COLLECTIVE DYNAMICS OF PROTEIN COMPLEX STRUCTURES: SCORING AND PREDICTION

by

Serdar Özsezen

B.S., Chemical Engineering, Anadolu University, 2012

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

Graduate Program in Chemical Engineering

Boğaziçi University

2015

ACKNOWLEDGEMENTS

First of all I would like to thank my supervisor Türkan Haliloğlu for all of her help, guidance and motivation. She made it possible for me to see diverse research subjects which led me to where I am at the moment. I hope her positive mood and curiosity will bring her much more success. I am grateful for the experience that I have had in Polymer Research Center.

I would like to express my gratitude to my colleagues and friends Zeynep Erge Akbaş, Fidan Sümbül and Melis Yıldırım. It was a joy to work with them. My deepest appreciations are for PRC members Ece Acuner Özbabacan, Burçin Acar, Seren Soner, Zeynep Kürkçüoğlu Soner, Burcu Aykaç Fas, Arzu Uyar, Doğa Fındık, and for my friends Emre Demirel, Barış Burnak, Begüm Alaybeyoğlu, Serhat Erşahin, Coşar Doğa Demirhan, Ezgi Karaca for their support and their company.

My special thanks are to my professors Işıl Bozma for introducing me to Java, Elif Özkırımlı for her support and for making me like my research even more, to my thesis committee, Pemra Doruker and Uğur Sezerman for their valuable time.

In addition, I acknowledge Boğaziçi University Scientific Research Project (project code 14A05P4) and TUBITAK (project code 112T569) for funding of this research.

Finally, my gratitudes are for my family; to my mother for her motivation and to my father for his support, to my brother Metin and his wife Beste for their moral support, to Marks family for their trust in me, most importantly to my girlfriend Lisanne Marks who kept me in this path with her love and support. I dedicate this thesis to them.

ABSTRACT

COLLECTIVE DYNAMICS OF PROTEIN COMPLEX STRUCTURES: SCORING AND PREDICTION

Docking algorithms are widely used in determination of protein complex structures. In order to determine which structure is the most similar to native one, ranking algorithms are used in docking studies. Each year both docking and ranking algorithms are tested in a community-wide experiment called CAPRI (Critical Assessment for Prediction of Interactions). With the purpose of creating a ranking scheme, native protein complex structures were characterized according to their dynamic entities using Gaussian and Anistropic Network Models (GNM and ANM, respectively). Dynamic properties such as mean square residue fluctuations (MSF), positions of hinge residues, and the extent of the correlation of the direction of structural motion in unbound and bound states were investigated. Results display that there is a drop in MSF of interface residues and increase in MSF of non-interface residues upon binding. Also, there is a shift in ANM modes upon binding, which tells that dynamic characteristics of a binding partner in unbound state is contained in dynamic characteristics a of complex. In addition, existence of hinge residues at interface is significant for binding. Degree of collectivity, a property which is calculated by ANM modes, increases upon binding. Using these properties, two distinct software were developed. CLAMshell, a dynamics-based clustering tool which is able to cluster native-like structures based on the similarity of the dynamics of protein chains in unbound and bound states; HiPlane, a tool to draw hinge planes and obtain plane normals for dividing the structure into dynamic domains described by global modes for characterizing orientations of binding partners. With the help of the software that were developed, dynamic properties of predicted models from former CAPRI targets and native structures were also investigated in order to build a ranking scheme.

ÖZET

PROTEİN KOMPLEKSLERİNİN KOLLEKTİF DİNAMİKLERİ: PUANLAMA VE TAHMİN

Kenetlenme algoritmaları protein komplekslerinin yapılarının belirlenmesinde yaygın olarak kullanılmaktadır. Bu algoritmaların en önemli kısımlarından biri puanlamadır. Hangi yapının gerçek yapıya en yakın olduğunu görmek için puanlama algoritmaları kullanılır. Her yıl kenetlenme ve puanlama algoritmaları gruplar arası bir deney ile test edilir (CAPRI, Critical Assessment for Prediction of Interactions). Bu çalışma boyunca gerçek protein kompleksleri dinamik özelliklerine göre Gaussian ve Anizotropik Ağ Yapı modelleri (GNM ve ANM) kullanılarak karakterize edilmeye çalışılmıştır. Amaç, elde edilen bilgileri kullanarak bir puanlama fonksiyonu yaratmaktır. Aminoasitlerin kök ortalama kare sapmaları, eklem (hinge) aminoasitlerin pozisyonları, bağlı ve serbest haldeki yapısal hareketler araştırılmıştır. Sonuçlar, proteinlerin bağlandıktan sonra arayüzlerindeki hareketliliğin azaldığını ve arayüz dışındaki bölgelerde arttığını, bağlanma sonrasında ANM modlarında kayma goruldugunu ve serbest haldeki dinamiklerin kompleks haldeyken de korunduğunu, eklem aminoasitlerin arayüzlerde önem arz ettiğini göstermektedir. Bu özelliklerden yola çıkılarak iki yazılım geliştirilmiştir. CLAMshell, yapıları dinamik temelli özelliklerle sınıflandırma aracıdır; HiPlane, eklem aminoasitlerden geçen bir düzlem çizip bağlanan yapıların birbirlerine göre yönelimini karakterize eden bir araçtır. Geliştirilen yazılımların da yardımıyla, eski CAPRI deneylerinin sonuçlarını ve gerçek kompleks yapılarını kullanarak bir puanlama sistemi elde edilmeye çalışılmıştır.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS iii
ABSTRACT iv
ÖZET v
LIST OF FIGURES viii
LIST OF TABLES xv
LIST OF SYMBOLS xvii
LIST OF ACRONYMS/ABBREVIATIONS xix
1. INTRODUCTION 1
2. LITERATURE SURVEY 3
2.1. Protein-Protein Interactions 3
2.2. Protein-Protein Docking 4
2.3. Scoring Functions 6
2.3.1. Empirical Scoring Functions 6
2.3.2. Knowledge Based Scoring Functions 7
2.3.3. Force-Field Based Scoring Functions 8
3. MATERIALS AND METHODS 9
3.1. Gaussian Network Model (GNM) 9
3.1.1. Hinge Residues 11
3.1.2. Hinge Planes 11
3.2. Anisotropic Network Model (ANM) 14
3.2.1. Eigenvector Correlations 15
3.3. Information Entropy and Collectivity 17
4. RESULTS 19
4.1. Importance of Hinge Residues Positioned at the Interface 19
4.2. Angles between Hinge Planes 22
4.3. Mean Square Fluctuations of Interface and Non-Interface Residues . . . 24
4.4. ANM Mode Correlations 30
4.5. ANM Mode Correlations as a Metric of Dynamics Similarity 36
4.6. Information Entropy and Collectivity 42

4.7. CLAMshell and Ranking Algorithm 46
5. CONCLUSION 50
APPENDIX A: ANM Mode Correlation Clusters 53
APPENDIX B: Changing Number of ANM Mode Correlation Clusters 68
APPENDIX C: GNM Collectivities of Benchmark Complexes 75
APPENDIX D: ANM Collectivities of Benchmark Complexes 78
REFERENCES 85

LIST OF FIGURES

Figure 2.1. Protein-protein docking process. 4

Figure 3.1. Representation of elastic network model. Red spheres show alpha-carbon atoms and R_c states the cutoff radius. 9

Figure 3.2. Illustration of hinge plane for lysozime (PDB code 1BVK, chain F). 12

Figure 3.3. Eigenvector correlations (10×10 matrix) for human tissue factor/inhibitor complex (PDB code 1AHW, chains AB and C). . . . 16

Figure 3.4. Eigenvector correlations (10×20 matrix) for human tissue factor/inhibitor complex (PDB code 1AHW, chains AB and C). . . . 17

Figure 4.1. Hinge planes of humanized anti-lysozyme FV complex with lysozyme (PDB code 1BVK). 22

Figure 4.2. Distribution of hinge plane angles among Docking Benchmark complexes. 23

Figure 4.3. Distributions of hinge plane angles among CAPRI targets. 24

Figure 4.4. Mean square fluctuations difference between bound and unbound states of the interface residues of benchmark complexes. 25

Figure 4.5. Change in mean square fluctuations of C2 fragment of streptococcal protein G in complex with the FC domain of human IGG protein complex (1FCC) upon binding. 26

Figure 4.6. Mean square fluctuations difference between bound and unbound states of the non-interface residues of benchmark complexes. . . . 26

Figure 4.7. Overall mean square fluctuations difference between bound and unbound states of the benchmark complexes. 27

Figure 4.8. Number of unbound-bound mode correlations higher than 0.5. . . 31

Figure 4.9. Number of unbound-bound mode correlations for antibody/antigen complexes. 32

Figure 4.10. Number of unbound-bound mode correlations for antigen/bound antibody complexes. 33

Figure 4.11. Number of unbound-bound mode correlations for enzyme/inhibitor complexes. 34

Figure 4.12. Number of unbound-bound mode correlations for other complexes. 34

Figure 4.13. Combinations of unbound and bound ANM modes with highest correlations (U: Unbound mode, B: Bound mode). 35

Figure 4.14. Combinations of unbound and bound ANM modes of interface residues with highest correlations (U: Unbound mode, B: Bound mode). 36

Figure 4.15. Distributions and probability density functions of bound and unbound collectivities for first slowest ANM mode of first binding partners in Docking Benchmark. 44

Figure 4.16. Distributions and probability density functions of bound and unbound collectivities for first slowest ANM mode of first binding partners in Docking Benchmark. 44

Figure 4.17. Probability distribution functions of collectivity summations for first ten slowest modes of Docking Benchmark complexes. 45

Figure 4.18. Probability distribution functions of collectivity differences for first ten slowest modes of Docking Benchmark complexes. 45

Figure 4.19. Probability distribution functions of collectivity gains for average of first ten slowest modes of target models and Docking Benchmark complexes. 46

Figure 4.20. Ranking algorithm. 47

Figure B.1. Number of clusters versus in cluster probability for Target 30. . . 68

Figure B.2. Number of clusters versus in cluster probability for Target 32. . . 68

Figure B.3. Number of clusters versus in cluster probability for Target 35. . . 69

Figure B.4. Number of clusters versus in cluster probability for Target 36. . . 69

Figure B.5. Number of clusters versus in cluster probability for Target 37. . . 69

Figure B.6. Number of clusters versus in cluster probability for Target 40_CA. 70

Figure B.7. Number of clusters versus in cluster probability for Target 40_CB. 70

Figure B.8. Number of clusters versus in cluster probability for Target 41. . . 70

Figure B.9. Number of clusters versus in cluster probability for Target 42_AB. 71

Figure B.10. Number of clusters versus in cluster probability for Target 42_BC. 71

Figure B.11. Number of clusters versus in cluster probability for Target 46. . . 71

Figure B.12. Number of clusters versus in cluster probability for Target 47. . . 72

Figure B.13. Number of clusters versus in cluster probability for Target 48. . . 72

Figure B.14. Number of clusters versus in cluster probability for Target 49. . . 72

Figure B.15. Number of clusters versus in cluster probability for Target 50. . . 73

Figure B.16. Number of clusters versus in cluster probability for Target 51_1. . 73

Figure B.17. Number of clusters versus in cluster probability for Target 51_2. . 73

Figure B.18. Number of clusters versus in cluster probability for Target 53. . . 74

Figure B.19. Number of clusters versus in cluster probability for Target 59. . . 74

Figure C.1. Distributions and probability density functions of bound and unbound collectivities for GNM mode 1 of first binding partners in Docking Benchmark. 76

Figure C.2. Distributions and probability density functions of bound and unbound collectivities for GNM mode 1 of first second partners in Docking Benchmark. 76

Figure C.3. Distributions and probability density functions of bound and unbound collectivities for GNM mode 2 of first binding partners in Docking Benchmark. 76

Figure C.4. Distributions and probability density functions of bound and unbound collectivities for GNM mode 2 of first second partners in Docking Benchmark. 77

Figure C.5. Distributions and probability density functions of bound and unbound collectivities for GNM mode 3 of first binding partners in Docking Benchmark. 77

Figure C.6. Distributions and probability density functions of bound and unbound collectivities for GNM mode 3 of first second partners in Docking Benchmark. 77

Figure D.1. Distributions and probability density functions of bound and unbound collectivities for ANM mode 2 of first binding partners in Docking Benchmark. 78

Figure D.2. Distributions and probability density functions of bound and unbound collectivities for ANM mode 2 of first second partners in Docking Benchmark. 79

Figure D.3. Distributions and probability density functions of bound and unbound collectivities for ANM mode 3 of first binding partners in Docking Benchmark. 79

Figure D.4. Distributions and probability density functions of bound and unbound collectivities for ANM mode 3 of first second partners in Docking Benchmark. 79

Figure D.5. Distributions and probability density functions of bound and unbound collectivities for ANM mode 4 of first binding partners in Docking Benchmark. 80

Figure D.6. Distributions and probability density functions of bound and unbound collectivities for ANM mode 4 of first second partners in Docking Benchmark. 80

Figure D.7. Distributions and probability density functions of bound and unbound collectivities for ANM mode 5 of first binding partners in Docking Benchmark. 80

Figure D.8. Distributions and probability density functions of bound and unbound collectivities for ANM mode 5 of first second partners in Docking Benchmark. 81

Figure D.9. Distributions and probability density functions of bound and unbound collectivities for ANM mode 6 of first binding partners in Docking Benchmark. 81

Figure D.10. Distributions and probability density functions of bound and unbound collectivities for ANM mode 6 of first second partners in Docking Benchmark. 81

Figure D.11. Distributions and probability density functions of bound and unbound collectivities for ANM mode 7 of first binding partners in Docking Benchmark. 82

Figure D.12. Distributions and probability density functions of bound and unbound collectivities for ANM mode 7 of first second partners in Docking Benchmark. 82

Figure D.13. Distributions and probability density functions of bound and unbound collectivities for ANM mode 8 of first binding partners in Docking Benchmark. 82

Figure D.14. Distributions and probability density functions of bound and unbound collectivities for ANM mode 8 of first second partners in Docking Benchmark. 83

Figure D.15. Distributions and probability density functions of bound and unbound collectivities for ANM mode 9 of first binding partners in Docking Benchmark. 83

Figure D.16. Distributions and probability density functions of bound and unbound collectivities for ANM mode 9 of first second partners in Docking Benchmark. 83

Figure D.17. Distributions and probability density functions of bound and unbound collectivities for ANM mode 10 of first binding partners in Docking Benchmark. 84

Figure D.18. Distributions and probability density functions of bound and unbound collectivities for ANM mode 10 of first second partners in Docking Benchmark. 84

LIST OF TABLES

Table 4.1. Number of models in CAPRI dataset. 20

Table 4.2. Probabilities of finding acceptable models with filtering by occurrences of i-hinges. 21

Table 4.3. Probabilities of finding acceptable models after filtering by ΔMSF. 29

Table 4.4. Numbers of dynamic clusters. 38

Table 4.5. Statistics of clusters which hold the highest number of acceptable models. 39

Table 4.6. Number of models before and after filtering by clusters and i-hinge occurrences. 41

Table 4.7. Number of models before and after filtering by clusters and ΔMSF. 42

Table 4.8. Comparison of ranking algorithm performance with random picking. 48

Table A.1. Clusters of Target 30 models. 53

Table A.2. Clusters of Target 32 models. 54

Table A.3. Clusters of Target 35 models. 55

Table A.4. Clusters of Target 36 models. 56

Table A.5. Clusters of Target 37 models. 57

Table A.6. Clusters of Target 40_CA models. 58

Table A.7. Clusters of Target 40_CB models. 59

Table A.8. Clusters of Target 41 models. 60

Table A.9. Clusters of Target 42_AB models. 60

Table A.10. Clusters of Target 42_BC models. 61

Table A.11. Clusters of Target 46 models. 62

Table A.12. Clusters of Target 47 models. 63

Table A.13. Clusters of Target 48 models. 63

Table A.14. Clusters of Target 49 models. 64

Table A.15. Clusters of Target 50 models. 64

Table A.16. Clusters of Target 51_1 models. 65

Table A.17. Clusters of Target 51_2 models. 65

Table A.18. Clusters of Target 53 models. 66

Table A.19. Clusters of Target 59 models. 67

LIST OF SYMBOLS

A	Helmholtz free energy
C_B	Between sets variance
C_corr	Eigenvector correlation matrix
C_k	Cross-correlation matrix
C_T	Total variance
C_W	Within set variance
D_{ij}	Matrix norm of difference between matrices i and j
E	Total energy
ΔE	Change in total energy
G	Gibbs free energy
ΔG	Gibbs free energy change
$g(r)$	Radial distribution function
$\mathbf{H}$	Hessian matrix
K	Degree of collectivity
k	Boltzmann constant
K_d	Dissociation constant
r_i	Coordinates of ith residue
$\bar{r}$	Mean coordinates of all alpha-carbon atoms
R_c	Cutoff radius
$\mathbf{R}_i$	Position vector of alpha-carbon of ith residue
$\mathbf{\Delta R_i}$	Displacement vector of alpha-carbon of ith residue
q	Atomic charge
T	Temperature
$\mathbf{U}$	Eigenvector matrix
U	Internal energy
u_k	kth eigenvector
V	ANM potential energy
$Z(T)$	Partition function

ϵ	Dielectric constant
$\mathbf{\Gamma}$	Kircchoff matrix
γ	Spring constant of Hookean pairwise potential
$\mathbf{\Lambda}$	Eigenvalue matrix
λ_k	kth eigenvalue
$\rho(i)$	Probability density function

LIST OF ACRONYMS/ABBREVIATIONS

ANM	Anisotropic Network Model
CAPRI	Critical Assessment of Predicted Interactions
DNA	Deoxyribonucleic Acid
ENM	Elastic Network Model
FFT	Fast Fourier Transform
GNM	Gaussian Network Model
LDA	Linear Discriminant Analysis
MD	Molecular Dynamics
NMR	Nuclear Magnetic Resonance
PDB	Protein Data Bank
RMSD	Root Mean Square Deviation
RNA	Ribonucleic Acid

1. INTRODUCTION

Proteins are essential for life on earth. Biological functions and processes are possible through proteins interacting with each other and with other molecules. Although there are numerous experimental methods to understand the binding phenomena of proteins and their interaction states [1] in addition to available experimentally determined protein complex structures [2], protein-protein interactions still remain to be a mystery.

Due to various handicaps in experimental procedures such as high cost and technical difficulties, the ability to determine protein complex structures by experimental methods is still limited. As an alternative, computational tools such as docking are trending for using in drug discovery research where protein-protein interactions are targeted [3]. Through docking algorithms, probable set of structures are produced by using various computational approaches. Last part of a docking study involves scoring. The objective of scoring algorithms is to rank the model structures in order of similarity to native structure. A scoring function that performs ideally would rank the structure that is the most similar to native structure the highest [4].

There are numerous scoring functions developed by computational structural biology research groups which are tested by a community-wide experiment called CAPRI (Critical Assessment of Predicted Interactions) [5]. CAPRI experiment has two phases; docking and scoring. Predictors who are given the unbound conformations of binding partners, dock the unbound binding partners and upload up to 100 complexes for scorers to rank. Scorers rank these results and submit their top ten structures. Basically, scorers' success depend on how successful the predictions are.

The aim of this study is to develop a unique scoring algorithm based on the dynamics of protein complexes using two distinct elastic network models, namely Gaussian and Anisotropic Network Models. Earlier work has shown that integrating ANM along with collectivity in a scoring algorithm is proven to be successful in distinguish-

ing the native from the design complex structures [6]. Here, a more extensive analysis based on the dynamics by predictions of both GNM and ANM will be performed. Clustering analysis is also utilized in order to pave the way for a more successful ranking scheme.

2. LITERATURE SURVEY

2.1. Protein-Protein Interactions

One of the most crucial aspects of life is protein-protein interactions. Starting from the beginning of life, cellular processes evolved in a way that functions through protein interactions. Central dogma of biology revolves around protein interactions. By interacting, proteins perform numerous tasks inside the cell. They transfer information from DNA to RNA, carry molecules, catalyse reactions, regulate the channels on cell membrane and many more biological processes are made possible through interacting proteins. Therefore, importance of understanding of protein-protein interactions is significant for comprehension of biological processes [7].

Interaction of the proteins happens through the interface, from the surfaces of the two proteins by various specificity and affinity. Interfaces have rather large surface areas (700-1500 $Å^2$ per protein) and binding surfaces have a flat topology [8]. These interactions can be transient or obligatory. Proteins that are in interacting state are named as "protein complex". A protein interacts with another complementary protein when topological and physiochemical conditions are met. These conditions are promoted by hydrophobic effect, hydrogen bonds, salt bridges, van der Waals interactions and the flexibility of the surface residues. In addition, these conditions determine the stability of the complex [9]. While interacting, atoms of two binding partners that are located at the interface come into contact. Accessible surface area, which is the surface area that is reachable by solvent molecules, decreases on these territory.

Although many aspects that govern the protein-protein interactions are known, they are still not able to provide enough insight to predict these interactions. Detailed physiochemical information of protein surfaces are available but they are not sufficient for detecting an association with high probability [7].

Information of three dimensional structures are proven to be useful for under-

standing protein interactions. Advancements in X-ray crystallography and other imaging techniques such as solution NMR, let us search for near-atomic resolution information of protein complex structures. There are numerous high resolution structures of protein complexes available in Protein Data Bank [2]. However the experiments for predicting these interactions are quite cumbersome to perform [1]. Therefore, interest to computational approaches such as protein-protein docking processes that predict the three dimensional structure of protein-protein complexes is raising significantly. However, they also require a deep understanding of physical and biological determinants of binding phenomenon [10].

2.2. Protein-Protein Docking

The story of protein-protein docking starts in the late 1970s with leading work by Janin and Wodak [11]. Janin and Wodak's algorithm uses a coarse grain model where each residue is represented as a sphere. It creates orientations of two binding partners relative to each other and creates a contact between them with translational and rotational displacement. It was the first example of rigid-body docking as no flexibility was involved for binding partners [12]. Since then there have been many developments in docking algorithms along with increasing computational resources.

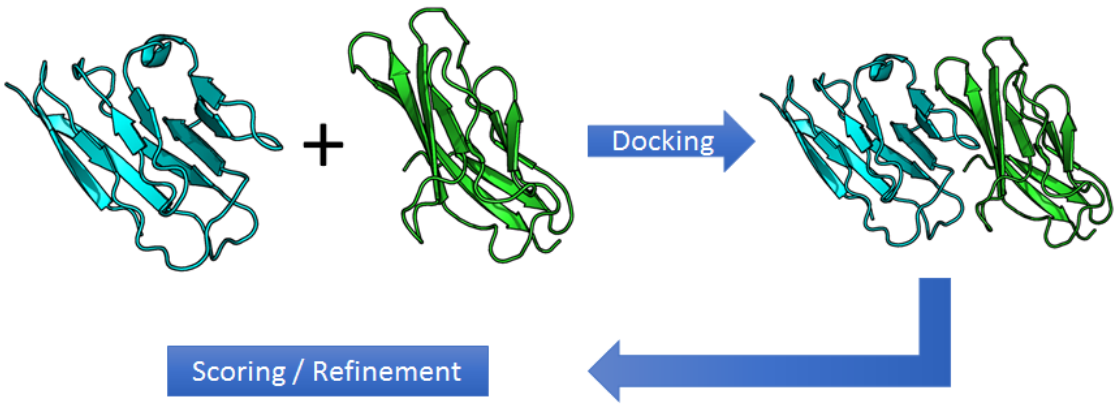

Figure 2.1. Protein-protein docking process.

Docking exercises involve two key parts (as shown in Figure 2.1): The search

algorithm that samples configurational and conformational degrees of freedom and the scoring function that ranks the solutions [13,14]. These parts do not have to take place subsequently. It is possible to use the scoring function during or at different stages of docking process [3]. Accuracy of docking predictions largely depends on the scoring model's ability to distinguish between native and non-native arrangements of atoms [1].

Conformational and configurational search methods can be classified into three groups in general. These are exhaustive global search, randomized search and local shape feature matching methods [3]. Exhaustive global search methods involve search for binding orientations which include six degrees of freedom; three dimensional translational in addition to three dimensional rotational degrees of freedom [15]. As this exhaustive search is computationally expensive, FFT-based (Fast Fourier Transform) algorithms are commonly used [3, 16]. Some of well known docking software that use exhaustive global search algorithms are ZDOCK [17], FTDOCK [18], DOT [19], GRAMM [20], MolFit [16], PIPER [21], F2DOCK [22]. In randomized search algorithms, one of the binding partners is fixed and the other protein is placed around the binding site (or all around the static protein) randomly. Randomized search algorithms do not exhaustively search the complete six dimensional space. RosettaDock [23], ATTRACT [24] and HADDOCK [25] are among the docking software which use randomized search algorithms. Local shape feature matching methods represent protein as molecular shapes. These molecular shapes are obtained by calculating solvent accessible surface. Local shape matching algorithms look for a local shape complementarity between two binding partners. Transformation matrices in six dimensions are only calculated if a shape complementarity is found between two proteins [3]. Some examples of docking software that use shape matching algorithms are DOCK [26], PatchDOCK [27], LZerD [28] and GAPDOCK [29].

For the last years, considerable amount of effort has been put forward into the development of docking algorithms. The success of docking algorithms has consistently improved as measured by CAPRI (Critical Assessment of Predicted Interactions) blind docking experiments [30]. CAPRI and also other relatively large scale docking studies represent very important steps towards understanding the capabilities and limitations

of current protein-protein docking methods [31].

2.3. Scoring Functions

Although scoring might be involved during the docking process, scoring and docking do not always share the same algorithm. Protein-protein scoring functions are similar to protein-ligand docking scoring functions. In general, scoring methods can be categorized into three groups; empirical, knowledge based and force-field based. A scoring method can use combinations of these methods as well [4].

2.3.1. Empirical Scoring Functions

Empirical scoring functions predict the binding affinity of a complex based on sum of weighted energy terms [4].

$$\Delta G = \sum_i w_i \Delta G_i \tag{2.1}$$

An example for empirical scoring function is used by HADDOCK group is presented in Equation 2.2.

$$-\log K_d = (w_a * E_{VDW} + w_b * E_{Elec} + w_c * E_{Hb} + w_d * \Delta G_{Desolv} + w_e * BSA) * F \tag{2.2}$$

Here w_b, w_a, w_c denote different weights for each contribution of the terms E_{Elec}, E_{VDW}, E_{Hb} which state the energetic contributions of intermolecular electrostatic, van der Waals and hydrogen bond interactions respectively. w_e is the weight of BSA which is the buried surface area in Å^2. ΔG_{Desolv} is the Lazardis-Karplus solvation term and w_d states its contribution weight. Finally, F is a scaling factor. Resulting value from this equation $-\log K_d$ is the theoretical dissociation constant, which can be used for calculating Gibbs free energy of binding.

$$\Delta G_r^0 = 2.303RT(-\ln K_d) \tag{2.3}$$

Therefore, it is possible to predict if a model complex structure has the native-like state through this theoretical dissociation constant [6, 32].

2.3.2. Knowledge Based Scoring Functions

Knowledge based scoring functions are created by using energy potentials that are derived using the structural information from 3D structures of proteins [4]. The idea of knowledge based potentials is derived from Boltzmann distribution.

$$\frac{n(i)}{N} = \rho(i) = \frac{e^{-E(i)/kT}}{Z(T)} \tag{2.4}$$

Here $n(i)$ is number of particles in an ensemble of states i with energy of $E(i)$, k is the Boltzmann constant, N is the total number of particles in the system, T is the absolute temperature. $Z(T)$ is the partition function and $\rho(i)$ is the probability density function. This equation is valid for a thermodynamic system with fixed temperature, volume and number of particles. It is possible to rearrange it as shown in Equation 2.5.

$$E(i) = -kT \ln \rho(i) - kT \ln Z(T) \tag{2.5}$$

For a given state it is possible to calculate the energy difference based on a reference state as stated in Equation 2.6.

$$\Delta E(i) = E(i) - E_{ref} = -kT \ln \rho(i) + kT \ln \rho_{ref} = -kT \ln \left(\frac{\rho(i)}{\rho_{ref}} \right) \tag{2.6}$$

Helmholtz free energy A_{12} of two particles 1 and 2 located in a homogeneous liquid solvent can be computed with Equation 2.7.

$$A_{12} = -kT \ln g(r) = U_{12} + \Delta G_{12} \tag{2.7}$$

where radial distribution function $g(r)$ is calculated as shown in Equation 2.8.

$$g(r) = \frac{\rho_{12}(r)}{\rho_{ref}(r)} = \frac{P_{12}(r)}{P_{ref}(r)} \tag{2.8}$$

Due to their interactions with each other (U_{12} term) and the surroundings (G_{12} term) summation of both interactions correspond to Helmholtz free energy (A_{12} term), which is also named as potential of mean force [33].

Statistical potentials are not energies, therefore it is possible to drop kT terms from Equation 2.6 in order to derive a scoring function.

$$score(i) = -\ln\left(\frac{\rho(i)}{\rho_{ref}}\right) \tag{2.9}$$

Equation 2.9 forms the basis of knowledge based scoring functions [33].

2.3.3. Force-Field Based Scoring Functions

As the name states force-field based scoring functions basically make use of atomic interactions such as van der Waals and electrostatic interactions in addition to stretching, bending and torsional forces that act on bonds between atoms. Parameters of these scoring functions are obtained either from experimental data or quantum mechanical calculations. An example for force-field based scoring function is called DOCK [4, 34].

$$E = \sum_i \sum_j \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} + \frac{q_i q_j}{\epsilon(r_{ij}) r_{ij}}\right) \tag{2.10}$$

Terms that are summed up in Equation 2.10 involve Lennard-Jones coefficients for van der Waals forces and atomic charges for electrostatic forces. Here r_{ij} is the distance between atoms i and j (of receptor and ligand). A_{ij} and B_{ij} are van der Waals parameters, q_i and q_j represent the atomic charges, $\epsilon(r_{ij})$ is a dielectric constant for solvent effect. These energy terms are obtained from Amber [4, 34].

3. MATERIALS AND METHODS

3.1. Gaussian Network Model (GNM)

The Gaussian Network Mode (GNM) [35] assumes the folded state of protein equivalent to a three dimensional elastic network. The nodes are identified with alpha-carbon atoms in the protein. Edges are defined by elastic springs that are situated between these atoms which are separated by a certain cutoff distance as shown in Figure 3.1. These atoms experience Gaussian distributed fluctuations where the model

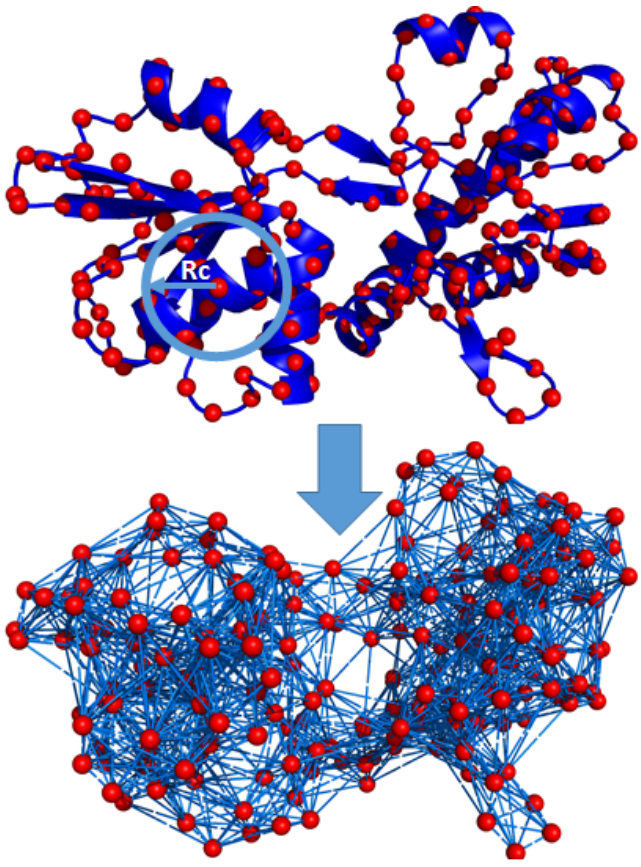

Figure 3.1. Representation of elastic network model. Red spheres show alpha-carbon atoms and R_c states the cutoff radius.

takes its name from. Equilibrium correlation between fluctuations $\Delta\mathbf{R}_i$ and $\Delta\mathbf{R}_j$ of

two alpha-carbons i and j is given by Equation 3.1.

$$\langle \Delta \mathbf{R}_i \cdot \Delta \mathbf{R}_j \rangle = \frac{3k_bT}{\gamma}[\mathbf{\Gamma}^{-1}]_{ij} \tag{3.1}$$

Here $\mathbf{\Gamma}$ is a symmetric matrix known as Kirchhoff or connectivity matrix, the subscript ij identifies the ijth element, $\mathbf{R}_i$ is the position vector of the ith alpha-carbon, k_b is the Boltzmann constant T is the absolute temperature and γ is the single parameter (force constant) of the Hookean pairwise potential for representing the interactions between the residues in the folded structure. The elements of $\mathbf{\Gamma}$ are given by

$$\mathbf{\Gamma}_{ij} = \begin{cases} -1 & \rightarrow i \neq j \ \wedge \ \mathbf{R}_{ij} \leq \ r_c, \\ 0 & \rightarrow i \neq j \ \wedge \ \mathbf{R}_{ij} > r_c, \\ \sum_{i,i\neq j}^{N} \mathbf{\Gamma}_{ij} & \rightarrow i = j. \end{cases} \tag{3.2}$$

where r_c is the cutoff distance where alpha-carbons are assumed to be interacting. $\mathbf{R}_{ij}$ is the distance between ith and jth alpha-carbons and N is the number of residues. It is possible to decompose $\mathbf{\Gamma}$ matrix into its eigenvalues and eigenvectors in order to take the inverse of $\mathbf{\Gamma}$.

$$\mathbf{\Gamma} = \mathbf{U}(\mathbf{\Lambda^{-1}})\mathbf{U}^T \tag{3.3}$$

Here $\mathbf{U}$ is the eigenvector matrix whereas $\mathbf{\Lambda}$ is the diagonal eigenvalue matrix of $\mathbf{\Gamma}$. For a given structure with N residues, it is possible to decompose $\mathbf{\Gamma}^{-1}$ into sum of $N-1$ modes. Nth eigenvalue is almost equal to zero, therefore it is not considered in the summation.

$$\mathbf{\Gamma}^{-1} = \mathbf{\Gamma}_1^{-1} + \mathbf{\Gamma}_2^{-1} + \cdots + \mathbf{\Gamma}_{N-1}^{-1} = \sum_{k=1}^{N-1} \lambda_k^{-1}\mathbf{u}_k\mathbf{u}_k^T \tag{3.4}$$

Here, $\mathbf{u}_k$ is defined as the eigenvector of kth matrix and λ_k is the corresponding eigenvalue. Diagonal of Γ^{-1} results in $\langle R_i^2 \rangle$ (mean square fluctuations) of residues in the structure.

Modes that correspond to lower eigenvalues are named as "slow modes" because of the low frequencies that result in slow motions.

3.1.1. Hinge Residues

Each term that is summed up in Equation 3.4 ($\mathbf{\Gamma}_k$) is a matrix that constitutes of $\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \rangle_k$ in off-diagonal elements and $\langle \Delta\mathbf{R}_i^2 \rangle_k$ in diagonal elements where k is the mode number. Both $\mathbf{\Gamma}_k$ and Γ are covariance matrices as they represent how fluctuations of one residue relates to the other. For a given correlation matrix Γ_k^{-1}, the elements of correlation matrix C_k is calculated as shown in Equation 3.5.

$$C_{k_{ij}} = \frac{\langle \Delta\mathbf{R}_i \cdot \Delta\mathbf{R}_j \rangle_k}{\sqrt{\langle \Delta\mathbf{R}_i^2 \rangle_k \langle \Delta\mathbf{R}_j^2 \rangle_k}} = \frac{\Gamma^{-1}_{k_{ij}}}{\sqrt{\Gamma^{-1}_{k_{ii}} \Gamma^{-1}_{k_{jj}}}} \tag{3.5}$$

When only one mode is considered, elements of C_k are either -1 or +1, which means that fluctuations of residues are either relatively correlated or not correlated. These correlations between the fluctuations define positively and negatively correlated regions on the structure. Along the structure, "hinge residues" are defined as the residues that are located in the border between positively and negatively correlated regions for a given slow mode [36].

In scope of this study, interface residues are defined as residues that are located in different chains which have less than 4.0 Å distance between each other. The hinges that are located at the interface of a protein complex (i-hinges) are hinge residues that have at least one atom in the interface.

3.1.2. Hinge Planes

Slowest GNM modes tend to divide the structures into two regions, namely positively and negatively correlated regions. These two sets of residues are obtained by the correlation matrix of the first slowest GNM mode. Hinge residues, which are located between these regions separate the structure into two regions as if a plane goes

through between them. The method for obtaining the normal of this plane (LDA, Linear Discriminant Analysis) is explained below.

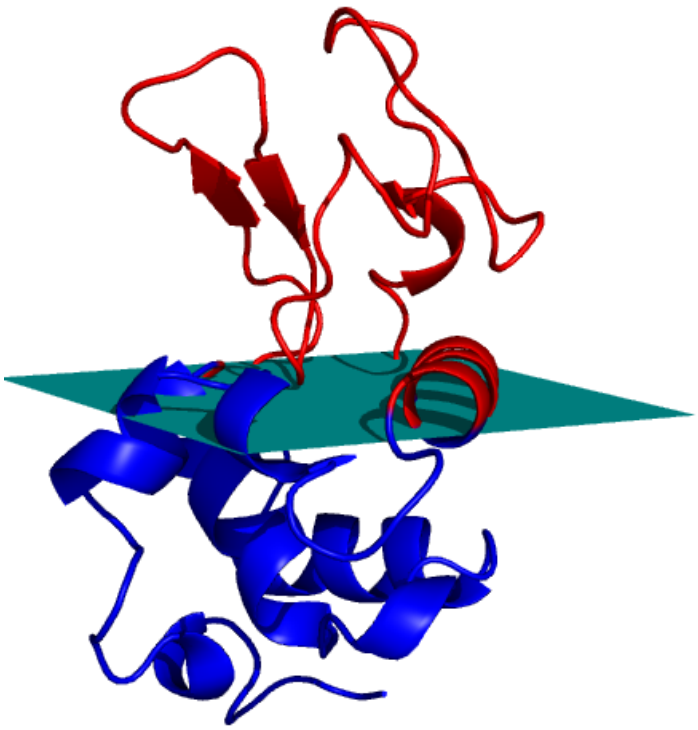

Figure 3.2. Illustration of hinge plane for lysozime (PDB code 1BVK, chain F).

$$C_B = \sum_c N_c (r_c - \bar{r})(r_c - \bar{r})^T \tag{3.6}$$

$$C_W = \sum_c \sum_{i \in c} (r_i - \bar{r_c})(r_i - \bar{r_c})^T \tag{3.7}$$

Equation 3.6 gives the between sets variance C_B and Equation 3.7 gives within the set variance C_W where sets are groups of positively and negatively correlated residues respectively. c states the set type, N_c is the number of alpha-carbons in given set, $\bar{r}$ is the mean coordinates of all alpha-carbons, $\bar{r_c}$ is the mean coordinates of alpha-carbons for given set, r_c is array of coordinates for all atoms in given set and r_i is the coordinates

of ith residue. Total variance is given as shown in Equation 3.8.

$$C_T = \sum_i (r_i - \bar{r})(r_i - \bar{r})^T \tag{3.8}$$

This shows that total variance is sum of between sets and within set variances respectively ($C_T = C_B + C_W$). The purpose is to find a scaling factor $\mathbf{n}$, which can make $\mathbf{n}^T C_W \mathbf{n} = 1$ or minimize $-\frac{1}{2}(\mathbf{n}^T C_B \mathbf{n})$. This is a constrained optimization problem and its Lagrangian L can be built as shown in Equation 3.9 [37].

$$L(\mathbf{n}) = -\frac{1}{2}\mathbf{n}^T C_B \mathbf{n} + \frac{1}{2}\lambda(\mathbf{n}^T C_W \mathbf{n} - 1) \tag{3.9}$$

First derivative of L disappears at extreme points, then Equation 3.9 becomes

$$C_W^{-1} C_B \mathbf{n} = \lambda \mathbf{n} \tag{3.10}$$

which cannot be solved eigenvalue decomposition since $C_W^{-1} C_B$ is not a positive definite matrix. In order to make it positive definite, C_B can be decomposed and scaled as shown in Equation 3.11.

$$C_B^{-1/2} = W \Lambda^{\frac{1}{2}} W^T \tag{3.11}$$

Let $m = C_B^{\frac{1}{2}} \mathbf{n}$ so that Equation 3.10 can be rewritten as

$$C_B^{\frac{1}{2}} C_W^{-1} C_B^{\frac{1}{2}} m = \lambda m \tag{3.12}$$

Now that $C_B^{\frac{1}{2}} C_W^{-1} C_B^{\frac{1}{2}}$ is a positive definite matrix λ and m can be solved via eigenvalue decomposition of Equation 3.12. Vector that is to be found ($\mathbf{n}_1$) is the one that gives the largest between-category (i.e. between C_B and C_W) variance. It is also the normal of the hinge plane that we want to obtain.

$$\mathbf{n}_1 = C_B^{-\frac{1}{2}} \mathbf{v}_1 \tag{3.13}$$

Here, $\mathbf{v}_1$ is the eigenvector which corresponds to the largest eigenvalue obtained from eigenvalue decomposition [37]. Using the normal, it is possible to obtain the hinge plane as illustrated in Figure 3.2. On the structure, positively correlated residues are showed in red and negatively correlated residues are shown in blue.

3.2. Anisotropic Network Model (ANM)

In GNM all fluctuations are implicitly assumed to be isotropic. Therefore, no directional information of motion is considered. However, in reality the fluctuations are mostly anisotropic and it is important to determine the directions of collective motions, as these can be directly related to biological function and mechanisms. ANM, which is an extension of GNM is derived to overcome this issue [38].

ANM potential energy for the system of N nodes is the summation of all harmonic interactions in the structure:

$$V = \frac{\gamma}{2} \sum_{j|j \neq i} (\mathbf{\Gamma}_{ij})(R_{ij} - R_{ij}^0)^2 \tag{3.14}$$

R_{ij} and R_{ij}^0 are instantaneous and equilibrium distances between nodes i and j, $\mathbf{\Gamma}_{ij}$ is the ijth element of Kirchhoff matrix $\mathbf{\Gamma}$ which is obtained from GNM (see Equation 3.2). Second derivatives of V with respect to residue positions result in $3N \times 3N$ shaped Hessian matrix $\mathbf{H}$ whose 3×3 shaped super elements are defined as:

$$\mathbf{H}_{ij} = \begin{cases} \frac{\gamma \mathbf{\Gamma}_{ij}}{(R_{ij}^0)^2} \begin{bmatrix} X_{ij}X_{ij} & X_{ij}Y_{ij} & X_{ij}Z_{ij} \\ Y_{ij}X_{ij} & Y_{ij}Y_{ij} & Y_{ij}Z_{ij} \\ Z_{ij}X_{ij} & Z_{ij}Y_{ij} & Z_{ij}Z_{ij} \end{bmatrix} & \rightarrow i \neq j \\ -\sum_{j|j \neq i} \mathbf{H}_{ij} & \rightarrow i = j \end{cases} \tag{3.15}$$

Here, X_{ij}, Y_{ij} and Z_{ij} are elements of R_{ij}^0 vector. Similarly to Kircchoff matrix, Hessian matrix is also not invertible and therefore a pseudo-inverse is constructed by decomposing it by eigenvalue decomposition. From resulting $3N-6$ non-zero eigenvalues and

eigenvectors, pseudo-inverse is created as shown in Equation 3.16, similar to Equation 3.4 [39].

$$\mathbf{H}^{-1} = \sum_{i=1}^{3N-6} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T \tag{3.16}$$

Here $\mathbf{u}_i$ is the eigenvector of ith mode and λ_i is the corresponding eigenvalue.

3.2.1. Eigenvector Correlations

ANM eigenvectors provide information about fluctuation directions of residues for a given mode. Therefore, for a given set of modes it is possible to compute cross-correlations of eigenvectors between two distinct states of a structure (i.e. bound and unbound states of a chain in a complex).

Let a protein complex with chains A and B has $M + N$ number of residues in total where number of residues in chain A is equal to M and number of residues in chain B is equal to N. After creating Hessian matrices for chain A, chain B and both chains together (complex), following their eigenvalue decomposition up to 10 slowest modes eigenvector matrices are obtained. Let $\mathbf{U}_A$, $\mathbf{U}_B$ and $\mathbf{U}_{AB}$ be eigenvector matrices of chain A, chain B and complex which have shapes of $3M \times 10$, $3N \times 10$ and $3(M + N) \times 10$ respectively. Bearing the chain sequence of the complex in mind, $\mathbf{U}_A$ and $\mathbf{U}_B$ are concatenated resulting in $\mathbf{U}_{A \cup B}$ whose shape is $3(M + N) \times 10$. $\mathbf{U}_{A \cup B}$ is normalized column-wise in order to obtain normalized eigenvectors.

$$\mathbf{C}_{corr} = |\mathbf{U}_{A \cup B}^T \cdot \mathbf{U}_{AB}| \tag{3.17}$$

$\mathbf{C}_{corr}$ is a 10×10 eigenvector correlation matrix which has elements that hold values between 0 and 1. An example heat map visualization of an eigenvector correlation matrix is provided in Figure 3.3. As fluctuations might occur in exact opposite directions of obtained eigenvectors, orientation is more essential. Thus, taking absolute values of dot products is beneficial for understanding the relative orientations of eigenvectors.

When eigenvector correlation matrix of only one chain is of interest (i.e. chain A), part of the eigenvector matrix that belongs to that chain is separated from $\mathbf{U}_{AB}$ and resulting matrix is multiplied with $\mathbf{U}_A$ similarly to Equation 3.17.

Using such a correlation matrix as shown in Equation 3.17 has a drawback. It is not possible to consider combinations of modes from each binding partner. Both binding partners are calculated in same mode. If these combinations are of interest, it is possible to extend the dimensions of correlation matrix to 10×20. Simply by separating the relevant part from eigenvectors of complex two 10×10 matrices can be calculated as mentioned above. By appending these two matrices horizontally, it is possible to create a matrix as shown in Figure 3.4.

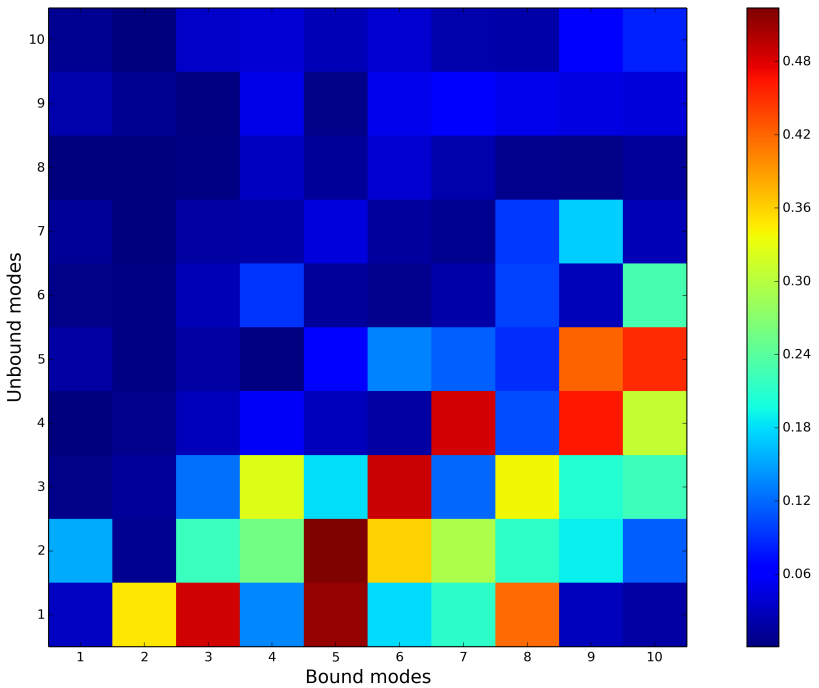

Figure 3.3. Eigenvector correlations (10×10 matrix) for human tissue factor/inhibitor complex (PDB code 1AHW, chains AB and C).

For a given set of models, in order to group incorrect models and native-like models in different groups, clustering analysis is done. For this purpose, k-means clustering algorithm is used. Distance metric for clustering is chosen as the distances

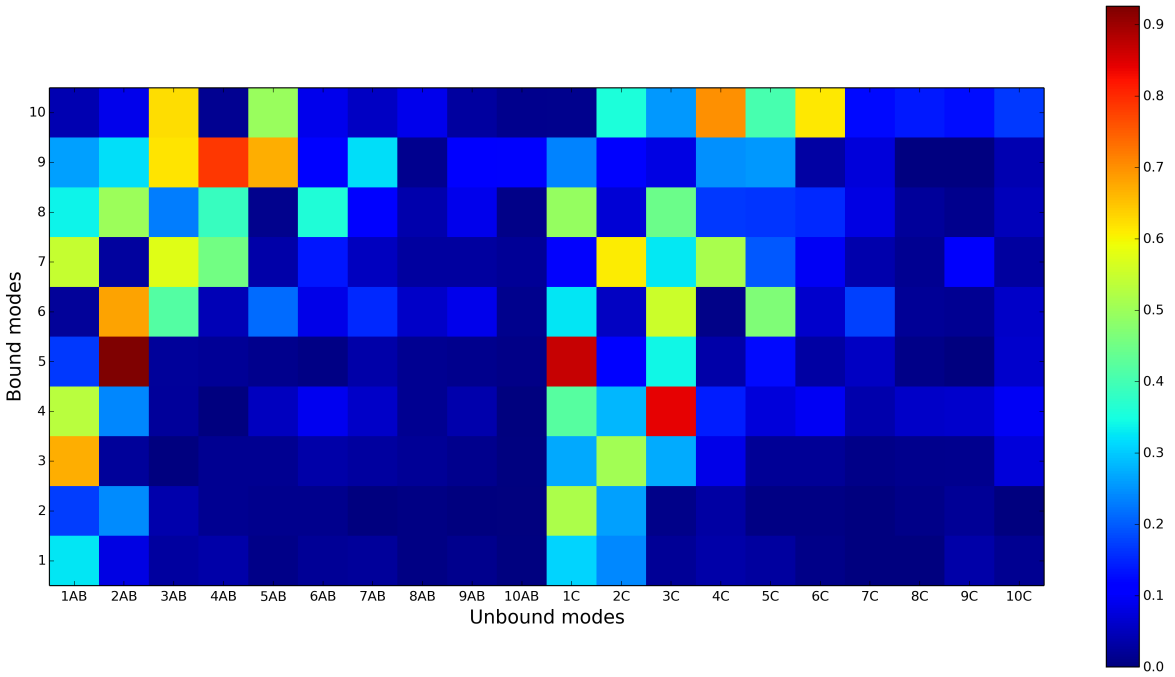

Figure 3.4. Eigenvector correlations (10×20 matrix) for human tissue factor/inhibitor complex (PDB code 1AHW, chains AB and C).

between eigenvector correlation matrices. It is computed as:

$$D_{ij} = ||\mathbf{C}_i - \mathbf{C}_j|| \tag{3.18}$$

Here, D_{ij} is the matrix norm of difference between eigenvector correlation matrix of ith model ($\mathbf{C}_i$) and jth model ($\mathbf{C}_j$) in the set. For a model set with N number of models, $N \times N$ shaped pairwise distance matrix is created. Pairwise distance matrix constitutes the input for k-means clustering algorithm.

3.3. Information Entropy and Collectivity

Idea of collectivity was firstly proposed by Brüschweiler [40, 41]. Degree of collectivity is a measure of how collectively a protein structure moves. It is proportional with exponential of information entropy, which is the vector that includes displacement vectors of all nodes. When ENM is used to determine the motion, this vector

corresponds to eigenvector of a certain mode of interest.

$$\kappa_k = \frac{1}{N} \exp\left(-\sum_i^N \alpha \Delta \mathbf{R}_i^2 \ln\left(\alpha \Delta \mathbf{R}_i^2\right)\right)_k \tag{3.19}$$

Here $\Delta \mathbf{R}_i^2$ is the square of elements of kth eigenvector. α is a normalization factor defined as:

$$\alpha = \frac{1}{\sum_i^N \Delta \mathbf{R}_i^2} \tag{3.20}$$

When structure demonstrates the most collective motion, all elements of the information entropy vector are expected to be identical. In case of the most collective motion κ approaches to 1, whereas in case of extreme local motion it approaches to $\frac{1}{N}$ as this can be seen from Equation 3.19 [6].

Within the scope of this study, degrees of collectivities are calculated for bound and unbound states. For both ANM and GNM eigenvectors are computed for both bound and unbound (isolated from complex) chains. In order to compare degrees of collectivities between bound and unbound states for each chain, eigenvectors in bound state are divided into two eigenvectors (one for each chain). Upon normalizing these eigenvectors, degrees of collectivities are obtained for bound states for each chain.

4. RESULTS

4.1. Importance of Hinge Residues Positioned at the Interface

By taking isolated binding partners (co-crystals) into account, number of i-hinge residues are obtained by GNM calculations. Among Docking Benchmark 4.0 [42] complexes (total of 176 structures), 82% contains at least one hinge residue at the interface in either one or both of the binding partners in the slowest mode. This ratio goes up to 95% when only second slowest mode is taken into account. When both slowest and second slowest modes are considered, same ratio goes as high as 97%. These ratios are valid for occurrence of at least one hinge residue at the interface from either one of the binding partners. This explains that occurrence of hinge residues at the interface for first two slowest modes is significant. If both binding partners should have i-hinges in either one of the first two modes then such complexes constitute 75% of the dataset. Such ratios are helpful to select probable incorrect models and disqualify them from target datasets.

In order to test this finding, number of hinges at the interface are calculated for all models in 19 targets from former CAPRI rounds [43], number of models are tabulated in Table 4.1. Results are tabulated in Table 4.2. In order to achieve a simple ranking system, medium, acceptable and high quality models are classified under the name of "acceptable" models along the study.

As results state in Table 4.2, among 19 targets if we filter the models by occurrences of hinge residues in the interface, our chances increase insignificantly. There is only one case that this filter eliminates acceptable models (Target 42, AB). With this approach, it is possible to eliminate 2% to 15% of the incorrect models depending on the number of incorrect and acceptable models for a given target. This method is a straightforward and fast way to decrease the amount of incorrect models in the first step. However, it could only be used for first tier filtering as vast amount of incorrect models still remain as candidates.

Table 4.1. Number of models in CAPRI dataset.

Targets	**Number of models**	**Acceptable**	**Incorrect**
T30	350	2	348
T32	354	34	320
T35	339	1	338
T36	319	1	318
T37	390	11	379
T40_CA	368	102	266
T40_CB	368	75	293
T41	327	153	174
T42_AB	278	15	263
T42_BC	278	5	273
T46	397	13	384
T47	277	217	60
T48	158	21	137
T49	158	12	146
T50	357	52	305
T51_1	287	3	284
T51_2	274	1	273
T53	384	46	338
T59	339	23	316

Table 4.2. Probabilities of finding acceptable models with filtering by occurrences of i-hinges.

Targets	**Acceptable***	**Incorrect****	**Probability**†	**Probability**‡
T30	0	22	0.006	0.006
T32	0	4	0.096	0.097
T35	0	5	0.003	0.003
T36	0	5	0.003	0.003
T37	0	11	0.028	0.029
T40_CA	0	2	0.277	0.279
T40_CB	0	2	0.204	0.205
T41	0	7	0.468	0.478
T42_AB	3	36	0.054	0.050
T42_BC	0	39	0.018	0.021
T46	0	8	0.033	0.033
T47	0	1	0.783	0.786
T48	0	2	0.133	0.135
T49	0	0	0.076	0.076
T50	0	7	0.146	0.149
T51_1	0	40	0.010	0.012
T51_2	0	47	0.004	0.004
T53	0	6	0.120	0.122
T59	0	2	0.068	0.068

* Number of acceptable models with no i-hinges in 1st and 2nd modes.

** Number of incorrect models with no i-hinges in 1st and 2nd modes.

† Probability of finding an acceptable model before filtering.

‡ Probability of finding an acceptable model after eliminating the models without i-hinges in 1st and 2nd modes.

4.2. Angles between Hinge Planes

Hinges in the slowest mode tend to divide the structure into two parts as if a plane goes through the structure. Through this approach, planes for each binding partners are drawn in order to see how these planes are positioned against each other (see Section 3.1.2). An example of these planes for humanized anti-lysozyme FV complex with lysozyme (PDB code 1BVK) is presented in Figure 4.1. Here, red and blue parts show positively and negatively correlated regions of receptor (chains D and E), orange and marine blue parts show negatively and positively correlated regions of ligand (chain F).

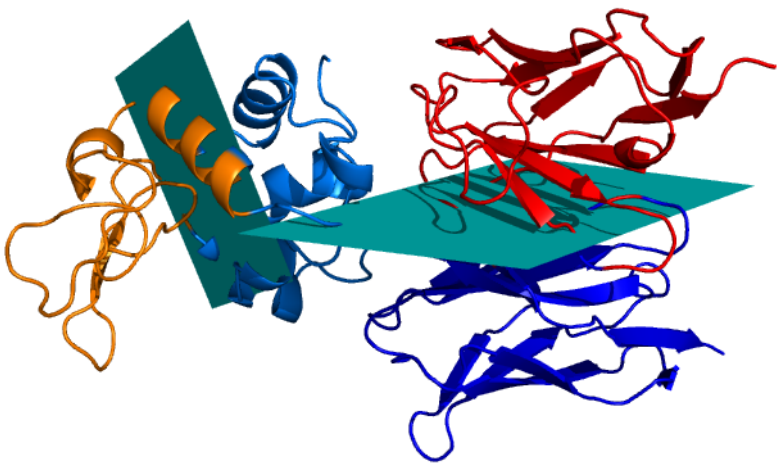

Figure 4.1. Hinge planes of humanized anti-lysozyme FV complex with lysozyme (PDB code 1BVK).

These planes are presumed to be oriented in such a manner that allows only certain ranges of angles to be present in between. In order to determine what range of plane angles are probable, the plane angles of each complex in Docking Benchmark [15] are calculated and a distribution is obtained. This distribution shows that preferred hinge plane angles are between 30 to 90 degrees (see Figure 4.2). For drawing hinge planes, a software is developed using Python programming language. This software (named as "HiPlane") creates the PyMOL script for coloring dynamic domains and

drawing the hinge planes by using LDA.

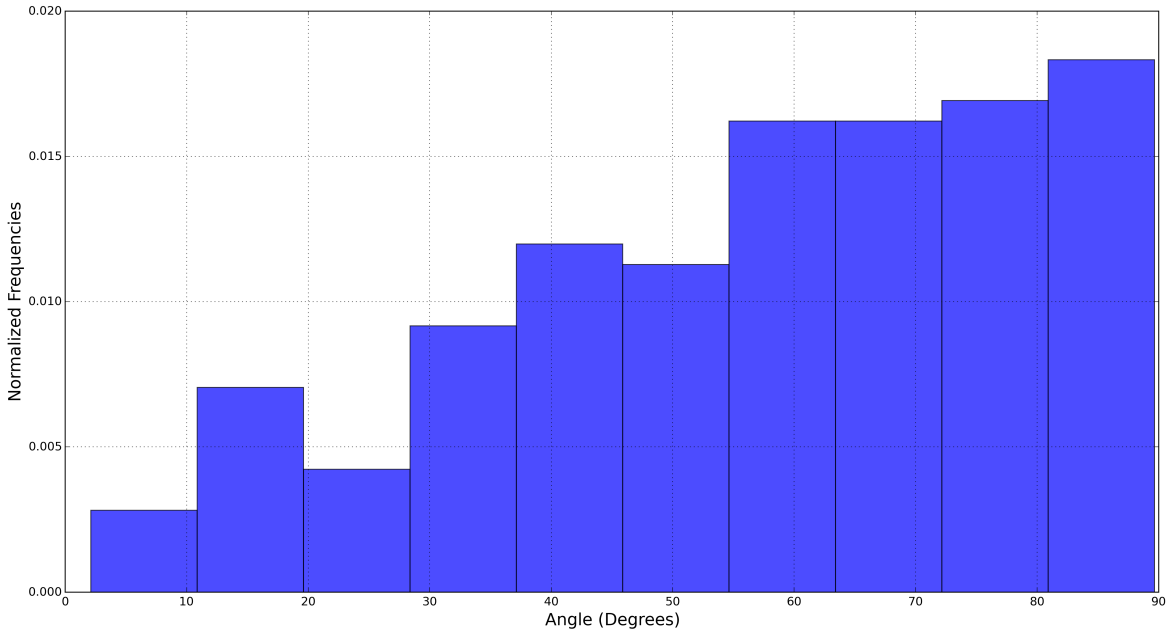

Figure 4.2. Distribution of hinge plane angles among Docking Benchmark complexes.

Though there is a distinct distribution of angles between hinge planes, information is not sufficient enough to distinguish incorrect models from acceptable ones only by looking at these angles. On the other hand, it gives us a clue about the preferred ways of orientation of hinge residues when two proteins bind. Figure 4.3 presents the histograms of hinge angles for every model of each CAPRI target. Similar to Figure 4.2, there are less models with angles between 10 and 30 degrees. Yet these distributions do not necessarily help for the separation of acceptable models from incorrect ones.

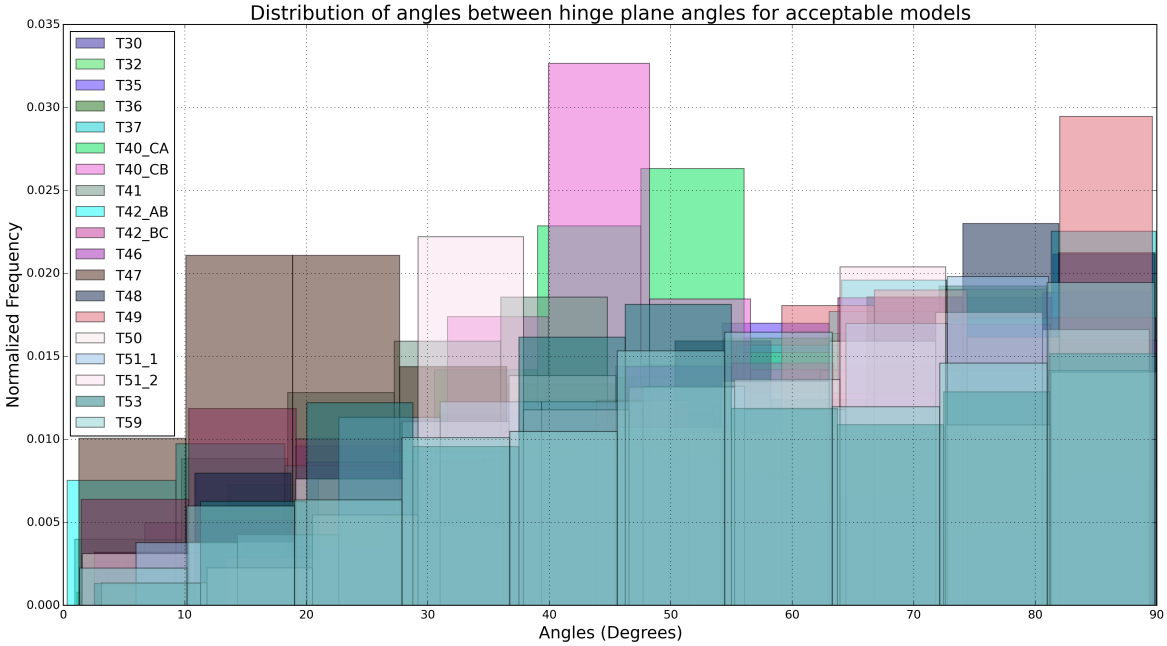

Figure 4.3. Distributions of hinge plane angles among CAPRI targets.

4.3. Mean Square Fluctuations of Interface and Non-Interface Residues

Information about mean square fluctuations of atoms is one of the most important data that reflect the dynamics of the molecule. Mean square fluctuations of alpha-carbon atoms, $\langle \Delta \mathbf{R}_i^2 \rangle$, can be calculated using GNM (see Equation 3.4). Mean square fluctuations are calculated for the interface versus the rest of residues including all modes of motion. Total sum of difference of normalized mean square fluctuations (shown with ΔMSF) between bound and unbound states (bound MSF – unbound MSF), of each complex are calculated and presented in Figure 4.4. MSF values are dimensionless as they are normalized. Except seven out of 176 structures, mean square fluctuations of interface residues decrease with binding. These exceptional complexes were seen to have domains connected with long loops or very densely packed residues with alpha helices. This means that interface residues lose their mobility upon binding. This is expected as interface residues make more neighbors with binding, which hinder the fluctuations of these residues. However, this hindrance may cause other residues outside the interface to become more mobile. To see this more clearly, as an example, the overall change in mean square fluctuations upon binding of C2 fragment of strep-

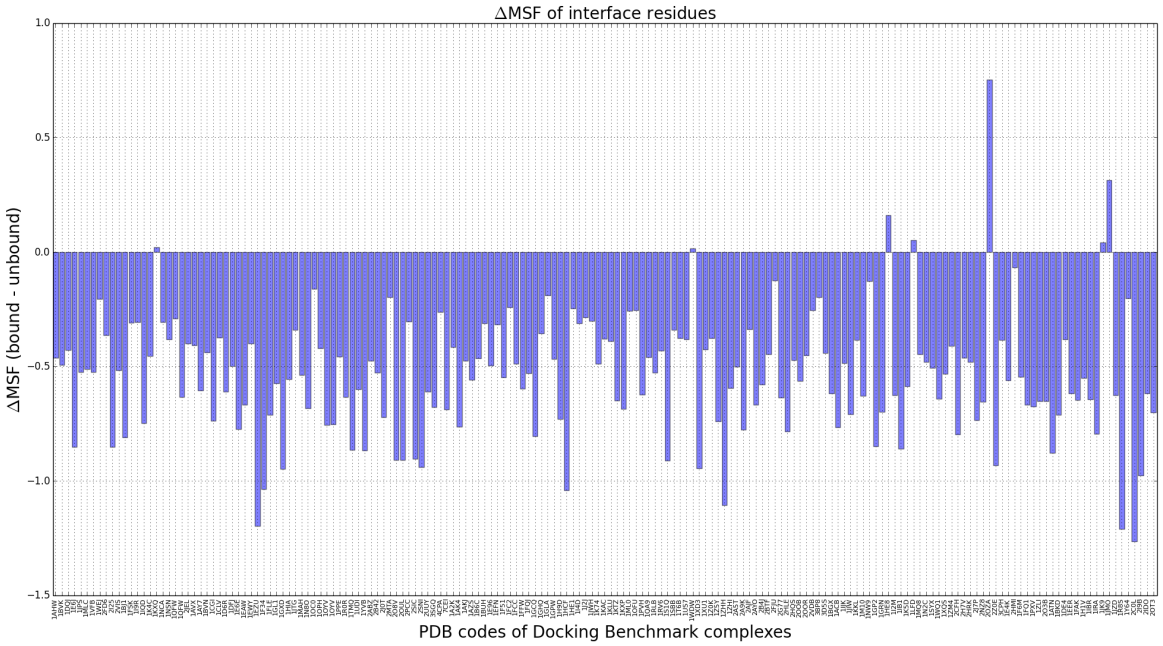

Figure 4.4. Mean square fluctuations difference between bound and unbound states of the interface residues of benchmark complexes.

tococcal protein G in complex with the FC domain of human IGG (PDB code 1FCC), which is one of the Docking Benchmark complexes, is focused. Results are presented in Figure 4.5. Down-pointing and up-pointing triangles show fluctuation values of interface residues upon and before binding. The increase in the fluctuations of residues between 256B and 346B are clearly visible.

This behavior that we see in C2 fragment of streptococcal protein G in complex with the FC domain of human IGG is also visible for the rest of the benchmark protein complexes. Overall difference in fluctuations of residues that are not interface residues are presented in Figure 4.6.

With exception of three complexes as seen in Figure 4.6, there is an increase in fluctuations of residues that are not located at interfaces. These exceptional structures contain domains that are connected with long loops. This could be the reason for decreasing fluctuations. In addition to that, 10 Å cutoff distance is taken for all calculations. This distance might not be an optimum cutoff distance for these complexes as well.

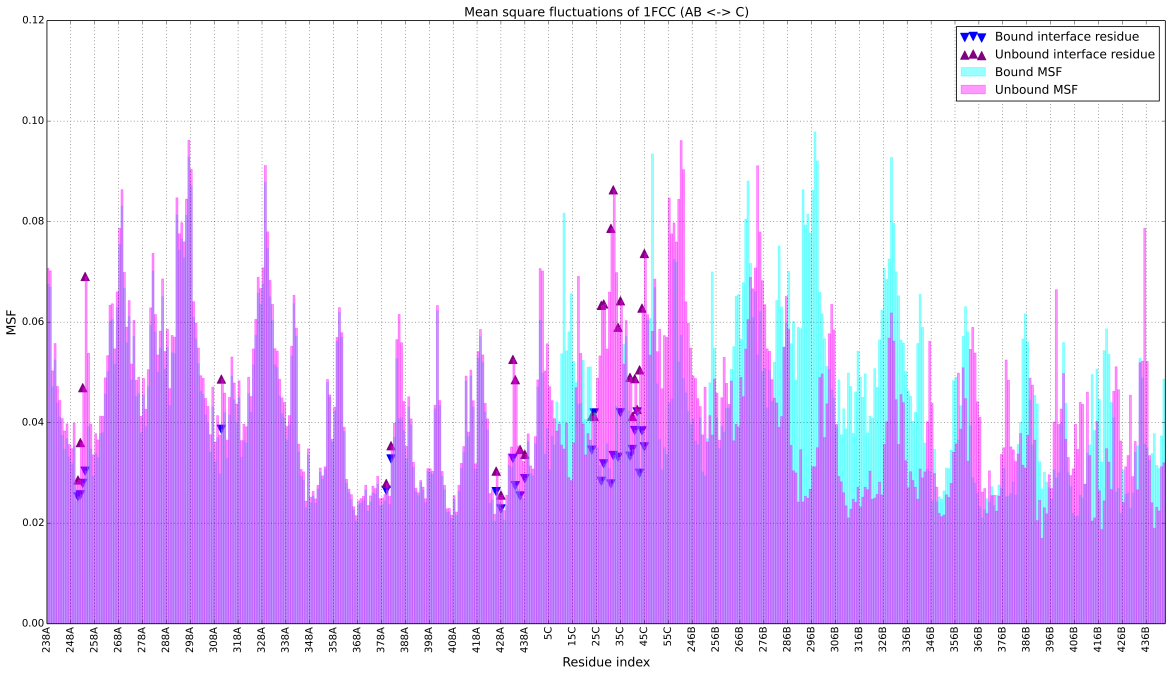

Figure 4.5. Change in mean square fluctuations of C2 fragment of streptococcal protein G in complex with the FC domain of human IGG protein complex (1FCC) upon binding.

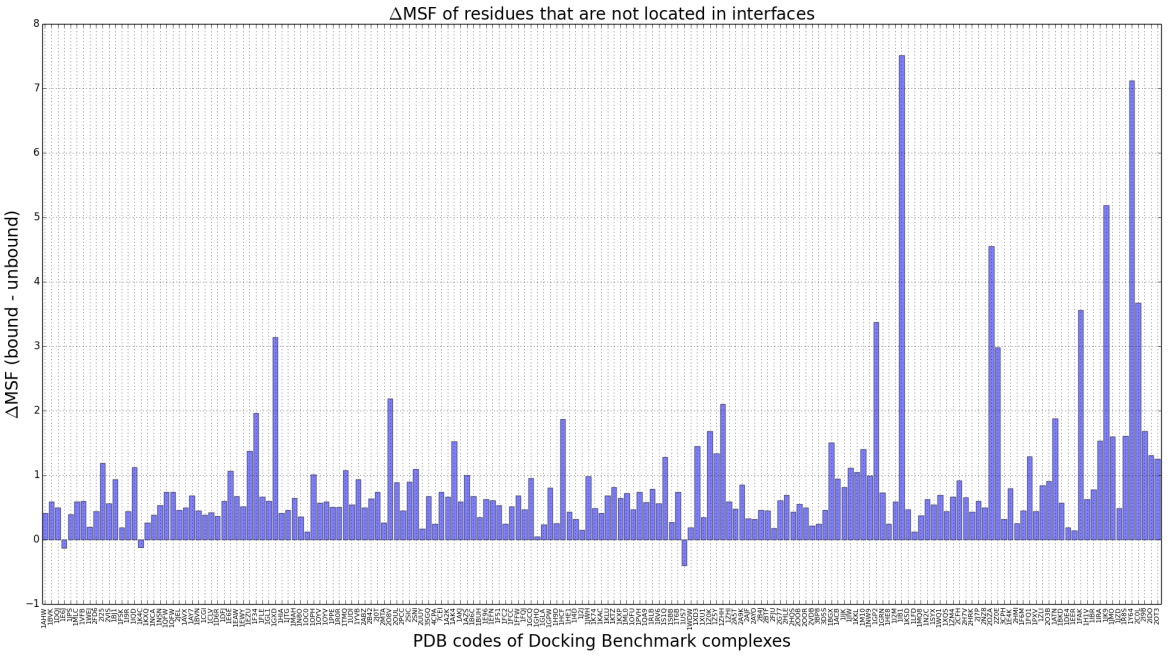

Figure 4.6. Mean square fluctuations difference between bound and unbound states of the non-interface residues of benchmark complexes.

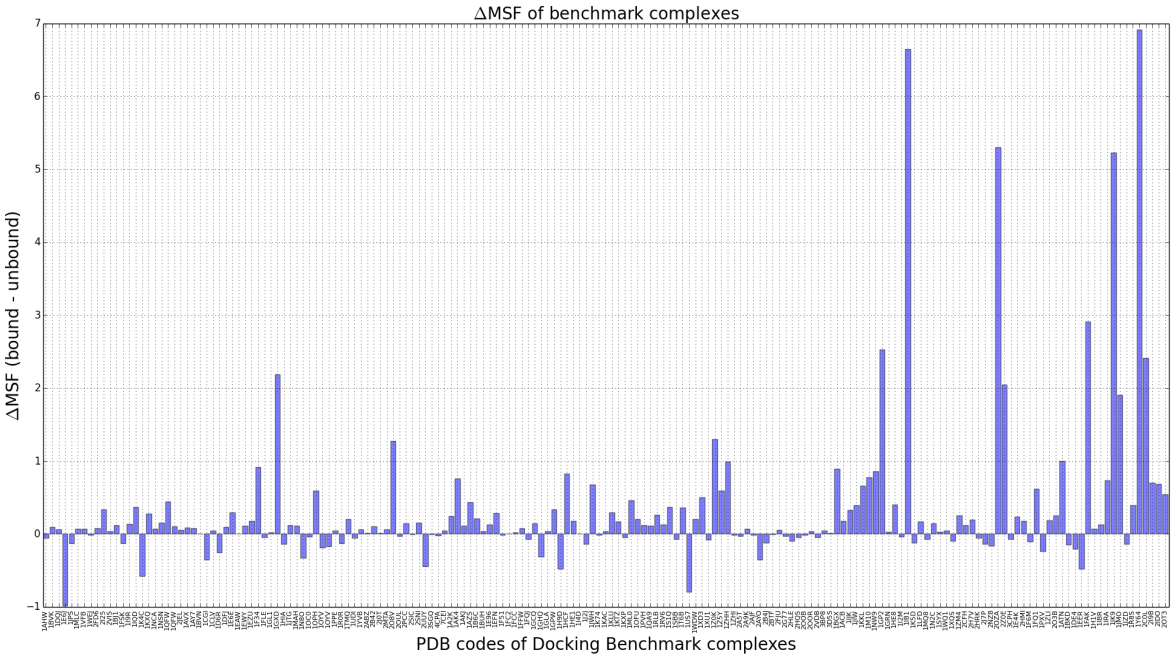

Figure 4.7. Overall mean square fluctuations difference between bound and unbound states of the benchmark complexes.

When the overall differences in sum of mean square fluctuations are analyzed, it is not possible to see a clear trend. As presented in Figure 4.7, there is a diversity of changes in mean square fluctuations upon binding. Therefore, it is hard to characterize the complexes according to overall changes in mean square fluctuations of their residues. Changes in fluctuations of residues could be because of changes in Gibbs free energy. As we know from thermodynamics, Gibbs free energy is also related to enthalpy. Diversification of the differences in sums of mean square fluctuations could be a result of enthalpy term which we are not able to take into account during GNM calculations.

If the differences in mean square fluctuations are used to filter the models, it is possible to set an interval for fluctuation of residues located in and outside of the interface. In 95% of Docking Benchmark complexes, ΔMSF is higher than -1 for interface residues (see Figure 4.4). Also, 76% of complexes have ΔMSF less than 1 for non-interface residues (see Figure 4.6). Table 4.3 represents results of a filtering that disqualifies models if ΔMSF of residues in the interface is less than -1 and ΔMSF of non-interface residues is higher than 1. Using this filter it is possible to discard up to 9%

of incorrect models for a given target. Understandably, this is an insufficient method by itself for finding the acceptable models as it does not increase the probability of finding the acceptable models significantly. In addition, there are some losses of acceptable models. Still, it is a fast method to disqualify incorrect models and among 19 targets, it never disqualifies all of the acceptable models. Therefore it is a promising method for preprocessing of the models dataset before prediction of the acceptable models.

Table 4.3. Probabilities of finding acceptable models after filtering by ΔMSF.

Targets	**Acceptable**▲	**Incorrect**▼	**Probability**■	**Probability**♦
T30	2	324	0.006	0.006
T32	33	299	0.096	0.099
T35	1	278	0.003	0.004
T36	1	243	0.003	0.004
T37	11	339	0.028	0.031
T40_CA	100	242	0.277	0.292
T40_CB	65	276	0.204	0.191
T41	153	169	0.468	0.475
T42_AB	15	262	0.054	0.054
T42_BC	5	272	0.018	0.018
T46	8	347	0.033	0.023
T47	217	58	0.783	0.789
T48	21	135	0.133	0.135
T49	12	145	0.076	0.076
T50	52	305	0.146	0.146
T51_1	3	234	0.010	0.013
T51_2	1	221	0.004	0.005
T53	46	329	0.120	0.123
T59	21	260	0.068	0.075

▲ Number of acceptable models after filtering.

▼ Number of incorrect models after filtering.

■ Probability of finding an acceptable model before filtering.

♦ Probability of finding an acceptable model after filtering.

4.4. ANM Mode Correlations

ANM analysis was done for each interaction partner of each protein complex in two different ways. Firstly, interaction partners were considered solitarily (unbound) without the other partner and normalized eigenvectors of ten slowest modes are recorded. Secondly, complexes were considered as a whole (bound) and normalized eigenvectors of ten slowest modes were recorded for the interaction partner of interest. Dot products of these eigenvectors result in correlations of modes between unbound and bound states (see Section 3.2.1). Cross correlations of all bound modes against unbound modes were calculated and following graphics were plotted. Cross correlation matrix is a 10x10 matrix. Correlations equal to 0.5 and above 0.5 are considered. Reason for that is analysis was repeated with cutoffs 0.6, 0.7 and 0.8 and areas of strongly correlated regions do not change significantly when cutoff is taken higher than 0.5. Therefore, only correlations above 0.5 are presented. Note that this analysis includes 352 structures, given 176 complexes with two binding partners each. Frequencies of complexes with high correlations are visualized using heat maps. Color bars show the numbers of structures.

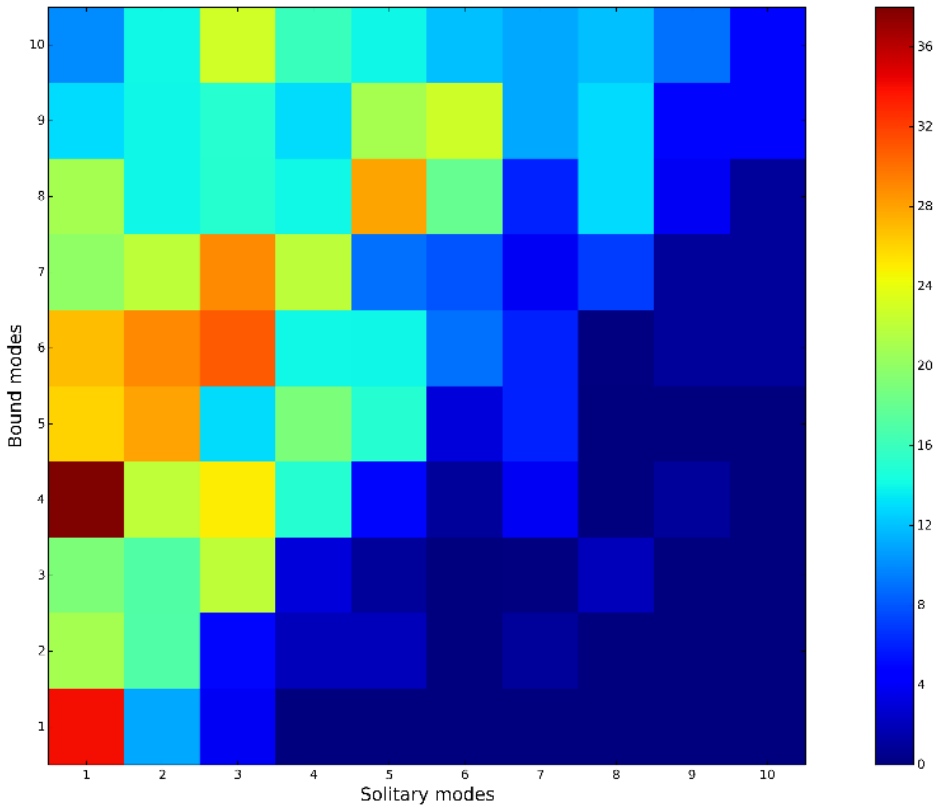

Figure 4.8. Number of unbound-bound mode correlations higher than 0.5.

In general, among all combinations of correlations in all modes, unbound mode 1 – bound mode 4 correlations is the most frequent one (see Figure 4.8). Second most frequent correlation is unbound mode 2 – bound mode 5 correlation, among all correlations. Finally, third most frequent correlation is unbound mode 3 – bound mode 6 correlation. Therefore, it is valid to say that upon involvement of rest of the complex structure, there is a shift of three modes among the slow modes.

Benchmark proteins are classified into three subgroups. These subgroups are antibody/antigen (13 structures), antigen/bound antibody (12 structures), enzyme inhibitor or enzyme substrate (52 structures) and others (99 structures). Same analysis is carried out for these subgroups separately.

Same three mode shift is observed on antibody/antigen group and others group (see Figures 4.9, 4.12) whereas there is no significant trend in mode correlations of

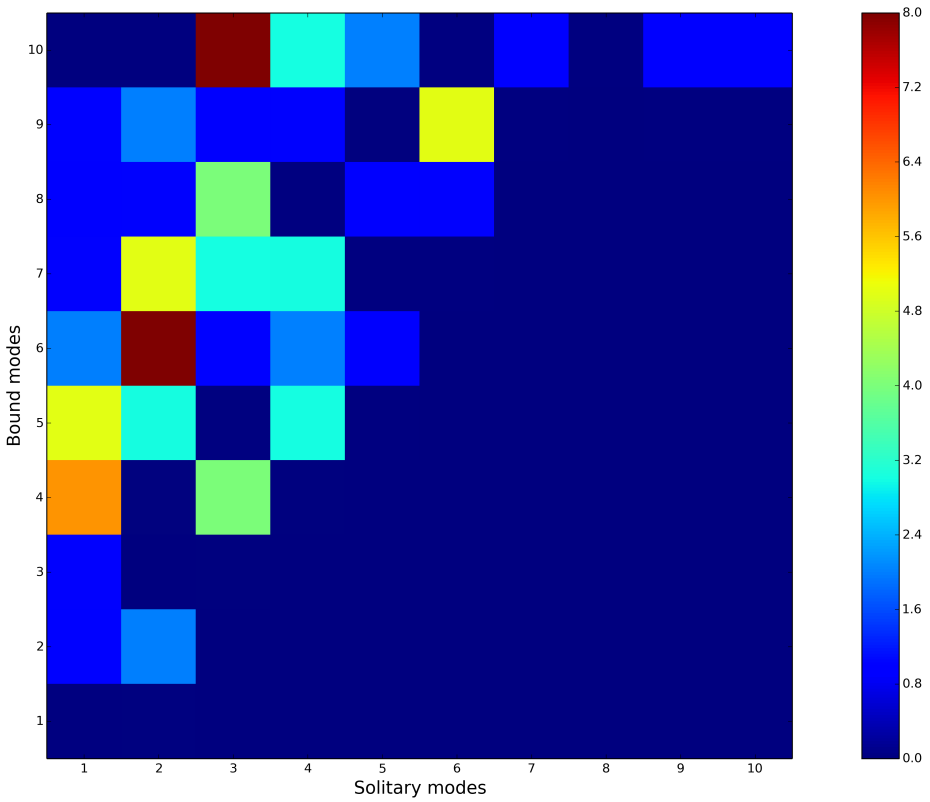

Figure 4.9. Number of unbound-bound mode correlations for antibody/antigen complexes.

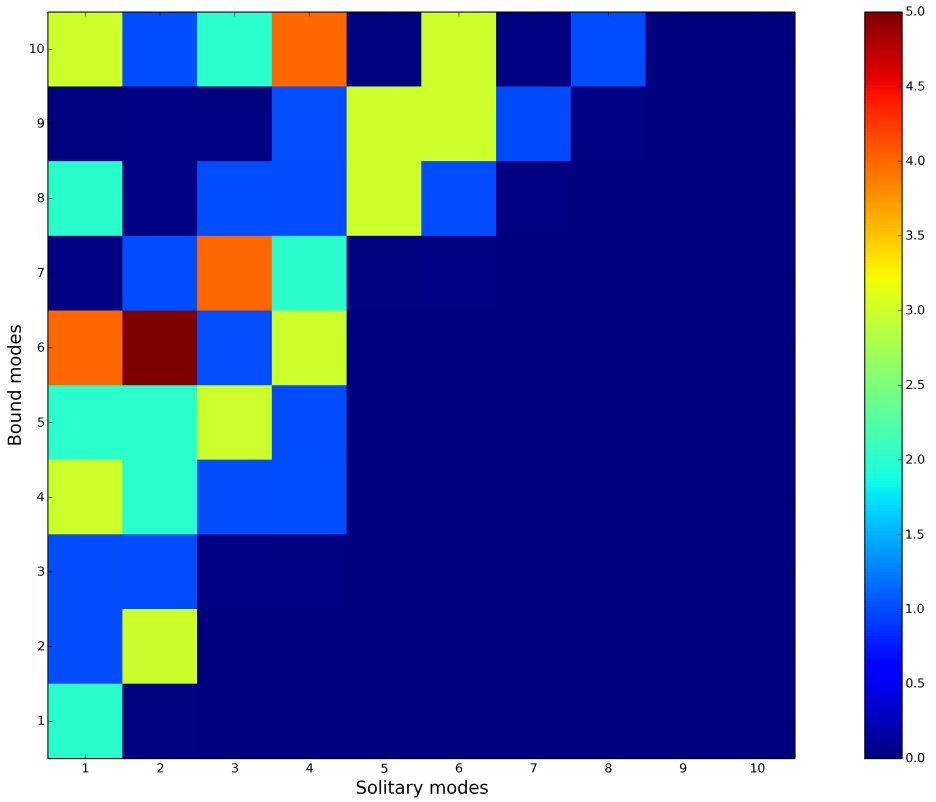

Figure 4.10. Number of unbound-bound mode correlations for antigen/bound antibody complexes.

antigen/bound antibody group and enzyme complexes group (see Figures 4.10, 4.11). Reason for missing significant trend in mode correlations of antigen/bound antibody group may be the lack of enough number of samples, since there are only 12 samples in the group. On the other hand, this is not the case for enzyme complexes because number of samples are 52 in the group and frequencies are rather evenly distributed. Trends are clearly visible in subgroup others which includes 99 complexes (see Figure 4.12). Unbound mode 1 – bound mode 4 correlation and unbound mode 3 – bound mode 4 correlations are the most frequent correlations and their frequencies are equal. Looking at the overall picture, we can see that unbound mode 1 – bound mode 4 is an important correlation for all subgroups.

Another approach that was used to check the correlations of eigenvectors between bound and unbound states is to consider complex as a whole. Specifically for unbound mode, eigenvectors of solitary modes are merged into one vector and this vector is normalized. Bound mode eigenvectors are taken as they are and correlations are calculated

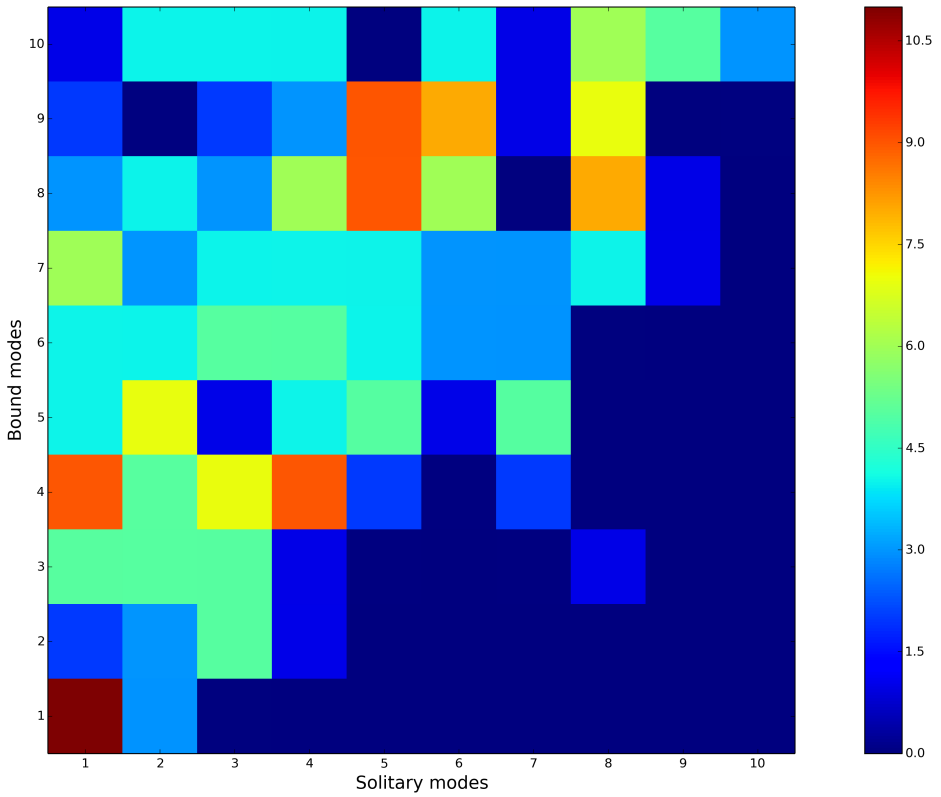

Figure 4.11. Number of unbound-bound mode correlations for enzyme/inhibitor complexes.

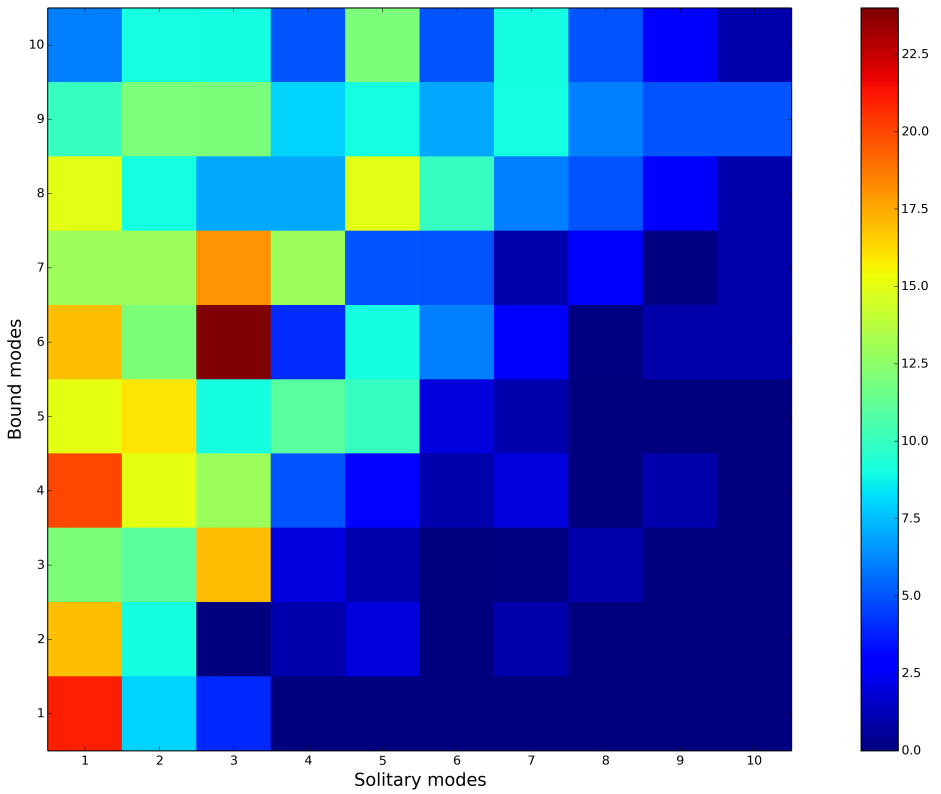

Figure 4.12. Number of unbound-bound mode correlations for other complexes.

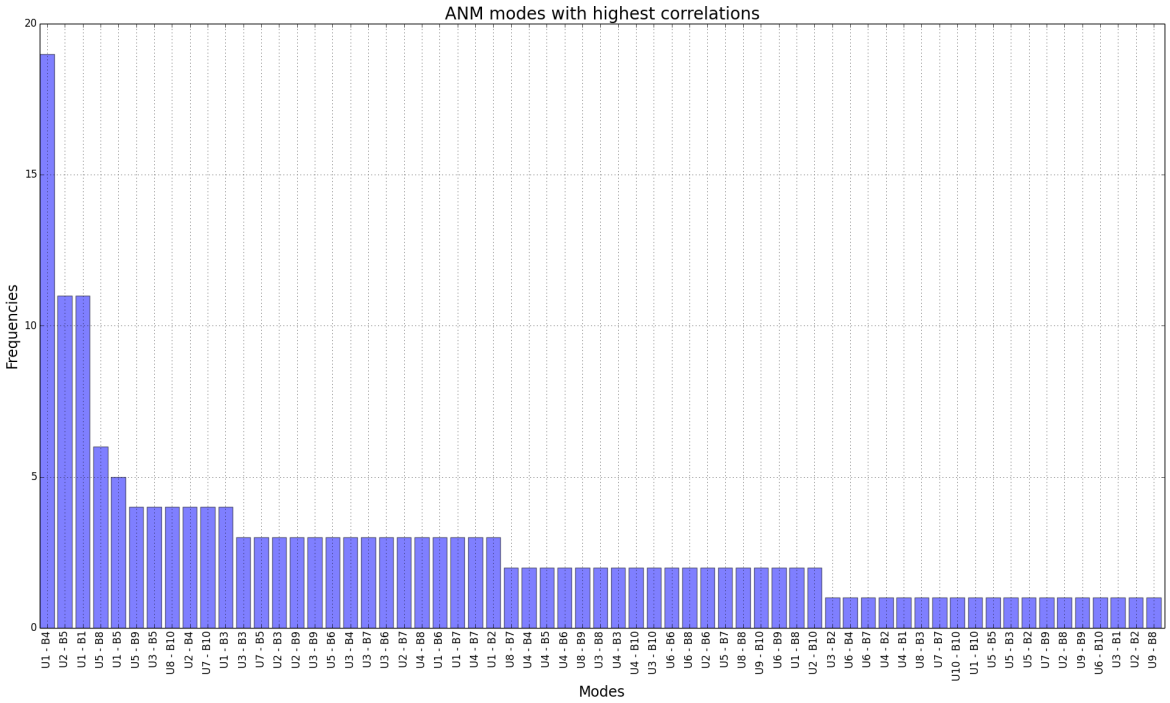

Figure 4.13. Combinations of unbound and bound ANM modes with highest correlations (U: Unbound mode, B: Bound mode).

by dot products of these two vectors (see Equation 3.17).

In order to see combinations of modes with highest correlations, from each correlation matrix highest correlation values along with names of the combinations are recorded. According to combinations of modes results are plotted on Figure 4.13.

As observed in solitary mode – bound mode correlations, similarly there is a three mode shift between bound and unbound modes. Number of highest correlations are between unbound mode 1 – bound mode 4 and unbound mode 2 and bound mode 5 as well as unbound mode 5 – bound mode 8. If we do the same analysis using eigenvectors of only interface residues, we get the distribution that is shown in Figure 4.11. Then again, unbound mode 1 – bound mode 4 and unbound mode 2 – bound mode 5 combinations are among the mostly occurring correlations. Hitherto, what we understand is dynamic information that a complex contains in its lower modes correspond to globular (slow modes of) motions of each unit. The correlation is thus mostly between a lower mode of complex and a higher mode of unbound structure as

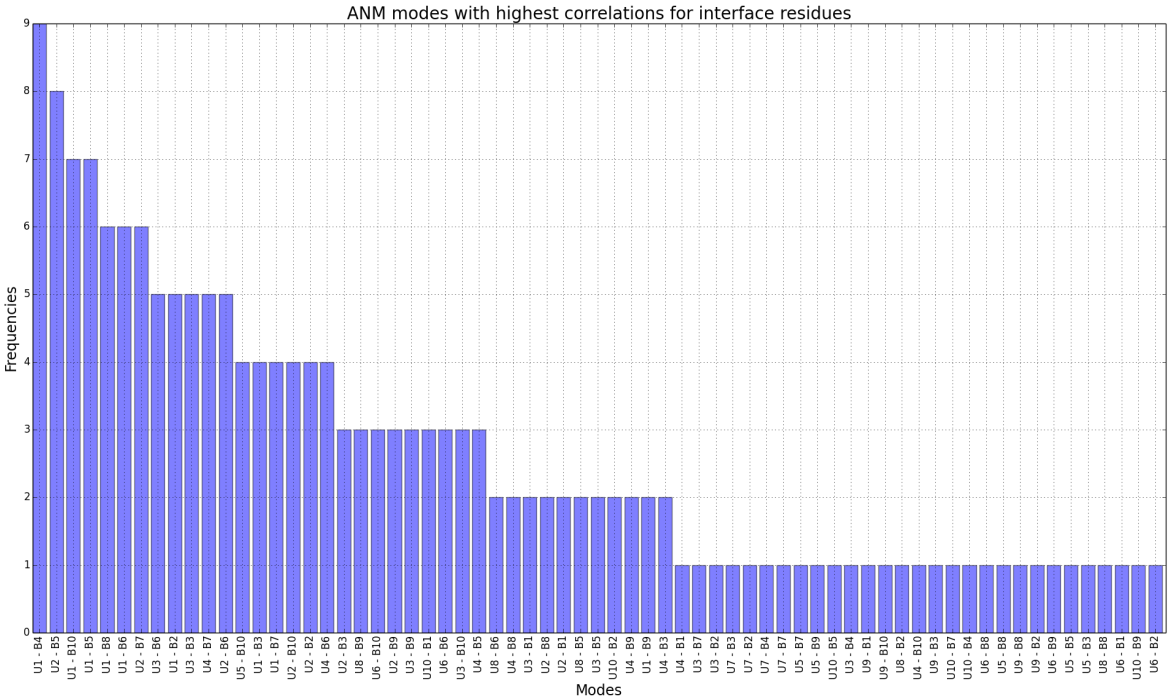

Figure 4.14. Combinations of unbound and bound ANM modes of interface residues with highest correlations (U: Unbound mode, B: Bound mode).

can be seen in Figures 4.13 and 4.14.

Though it is not possible to distinguish the acceptable models from the incorrect ones only by looking at highly correlated modes, the mode correlation matrices give us some information about how we can compare structures with each other. Instead of aligning models to each other, it is possible to create a metric for similarity of the dynamics for protein complexes.

4.5. ANM Mode Correlations as a Metric of Dynamics Similarity

Through the dynamics of bound and unbound states of protein complex structures via ANM calculations, we are able to calculate correlation matrices of bound and unbound modes (see Section 3.2.1). As purpose is to find the few native-like structures among a set of decoys, we need to understand how similar a model is to another one. Since acceptable models are similar to each other, we might be able to distinguish acceptable models from incorrect ones by using the dynamic information in the form

of a correlation matrix as a metric of similarity for clustering.

Within the scope of this study, ten slowest individual modes were chosen for understanding the binding dynamics of complexes. 10×20 sized matrices for each complex structure were obtained. For each target a distance matrix is created using this metric and matrices are clustered using k-means clustering algorithm (see Section 3.2.1).

Number of clusters were determined by the square root of half of number of samples. This is a basic rule of thumb for k-means clustering algorithm. Among all of the clusters for each target, the cluster which contains the highest number of acceptable models is chosen. Then, the probability of finding the acceptable structure from this cluster is calculated, the results of which are tabulated in Tables 4.4 and 4.5.

Information of all clusters for all available targets are tabulated in Tables A.1-19. It is observed that clustering according to unbound – bound ANM modes correlation matrices result in at least one cluster which gives significantly higher probability on choosing an acceptable model. Moreover, in order to see how number of clusters affect the in cluster probability, clustering was done multiple times with different numbers of clusters. Results are presented in Figures A.1-19. It is seen that for almost all targets, it is possible to obtain clusters that contain only acceptable models if number of clusters are roughly between 60 and 80. In addition, it can be seen that clustering helps more if there are multiple acceptable models within the model set. As it is seen on Figure A.3 increasing the number of clusters does not really help to purpose of finding acceptable models.

Choosing the right cluster with the high number of acceptable models is still challenging, especially when number of clusters are high. Merging clustering with filtering methods according to i-hinge occurrences, it might be possible to decrease the amount of incorrect models in order to find the cluster with highest number of acceptable models. When i-hinge filter is applied (see Section 4.1) by removing clusters which have at least three models that cannot pass the filter, it is possible to decrease

Table 4.4. Numbers of dynamic clusters.

Targets	**Number of clusters**	**Number of clusters containing acceptable models**
T30	13	2
T32	13	7
T35	13	1
T36	13	1
T37	14	4
T40_CA	14	9
T40_CB	14	7
T41	13	13
T42_AB	12	5
T42_BC	12	3
T46	14	3
T47	12	12
T48	9	2
T49	9	4
T50	13	3
T51_1	12	1
T51_2	12	1
T53	14	10
T59	13	8

Table 4.5. Statistics of clusters which hold the highest number of acceptable models.

Targets	**Acceptable**$^{\triangle}$	**Incorrect**$^{\triangledown}$	**Probability**$^{\square}$	**Probability**$^{\lozenge}$
T30	1	127	0,006	0,008
T32	18	20	0,096	0,474
T35	1	57	0,003	0,017
T36	1	32	0,003	0,030
T37	5	25	0,028	0,167
T40_CA	32	4	0,277	0,889
T40_CB	61	5	0,204	0,924
T41	67	22	0,468	0,753
T42_AB	8	60	0,054	0,118
T42_BC	2	44	0,018	0,043
T46	8	9	0,033	0,471
T47	92	39	0,783	0,702
T48	20	23	0,133	0,465
T49	5	18	0,076	0,217
T50	35	94	0,146	0,271
T51_1	3	189	0,010	0,016
T51_2	1	16	0,004	0,059
T53	13	88	0,120	0,129
T59	12	130	0,068	0,085

$^{\triangle}$ Number of acceptable models in the cluster.

$^{\triangledown}$ Number of incorrect models in the cluster.

$^{\square}$ Probability of finding an acceptable model before clustering.

$^{\lozenge}$ Probability of finding an acceptable model in the cluster.

the number of incorrect models down to a third (see Table 4.6). Though it is not always possible to get rid of incorrect models with this kind of filtering, for some cases it is possible to disqualify vast amount of incorrect models. Still, this method needs optimization for a number of clusters since in some cases acceptable models are disqualified as well.

Similarly, it is possible to filter incorrect models by combining the clustering with MSF. If target datasets are filtered with the purpose of having models within these ΔMSF intervals, the results are obtained as shown in Table 4.7. This filter is likely to drop the number of incorrect models significantly, however in some cases it disqualifies all of the acceptable models as well. Therefore, the parameters need further adjustments for both of the filtering methods.

Table 4.6. Number of models before and after filtering by clusters and i-hinge occurrences.

	Number of models			
	Before filtering		**After filtering**	
Targets	**Acceptable**	**Incorrect**	**Acceptable**	**Incorrect**
T30	2	348	2	138
T32	34	320	34	320
T35	1	338	0	74
T36	1	318	1	318
T37	11	379	7	274
T40_CA	102	266	102	266
T40_CB	75	293	75	293
T41	153	174	153	174
T42_AB	15	263	4	140
T42_BC	5	273	1	147
T46	13	384	12	232
T47	217	60	217	60
T48	21	137	21	137
T49	12	146	12	146
T50	52	305	42	256
T51_1	3	284	0	34
T51_2	1	275	0	79
T53	46	338	40	272
T59	23	316	23	316

Table 4.7. Number of models before and after filtering by clusters and ΔMSF.

	Number of models			
	Before filtering		**After filtering**	
Targets	**Acceptable**	**Incorrect**	**Acceptable**	**Incorrect**
T30	2	348	0	14
T32	34	320	30	143
T35	1	338	0	6
T36	1	318	0	47
T37	11	379	8	201
T40_CA	102	266	10	29
T40_CB	75	293	4	93
T41	153	174	76	60
T42_AB	15	263	10	175
T42_BC	5	273	5	174
T46	13	384	0	41
T47	217	60	139	32
T48	21	137	21	123
T49	12	146	12	146
T50	52	305	52	305
T51_1	3	284	0	44
T51_2	1	275	0	70
T53	46	338	7	46
T59	23	316	1	39

4.6. Information Entropy and Collectivity

Information entropy and collectivity can be calculated using both GNM and ANM (see Materials and Methods, Information Entropy and Collectivity). With the purpose of characterizing the native complex structures, information entropies of benchmark

complexes were calculated in order to see the change in collectivity of the complex structures with binding. For GNM, first three slowest modes were considered whereas for ANM first ten slowest modes were taken into account. Reason for that is, these modes correspond to significant part of the global movements that structures are capable of performing. Firstly, information entropies are calculated for first binding partner for each complex structure in bound and unbound state. Then, same was done for the second partners. Results for GNM collectivities are shown in Figures C.1-C.6.

According to Figure C.1 and C.2 there is a slight increase in collectivity in first mode for both partners, but it is more clearly visible on second binding partner. Although distributions are a little skewed for bound collectivities, not all the structures have increasing collectivities. Same trends are valid for the second slowest mode of GNM (see Figure C.3-C.4). GNM collectivities show that there is a slight increase in collectivity upon binding but this is not clear enough. This might be because information about directions of movement are significant for information entropy calculations.

Same analysis was done using ANM and results for the first slowest mode are presented in Figures 4.15 and 4.16. For both binding partners, distribution shift towards higher values of collectivity. This means, in total there is an increase in collectivity upon binding for both binding partners of all Docking Benchmark complexes. This is valid for first ten slowest modes. Results for rest of the modes are presented in Figures D.1-D.18.

Given K_{1u}, K_{2u} are the unbound collectivities of first and second binding partners respectively and K_{1b}, K_{2b} are the bound collectivities of first and second binding partners, overall change in difference and sum of collectivities are calculated as $(K_{1b} - K_{1u})$–$(K_{2b} - K_{2u})$ and $(K_{1b} - K_{1u}) + (K_{2b} - K_{2u})$. Probability distribution functions of sum and differences of collectivities calculated as such are presented in Figures 4.17 and 4.18 for first ten slowest modes. All of the peaks are gathered around zero for collectivity differences of first ten modes. This shows that collectivity earnings of both partners are rather similar in general. Peaks of collectivity sums (collectivity gains) gather around 0.5.

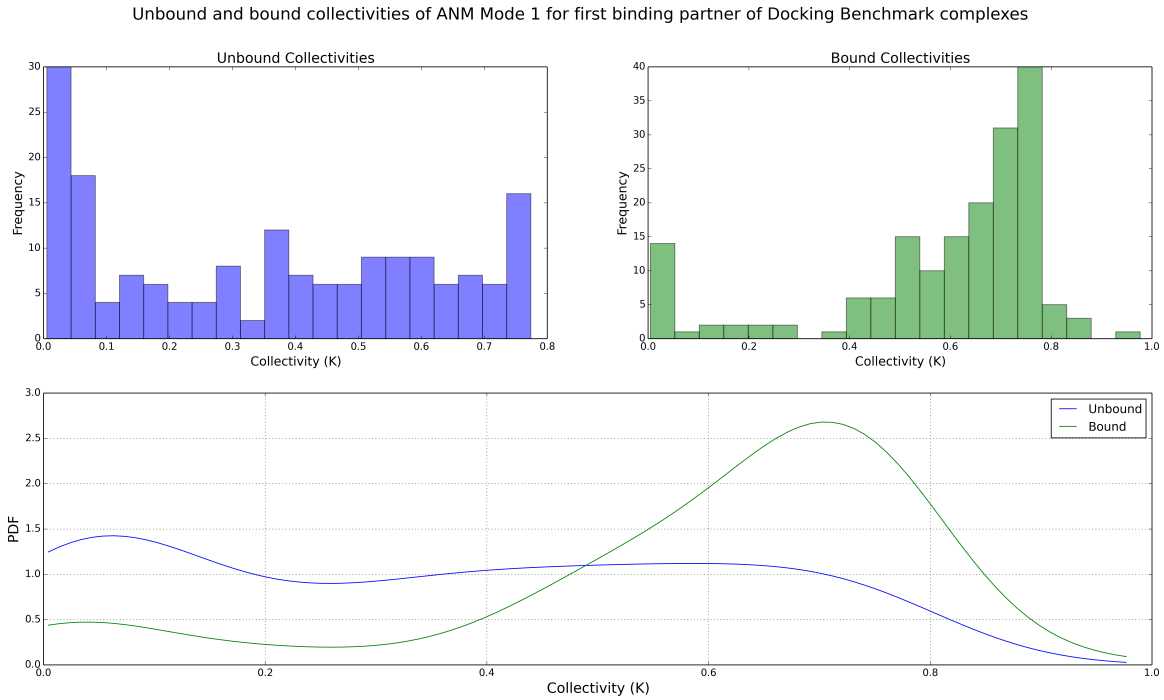

Figure 4.15. Distributions and probability density functions of bound and unbound collectivities for first slowest ANM mode of first binding partners in Docking Benchmark.

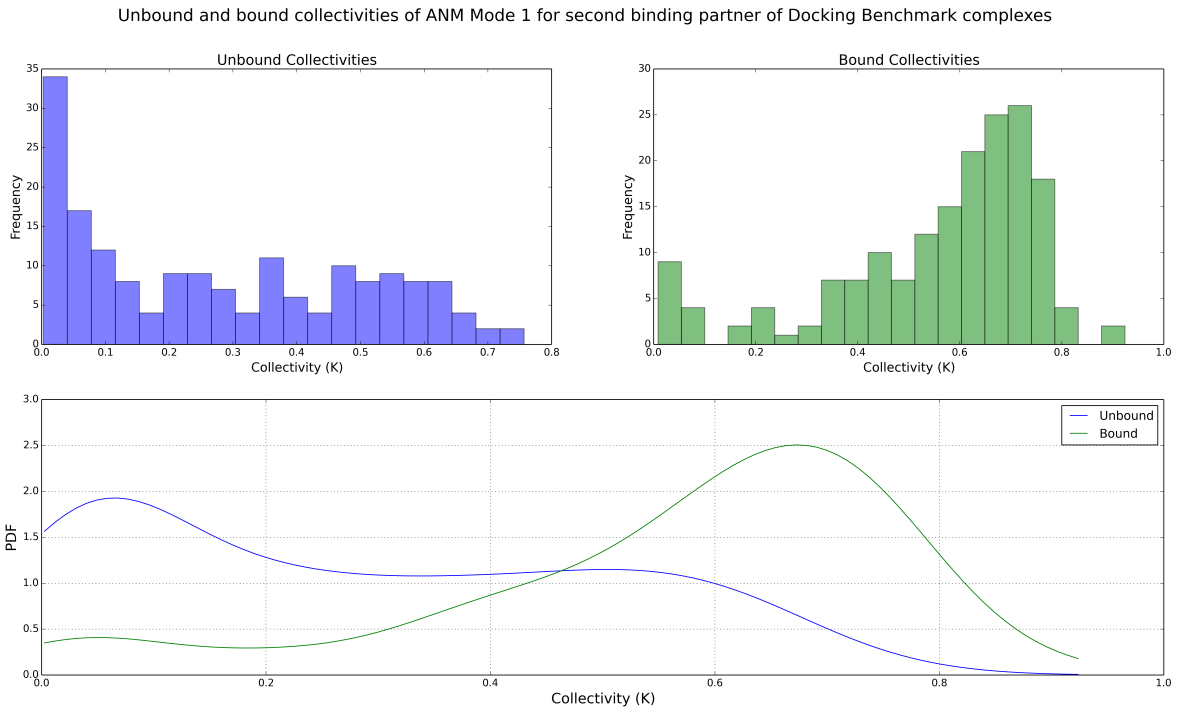

Figure 4.16. Distributions and probability density functions of bound and unbound collectivities for first slowest ANM mode of first binding partners in Docking Benchmark.

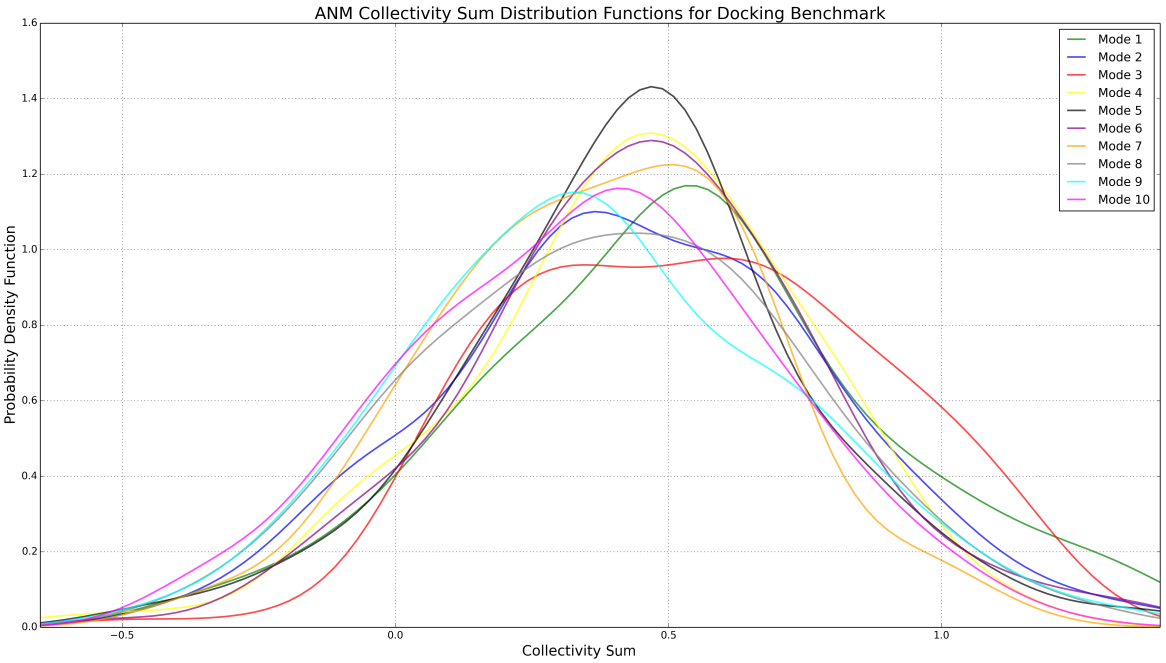

Figure 4.17. Probability distribution functions of collectivity summations for first ten slowest modes of Docking Benchmark complexes.

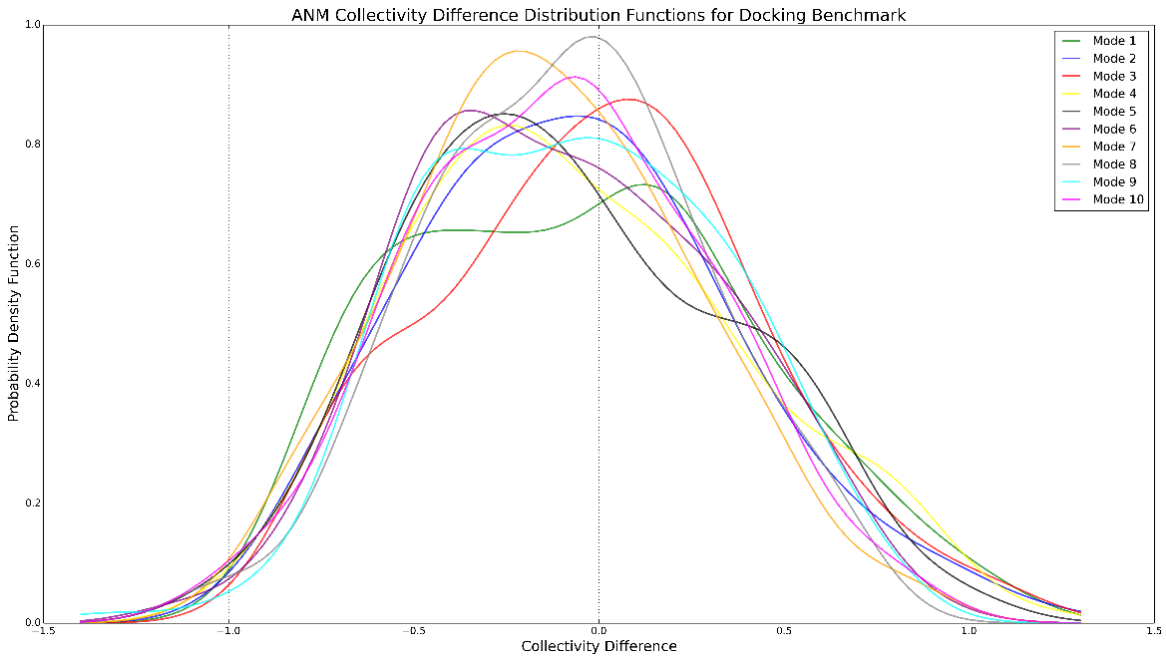

Figure 4.18. Probability distribution functions of collectivity differences for first ten slowest modes of Docking Benchmark complexes.

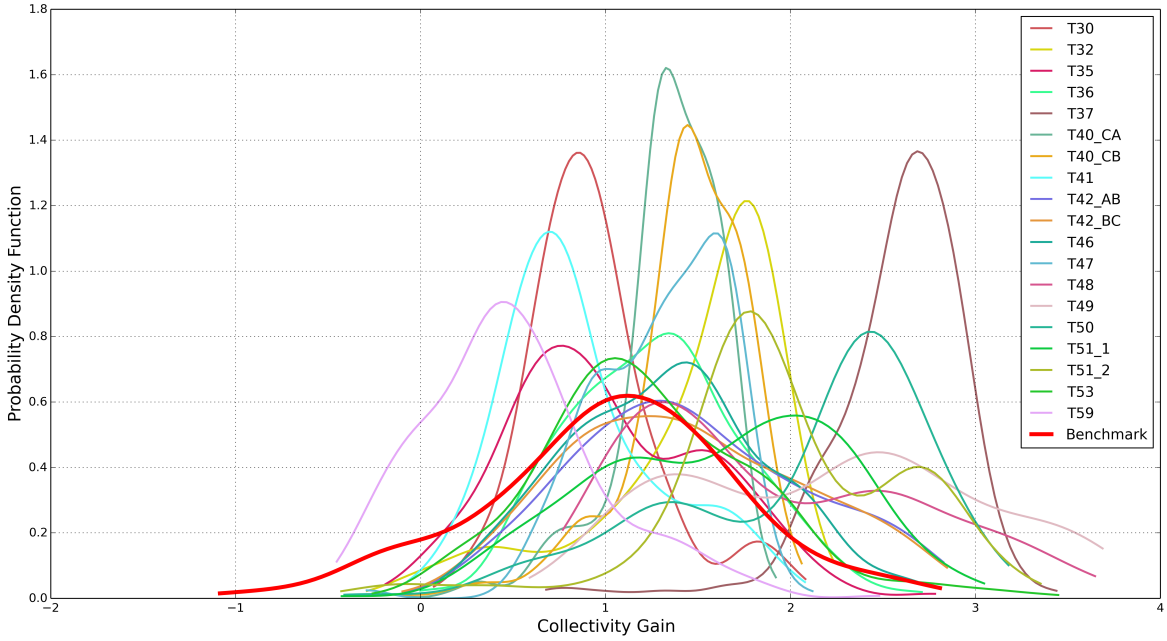

Figure 4.19. Probability distribution functions of collectivity gains for average of first ten slowest modes of target models and Docking Benchmark complexes.

In order to see where collectivity gains lie for Docking Benchmark complexes and target models, probability distribution functions of all of them are plotted and presented in Figure 4.16. Average of first ten modes are considered for all of the structures. As observed in this plot, it is hard to distinguish a clear border between native complexes and targets most of which consist of incorrect models. Also, because of diverse types of complexes Benchmark curve has a wider range in negative gain (collectivity loss). Except Target 48, number of models with collectivity loss are very less.

4.7. CLAMshell and Ranking Algorithm

In order to simplify clustering, a dynamics clustering tool "CLAMshell" (**CL**ustering by **A**NM **M**odes in IPython **shell**) was developed in Python programming language with the help of NumPy, SciPy [44], scikit-learn [45] and Pandas [46] third party libraries. This in-house software includes an ANM module where eigenvectors and eigenvector correlation matrices are calculated, pairwise distance module where pairwise distance matrix is created from eigenvector correlation matrices and finally k-means

clustering module. Resulting property tables are presented as Pandas data frame (table) which provides an easy to use table view. These tables can also be converted to Microsoft Excel spreadsheets.

The ranking algorithm that was tried, consists of the modules which calculate the properties mentioned earlier. Namely, occurrences of i-hinges, collectivity and dynamics similarity. MSF filtering is not utilized because of the lack of sensitivity to acceptable models (see Table 4.7). The algorithm is presented below in detail.

Input PDB files of target models with chain names
Result List of PDB file names
Initialize;
N ← Number of models
C ← Number of clusters
pw ← Pairwise distance matrix ($N \times N$)
Model List ← List for holding names of models for result
k-means ← K-Means clustering function
$C = \sqrt{\frac{N}{2}}$
Calculate eigenvector correlation matrices for models with Equation a
Calculate pw from correlation matrices using Equation b
Clusters = k-means(pw,C)
for cluster in clusters **do**
  if number of models without i-hinges in first two modes > 3 **then**
    remove cluster
  end if
end for
for cluster in clusters **do**
  add the model with highest collectivity to Model List
end for
return Model List
End.

Figure 4.20. Ranking algorithm.

For the target dataset number of clusters vary between from 9 to 14. This is also the number of models obtained from the procedure. Once PDB files are provided, CLAMshell returns the list of structures with according to their clusters as a table. Using this table, rest of the procedure is executed. Besides CLAMshell, rest of the procedure is also coded in Python.

Table 4.8. Comparison of ranking algorithm performance with random picking.

	Ranking algorithm			**Random Picking**		
Target	Acceptable	Incorrect	Probability	Acceptable	Incorrect	Probability
T30	0	10	0	0	10	0.006
T32	3	10	0.231	2	11	0.096
T35	0	10	0	0	10	0.003
T36	0	13	0	0	13	0.003
T37	0	13	0	0	13	0.028
T40_CA	5	9	0.357	4	10	0.277
T40_CB	1	13	0.071	0	14	0.204
T41	3	9	0.25	4	8	0.468
T42_AB	0	9	0	0	9	0.054
T42_BC	0	8	0	0	8	0.018
T46	4	8	0.333	0	12	0.033
T47	5	7	0.417	9	3	0.783
T48	2	7	0.222	2	7	0.133
T49	3	6	0.333	0	9	0.076
T50	0	12	0	2	10	0.146
T51_1	0	7	0	0	7	0.01
T51_2	0	6	0	0	6	0.004
T53	2	11	0.154	3	10	0.12
T59	0	13	0	1	12	0.068

Table 4.8 compares the algorithm success with random picking. Ranking algorithm probability states the odds of finding an acceptable model among the models that are tabulated whereas random picking probability states the total number of acceptable models divided by total number of models. From Table 4.8, it seems like there is no use in using this ranking algorithm because in some cases random picking gives better results and numbers of cases that is possible to find acceptable models are almost equal. One of the reasons of that is because ranking algorithm takes one representative from each cluster whereas there might be number of clusters without any acceptable model. Finding the right cluster and choosing representatives from that cluster is still a challenge because none of the properties that were investigated were able to characterize the dynamics clusters. At this step, further parameter optimization is needed. Especially by further investigating the MSF filtering boundaries statistically per cluster, it might be possible to achieve a more successful ranking system.

5. CONCLUSION

Both ANM and GNM are computationally inexpensive, therefore they are able to provide the computational efficiency that a ranking algorithm for protein-protein interaction predictions might need. It is possible to process hundreds of structures within minutes on a modern desktop (or laptop) computer for the collecting dynamics information of the structures. In addition to that, both models are able to characterize dynamic properties of protein complexes.

The GNM analysis suggests that occurrences of i-hinges in first two slowest modes are crucial in benchmark complexes and it is possible to identify some of the incorrect models by this property. Moreover, MSF decreases at interface region whereas it increases at non-interface regions. This is yet another property for disqualifying some of the incorrect models; but a further adjustment is needed for setting boundaries of disqualification by means of MSF. GNM solitarily is not enough for the ranking but it is possible to embed it in a scoring function to perform a preliminary low quality model disqualification filtering.

Linear discriminant analysis paves the way for defining dynamic domains by using GNM. Hinge plane is oriented such a way that variances (of positions) of positively (or negatively) correlated fluctuating regions –based on the global modes- is minimized and variance between these regions is maximized. The distribution of preferred orientations of hinge planes show that preferred angles between hinge planes are mostly between 30 and 90 degrees for benchmark complexes. However, this distribution is shown to be futile for the discrimination of incorrect models from correct ones. On the other hand, the orientation of hinge planes might be useful for studying the global behavior of two monomers with MD (molecular dynamics). As the method is computationally inexpensive, it is possible to apply it to a whole MD trajectory. In the course of this thesis, “HiPlane”, an in house software is developed for this purpose. That makes it possible to obtain coefficients of plane equation (or plane normal vector). It also generates a PyMOL script for visualizing the planes on the structure.

From ANM analysis, it is seen that native complexes gain collectivity in different levels upon binding. Although that is clear but it is still not possible to distinguish models according to their quality just by looking at their collectivity gains because some of the high quality models also have negative collectivity gains.

The ANM analysis was seen to be more expressive for using the predicted fluctuations in information entropy calculations. The anisotropy information in the motion appears important for adequate collectivity calculations compared to predictions by GNM. Also through ANM, eigenvector correlation matrices (see Section 3.2.1) made it possible to understand that at one point dynamics information of binding partners are embedded in the dynamics of the complex. Namely, there is a three mode shift observed upon binding of a structure. Eigenvector of a mode in unbound state is correlated with three level higher mode in bound state (i.e. bound mode 4 – unbound mode 1). In scope of this study, the correlation matrices also constitute the blue prints of normal mode dynamics of complex structures and the basis of dynamics similarity. Dynamics similarity metric that was created led the way to the development of CLAMshell.

Throughout this study, CLAMshell is used for grouping similar predicted models and it is proven to be successful with grouping high quality models with each other. This approach shines the light on a different kind of classification technique. Although, only ten slow modes used in CLAMshell, it is also possible to use fast modes. Therefore, it is also possible to create a classification model based on high frequency fluctuations of surface residues as an example. Moreover, the number of modes can be increased. This would create a larger correlation matrix, which would provide a more sensitive similarity metric. Elements of each cluster created by the software may be re-clustered by using a larger correlation matrix which includes more detailed information. This would steal computational resources, since procedure is repeated, but might characterize the clusters in better way. Just like filtering methods that are discussed, CLAMshell can be embedded into another scoring algorithm to characterize clusters and do the ranking based on these clusters. As an example, empirical scoring functions might be able to provide further information for the characterization of binding energies in predicted models within a cluster. It might thus be possible to choose the best cluster

(cluster of native-like models) in this case.

CLAMshell might also be useful in conformational search of complex structures. Conformations generated by conformational search algorithms are mostly clustered by RMSD. However, structures with different conformations might share similar dynamic information with respect to the dynamics off the binding partner. Thus, it might be possible to use dynamics clustering in conformation search, which may lead to identifying main conformers with a dynamic behavior. Different from complex structures, for cases of monomers, matrix norm of eigenvector correlation matrix (created by dot products of eigenvectors of two structures) can as well be a distance metric. To this end, the clustering performance and resulting clusters based on dynamics at first should be compared those by RMSD clustering.

APPENDIX A: ANM Mode Correlation Clusters

Table A.1. Clusters of Target 30 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	0	25
Cluster 2	0	4
Cluster 3	0	8
Cluster 4	1	127
Cluster 5	0	2
Cluster 6	1	18
Cluster 7	0	5
Cluster 8	0	23
Cluster 9	0	91
Cluster 10	0	8
Cluster 11	0	30
Cluster 12	0	3
Cluster 13	0	4

Table A.2. Clusters of Target 32 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	0	1
Cluster 2	1	48
Cluster 3	1	21
Cluster 4	0	18
Cluster 5	1	32
Cluster 6	3	32
Cluster 7	0	8
Cluster 8	0	8
Cluster 9	0	27
Cluster 10	2	42
Cluster 11	0	26
Cluster 12	18	20
Cluster 13	8	37

Table A.3. Clusters of Target 35 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	1	57
Cluster 2	0	5
Cluster 3	0	10
Cluster 4	0	1
Cluster 5	0	224
Cluster 6	0	4
Cluster 7	0	17
Cluster 8	0	2
Cluster 9	0	2
Cluster 10	0	1
Cluster 11	0	3
Cluster 12	0	2
Cluster 13	0	10

Table A.4. Clusters of Target 36 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	1	32
Cluster 2	0	1
Cluster 3	0	2
Cluster 4	0	5
Cluster 5	0	12
Cluster 6	0	181
Cluster 7	0	2
Cluster 8	0	3
Cluster 9	0	6
Cluster 10	0	16
Cluster 11	0	11
Cluster 12	0	46
Cluster 13	0	1

Table A.5. Clusters of Target 37 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	4	34
Cluster 2	0	14
Cluster 3	0	55
Cluster 4	0	5
Cluster 5	1	59
Cluster 6	0	14
Cluster 7	5	25
Cluster 8	0	18
Cluster 9	0	27
Cluster 10	0	76
Cluster 11	0	10
Cluster 12	1	0
Cluster 13	0	34
Cluster 14	0	8

Table A.6. Clusters of Target 40_CA models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	15	17
Cluster 2	0	1
Cluster 3	19	15
Cluster 4	0	7
Cluster 5	0	61
Cluster 6	0	1
Cluster 7	2	51
Cluster 8	3	34
Cluster 9	1	0
Cluster 10	0	27
Cluster 11	28	9
Cluster 12	32	4
Cluster 13	1	29
Cluster 14	1	10

Table A.7. Clusters of Target 40_CB models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	4	30
Cluster 2	0	1
Cluster 3	61	5
Cluster 4	1	0
Cluster 5	0	11
Cluster 6	0	31
Cluster 7	0	37
Cluster 8	0	36
Cluster 9	1	6
Cluster 10	0	1
Cluster 11	1	36
Cluster 12	0	23
Cluster 13	6	46
Cluster 14	1	30

Table A.8. Clusters of Target 41 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	32	47
Cluster 2	4	3
Cluster 3	2	6
Cluster 4	5	7
Cluster 5	1	2
Cluster 6	23	27
Cluster 7	4	0
Cluster 8	7	45
Cluster 9	3	12
Cluster 10	3	0
Cluster 11	67	22
Cluster 12	1	3
Cluster 13	1	0

Table A.9. Clusters of Target 42_AB models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	1	18
Cluster 2	0	2
Cluster 3	0	15
Cluster 4	0	3
Cluster 5	0	20
Cluster 6	2	45
Cluster 7	8	60
Cluster 8	0	15
Cluster 9	0	1
Cluster 10	0	4
Cluster 11	1	45
Cluster 12	3	35

Table A.10. Clusters of Target 42_BC models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	2	43
Cluster 2	0	1
Cluster 3	0	11
Cluster 4	0	46
Cluster 5	0	3
Cluster 6	0	22
Cluster 7	2	44
Cluster 8	0	2
Cluster 9	0	4
Cluster 10	0	19
Cluster 11	0	38
Cluster 12	1	40

Table A.11. Clusters of Target 46 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	0	40
Cluster 2	0	2
Cluster 3	4	4
Cluster 4	0	116
Cluster 5	0	2
Cluster 6	8	9
Cluster 7	0	1
Cluster 8	0	1
Cluster 9	1	21
Cluster 10	0	67
Cluster 11	0	60
Cluster 12	0	12
Cluster 13	0	46
Cluster 14	0	3

Table A.12. Clusters of Target 47 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	92	39
Cluster 2	10	0
Cluster 3	1	2
Cluster 4	10	1
Cluster 5	8	1
Cluster 6	7	1
Cluster 7	6	1
Cluster 8	4	0
Cluster 9	2	0
Cluster 10	7	4
Cluster 11	69	11
Cluster 12	1	0

Table A.13. Clusters of Target 48 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	1	36
Cluster 2	0	9
Cluster 3	0	4
Cluster 4	20	23
Cluster 5	0	48
Cluster 6	0	7
Cluster 7	0	4
Cluster 8	0	1
Cluster 9	0	5

Table A.14. Clusters of Target 49 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	5	18
Cluster 2	2	28
Cluster 3	0	1
Cluster 4	0	4
Cluster 5	0	24
Cluster 6	3	36
Cluster 7	0	1
Cluster 8	0	19
Cluster 9	2	15

Table A.15. Clusters of Target 50 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	10	64
Cluster 2	0	13
Cluster 3	0	1
Cluster 4	0	2
Cluster 5	0	19
Cluster 6	35	94
Cluster 7	0	4
Cluster 8	0	46
Cluster 9	0	11
Cluster 10	0	11
Cluster 11	7	34
Cluster 12	0	4
Cluster 13	0	2

Table A.16. Clusters of Target 51_1 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	3	189
Cluster 2	0	13
Cluster 3	0	1
Cluster 4	0	10
Cluster 5	0	7
Cluster 6	0	3
Cluster 7	0	42
Cluster 8	0	4
Cluster 9	0	1
Cluster 10	0	4
Cluster 11	0	9
Cluster 12	0	1

Table A.17. Clusters of Target 51_2 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	0	5
Cluster 2	0	206
Cluster 3	0	12
Cluster 4	0	6
Cluster 5	0	2
Cluster 6	0	2
Cluster 7	0	9
Cluster 8	0	4
Cluster 9	0	1
Cluster 10	0	6
Cluster 11	1	16
Cluster 12	0	2

Table A.18. Clusters of Target 53 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	7	58
Cluster 2	0	10
Cluster 3	0	1
Cluster 4	0	13
Cluster 5	2	10
Cluster 6	0	2
Cluster 7	8	26
Cluster 8	1	1
Cluster 9	13	88
Cluster 10	2	33
Cluster 11	1	7
Cluster 12	1	11
Cluster 13	1	2
Cluster 14	10	76

Table A.19. Clusters of Target 59 models.

Clusters	**Acceptable Models**	**Incorrect Models**
Cluster 1	3	18
Cluster 2	12	130
Cluster 3	1	6
Cluster 4	1	19
Cluster 5	2	6
Cluster 6	0	14
Cluster 7	0	3
Cluster 8	2	20
Cluster 9	1	42
Cluster 10	1	31
Cluster 11	0	7
Cluster 12	0	14
Cluster 13	0	6

APPENDIX B: Changing Number of ANM Mode Correlation Clusters

In order to see how probability of finding an acceptable structure changes with number of clusters, target datasets were re-clustered multiple times with changing number of clusters. Following bar plots show how probability of finding acceptable clusters (within the cluster that has highest probability of finding acceptable models) changes with cluster amount. Number of clusters are increased by one starting from two until 100 clusters.

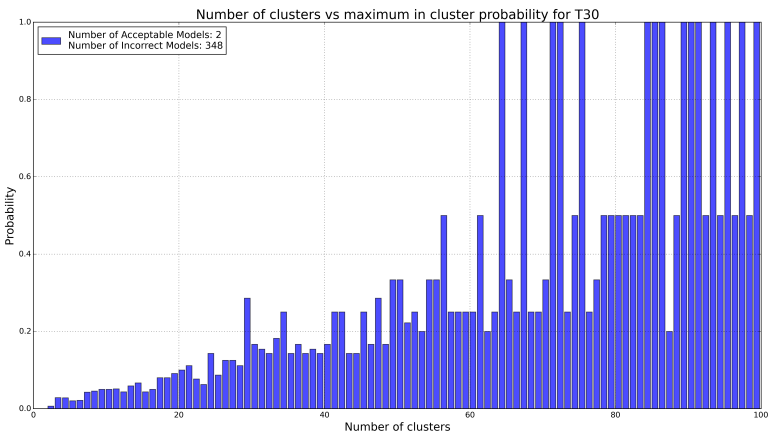

Figure B.1. Number of clusters versus in cluster probability for Target 30.

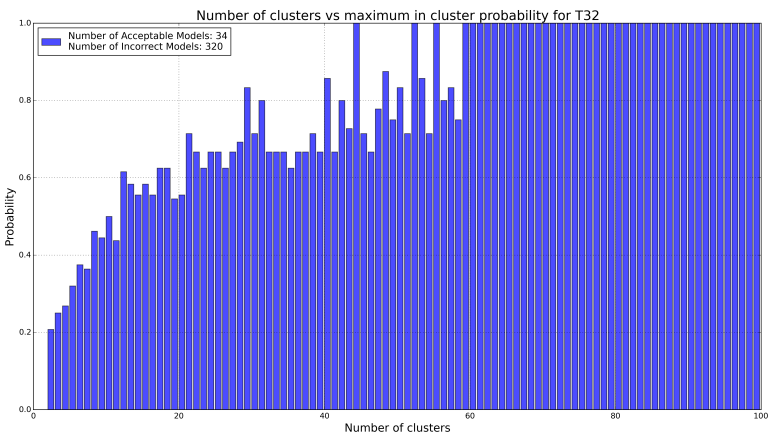

Figure B.2. Number of clusters versus in cluster probability for Target 32.

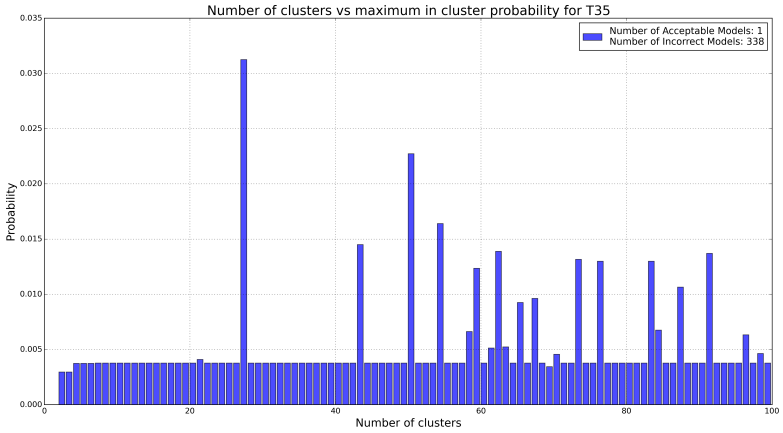

Figure B.3. Number of clusters versus in cluster probability for Target 35.

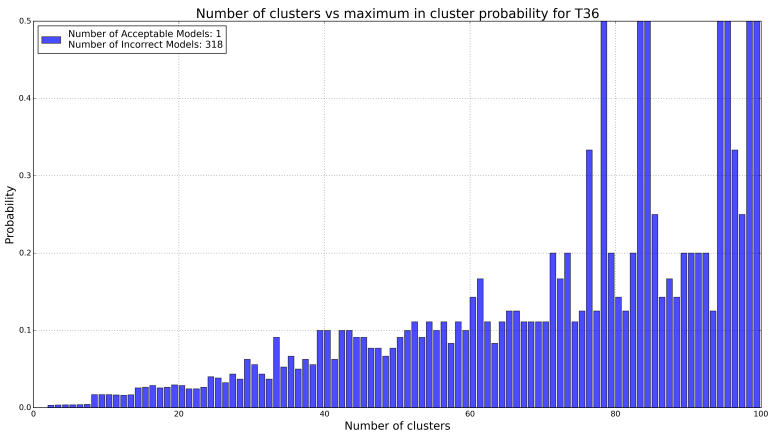

Figure B.4. Number of clusters versus in cluster probability for Target 36.

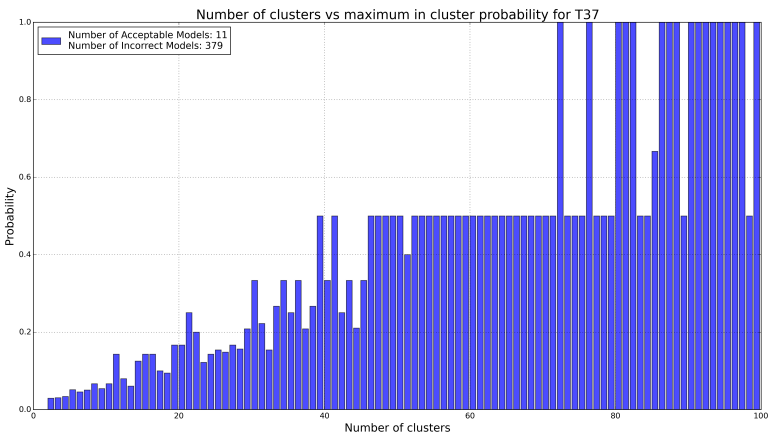

Figure B.5. Number of clusters versus in cluster probability for Target 37.

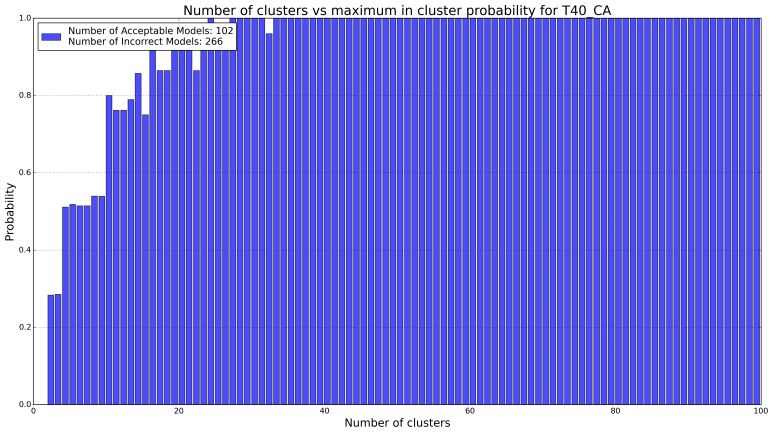

Figure B.6. Number of clusters versus in cluster probability for Target 40_CA.

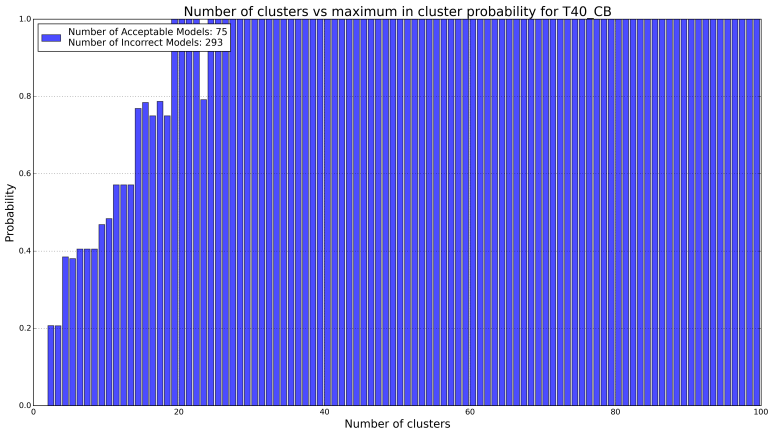

Figure B.7. Number of clusters versus in cluster probability for Target 40_CB.

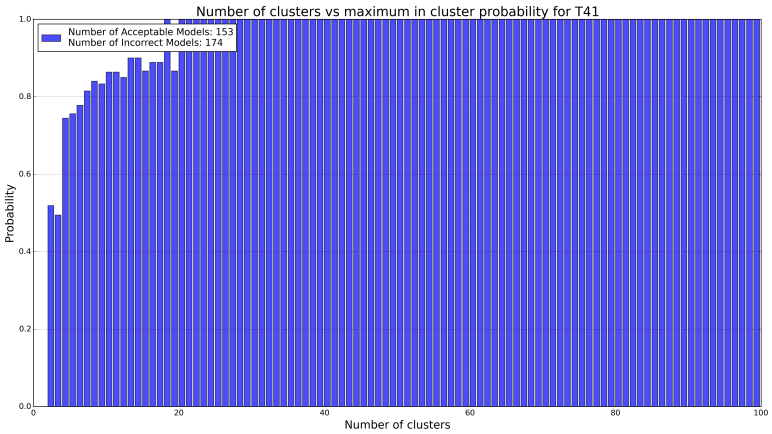

Figure B.8. Number of clusters versus in cluster probability for Target 41.

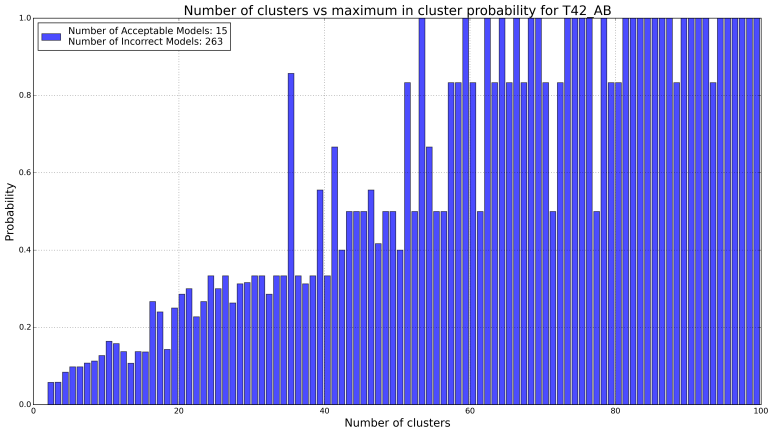

Figure B.9. Number of clusters versus in cluster probability for Target 42_AB.

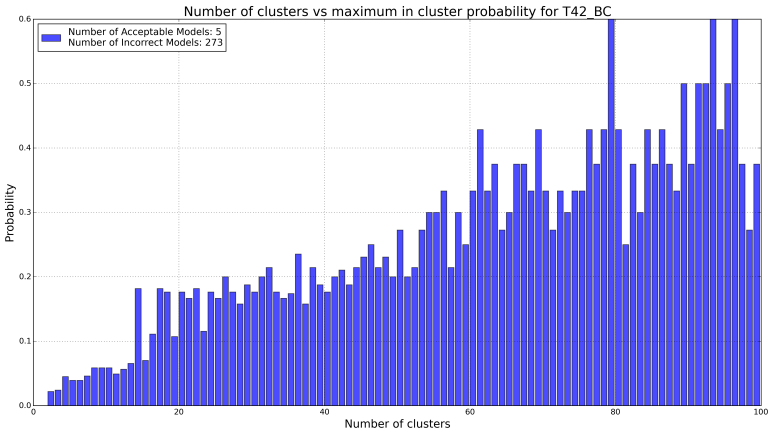

Figure B.10. Number of clusters versus in cluster probability for Target 42_BC.

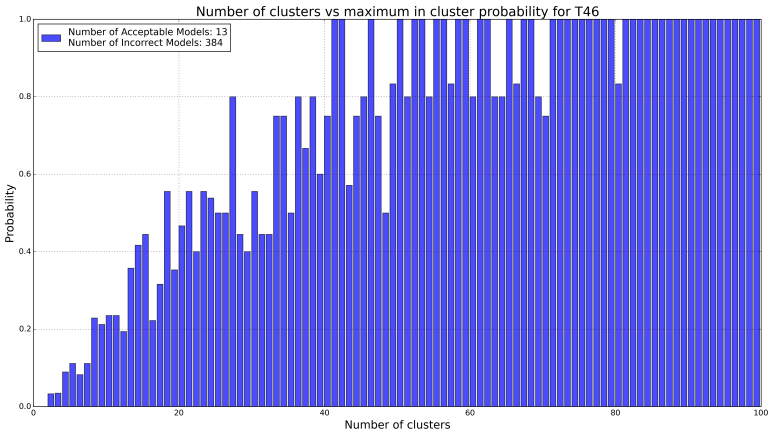

Figure B.11. Number of clusters versus in cluster probability for Target 46.

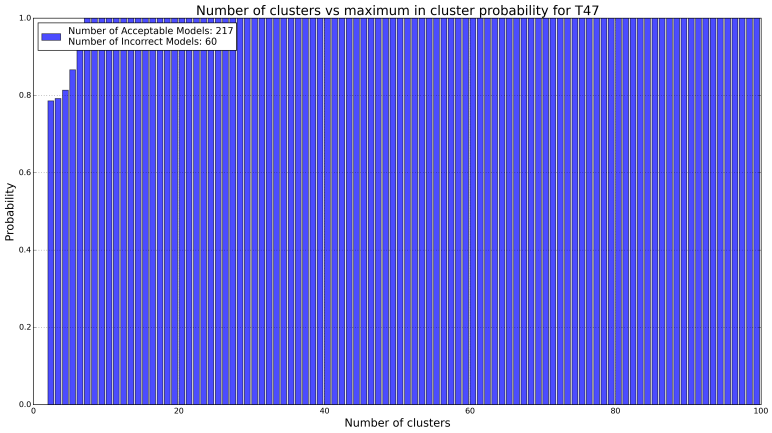

Figure B.12. Number of clusters versus in cluster probability for Target 47.

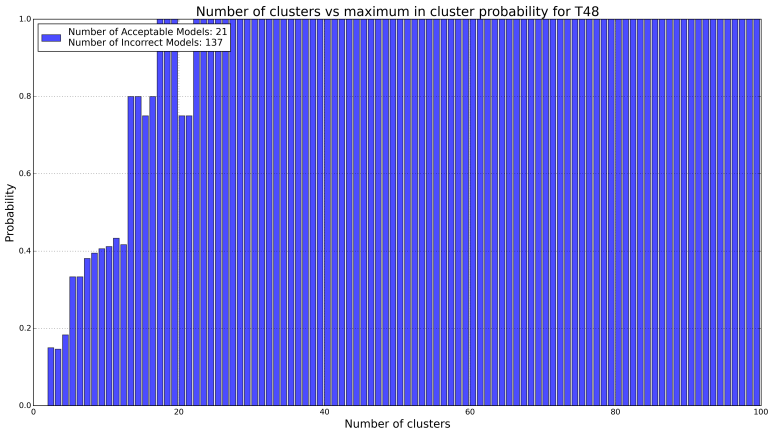

Figure B.13. Number of clusters versus in cluster probability for Target 48.

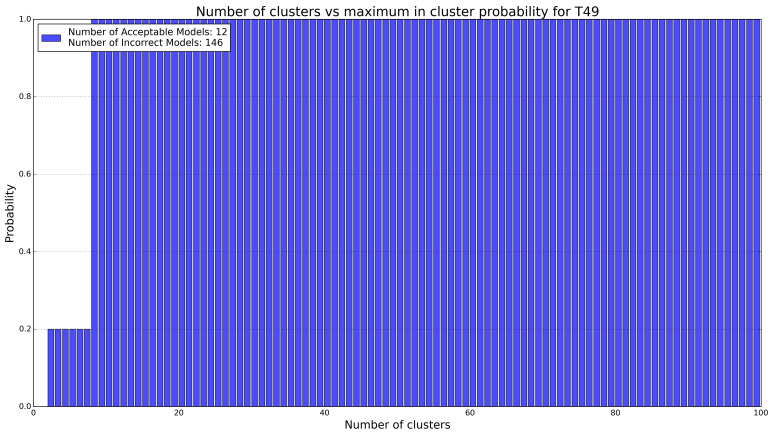

Figure B.14. Number of clusters versus in cluster probability for Target 49.

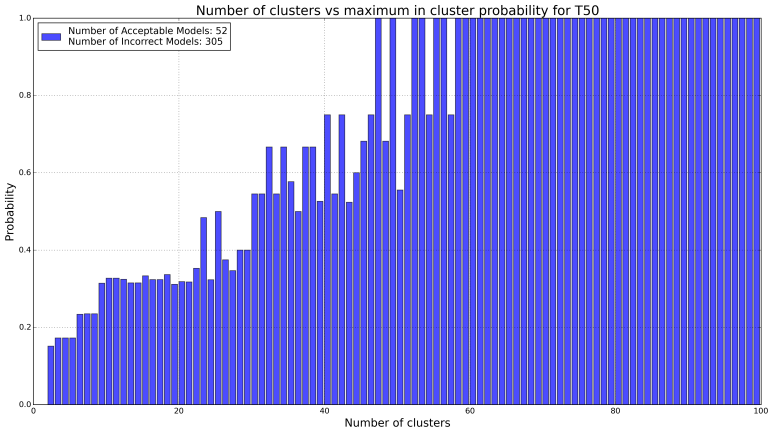

Figure B.15. Number of clusters versus in cluster probability for Target 50.

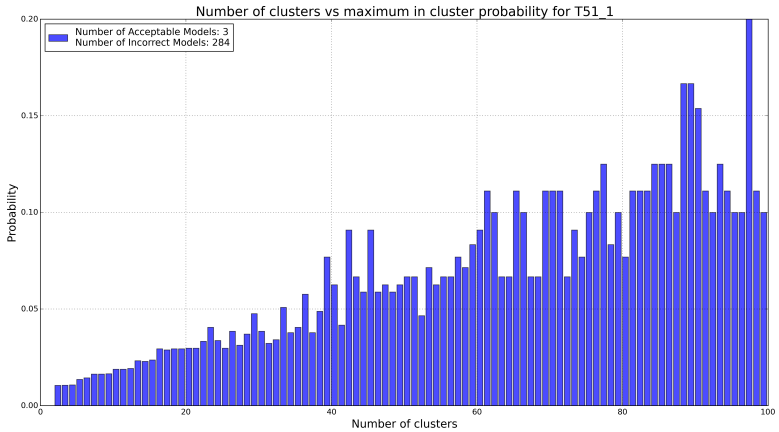

Figure B.16. Number of clusters versus in cluster probability for Target 51_1.

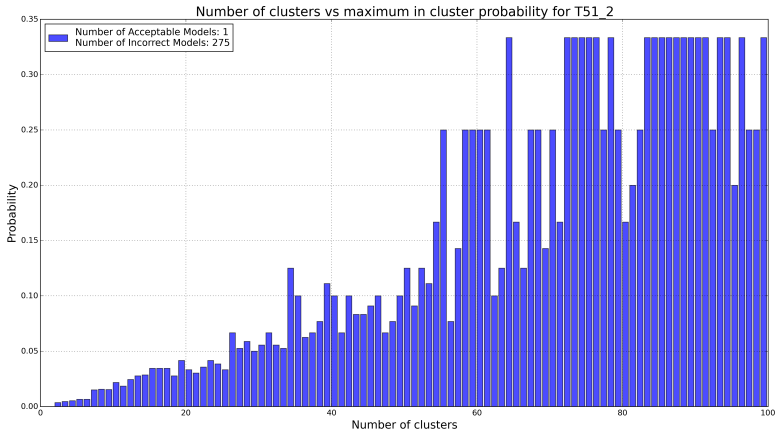

Figure B.17. Number of clusters versus in cluster probability for Target 51_2.

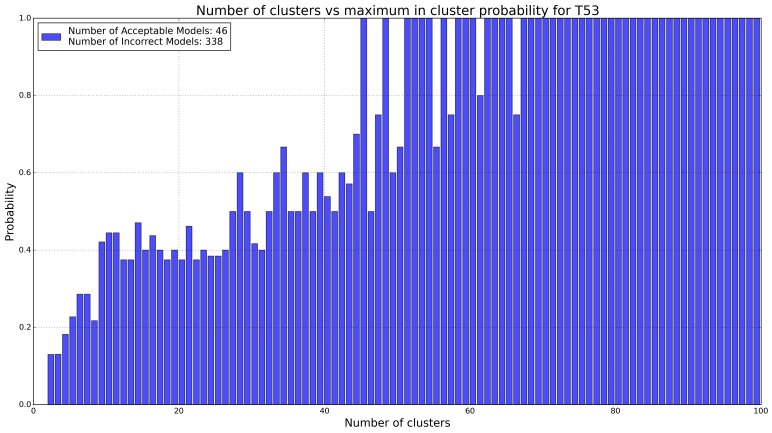

Figure B.18. Number of clusters versus in cluster probability for Target 53.

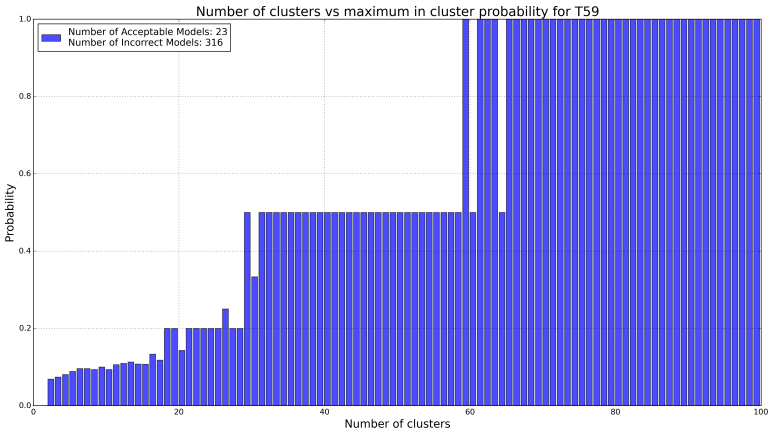

Figure B.19. Number of clusters versus in cluster probability for Target 59.

APPENDIX C: GNM Collectivities of Benchmark Complexes

Figures below present the distribution of collectivities of first and second binding partners of complexes in Docking Benchmark. Bound and unbound collectivities of binding partners are given as histograms and their probability density functions (using kernel density estimation) are presented.

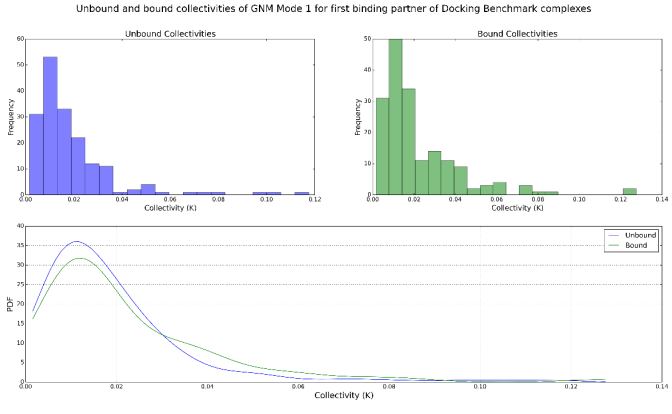

Figure C.1. Distributions and probability density functions of bound and unbound collectivities for GNM mode 1 of first binding partners in Docking Benchmark.

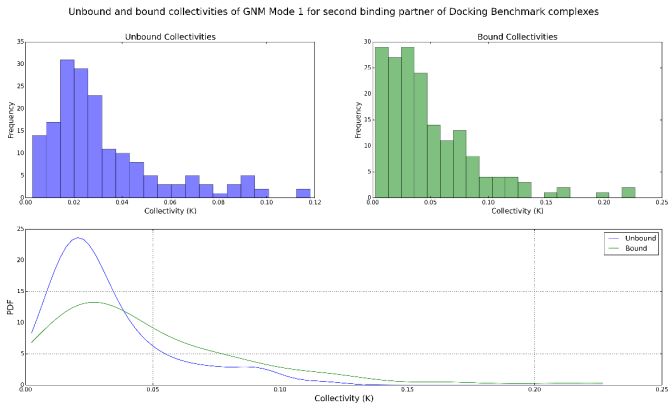

Figure C.2. Distributions and probability density functions of bound and unbound collectivities for GNM mode 1 of first second partners in Docking Benchmark.

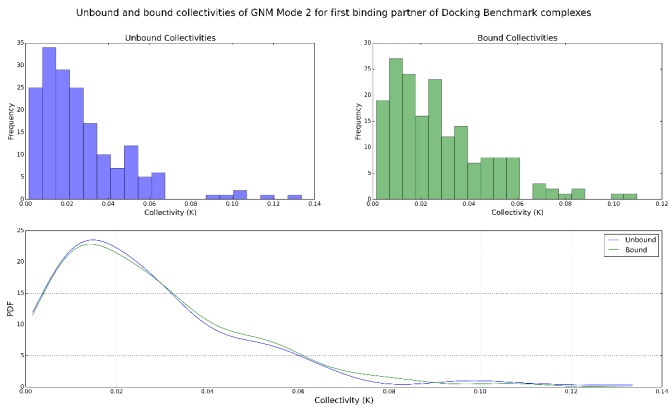

Figure C.3. Distributions and probability density functions of bound and unbound collectivities for GNM mode 2 of first binding partners in Docking Benchmark.

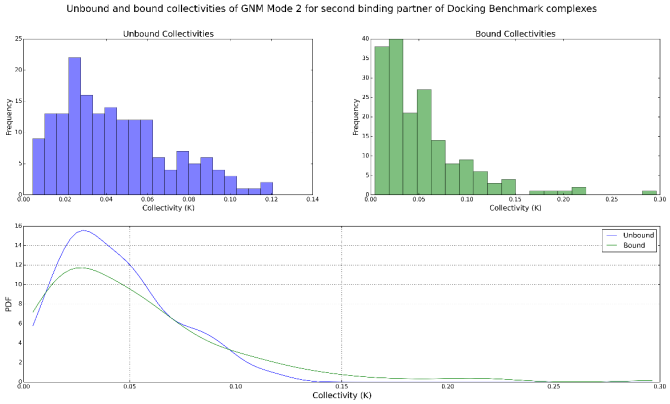

Figure C.4. Distributions and probability density functions of bound and unbound collectivities for GNM mode 2 of first second partners in Docking Benchmark.

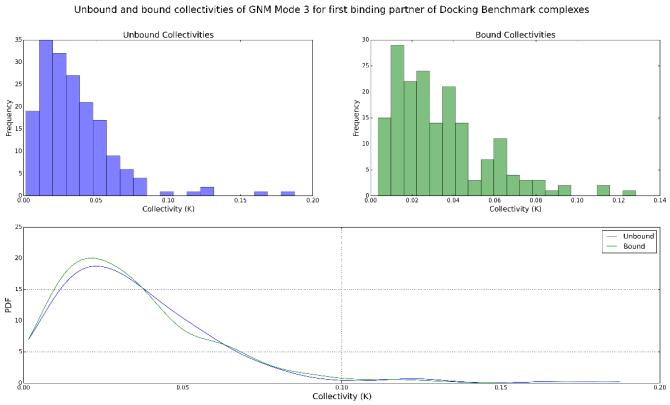

Figure C.5. Distributions and probability density functions of bound and unbound collectivities for GNM mode 3 of first binding partners in Docking Benchmark.

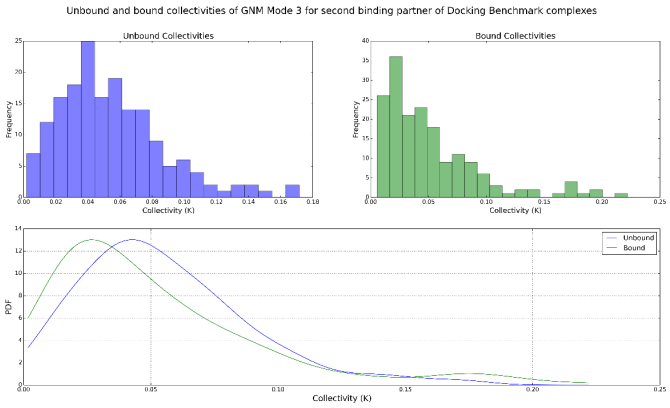

Figure C.6. Distributions and probability density functions of bound and unbound collectivities for GNM mode 3 of first second partners in Docking Benchmark.

APPENDIX D: ANM Collectivities of Benchmark Complexes

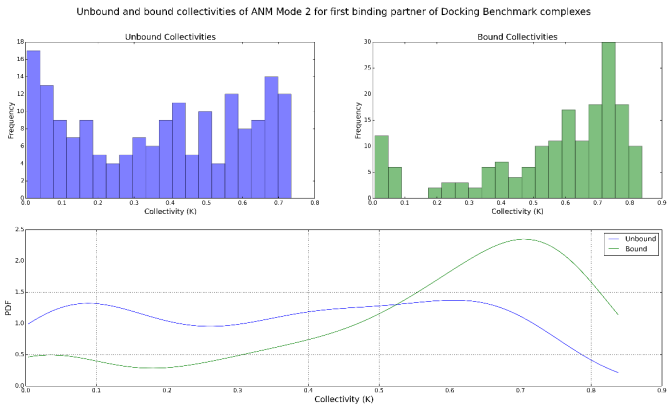

Figure D.1. Distributions and probability density functions of bound and unbound collectivities for ANM mode 2 of first binding partners in Docking Benchmark.

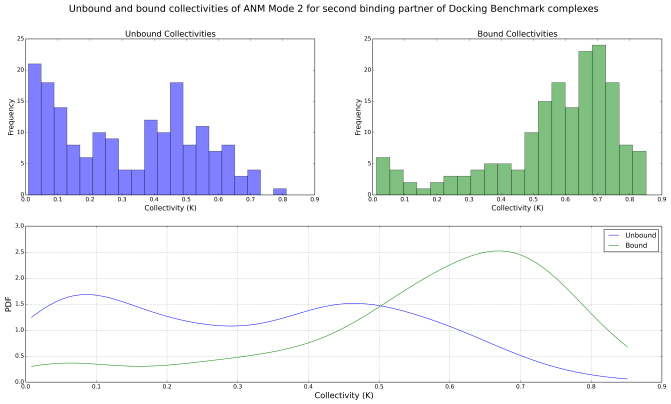

Figure D.2. Distributions and probability density functions of bound and unbound collectivities for ANM mode 2 of first second partners in Docking Benchmark.

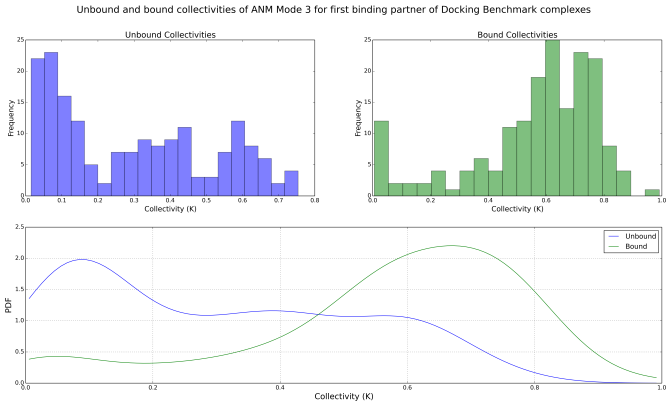

Figure D.3. Distributions and probability density functions of bound and unbound collectivities for ANM mode 3 of first binding partners in Docking Benchmark.

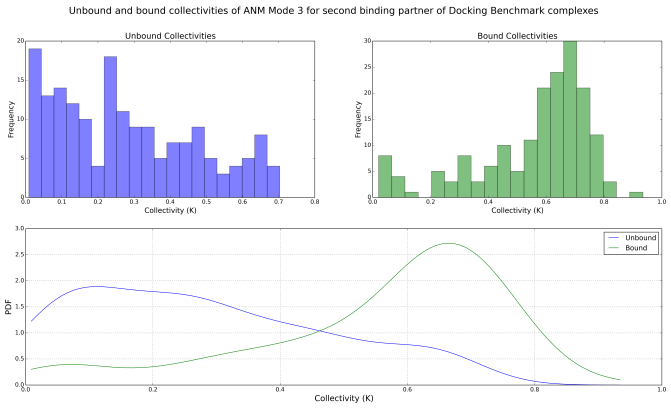

Figure D.4. Distributions and probability density functions of bound and unbound collectivities for ANM mode 3 of first second partners in Docking Benchmark.

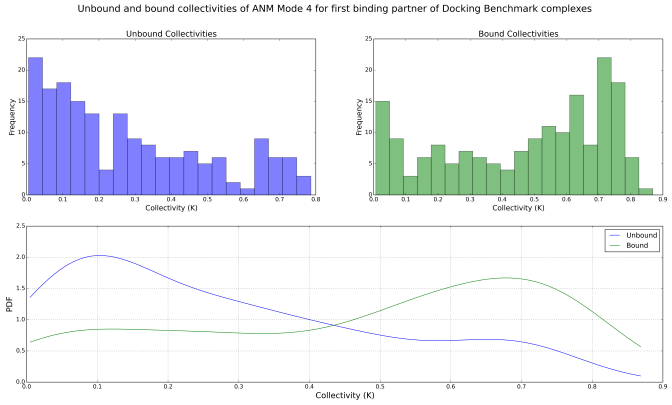

Figure D.5. Distributions and probability density functions of bound and unbound collectivities for ANM mode 4 of first binding partners in Docking Benchmark.

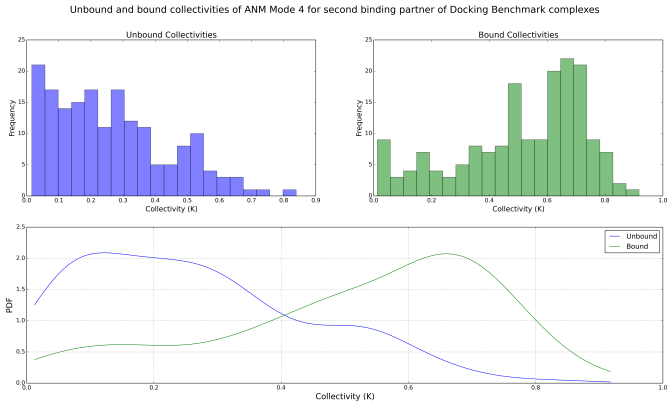

Figure D.6. Distributions and probability density functions of bound and unbound collectivities for ANM mode 4 of first second partners in Docking Benchmark.

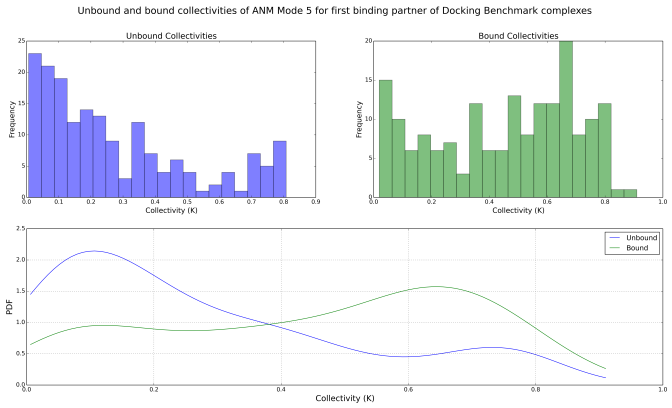

Figure D.7. Distributions and probability density functions of bound and unbound collectivities for ANM mode 5 of first binding partners in Docking Benchmark.

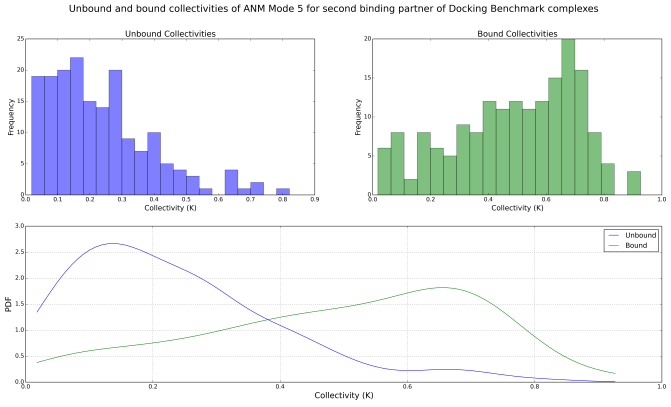

Figure D.8. Distributions and probability density functions of bound and unbound collectivities for ANM mode 5 of first second partners in Docking Benchmark.

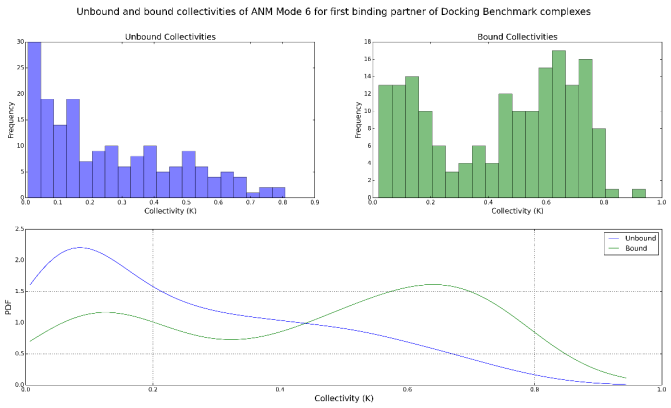

Figure D.9. Distributions and probability density functions of bound and unbound collectivities for ANM mode 6 of first binding partners in Docking Benchmark.

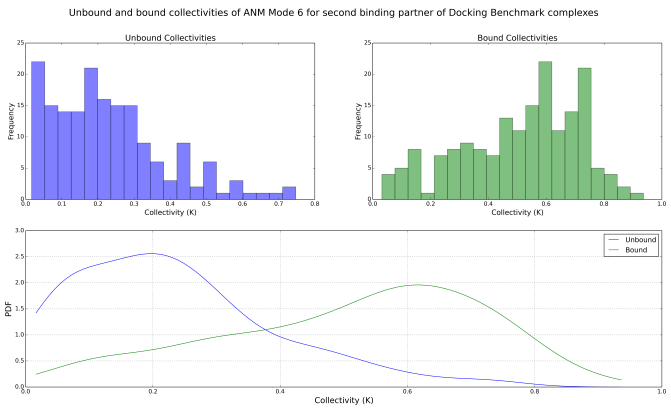

Figure D.10. Distributions and probability density functions of bound and unbound collectivities for ANM mode 6 of first second partners in Docking Benchmark.

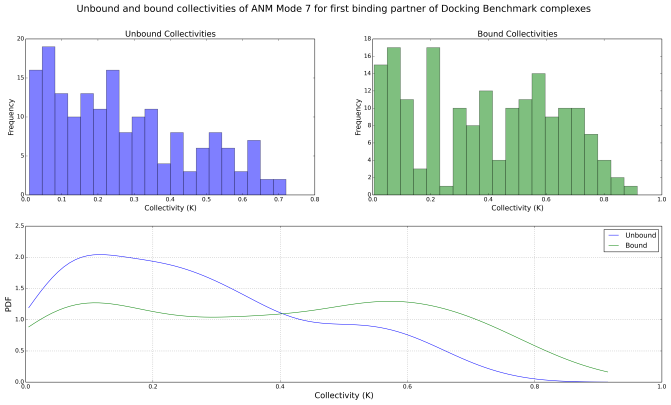

Figure D.11. Distributions and probability density functions of bound and unbound collectivities for ANM mode 7 of first binding partners in Docking Benchmark.

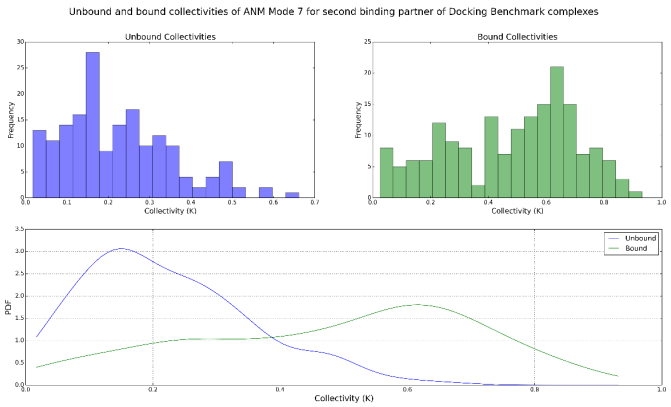

Figure D.12. Distributions and probability density functions of bound and unbound collectivities for ANM mode 7 of first second partners in Docking Benchmark.

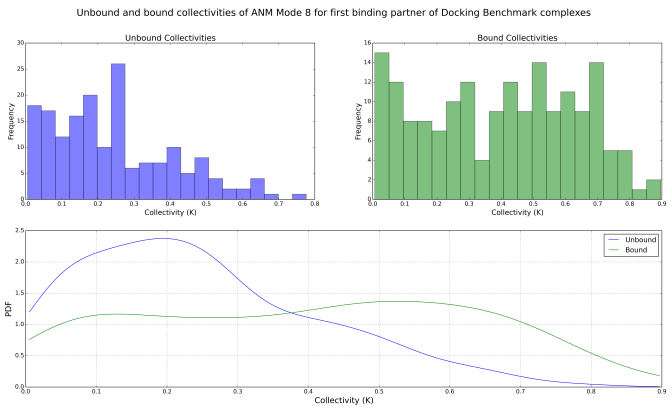

Figure D.13. Distributions and probability density functions of bound and unbound collectivities for ANM mode 8 of first binding partners in Docking Benchmark.

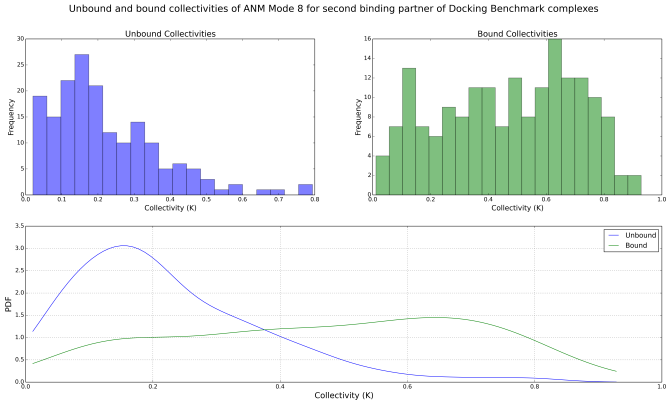

Figure D.14. Distributions and probability density functions of bound and unbound collectivities for ANM mode 8 of first second partners in Docking Benchmark.

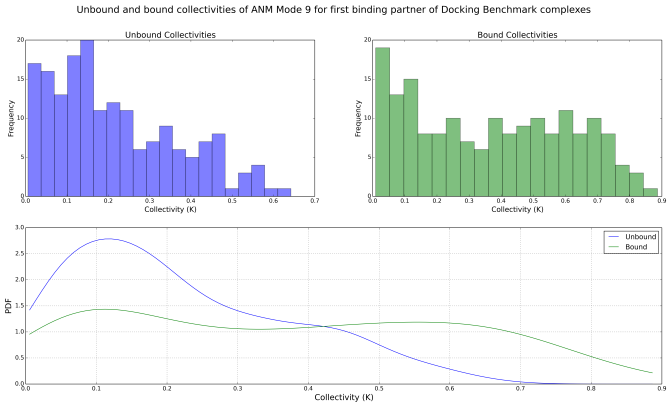

Figure D.15. Distributions and probability density functions of bound and unbound collectivities for ANM mode 9 of first binding partners in Docking Benchmark.

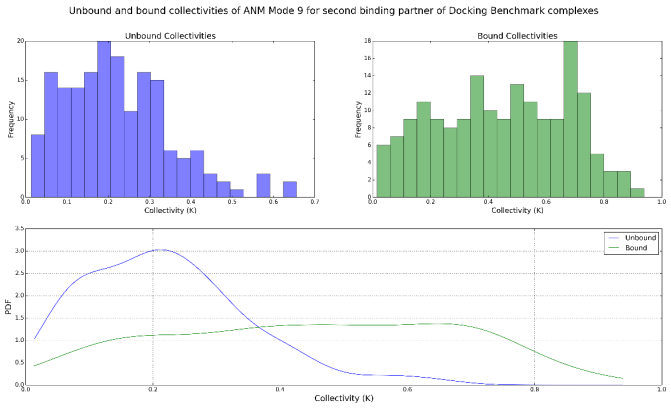

Figure D.16. Distributions and probability density functions of bound and unbound collectivities for ANM mode 9 of first second partners in Docking Benchmark.

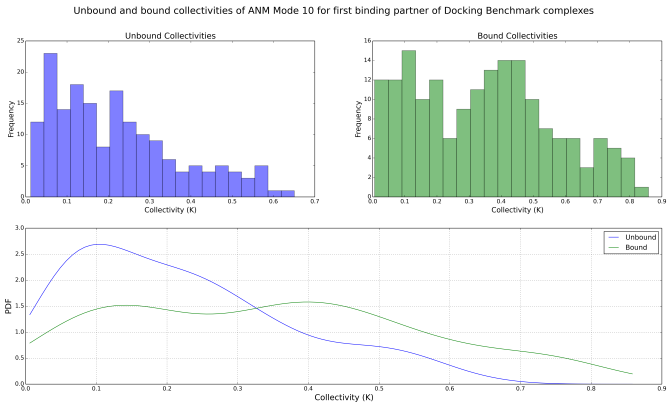

Figure D.17. Distributions and probability density functions of bound and unbound collectivities for ANM mode 10 of first binding partners in Docking Benchmark.

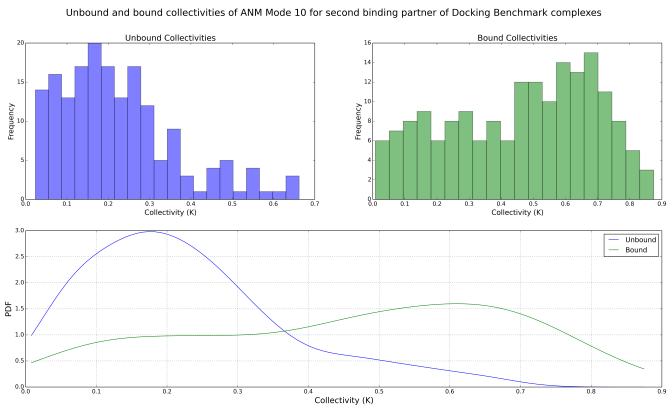

Figure D.18. Distributions and probability density functions of bound and unbound collectivities for ANM mode 10 of first second partners in Docking Benchmark.

REFERENCES

1. Chowdhury, R., M. Rasheed, D. Keidel, M. Moussalem, A. Olson, M. Sanner and C. Bajaj, "Protein-Protein Docking with F(2)Dock 2.0 and GB-rerank", *PloS one*, Vol. 8, No. 3, p. e51307, 2013.

2. Rose, P. W., A. Prlić, C. Bi, W. F. Bluhm, C. H. Christie, S. Dutta, R. K. Green, D. S. Goodsell, J. D. Westbrook, J. Woo, J. Young, C. Zardecki, H. M. Berman, P. E. Bourne and S. K. Burley, "The RCSB Protein Data Bank: Views of Structural Biology for Basic and Applied Research and Education", *Nucleic Acids Research*, Vol. 43, No. Database issue, pp. D345–56, 2014.

3. Huang, S.-Y., "Search Strategies and Evaluation in Protein-Protein Docking: Principles, Advances and Challenges", *Drug Discovery Today*, Vol. 19, No. 8, pp. 1081–1096, 2014.

4. Huang, S.-Y., S. Z. Grinter and X. Zou, "Scoring Functions and Their Evaluation Methods for Protein-Ligand Docking: Recent Advances and Future Directions", *Physical Chemistry Chemical Physics: PCCP*, Vol. 12, No. 40, pp. 12899–12908, 2010.

5. Janin, J., K. Henrick, J. Moult, L. T. Eyck, M. J. E. Sternberg, S. Vajda, I. Vakser and S. J. Wodak, "CAPRI: a Critical Assessment of PRedicted Interactions", *Proteins*, Vol. 52, No. 1, pp. 2–9, 2003.

6. Fleishman, S. J., T. a. Whitehead, E. M. Strauch, J. E. Corn, S. Qin, H. X. Zhou, J. C. Mitchell, O. N. a. Demerdash, M. Takeda-Shitaka, G. Terashi, I. H. Moal, X. Li, P. a. Bates, M. Zacharias, H. Park, J. S. Ko, H. Lee, C. Seok, T. Bourquard, J. Bernauer, A. Poupon, J. Azé, S. Soner, S. K. Ovali, P. Ozbek, N. B. Tal, T. Haliloglu, H. Hwang, T. Vreven, B. G. Pierce, Z. Weng, L. Pérez-Cano, C. Pons, J. Fernández-Recio, F. Jiang, F. Yang, X. Gong, L. Cao, X. Xu, B. Liu, P. Wang, C. Li, C. Wang, C. H. Robert, M. Guharoy, S. Liu, Y. Huang,

L. Li, D. Guo, Y. Chen, Y. Xiao, N. London, Z. Itzhaki, O. Schueler-Furman, Y. Inbar, V. Potapov, M. Cohen, G. Schreiber, Y. Tsuchiya, E. Kanamori, D. M. Standley, H. Nakamura, K. Kinoshita, C. M. Driggers, R. G. Hall, J. L. Morgan, V. L. Hsu, J. Zhan, Y. Yang, Y. Zhou, P. L. Kastritis, A. M. J. J. Bonvin, W. Zhang, C. J. Camacho, K. P. Kilambi, A. Sircar, J. J. Gray, M. Ohue, N. Uchikoga, Y. Matsuzaki, T. Ishida, Y. Akiyama, R. Khashan, S. Bush, D. Fouches, A. Tropsha, J. Esquivel-Rodríguez, D. Kihara, P. B. Stranges, R. Jacak, B. Kuhlman, S. Y. Huang, X. Zou, S. J. Wodak, J. Janin and D. Baker, "Community-wide Assessment of Protein-Interface Modeling Suggests Improvements to Design Methodology", *Journal of Molecular Biology*, Vol. 414, No. 2, pp. 289–302, 2011.

7. Keskin, O., A. Gursoy, B. Ma and R. Nussinov, "Principles of Protein-Protein Interactions: What are the Preferred Ways for Proteins to Interact?", *Chemical Reviews*, Vol. 108, No. 4, pp. 1225–1244, 2008.

8. Reichmann, D., O. Rahat, M. Cohen, H. Neuvirth and G. Schreiber, "The Molecular Architecture of Protein-Protein Binding Sites", *Current Opinion in Structural Biology*, Vol. 17, No. 1, pp. 67–76, 2007.

9. Nooren, I. M. and J. M. Thornton, "Structural Characterisation and Functional Significance of Transient Protein–Protein Interactions", *Journal of Molecular Biology*, Vol. 325, No. 5, pp. 991–1018, 2003.

10. Fink, F., S. Ederer and W. Gronwald, "Protein-Protein Interaction Analysis by Docking", *Algorithms*, Vol. 2, No. 1, pp. 429–436, 2009.

11. Wodak, S. J. and J. Janin, "Computer Analysis of Protein-Protein Interaction", *Journal of Molecular Biology*, Vol. 124, No. 2, pp. 323–342, 1978.

12. Janin, J., "Protein-Protein Docking Tested in Blind Predictions: The CAPRI Experiment", *Molecular Biosystems*, Vol. 6, No. 12, pp. 2351–2362, 2010.

13. Kastritis, P. L. and A. M. J. J. Bonvin, "Are Scoring Functions in Protein-Protein

Docking Ready to Predict Interactomes? Clues from a Novel Binding Affinity Benchmark", *Journal of Proteome Research*, Vol. 9, No. 5, pp. 2216–2225, 2010.

14. Halperin, I., B. Ma, H. Wolfson and R. Nussinov, "Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions", *Proteins*, Vol. 47, No. 4, pp. 409–43, 2002.

15. Hwang, H., T. Vreven, B. G. Pierce, J. H. Hung and Z. Weng, "Performance of ZDOCK and ZRANK in CAPRI rounds 13-19", *Proteins: Structure, Function and Bioinformatics*, Vol. 78, No. 15, pp. 3104–3110, 2010.

16. Katchalski-Katzir, E., I. Shariv, M. Eisenstein, a. a. Friesem, C. Aflalo and I. a. Vakser, "Molecular Surface Recognition: Determination of Geometric Fit Between Proteins and Their Ligands by Correlation Techniques", *Proceedings of the National Academy of Sciences of the United States of America*, Vol. 89, No. 6, pp. 2195–2199, 1992.

17. Chen, R. and Z. Weng, "Docking Unbound Proteins Using Shape Complementarity, Desolvation and Electrostatics", *Proteins: Structure, Function and Genetics*, Vol. 47, No. 3, pp. 281–294, 2002.

18. Gabb, H. a., R. M. Jackson and M. J. Sternberg, "Modelling Protein Docking Using shape Complementarity, Electrostatics and Biochemical Information", *Journal of Molecular Biology*, Vol. 272, No. 1, pp. 106–120, 1997.

19. Mandell, J. G., V. a. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny and L. F. Ten Eyck, "Protein Docking Using Continuum Electrostatics and Geometric Fit", *Protein Engineering*, Vol. 14, No. 2, pp. 105–113, 2001.

20. Vakser, I. a., "Evaluation of GRAMM Low-resolution Docking Methodology on the Hemagglutinin-Antibody Complex", *Proteins: Structure, Function and Genetics*, Vol. 29, No. SUPPL. 1, pp. 226–230, 1997.

21. Kozakov, D., R. Brenke, S. R. Comeau and S. Vajda, "PIPER: an FFT-based Protein Docking Program with Pairwise Potentials", *Proteins*, Vol. 65, No. 2, pp. 392–406, 2006.

22. Bajaj, C. L., R. Chowdhury and V. Siddahanavalli, "F2dock: Fast Fourier Protein-Protein Docking", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 8, No. 1, pp. 45–58, 2011.

23. Gray, J. J., S. Moughon, C. Wang, O. Schueler-Furman, B. Kuhlman, C. A. Rohl and D. Baker, "Protein–Protein Docking with Simultaneous Optimization of Rigid-body Displacement and Side-chain Conformations", *Journal of Molecular Biology*, Vol. 331, No. 1, pp. 281–299, 2003.

24. Zacharias, M., "Protein-Protein Docking with a Reduced Protein Model Accounting for Side-chain Flexibility", *Protein Science: A Publication of the Protein Society*, Vol. 12, No. 6, pp. 1271–82, 2003.

25. Dominguez, C., R. Boelens and A. M. J. J. Bonvin, "HADDOCK: a Protein-Protein Docking Approach Based on Biochemical or Biophysical Information", *Journal of the American Chemical Society*, Vol. 125, No. 7, pp. 1731–7, 2003.

26. Kuntz, I. D., J. M. Blaney, S. J. Oatley, R. Langridge and T. E. Ferrin, "A Geometric Approach to Macromolecule-Ligand Interactions", *Journal of Molecular Biology*, Vol. 161, No. 2, pp. 269–288, 1982.

27. Schneidman-Duhovny, D., Y. Inbar, R. Nussinov and H. J. Wolfson, "PatchDock and SymmDock: Servers for Rigid and Symmetric Docking", *Nucleic Acids Research*, Vol. 33, No. Web Server issue, pp. W363–7, 2005.

28. Venkatraman, V., Y. D. Yang, L. Sael and D. Kihara, "Protein-Protein Docking Using Region-based 3D Zernike Descriptors", *BMC Bioinformatics*, Vol. 10, No. 1, p. 407, 2009.

29. Gardiner, E. J., P. Willett and P. J. Artymiuk, "Protein Docking Using a Genetic Algorithm", *Proteins*, Vol. 44, No. 1, pp. 44–56, 2001.

30. Janin, J., "Assessing Predictions of Protein-Protein Interaction: The CAPRI Experiment", *Protein Science: A Publication of the Protein Society*, Vol. 14, No. 2, pp. 278–283, 2005.

31. Vajda, S. and D. Kozakov, "Convergence and Combination of Methods in Protein-Protein Docking", *Current Opinion in Structural Biology*, Vol. 19, No. 2, pp. 164–170, 2009.

32. Kastritis, P. L., I. H. Moal, H. Hwang, Z. Weng, P. a. Bates, A. M. J. J. Bonvin and J. Janin, "A Structure-Based Benchmark for Protein-Protein Binding Affinity", *Protein Science*, Vol. 20, No. 3, pp. 482–491, 2011.

33. Neudert, G. and G. Klebe, "DSX: A Knowledge-Based Scoring Function for the Assessment of Protein-Ligand Complexes", *Journal of Chemical Information and Modeling*, Vol. 51, No. 10, pp. 2731–2745, 2011.

34. Adcock, S. A. and J. A. McCammon, "Molecular Dynamics: Survey of Methods for Simulating the Activity of Proteins", *Chemical Reviews*, Vol. 106, No. 5, pp. 1589–615, 2006.

35. Haliloglu, T., I. Bahar and B. Erman, "Gaussian Dynamics of Folded Proteins", *Physical Review Letters*, Vol. 79, No. 16, pp. 3090–3093, 1997.

36. Emekli, U., D. Schneidman-Duhovny, H. J. Wolfson, R. Nussinov and T. Haliloglu, "HingeProt: Automated Prediction of Hinges in Protein Structures", *Proteins*, Vol. 70, No. 4, pp. 1219–27, 2008.

37. Li, H., S. Sakuraba, A. Chandrasekaran and L.-W. Yang, "Molecular Binding Sites are Located Near the Interface of Intrinsic Dynamics Domains (IDDs)", *Journal of Chemical Information and Modeling*, Vol. 54, No. 8, pp. 2275–85, 2014.

38. Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin and I. Bahar, "Anisotropy of Fluctuation Dynamics of Proteins with an Elastic Network Model", *Biophysical Journal*, Vol. 80, No. 1, pp. 505–15, 2001.

39. Eyal, E., L.-W. Yang and I. Bahar, "Anisotropic Network Model: Systematic Evaluation and a New Web Interface", *Bioinformatics*, Vol. 22, No. 21, pp. 2619–2627, 2006.

40. Tama, F. and Y. H. Sanejouand, "Conformational Change of Proteins Arising from Normal Mode Calculations", *Protein Engineering*, Vol. 14, No. 1, pp. 1–6, 2001.

41. Ming, D. and R. Brüschweiler, "Reorientational Contact-Weighted Elastic Network Model for the Prediction of Protein Dynamics: Comparison with NMR Relaxation", *Biophysical Journal*, Vol. 90, No. 10, pp. 3382–3388, 2006.

42. Hwang, H., T. Vreven, J. Janin and Z. Weng, "Protein-Protein Docking Benchmark Version 4.0", *Proteins*, Vol. 78, No. 15, pp. 3111–4, 2010.

43. *CAPRI Web Page EBI*, 2015, `http://www.ebi.ac.uk/msd-srv/capri/`, June 2015.

44. van der Walt, S., S. C. Colbert and G. Varoquaux, "The NumPy Array: A Structure for Efficient Numerical Computation", *Computing in Science & Engineering*, Vol. 13, No. 2, pp. 22–30, 2011.

45. Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg and Others, "Scikit-learn: Machine Learning in Python", *The Journal of Machine Learning Research*, Vol. 12, pp. 2825–2830, 2011.

46. McKinney, W., "Data Structures for Statistical Computing in Python", S. van der Walt and J. Millman (Editors), *Proceedings of the 9th Python in Science Conference*, pp. 51–56, 2010.