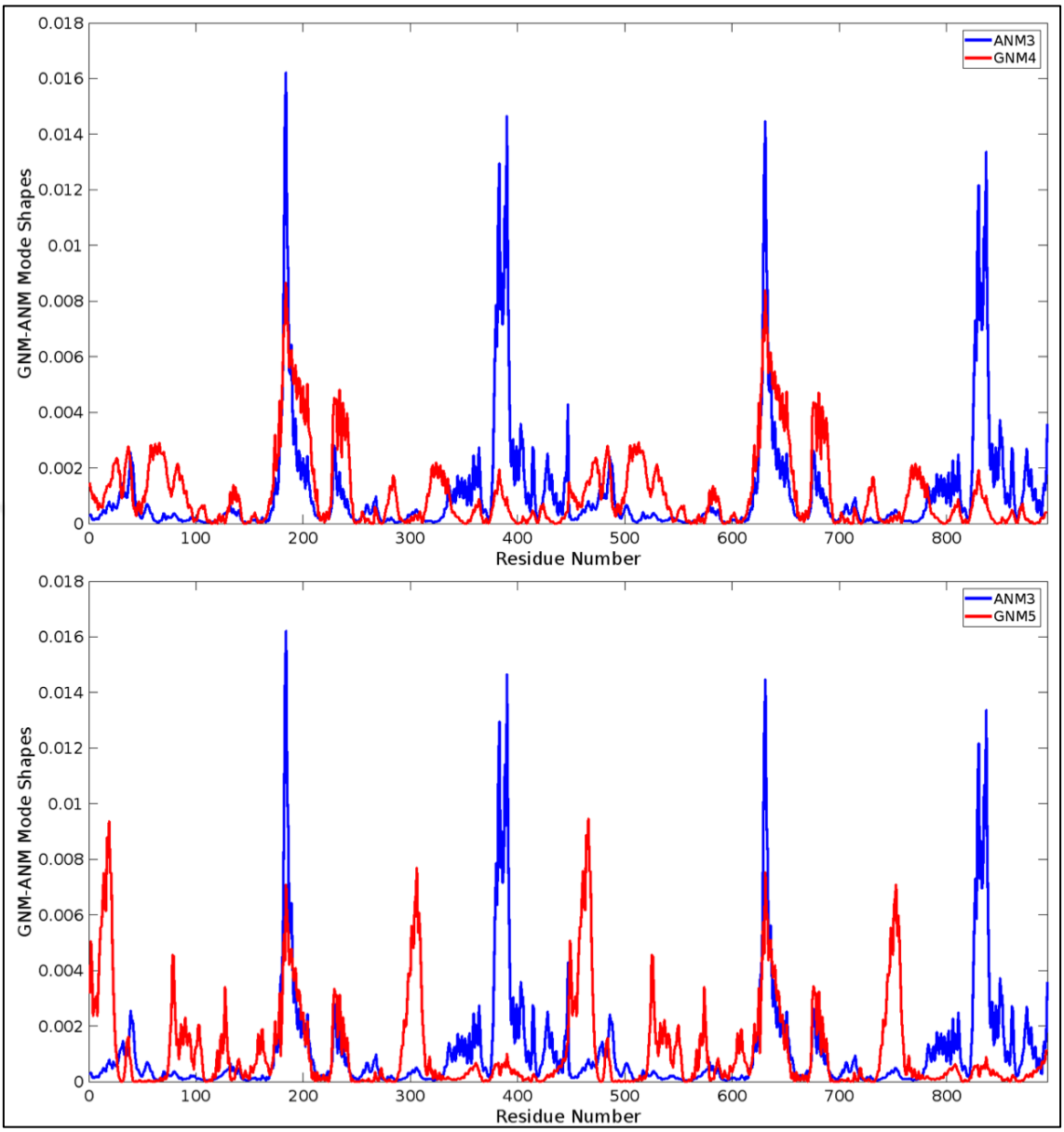

Figure C.6. ANM3 overlaps with GNM4 & GNM5 for N48-N15

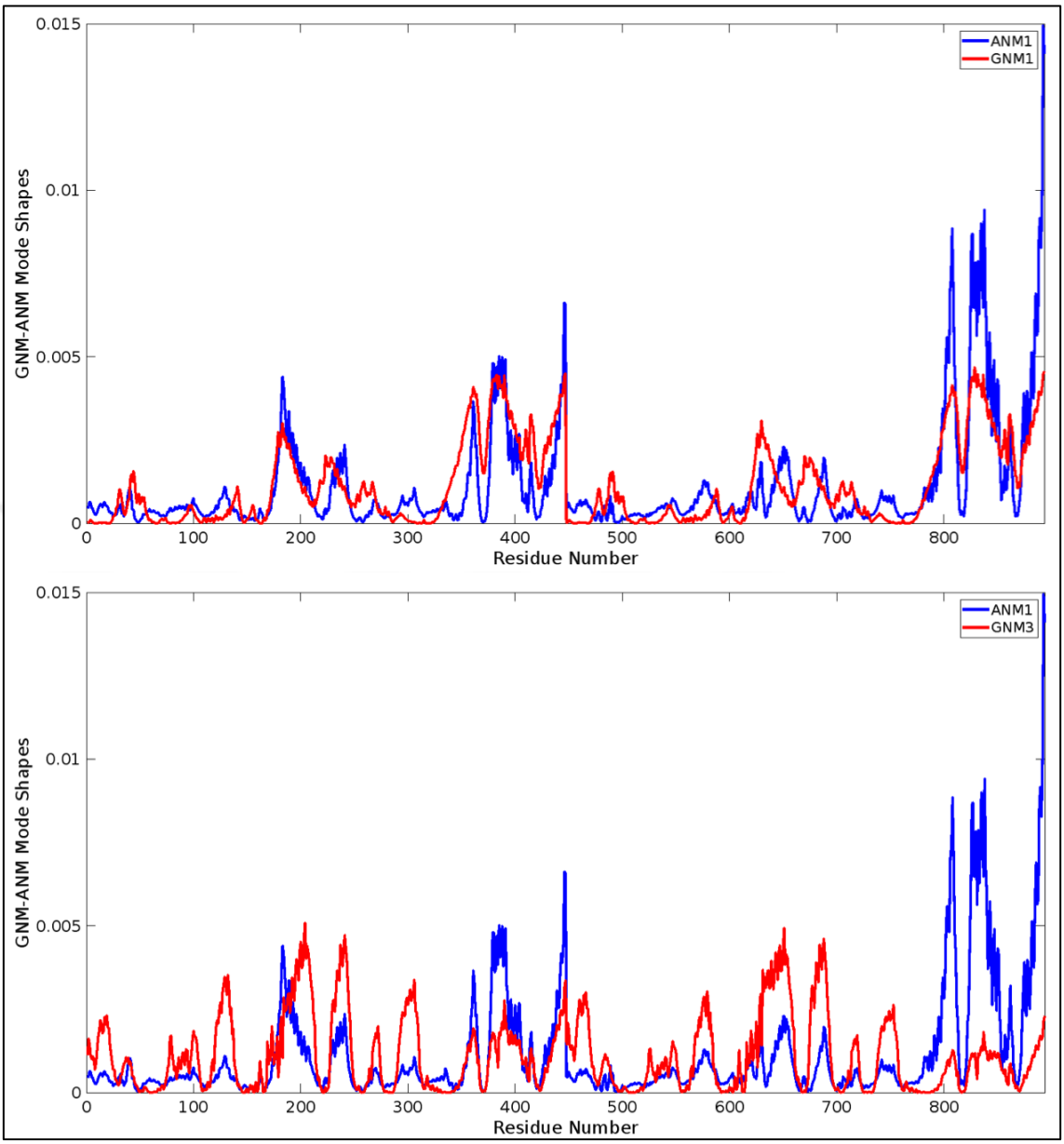

Figure C.7. ANM1 overlaps with GNM1 & GNM3 for N15-N6

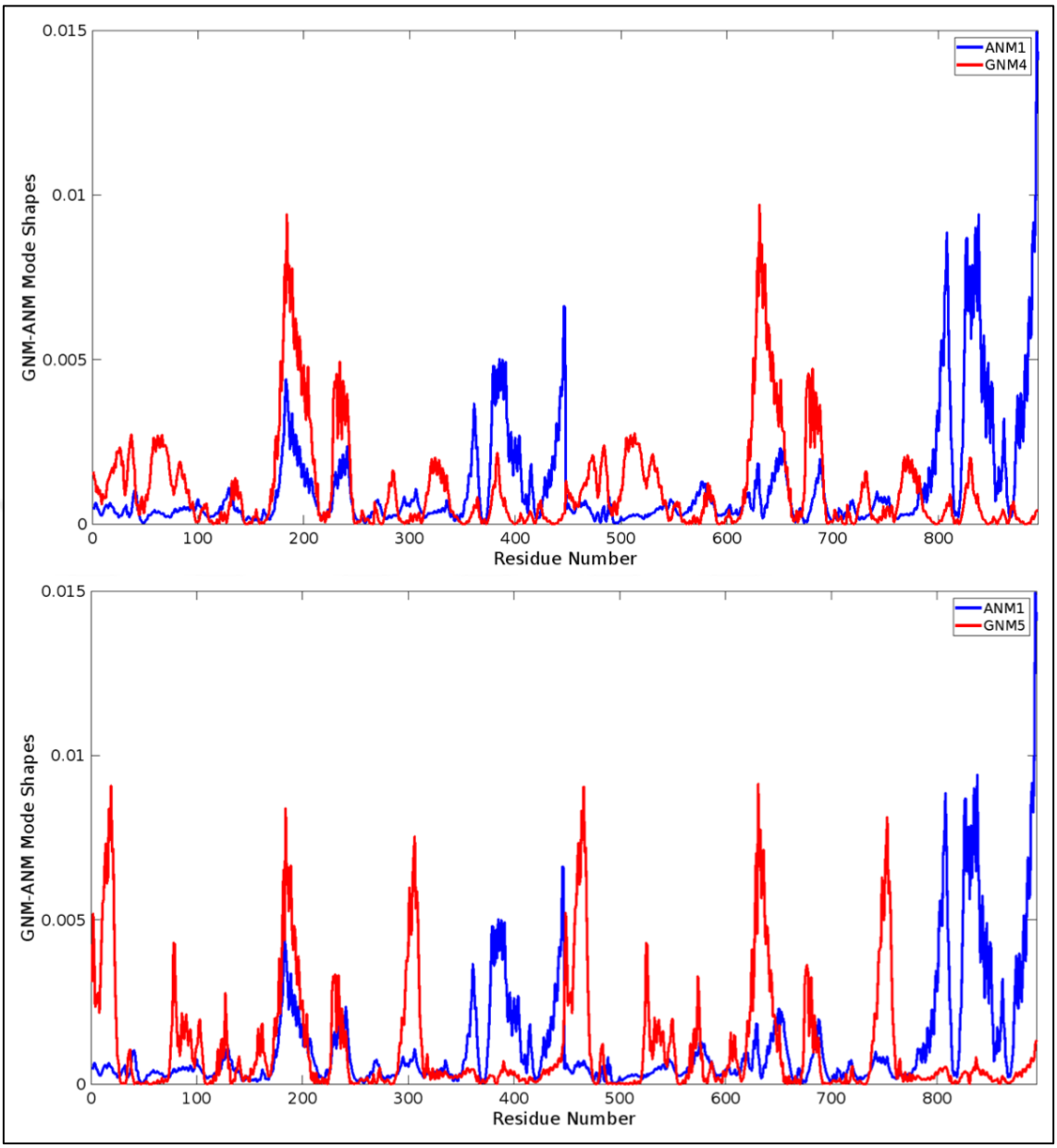

Figure C.8. ANM1 overlaps with GNM4 & GNM5 for N15-N6

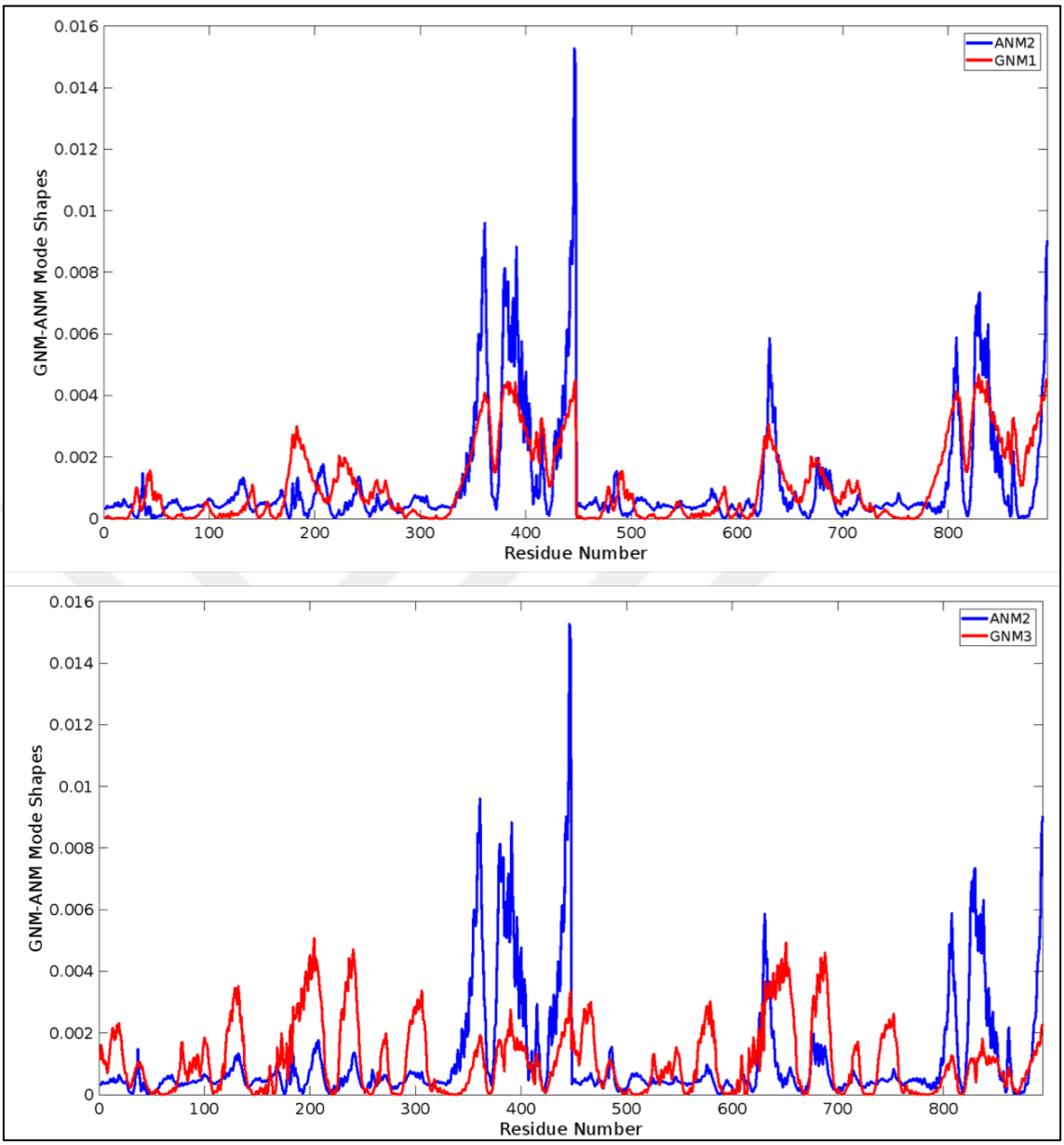

Figure C.9. ANM2 overlaps with GNM1 & GNM3 for N15-N6

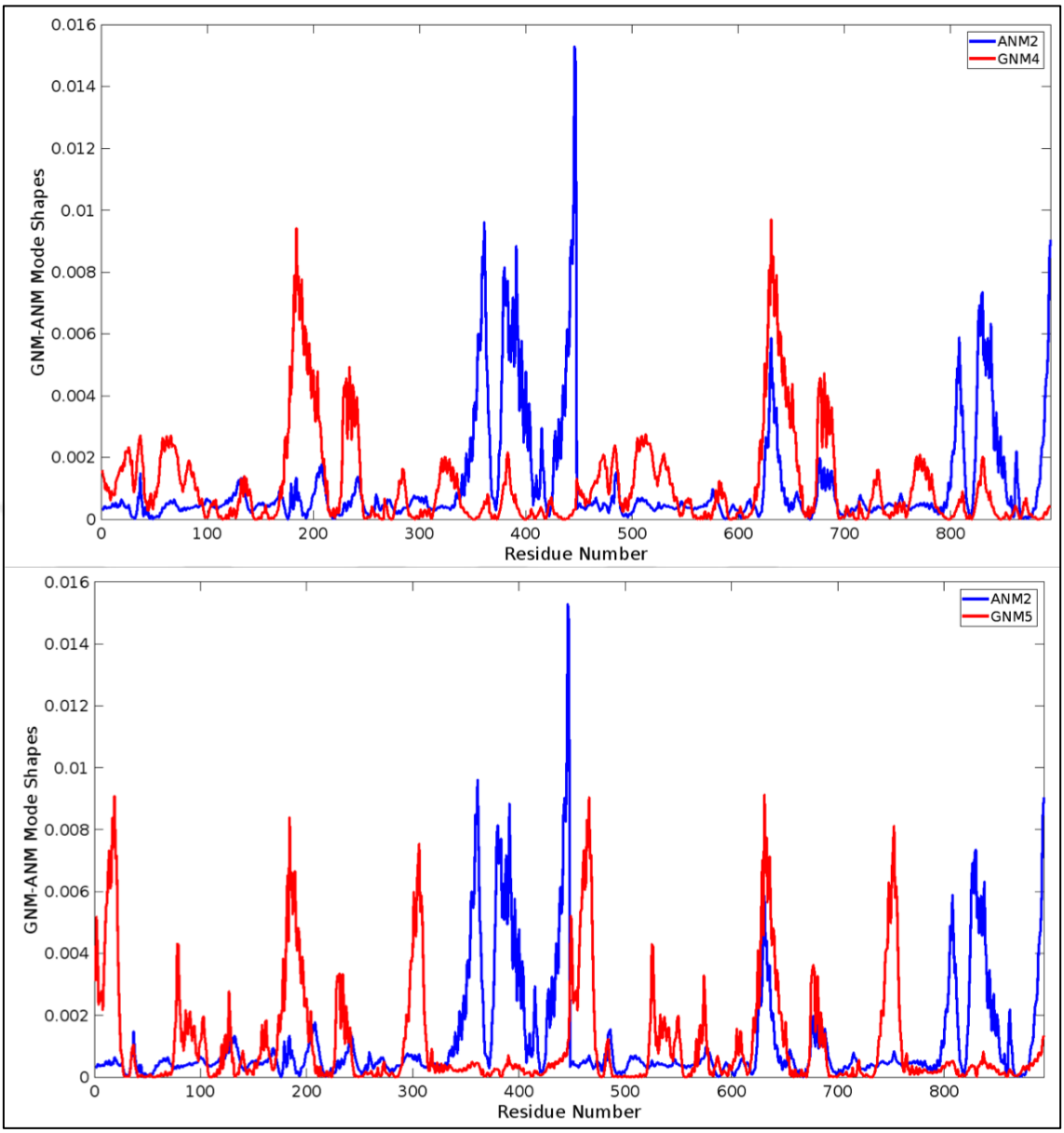

Figure C.10. ANM2 overlaps with GNM4 & GNM5 for N15-N6

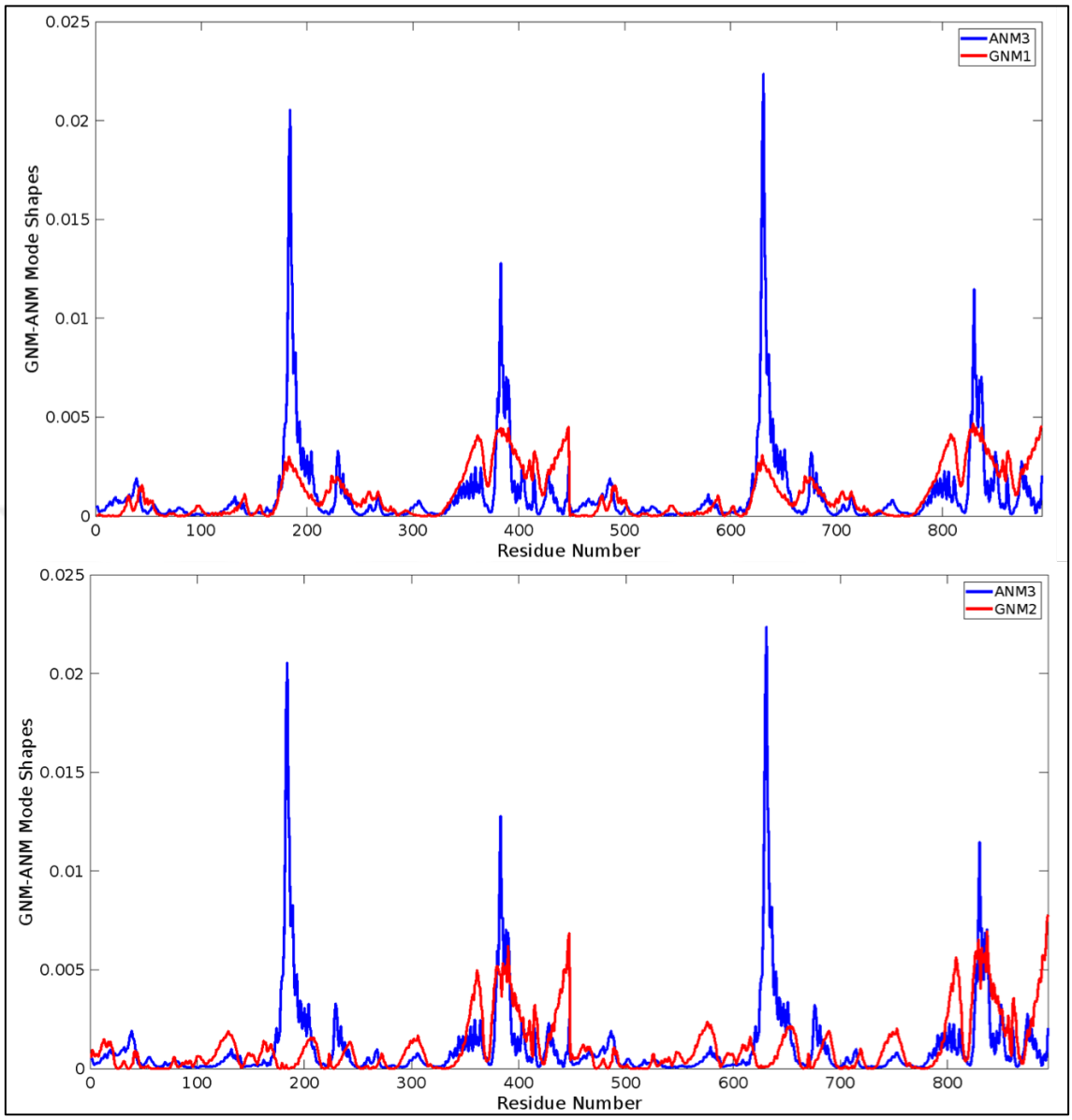

Figure C.11. ANM3 overlaps with GNM1 & GNM2 for N15-N6

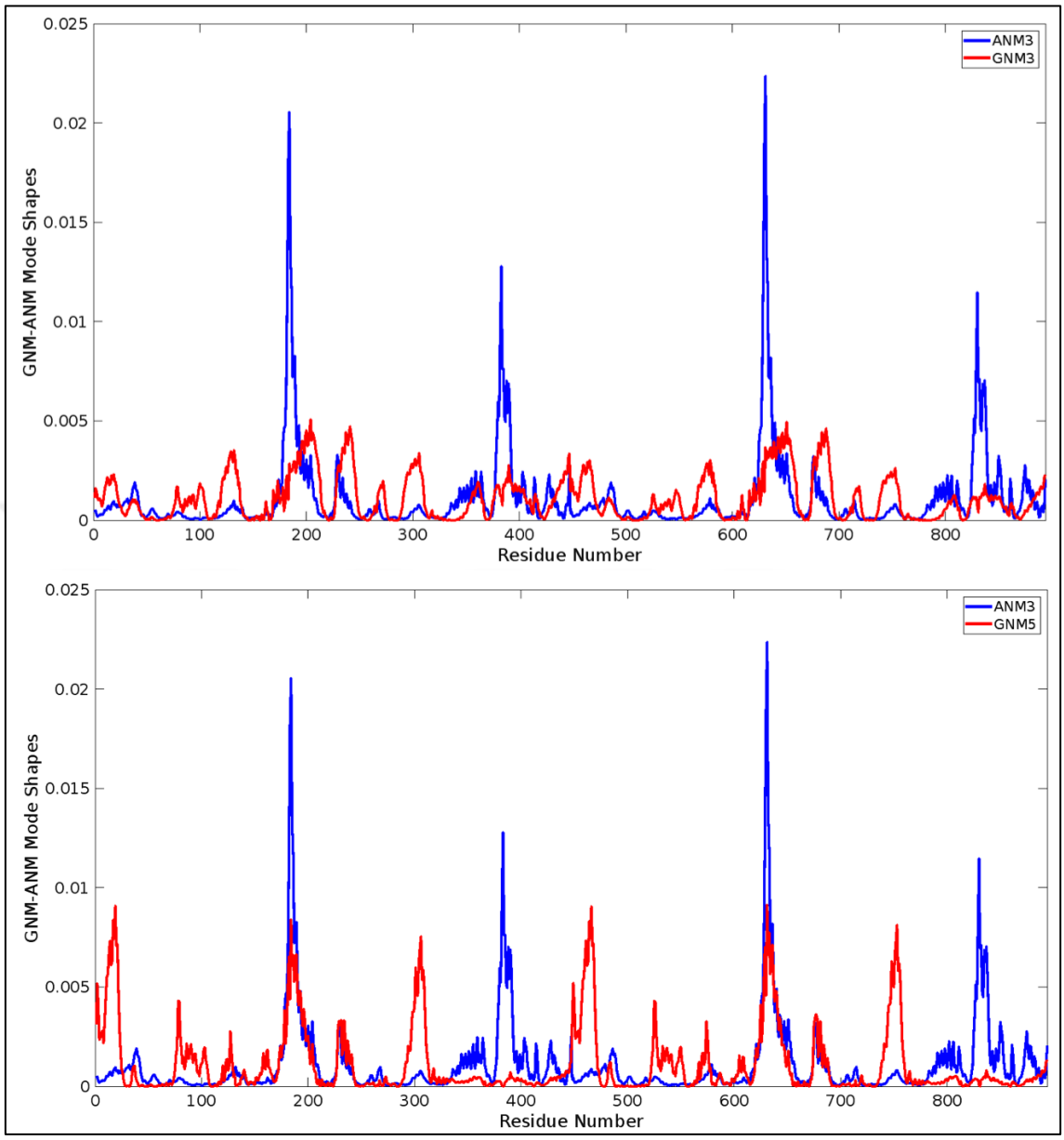

Figure C.12. ANM3 overlaps with GNM3 & GNM5 for N15-N6

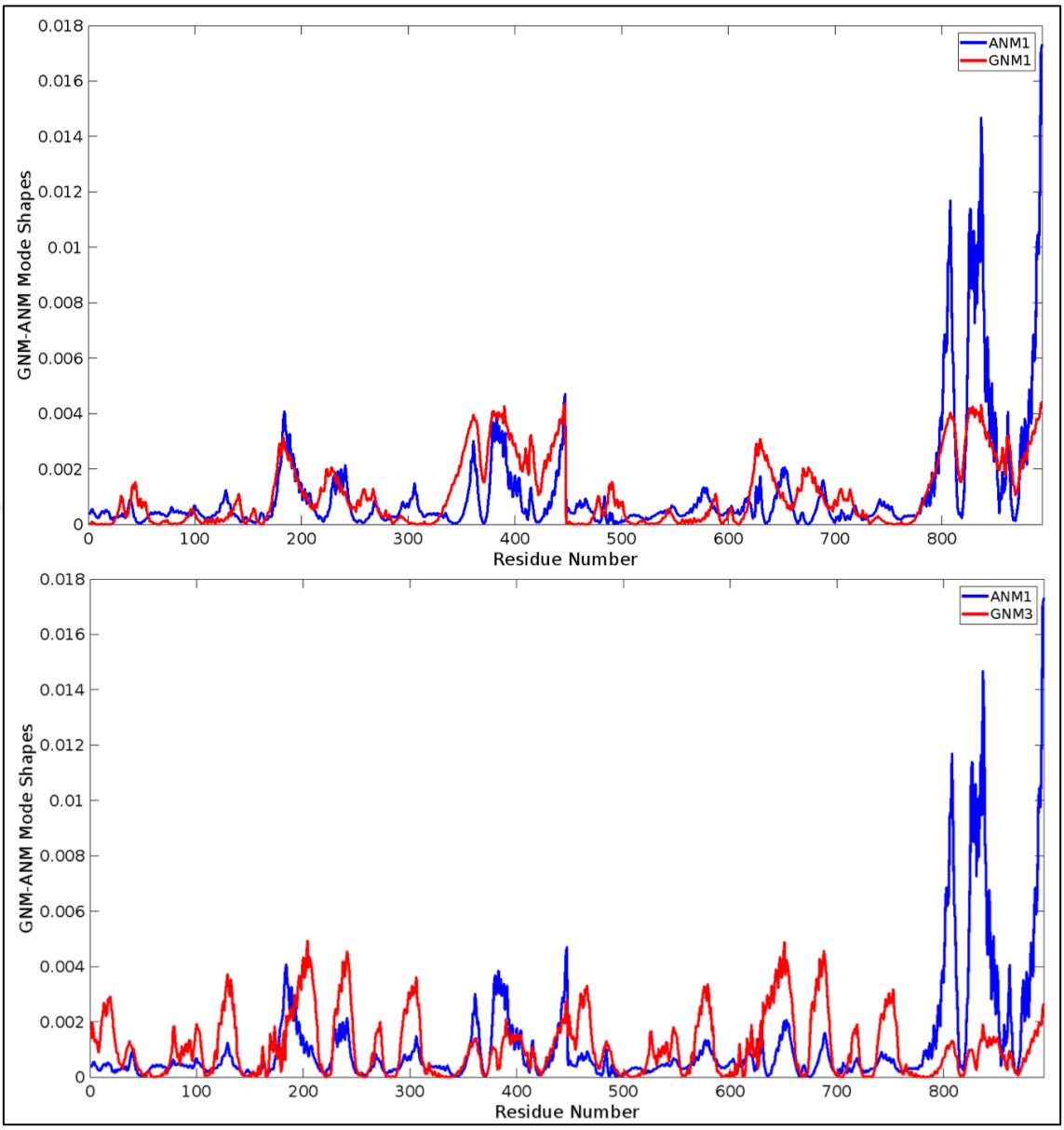

Figure C.13. ANM1 overlaps with GNM1 & GNM3 for N6-KES23360

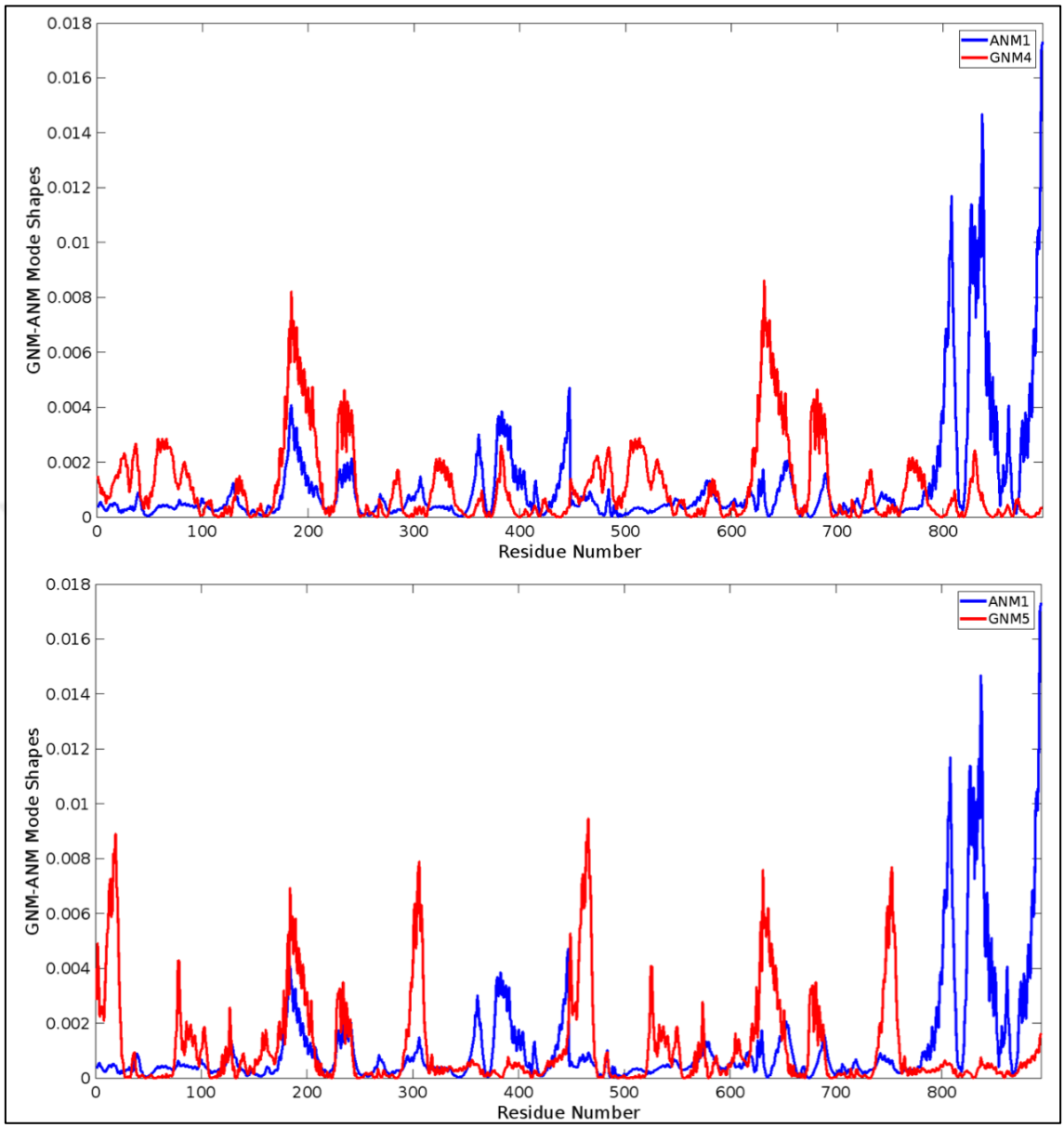

Figure C.14. ANM1 overlaps with GNM4 & GNM5 for N6-KES23360

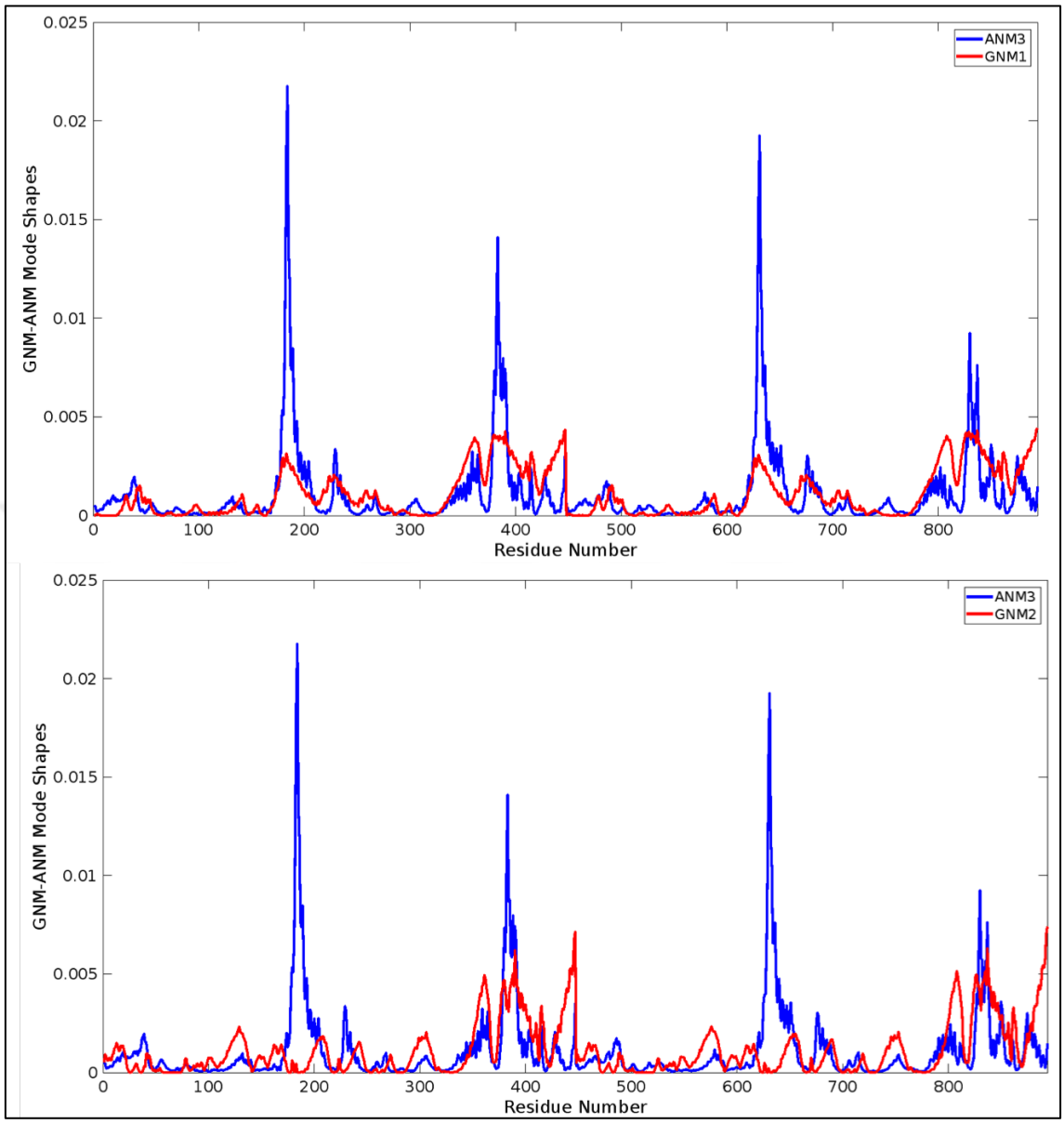

Figure C.15. ANM3 overlaps with GNM1 & GNM2 for N6-KES23360

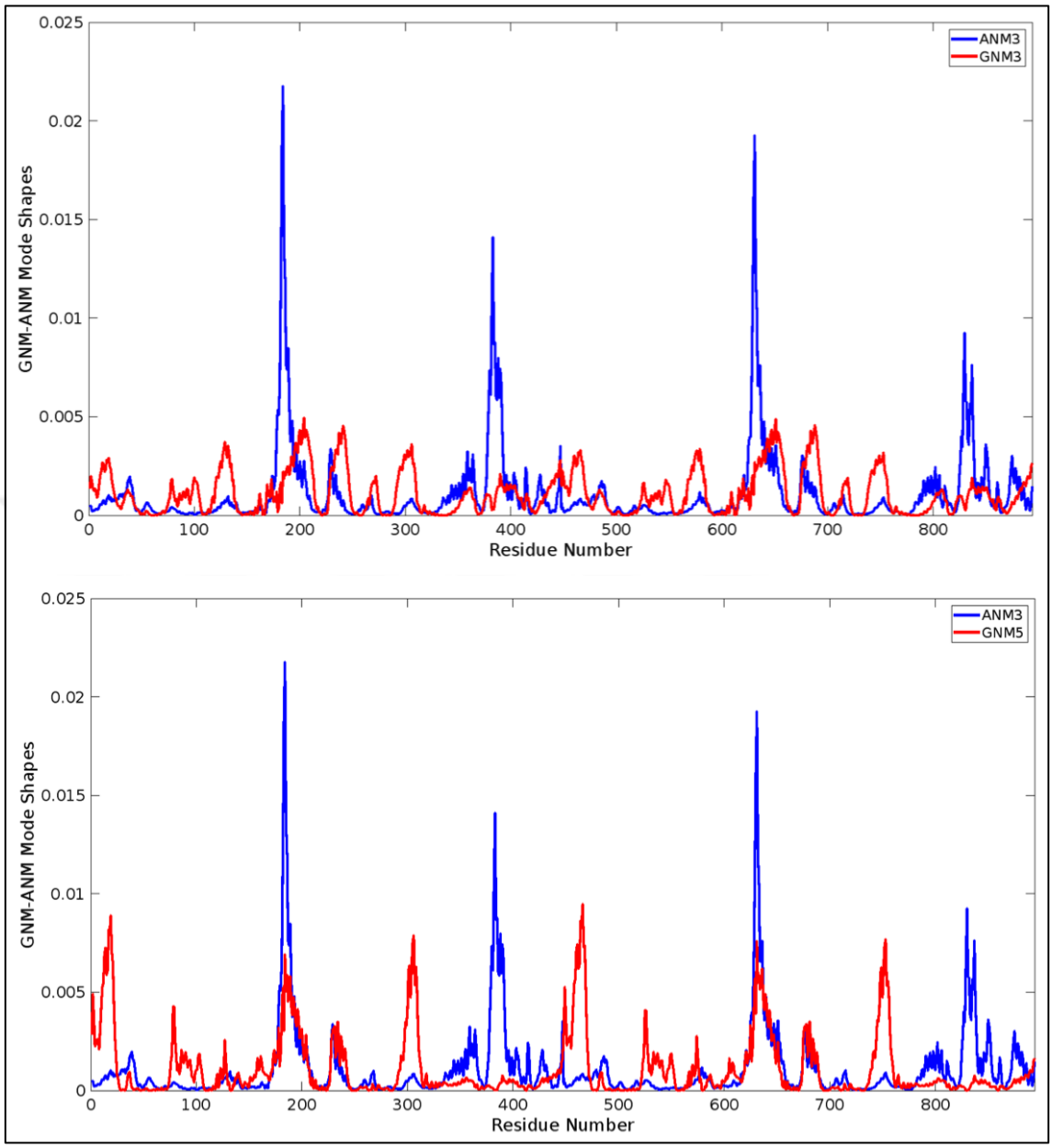

Figure C.16. ANM3 overlaps with GNM3 & GNM5 for N6-KES23360

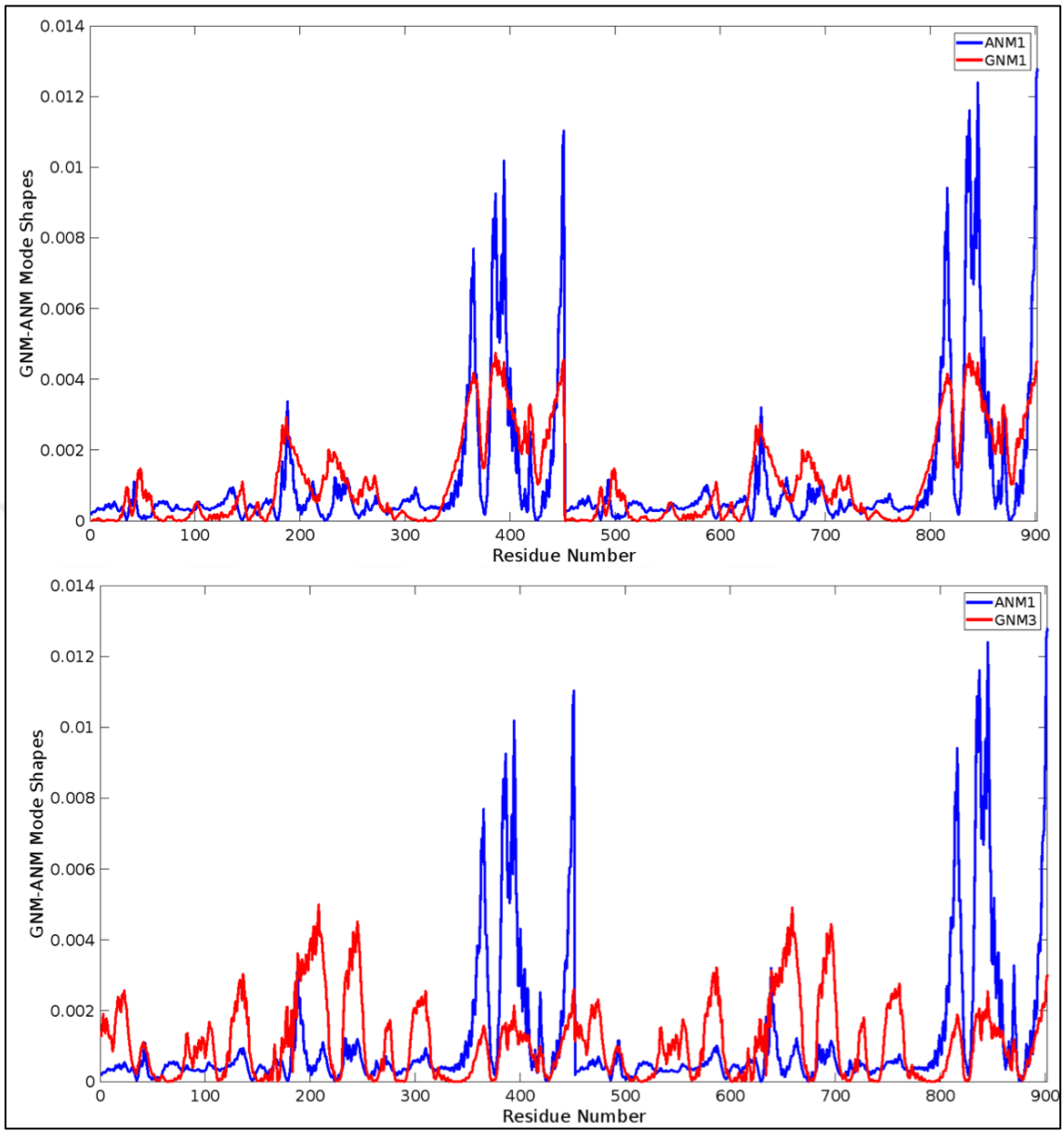

Figure C.17. ANM1 overlaps with GNM1 & GNM3 for N48-KES23360

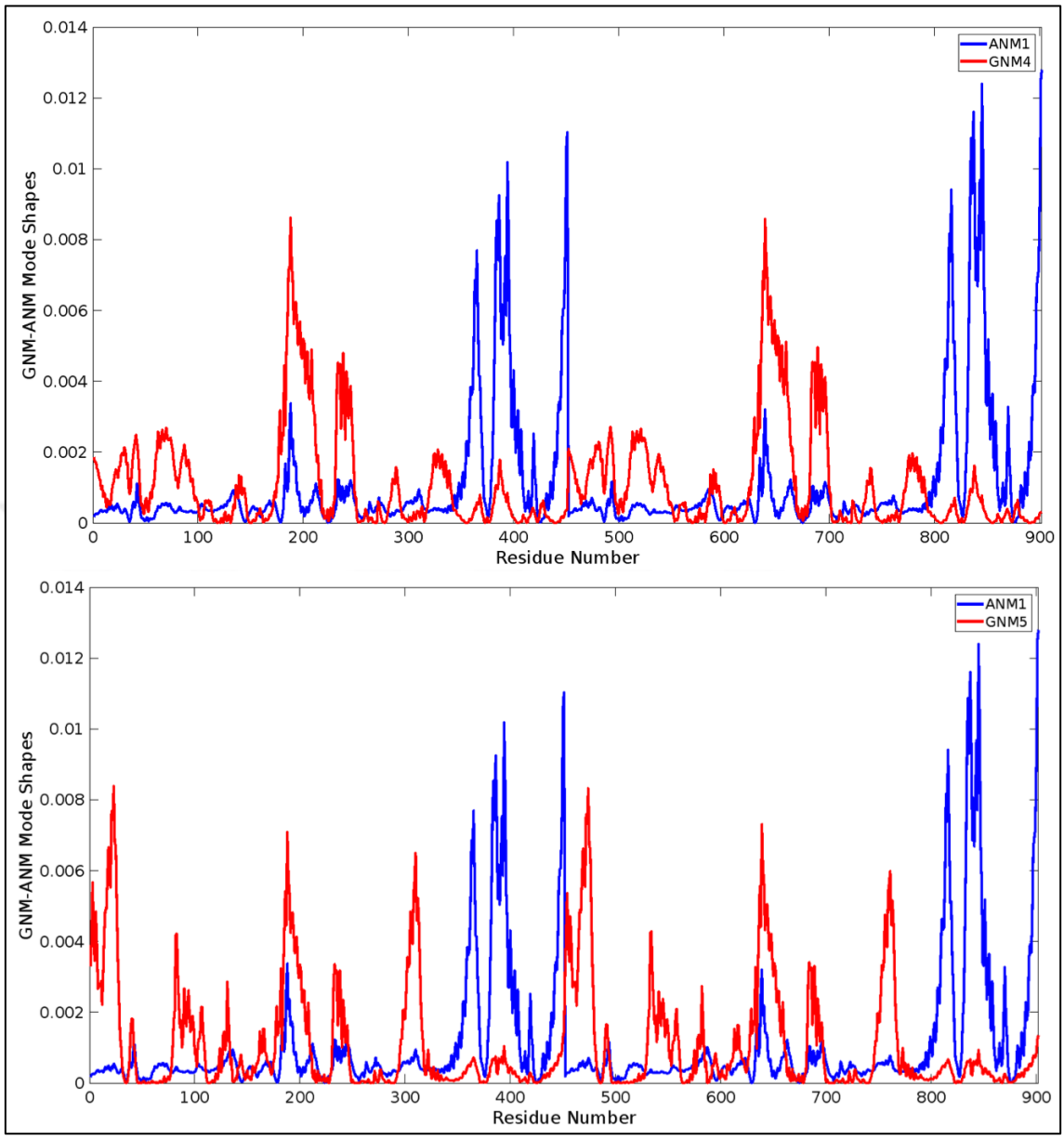

Figure C.18. ANM1 overlaps with GNM4 & GNM5 for N48-KES23360

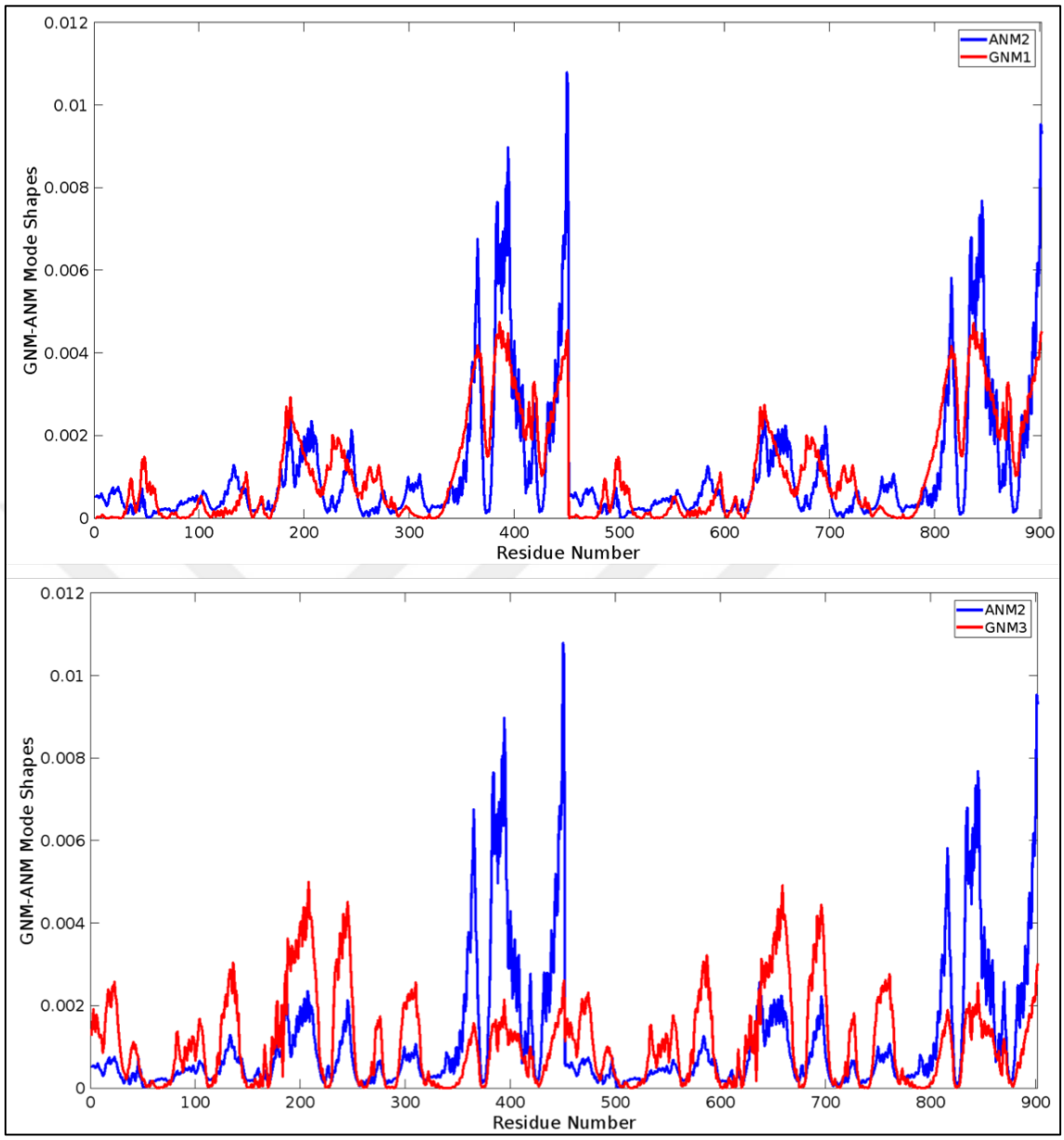

Figure C.19. ANM2 overlaps with GNM1 & GNM3 for N48-KES23360

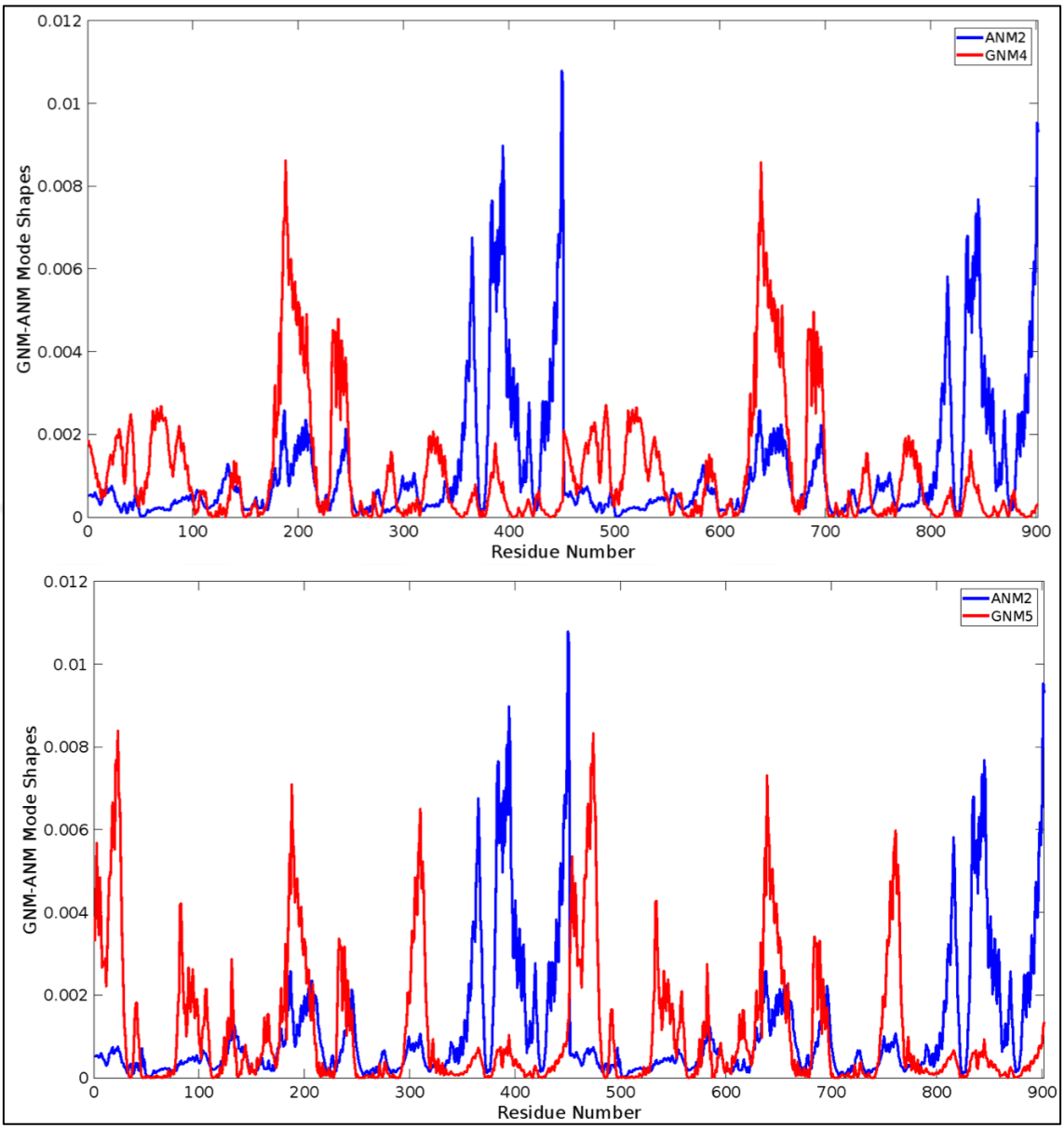

Figure C.20. ANM2 overlaps with GNM4 & GNM5 for N48-KES23360

APPENDIX D: CROSS-MAP DIFFERENCES OF AVERAGE FIRST FIVE GNM MODES

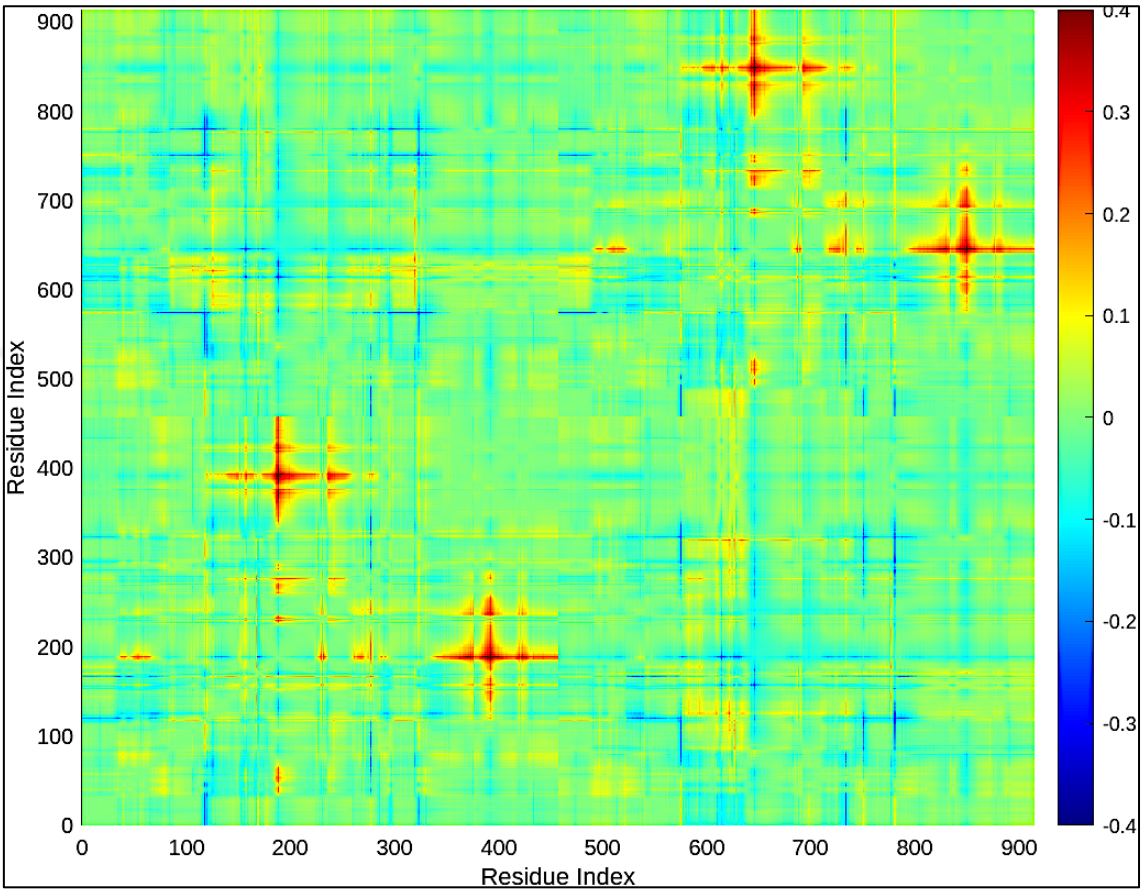

Figure D.1. Cross-map Difference of N15-N6 for the averages of the first five slowest GNM modes

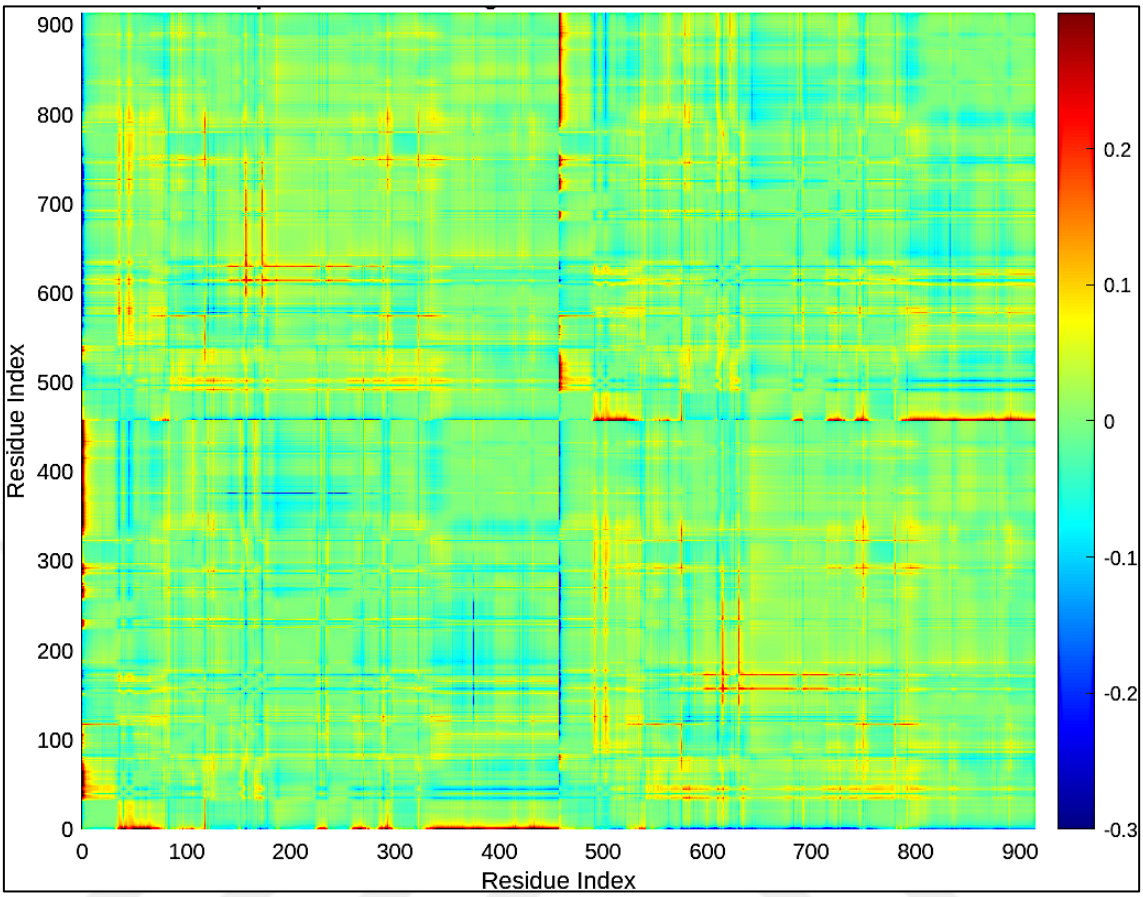

Figure D.2. Cross-map Difference of N6-KES23360 for the averages of the first five slowest GNM modes