A Computational Study Investigating the Communication Network within Proteins and the Change in their Dynamics upon Ligand Binding, Mutation and Post-translational Modifications

by

Aysima Hacısüleyman

A Dissertation Submitted to the Graduate
School of Sciences and Engineering in Partial
Fulfillment of the Requirements for
the Degree of

Doctor of Philosophy

in

Chemical and Biological Engineering

KOÇ ÜNİVERSİTESİ

December 5, 2019

A Computational Study Investigating the communication network within proteins and the change in their Dynamics upon ligand binding, mutation and post translational modifications

Koç University

Graduate School of Sciences and Engineering

This is to certify that I have examined this copy of a doctoral dissertation by

Aysima Hacısüleyman

and have found that it is complete and satisfactory in all respects, and
that any and all revisions required by the final
examining committee have been made.

Committee Members:

____

Prof. Burak Erman

____

Prof. Halil Kavaklı

____

Assoc. Prof. Alkan Kabakçıoğlu

____

Prof. Türkan Haliloğlu

____

Asst. Prof. Mert Gür

Date: ________________________________

ABSTRACT

A Computational Study Investigating the Communication Network within Proteins and the Change in their Dynamics upon Ligand Binding, Mutation and Post-translational Modifications

Aysima Hacısüleyman

Doctor of Philosophy in Chemical and Biological Engineering

December 5, 2019

The thesis analyses the dynamic and structure changes in biomolecules upon ligand or protein binding, mutation and post translational modifications such as phosphorylation or methylation by using computational methods. The computational methods utilized throughout this study are biophysics based; harmonic interaction and molecular dynamics (MD)-based approaches.

The entropy transfer concept is used to detect allosteric communication network in proteins in first chapter and the driver-driven relationships between residue pairs are introduced, by using Molecular Dynamics based and Gaussian Network Model based approaches.

In second chapter, a special derivative of antibodies produced by camelids named as nanobodies are optimized for humans and specific design criteria and methods are introduced to optimize a given nanobody to a specific antigen.

Longer simulations must be conducted in order to get reliable results from a Molecular dynamics simulation. To enhance sampling by running several short MD simulations, most important motions must be filtered out by removing the noise in the simulations by a Latent Semantic Indexing analysis. This method is used to enhance conformational sampling quality from three short MD simulations of MEK1 protein and residue fluctuations of the filtered and unfiltered trajectories are compared.

In the following chapter the activation mechanism of MEK1 protein upon Raf binding and following phosphorylation are investigated. The effect of ligand binding and mutation on MEK1 dynamics are revealed and the effects of these events on MEK1 and dual phosphorylated MEK1 are compared.

In the final chapter the effect of DNA methylation on methyl binding protein affinities are investigated and the effect of hydration of the binding site is explained. MeCP2 and MBD2 proteins are used to investigate these effects and the binding affinities are compared by using Steered Molecular Dynamics simulations.

Biology adapted itself to computer and computational methods makes biological concepts easily testable. With the integration of mathematics, physics, statistics and computational methods the ideas will shape into hypotheses and large amount of data produced by computational methods will direct and shape experimental studies.

ÖZETÇE

Proteinler İçindeki İletişim Ağını ve Ligand Bağlama, Mutasyon ve Çeviri Sonrası Değişiklikler Üzerindeki Dinamiklerindeki Değişimi İnceleyen Hesaplamalı Bir Çalışma

Aysima Hacısüleyman

Kimya ve Biyoloji Mühendisliği, Doktora
5 Aralık 2019

Bu tez, ligand veya protein bağlanması, mutasyon ve fosforilasyon veya metilasyon gibi translasyon sonrası değişiklikler üzerine hesaplanmış metotlar kullanılarak biyomoleküllerdeki dinamik ve yapı değişikliklerini analiz eder. Bu çalışma boyunca kullanılan hesaplama yöntemleri biyofizik tabanlıdır; harmonik etkileşim ve moleküler dinamik (MD) temelli yaklaşımlar kullanılmıştır.

Birinci bölümde, entropi transfer kavramı, proteinlerdeki allosterik iletişim ağını saptamak için kullanılmıştır ve Moleküler Dinamik tabanlı ve Gaussian Ağ Modeli tabanlı yaklaşımlar kullanılarak artık çiftler arasındaki sürücü odaklı ilişkiler tanıtılmıştır.

İkinci bölümde, nano gövdeler olarak adlandırılan devegiller tarafından üretilen antikorların özel bir türevi olan nanobodyler insanlar için optimize edilmiştir ve belirli bir nanobody'i belirli bir antijene optimize etmek için spesifik tasarım kriterleri ve yöntemleri getirilmiştir.

Moleküler dinamik bir simülasyondan güvenilir sonuçlar almak için daha uzun simülasyonlar yapılmalıdır. Birkaç kısa MD simülasyonu çalıştırarak örneklemeyi geliştirmek için, simülasyonlardaki gürültüyü bir Latent Semantic Indexing analizi ile gidererek en önemli hareketlerin filtrelenmesi gerekir. Bu yöntem MEK1 proteininin üç kısa MD simülasyonundan konformasyonel örnekleme kalitesini arttırmak için kullanılır ve filtrelenmiş ve filtrelenmemiş yörüngelerin kalıntı dalgalanmaları karşılaştırılır.

Takip eden bölümde Raf bağlama ve takip eden fosforilasyon olayından sonra MEK1 proteininin aktivasyon mekanizması incelenmiştir. Ligand bağlanmasının ve mutasyonun MEK1 dinamiği üzerindeki etkisi ortaya çıkarılmış ve bu olayların MEK1 ve çift fosforile edilmiş MEK1 üzerindeki etkileri karşılaştırılmıştır.

Son bölümde, DNA metilasyonunun metil bağlayıcı protein afiniteleri üzerindeki etkisi araştırılmış ve bağlama bölgesinin hidrasyonunun etkisi açıklanmıştır. MeCP2 ve MBD2 proteinleri bu etkileri araştırmak için kullanılır ve bağlanma afiniteleri Steered Molecular Dynamics simülasyonları kullanılarak karşılaştırılır.

Biyoloji kendini bilgisayara adapte etmiştir ve hesaplamalı yöntemler biyolojik kavramları kolayca test edilebilir hale getirmiştir. Matematik, fizik, istatistik ve hesaplama yöntemlerinin entegrasyonu ile fikirler hipotezler halinde şekillenecek ve hesaplama yöntemleri ile üretilen büyük miktarda veri deneysel çalışmaları yönlendirecek ve şekillendirecektir.

To my family

and

my caring advisor

ACKNOWLEDGMENTS

The work in this thesis and the knowledge I have gained within the past few years would have been impossible without the help of my advisor Burak Erman. It is a pleasure to express my gratitude to my guide and mentor. His dedication, knowledge, experience and consistent attitude to help everyone around him had been the major support for me to complete my work. I feel grateful and very lucky to be one of his students and trained by him. I have grown in many different aspects of my life other than my scientific and scholar aspects thanks to him. I feel honored to be his last PhD student and I am very lucky to take the Drug Design course in 2013 and met this topic. Burak Erman's lecture inspired me to follow his footsteps and continue in this path for the rest of my life. He really is an inspiration of a lifetime.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Å : : ångström 14, 15, 35, 75, 87, 88, 95, 97, 99, 100, 120, 121, 124, 125, 126, 127, 129, 149, 154, 157, 168, 169, 170, 172, 173, 174, 175, 177, 178, 179, 180, 181, 182, 183, 184, 185

CAC : : Cytosine - Adenine - Cytosine xi, xii, 26, 167, 168, 169, 170, 171, 176, 181, 183, 184, 185, 186

CD : : Cooper and Dryden 29

CG : : Cytosine - Guanine xi, xii, 16, 25, 26, 167, 168, 169, 170, 171, 173, 178, 181, 182, 183, 184, 185, 186

CpG : : 5'—C—phosphate—G—3 165

CPMG : :Carr-Purcell-Meiboom-Gill 26

DFG : : Asp-Phe-Gly 146, 148, 162, 164

DFMD : : Digitally Filtered Molecular Dynamics 20

FBP : : Fructose 1,6-bisphosphate 74

HTMD : : High-Temperature Molecular Dynamics 20

Ig : : Immunoglobulin 107

LSI : : Latent Semantic Indexing 24, 136, 137

MEK : : Mitogen-activated protein kinase kinase 24, 143, 144, 145, 147, 152, 153, 155, 160, 162, 165

MIST : : Mutual Information Spanning Trees 43

MTS : : Multiple time step 20

NAMD : : Nanoscale Molecular Dynamics 35, 112, 148, 149, 168, 169

NMR : Nuclear magnetic resonance 20, 26, 34, 55, 56, 58, 69, 86

NPT : : isobaric-isothermic ensemble 35, 149

REMD : : Replica Exchange Molecular Dynamics 20

RNA : : Ribonucleic acid 21

sdAb : : single domain Antibody database 111

SMD : : Steered Molecular Dynamics 26, 112, 115, 120, 124, 126, 130, 131, 132, 133, 134, 149, 168, 169, 170, 171, 172, 173, 174, 175, 177, 178, 179, 180, 184, 185

SVD : : Singular value decomposition 24, 136, 137, 139

TRD : : Transcription repression domain 25, 166

UBQ : : Ubiquitin 59

VdW : : van der Waals 112

WT : : Wild type xii, 14, 15, 121, 125, 127, 128, 129, 130, 131, 132, 133, 134, 147, 161, 162

CHAPTER 1.
INTRODUCTION

Proteins are the complex dynamic building blocks of a biological system and unraveling how these building blocks are organized and function is the key to understand the working pattern of the system as a whole. The real complexity of proteins arises from their dynamics. They consist of a combination of amino acid residues, where their tertiary structure determines their function, but there is not a single conformation encoded in their primary amino acid sequence that determines their function, instead, proteins can adopt an ensemble of conformations since they consist of amino acid residues that are in constant communication. (Bettati, Luque, & Viappiani, 2011; Jha & Udgaonkar, 2010; Münz & University of Oxford, 2012) Proteins are flexible systems that can constantly vary their shape by their atomic fluctuations or large domain motions under varying environmental conditions. The ability to adopt different shapes is a key in the function of some enzymes, signaling and transport proteins. After synthesis, different conformations of a protein exist in dynamical equilibrium with each other where all sub-states reside in different energy well. Transition rates between different conformations depend on the height of each energy well, which is also called the energy barrier. Small and large amplitude transitions between energy wells are described as local motions observed on pico to nanosecond timescales and collective motions observed on micro to millisecond timescales respectively.(Jha & Udgaonkar, 2010) The transition is possible via atomic fluctuations of the protein which plays a vital role in intra and inter signaling. Fluctuations without a significant conformational change by mutation, ligand binding, interaction with different binding partners or post-translational modifications such as acetylation, phosphorylation and methylation etc., may alter the dynamics and change the signal propagation within the protein. This is an intrinsic spatiotemporal property of all proteins called allostery.(Gunasekaran, Ma, & Nussinov, 2004; Münz & University of Oxford, 2012; Shin et al., 2011)

Experimental and computational methods allow us to study conformational dynamics and functional mechanisms of biomolecules. Experimental methods that allow us to study different conformations of biomolecules as snapshots are nuclear magnetic resonance (NMR) spectroscopy, cryo-electron microscopy (cryo-EM), X-ray crystallography and X-ray scattering. In addition to experimental techniques, computational methods allow us to explore protein dynamics in a more detailed way. Molecular dynamics (MD) simulations which produce continuous trajectories of biomolecules are extensively used to explore conformational dynamics with femtosecond time resolution. But since MD simulations are computationally intensive, instead of running longer simulations, several other simulation techniques are being used to accelerate and improve ensemble sampling. These methods are Replica Exchange Molecular Dynamics (REMD), Multiple Time-step methods (MTS), Digitally Filtered Molecular Dynamics (DFMD), Umbrella Sampling and High-Temperature Molecular Dynamics (HTMD). In addition to all these techniques the system can be further simplified by considering a group of atoms as a single particle. This approach is called coarse grained MD. This approach enables the simulation of larger biomolecular systems, but it is not as accurate as all atom MD simulations.

Intrinsic dynamics of proteins are encoded in their three-dimensional structure. There are methods available to identify functional motions of proteins based on normal mode analysis and elastic network models. In these methods proteins are represented as a network of beads connected with a spring. Gaussian network models are also classified as an elastic network model, studied at the residue level. Each residue is represented by its α-carbon atom (see a more detailed description of the model in Chapter 2).

Chapter 2 focuses on entropy transfer. Entropy transfer is a concept based on information theory, useful for quantifying correlation, causality relationship and coherence between parts of a dynamic system. This concept has been extensively used for estimating connectivity of neurons in neuroscience and dynamics of group and individual behaviors in social networks. Proteins are dynamic entities which fluctuate between different conformations. Each conformation holding a different functional significance, varying in their individual atomic fluctuations and side chain rotations. Quantifying and determining the communication patterns and amount of signal transmitted between parts of the protein are able to be distinguished by an information theoretic measure, entropy transfer. A model for quantifying the entropy transfer between residue pairs of proteins is developed.

Chapter 3 focuses on the design of nanobodies. Nanobodies are fragments of antibodies which consist mostly of the heavy chains. They are classified as therapeutic proteins. Their specificity and stability against heat over a pH range, are higher than the antibodies, these advantages encourage nanobody engineering for use as a promising research tool in drug design. Extracted shark or camelid RNAs can be easily expressed in microorganisms to generate over 10^{11} nanobodies. There are several methods available to screen nanobody libraries by phage display or any other protocol involving ribosome display, yeast display, bacterial display, intracellular 2 hybrid selection. The entire experimental procedure for screening of a nanobody library takes less than two months. Hundreds of antigens can be found each year and the experimental procedures are robust but there is no way of regulating or identifying which epitope will be targeted on any given antigen. There are computational methodologies available that will enable the rapid design and screening of custom nanobodies for any antigen targeted. Current computational studies focus on changing the complementary determining regions (CDR), which can allow binding and penetration or non-CDR parts on the nanobody. There is no certain rule or criterion on deciding which properties of the nanobodies to change and optimize their binding with their antigens. Combining the computational and experimental studies will provide a basis for the discovery of nanobody based pharmaceuticals and automation of the design procedure is required to generate unique and reliable antigen specific nanobodies. The design approach that we propose is based on improving the a given nanobody by mutating its important CDR residues, with computational techniques and optimizing the interaction with its antigen and improving its stability. As a result of the fast-computational protocol a library of several optimized structures will be proposed and characterized with ELISA based diagnostic assays. The techniques and approaches used in this design protocol will provide a rapid way of discovering nanobody-based biopharmaceuticals.

Chapter 4 focuses on the superposition of the molecular dynamics trajectories. Molecular dynamics simulations are functional tools for understanding the conformational patterns and properties of protein behavior, receptor-ligand interactions and the conformational changes that a biomolecule may undergo under various conditions. In vivo study of protein dynamics is very complicated and time consuming, therefore making use of simulations at different levels of detail and integrating them into the life sciences is a useful and powerful approach. A single MD simulation for a study is not enough to get a reliable result and several duplicate simulations of the same system must be conducted to replicate the same interactions and motions observed in a single MD run. In principle, MD is a method that samples the response of a biomolecule under a particular set of conditions, including random motions caused by thermal fluctuations. A simulation trajectory represents a sample of protein's behavior in its dynamics under a particular condition. The motions of biomolecules are observed to be reproducible, dominant patterns in conformational dynamics are frequently visited. Progression of motions in one MD simulation is random, duplicate trajectories of the same system may not yield in the same results. A common approach for comparing the trajectories of the same system is the cluster analysis, but setting up a suitable distance function, the routines for delimiting the clusters might cause problems in assessment of reliability. Several simulations may converge at different time steps due to the randomness of starting velocities and motions. This randomness causes a noise in the generated data, but dominant collective modes of the trajectories resulting from different MD runs will be similar since the statistical significance of local dynamics of the system stays the same. Main patterns of the data can be determined by reducing the dimension to the few, largest principal components. By applying an orthogonal linear transformation and transforming it to a new coordinate system, biologically meaningful signals can be extracted from the data. In order to be able to compare different trajectories, they need to be brought to a common ground by using a technique called Latent Semantic Indexing (LSI). It is an information retrieval method that uses singular value decomposition (SVD) to identify hidden patterns in the relationships between the concepts contained in a collection of terms by calculating the best rank-l approximation of the term-document matrix and reduces the dimensional space to eigen-term and eigen-document. Therefore, quantifying and comparing the similarity of local dynamics in a mode-based approach will be more precise to determine the quality and the accuracy of the simulation results, and will allow determining the most significant dynamic patterns of the system. This analysis method will allow us to detect document similarity, project different trajectories onto each other without perturbing the patterns of the system, determine the common dominant motions and filter out outlier motions caused by the randomness of protein motions and increase ensemble sampling.

Chapter 5 focuses on MEK1 protein involving in Ras/Raf/MEK/Erk pathway. Ras/Raf/MEK/Erk pathway is the most commonly activated cascade in cancers. Ras inhibitors are challenging to identify. More attention is being focused on the study of MEK inhibitors. The activation of MEK proteins occur upon phosphorylation by Raf proteins. Phosphorylation, ligand binding and mutation induces a conformational change and these events cause a population shift of fluctuations. This change plays a pivotal role in the regulation of protein activity. The biophysical characterization of these events is still not clearly understood. We investigate the effect of phosphorylation, mutation and ligand binding on the fluctuations, correlations, information transfer network changes and the activity of MEK, specifically MEK1, in complex with MgATP by incorporating GNM and MD.

Chapter 6 focuses on the effects of DNA methylation and mutation on protein-DNA binding. Methylation is a mechanism that occurs by the methyl group addition on the DNA molecule. This epigenetic mechanism is related to the gene regulation and disease. Methyl binding domain (MBD) proteins play an important role in DNA methylation process. Two MBD proteins are studied in human neoplasia, MeCP2 and MBD2. MeCP2 is the mostly studied MBD protein with 2 domains, 1 MBD and one transcriptional repression domain (TRD). MeCP2 overexpression causes Rett syndrome. It is a neurological syndrome caused by the mutations on the gene coding for methyl-CG binding protein 2, MeCP2. It is a transcriptional repressor which inhibits gene expression when there are DNA methylation and histone acetylation present. Methyl binding proteins are critical in the maintenance of central nervous system function. Mutations in MeCP2, lead to abnormal behavioral and neurophysiological phenotypes. MeCP2 is characterized by binding to the symmetrically methylated cytosines, mCG. It is also reported that MeCP2 binds to mCAC trinucleotides. Binding of MeCP2 to chromosal DNA is proportional to the mCAC + mCG density. DNA hypermethylation is associated with cancer and MBD proteins can bind methylated DNA and can aid in gene silencing. MBD2 preferentially bind to mCG-DNA from its TRD domain. The exact mechanism of MBD binding, the changes upon methylation are still not fully understood. (Stirzaker et al., 2017)

The aim of this study is to understand the changes in the affinity of MeCP2 and MBD2 binding to methylated and non-methylated CG and CAC portions of the DNA by incorporating MD simulation methods. The change in the binding affinities of these methyl binding proteins will be investigated by Steered Molecular dynamics (SMD) simulations and the effect of water molecules within the methyl binding protein binding region will be investigated.

CHAPTER 2.
ENTROPY TRANSFER

2.1 OVERVIEW

Allosteric communication is the process where action at one site of a protein is affected by an action at a distant site. This type of long-range communication is essential for protein function. It is known that the disruption of allosteric sites in a protein by a single point mutation results in cancer. The basic problem in determining the allosteric communication is to identify the residues that participate in allosteric communication. There are several approaches available for solving the problem(Dror et al., 2013; Hacisuleyman & Erman, 2017a, 2017b; Hardy & Wells, 2004; Kaya, Armutlulu, Ekesan, & Haliloglu, 2013; Lu, Huang, & Zhang, 2014; McClendon, Friedland, & Jacobson, 2010; Novinec et al., 2014; Panjkovich & Daura, 2014; Shukla, Meng, Roux, & Pande, 2014; Tang et al., 2007; Tsai & Nussinov, 2014). Allostery is an intrinsic spatiotemporal property of all proteins and it results from dynamic redistributions over different conformations. Carr–Purcell–Meiboom–Gill(CPMG) pulse sequence NMR spectroscopy measurements showed that dynamic redistribution controls allostery in a protein(Capdevila, Braymer, Edmonds, Wu, & Giedroc, 2017; Grutsch, Bruschweiler, & Tollinger, 2016; Kern & Zuiderweg, 2003). There is no first principle statistical mechanical model that describes the relationship between entropy transfer and dynamic transitions that result in allosteric communication. The aim of this study is to develop a rapid computational technique to detect interacting allosteric residue pairs based on information transfer and entropy concepts. As Feynman stated "everything that living things do can be understood in terms of the jiggling and wiggling of atoms" [Feynman, 1963]. By starting from this point of view we can say that communication patterns of atoms can be understood from their fluctuations. In theory, the fluctuations of two points that are actively communicating are correlated and the uncertainty in one trajectory decreases due to this correlation(Hacisuleyman & Erman, 2017a). This leads to transfer of entropy from one point to the other. Due to the transient nature of the correlations, entropy transfer is a function of time. In this study, we present two models that use 1) molecular dynamics trajectories(Hacisuleyman & Erman, 2017b) and 2) Gaussian Network Model, GNM(Hacisuleyman & Erman, 2017a), to characterize entropy transfer in proteins by employing the concept of entropy transfer introduced by Schreiber.(Schreiber, 2000b)

2.2 MEASURING ENTROPY TRANSFER WITH MOLECULAR DYNAMICS

2.2.1 INTRODUCTION

Allosteric communication describes the process in which action at one site of a protein is transmitted to another at which the protein performs its activity. The importance of allostery in biological systems has generated significant experimental and computational research. The basic problem is to identify residues that participate in allosteric communication in the hope of controlling their behavior related to protein function. Allosteric communication first requires the identification of two sites, the effector site, i.e., the site that is acted upon, and the regulatory site where protein's activity is regulated. Although more than 1000 allosteric sites are known (Huang et al., 2011) many more need to be characterized. In fact, several pairs of allosteric endpoints may exist in a protein (Barbany et al., 2015) which increases the number of candidate pairs that communicate. This problem becomes even more important when one considers the fact that most known cancers result from disruption of allosteric communication as a result of single mutations (Cerami et al., 2012; Henry, Bandrowski, Pepin, Gonzalez, & Desfeux, 2014) and the number of proteins associated with this phenomenon is very large. Expressed in simple terms, the solution of the problem reduces to finding whether two given residues communicate with each other, and if so what the consequences of this communication are. Various approaches to solve the problem may be found in References (Dror et al., 2013; Hacisuleyman & Erman, 2017a, 2017b; Hardy & Wells, 2004; Kaya et al., 2013; Lu et al., 2014; McClendon et al., 2010; Novinec et al., 2014; Panjkovich & Daura, 2014; Shukla et al., 2014; Tang et al., 2007; Tsai & Nussinov, 2014). The specific aim of the present paper is to develop a rapid computational technique that identifies interaction of residue pairs based on concepts of information transfer and entropy, to scan a given protein and identify pairs of sites that communicate and to determine whether these communicating pairs may be candidates for allosteric activity.

The present work departs from the approaches outlined in the preceding paragraph. We do not focus neither on single allosteric sites nor on allosteric paths. We consider the time trajectory of the fluctuations of two residues, which may be spatially distant, and search for information transfer from the trajectory of one residue to that of the other. The trajectories are obtained from long molecular dynamics (MD) equilibrium simulations that give the fluctuation of each atom at constant temperature. The first requirement for information to be transferred from an atom i to another atom, j, is that their trajectories should be correlated. The second requirement is that this transfer should be asymmetric, i.e., information going from i to j should not be equal to information from j to i. This requires the use of time delayed correlations of fluctuations which may be asymmetric in contrast to time independent correlations which are symmetric by definition and therefore lack information on directionality. If $C_{ij}(t,t+\tau)$ denotes the correlation of fluctuations of i at time t with those of j at time $t+\tau$, then asymmetry requires that $C_{ij}(t,t+\tau) \neq C_{ji}(t,t+\tau)$. This introduces directionality and therefore causality into the problem. If time delayed correlations are asymmetric, then can we quantify the net information that is transferred? The answer is yes if we pose the problem in terms of entropy transfer.

Before going into the discussion of entropy, it is worth pointing out that information transfer is exclusively based on the changes in the amplitudes and frequencies of fluctuations in the system. This was first suggested and modelled by Cooper and Dryden (CD) (Cooper & Dryden, 1984) and reached larger dimensions by the work of Gunasekaran (Gunasekaran et al., 2004) which suggests that since allosteric communication is a result of correlated fluctuations then allostery should be an intrinsic dynamic property of all proteins. The dynamics aspect of proteins resides in the fluctuations of atoms which may be evaluated by experimentally measuring the B-factors of the atoms. The B-factor of the i^{th} atom is related to its time independent autocorrelation of fluctuations, $\langle(\Delta R_i)^2\rangle$, by $B_i = \frac{8\pi^2}{3}\langle(\Delta R_i)^2\rangle$. However, knowledge of them is not sufficient for predicting allosteric communication and cross correlations, $\langle \Delta R_i \Delta R_j \rangle$, between the fluctuations of different atoms are needed. Allosteric activity requires not only the modulation of the cross correlations in the system but also on time delayed cross correlations, $\langle \Delta R_i(0) \Delta R_j(t) \rangle$, as will be described below in detail. The CD model is referred to as 'Allostery without conformational change'. In this respect, it goes beyond the classical Monod-Wyman-Changeux (MWC) (Monod, Wyman, & Changeux, 1965) model and its relative, the Koshland Nemethy Wyman (KNW) model (Koshland, Nemethy, & Filmer, 1966) both of which relate allostery to discrete conformational changes at the regulatory site. Sending information by changing the amplitude and frequencies of fluctuations is entropic (Motlagh, Wrabl, Li, & Hilser, 2014) and depends not only on the value of the entropy but also on the transfer of entropy from residue to residue during communication. Entropy as a source of information transfer is widely used in information theory (Schreiber, 2000b) which is only very recently used for a protein-DNA complex by Kamberaj and van der Vaart (Kamberaj & van der Vaart, 2009a). Through analysis of entropy transfer, they determined residues that act as drivers of the fluctuations of other residues, thereby determining causality that is inherent in the correlations. Determining residues that act as drivers and those that are driven is important especially from the point of view of drug design. Entropy transfer and causality is a new paradigm for studying allosteric communication in proteins, which we elaborate in detail in the present paper. On a broader scale, our findings show that all proteins may indeed exhibit allosteric communication and therefore supports the hypothesis by Gunasakaran, (Gunasekaran et al., 2004) which states that allostery is an intrinsic dynamic property of all proteins.

The quantitative measure of information flow between two correlated processes is introduced by Schreiber (Schreiber, 2000b) in 2000. In the present work, the processes are generated in the form of trajectories of atomic coordinates using MD simulations from which probabilities of atomic coordinate fluctuations required for evaluating transfer entropy are calculated. We calculate the entropies based on atoms and identify the entropy of a residue with the entropy of its alpha carbon. Denoting the probability of fluctuation of atom i by p_i, Callen showed (Callen, 1985) that the Shannon measure of disorder, $-k_B \sum_{i=1}^{N} p_i ln p_i$, with N and k_B denoting the number of elements of the system and the Boltzmann constant, is the entropy of the system which is maximized at constant energy (See Callen (Callen, 1985), Chapter 17. Entropy and disorder: Generalized canonical formulations). At this point we give a qualitative explanation of the relationship between information flow and a physical event such as fluctuations of atoms and continue this discussion on a quantitative way after we introduce the statistical mechanical basis of the model. Suppose we have two trajectories, one of atom *i* and the other of atom *j*. If the fluctuations of *i* and *j* are independent of each other, then knowledge of the fluctuations of *i* will not give us information on the fluctuations of *j* and the uncertainty associated with the two events will be a maximum. The total entropy of *i* and *j* will be the sum of the singlet entropies, S_i+S_j. If, on the other hand, *i* and *j* move in a correlated way, the fluctuations of *i* controlling the fluctuations of *j*, then we will have more information on the fluctuations of *j* than if they were uncorrelated. For example, if *i* and *j* were perfectly correlated, then we would know exactly what *j* will do if we know what *i* is doing. This extra information, I_{ij}, that we gain because of the physical coupling of *i* and *j* is obtained by the Shannon equation and is termed as the mutual information and is always positive. The total entropy, S_{ij} of *i* and *j* in this case is written as $S_{ij} = S_i + S_j - I_{ij}$ (see Eqs 12 and 13 and also Ref. (MacKay, 2004)). Thus, correlation of fluctuations, irrespective of whether they are negative or positive, always decreases the sum of the singlet entropies of *i* and *j*. These arguments and the Shannon equation have been used to obtain entropy changes in proteins at different levels of approximation(Ben-Naim, 2008; Fleck, Polyansky, & Zagrovic, 2016; Hnizdo, Tan, Killian, & Gilson, 2008; Karplus, Ichiye, & Pettitt, 1987; Karplus & Kushick, 1981; Killian, Kravitz, & Gilson, 2007; Killian et al., 2009; King, Silver, & Tidor, 2012; Motlagh et al., 2014; Numata & Knapp, 2012; Suarez, Diaz, Mendez, & Suarez, 2013; Suarez, Diaz, & Suarez, 2011; Suarez & Suarez, 2012; Zidek, Novotny, & Stone, 1999). However, we need to go beyond the Shannon equation in order to quantify allosteric communication in proteins which requires, as shown by Schreiber in 2000 (Schreiber, 2000b), the knowledge of time delayed conditional probabilities of two trajectories. In the interest of determining which residue drives the correlated motions and which residue responds, van der Vaart applied the Schreiber equation to determine information flow between Ets-1 transcription factor and its binding partner DNA (Kamberaj & van der Vaart, 2009a) (Also see references (A. Karolak & van der Vaart, 2012; van der Vaart, 2015) in similar context). Since this first work on entropy transfer in proteins there has been a limited number of studies on information transfer in proteins. Barr et al. (Barr et al., 2011) quantified entropy transfer among several residues in a molecular dynamics analysis of mutation effects on auto-phosphorylation of ERK2. Corrada et al. (Corrada, Morra, &

Colombo, 2013) analyzed entropy transfer in antibody antigen interactions. Perilla et al. (Perilla, Leahy, & Woolf, 2013) used the transfer entropy method to analyze barrier crossing transitions in epidermal growth factor receptors. Qi and Im (Qi & Im, 2013) quantified drive-response relations between residues during folding. Jo et al. (Jo, Qi, & Im, 2016) obtained a causality relationship between intramolecular hydrogen bonds and the conformational states of N-glycan core in glycoproteins. Zhang et al. (L. Q. Zhang, Centa, & Buck, 2014) applied the method to understand changes in the correlated motions in the Rho GTP-ase binding domain during dimerization. An extensive overview of similar techniques is given in reference (Aleksandra Karolak, 2015).

In the following section, we define the model on which we build the information theoretical basis of entropy. We then study the problem of time delayed correlation of fluctuations in proteins. Despite its importance in pointing to directionality of events in proteins, as has been shown recently for the allosteric activity of K-Ras (Vatansever, Gumus, & Erman, 2016), time delayed functions have not been studied in detail in the past. We then present a fast and accurate method of calculating entropy changes in proteins subject to pairwise interactions. Calculation of entropy of proteins is not new and has already been investigated by several authors (Karplus et al., 1987; Kassem, Ahmed, El-Sheikh, & Barakat, 2015; Numata, Wan, & Knapp, 2007; Suarez et al., 2013) at different levels of approximation. Our method of entropy calculation is motivated by the recent finding that the distribution functions for the magnitude of fluctuations of residues in globular proteins can be derived from a universal function (Erman, 2016). The method that we use for calculating the entropy is fast and accurate, based on histogramming the magnitude of fluctuations of each atom in a protein where the bin number is chosen according to the Sturges' rule of determining the widths of class intervals (Legg, Rosin, Marshall, & Morgan, 2013). We show that the use of Sturges' rule in our computational method leads to results that agree with earlier entropy calculations. We benchmark our method with calculations of Ubiquitin by Fleck et al (Fleck et al., 2016). The entropy change of Ubiquitin upon binding to human polymerase iota that we calculate agrees with the value obtained in reference (Fleck et al., 2016) using a different method of entropy estimation. The computational method that we adopt is efficient and plausible and may directly be applied for evaluating entropy transfer in proteins.

The association of Shannon equation with statistical mechanical definition of entropy and quantifying transfer entropy using the Schreiber equation allows us to interpret a wide range of events in proteins. If entropy transfer is considered in terms of changes in mobility, then transfer of entropy from *i* to *j* implies decrease in the mobility of *i* due to its correlation with *j*. Stated in another way, residue *j* extracts entropy from *i*. If binding is considered, one could then say that transfer of entropy from *i* to *j* would facilitate binding at *i* due to lowered mobility of *i,* although this may not be a general trend and may depend on several other factors. We use the model to study the directionality of information flow and entropy transfer in the 76 amino acid protein Ubiquitin which is known to propagate signals allosterically in the cell by binding to a vast number of substrates (Liu & Nussinov, 2013). Until the recent work of Smith et al., (Smith et al., 2016), the allosteric mechanisms of Ubiquitin were not widely recognized, and it was not generally classified as an allosteric protein. Using NMR relaxation dispersion measurements, Smith et al., identified a collective global motion that originates from a conformational switch spatially distant from the site where Ubiquitin binds to other proteins. The predictions of the model we present agree with observations of Smith et al. (Smith et al., 2016). The model goes one step further and predicts the direction of information transfer and therefore explain the underlying causes of the observed phenomenon. We discuss this in detail in the Discussion Section. In order to identify communication patterns leading to allosteric effects, we scanned the full Ubiquitin and identified the pairs of residues whose time delayed correlation functions are asymmetric, and we quantified the amount of entropy transferred between residue pairs. In order to have a feeling of the effects of entropy transfer on protein-protein interactions, we analyzed the behavior of Ubiquitin when complexed with the binding partner human polymerase iota, 2KTF.PDB. We observed that binding of Ubiquitin to iota modifies the fluctuation patterns on another site that may affect the binding of a third protein which may possibly affect the formation of a ternary complex (Garner et al., 2011).

2.2.2 METHODS

2.2.2.1 MOLECULAR DYNAMICS SIMULATIONS

All-atom Molecular Dynamics simulations were performed for unbound (PDB ID: 1UBQ) and bound states (PDB ID: 2KTF) of Ubiquitin, using NAMD 2.11 simulation program with CHARMM22 All-Hydrogen Parameter set for Proteins and Lipids. TIP3P water model was used to represent water molecules. Counter ions are placed to neutralize the system. Time step of simulations were 2 fs and periodic boundary conditions were applied in an isothermal-isobaric NPT ensemble with constant temperature of 300 K and constant pressure of 1 bar. Temperature and pressure are controlled by Langevin thermostat and Langevin piston barostat, respectively. System coordinates were saved every 1 ps. 1–4 scaling is applied to van der Waals interactions with a cutoff of 12.0 Å. Energy of the system was minimized, and the system is heated to 300 K for 50 ps and further subjected to MD production run for 600 ns. Frames in trajectories were aligned to the first frame of the simulation by using VMD 1.9.2 to eliminate all rotational and translational degrees of freedom and the analysis is done with the aligned Cartesian coordinates.

2.2.2.2 MOLECULAR DYNAMICS TRAJECTORIES

We perform molecular dynamics simulations for a protein in equilibrium and extract stationary trajectories for each atom. The trajectories for the atoms are expressed as

$$R(t_k) = R\big(R_1(t_k), R_2(t_k), R_3(t_k), \dots, R_N(t_k)\big) \qquad k = 1,2,3,\dots,n_T \qquad \text{EQUATION 1}$$

Here, $R_i(t_k)$ is the position vector of the i^{th} atom at the k^{th} time, t_k, expressed in terms of its Cartesian coordinates, $x_i(t_k)$, $y_i(t_k)$, and $z_i(t_k)$, N is the total number of atoms and t_k is the time at the k^{th} step. K ranges from 1 to n_T, the total number of steps in the simulation. If the total time is T, then the length ξ of each time step is $\xi = T/n_T$. Each atom has a unique equilibrium mean position defined by

$$\bar{R} = R(\overline{R_1}, \overline{R_2}, \overline{R_3}, \dots, \overline{R_N}) \qquad \text{EQUATION 2}$$

We denote the instantaneous state of fluctuation of a protein at time t_k by the vector

$$\Delta R(t_k) = R(t_k) - \bar{R} \qquad \text{EQUATION 3}$$

which reads in vector form as

$$\Delta R(t_k) = \Delta R\big(\Delta R_1(t_k), \Delta R_2(t_k), \Delta R_3(t_k), \dots, \Delta R_N(t_k)\big) \qquad \text{EQUATION 4}$$

For each t_k, $\Delta R(t_k) = \Delta R\; \Delta R_1(t_k), \Delta R_2(t_k), \Delta R_3(t_k), \dots, \Delta R_N(t_k)$ Equation 4 has *N* entries. For the purposes of the present study, we only need the magnitude of the fluctuations. In the following, we will let *$\Delta R_i(t_k)$* represent the magnitude of the fluctuation at time t_k.

2.2.2.3 EVALUATION OF PROBABILITIES

The most general expression for the probability of fluctuation *ΔR* is the joint probability, *$p(\Delta R) = p(\Delta R_1, \Delta R_2, \Delta R_3, \dots, \Delta R_N)$*. This expression contains information on all orders of dependence between atoms and is too general for use. In the other extreme, the simplest expression is the singlet probability function $p_i(\Delta R_i)$ which is obtained from the most general expression by

$$p(\Delta R_i) = \int_0^\infty \dots \int_0^\infty \dots \int_0^\infty \dots \int_0^\infty p\,(\Delta R_1, \Delta R_2, \Delta R_3, \dots, \Delta R_N)\boldsymbol{d}\Delta R_1, \dots, \Delta R_{i-1} \dots, \Delta R_{i+1}, \dots, \Delta R_N$$

EQUATION 5

N such functions define the probability of fluctuations of the N residues within the singlet approximation.

The next simplest probability is the pair probability $p_i(\Delta R_i, \Delta R_j)$ obtained from the most general expression by

$$p(\Delta R_i, \Delta R_j) =$$

$$\int_0^\infty \dots \int_0^\infty p\,(\Delta R_1, \Delta R_2, \Delta R_3, \dots, \Delta R_N)\boldsymbol{d}\Delta R_1, \dots, \Delta R_{i-1} \dots, \Delta R_{i+1}, \Delta R_{j-1} \dots, \Delta R_{j+1} \dots, \Delta R_N$$

EQUATION 6

For N atoms, there are $\frac{N(N-1)}{2}$ equations for pair probabilities.

In Eqs 5 and 6, *ΔR*'s are treated as continuous. In the remaining part of the paper, we will adopt a discrete representation for them in terms of histograms. The histograms will be expressed in terms of *n* bins. We refer to each bin as a state. The variables in the probabilities will then be functions of state variables. Thus, we write $p(\Delta R_i(k))$, where k goes from 1 to *n* where *n* is the number of states that define ΔR_i. Similarly, $p(\Delta R_i(k), \Delta R_j(l))$. In order to simplify the notation, we will suppress the state index, and write $p(\Delta R_i(k))$ as $p(\Delta R_i)$. Similarly, $p(\Delta R_i(k), \Delta R_j(l)) = p(\Delta R_i, \Delta R_j)$.

2.2.2.4 TIME DELAYED CORRELATION FUNCTIONS

We let $p(\Delta R_i(t), \Delta R_j(t+\tau))$ denote the joint probability of observing the fluctuation ΔR_i at time t and ΔR_j at time $t+\tau$. In this simplified notation, $\Delta R_i(t)$, represents the value of ΔR_i in state k at time t, which is identical to $\Delta R_i(k,t)$. $\Delta R_j(t+\tau)$ may be affected by the earlier fluctuations of $\Delta R_i(t)$. The extent of this effect may be quantified by the time delayed correlation function

$$C_{ij} = \frac{\left[\sum_{k=1}^{n_T-\tau/\xi} \Delta R_i(t_k)\Delta R_j(t_k+\tau)/(n_T-\frac{\tau}{\xi})\right]}{\langle(\Delta R_i)^2\rangle^{1/2}\langle\left(\Delta R_j\right)^2\rangle^{1/2}} \qquad \text{EQUATION 7}$$

This is a conditional correlation where ΔR_i comes before ΔR_j. In general, $p(\Delta R_i(t), \Delta R_j(t+\tau)) \neq p(\Delta R_j(t), \Delta R_i(t+\tau))$. This leads to directionality in the structure, known as causality, and consequently,

$$C_{ij}(\tau) \neq C_{ji}(\tau) \qquad \text{EQUATION 8}$$

If the fluctuations of residue *i* control the fluctuations of residue *j*, i.e., if residue *j* is driven by *i*, then the decay time for $C_{ij}(\tau)$ will be larger than that of $C_{ji}(\tau)$.

When τ=0, time independent Pearson correlation function is obtained as

$$C_{ij}(0) = \frac{\langle\Delta R_i \Delta R_j\rangle}{\langle(\Delta R_i)^2\rangle^{1/2}\langle\left(\Delta R_j\right)^2\rangle^{1/2}} \qquad \text{EQUATION 9}$$

2.2.2.5 ENTROPY

For a pair of subsytems, i and j, the entropy $S_{2,ij}$ is given as

$$S_{2,ij} = -k_B \sum_k \sum_l p_{ij}(k,l) ln\, p_{ij}(k,l) = -k_B \langle lnp_{ij} \rangle \qquad \text{EQUATION 10}$$

In Eq 10, the indices k and l denote the indices for the states of the respective subsystems, the states being obtained from the histograms described in the preceding section. For the case of pairwise interactions, the expression $p_{ij}(k,l)$ represents the joint probability where subsystem i is in state k and the subsystem j is in state l. We used the notation $p_{ij}(k,l) = p(\Delta R_i(k),\Delta R_j(l))$ for brevity of presentation.

In Eq 10, $S_{2,ij}$ signifies the joint entropy for two subsystems with pair probabilities.

We now divide and multiply the entropy expression by the singlet probabilities:

$$S_{2,ij} = -k_B \left\langle ln\left(\frac{p_{ij}}{p_i p_j} p_i p_j\right)\right\rangle \qquad \text{EQUATION 11}$$

which leads to the expression

$$S_{2,ij} = -k_B \langle lnp_i \rangle - k_B \langle lnp_j \rangle - k_B \left\langle ln\left(\frac{p_{ij}}{p_i p_j}\right)\right\rangle \qquad \text{EQUATION 12}$$

$$S_{2,ij} = S_{1,i} + S_{1,j} - I_{2,ij} \qquad \text{EQUATION 13}$$

where, $S_{1,i} = -k_B \langle lnp_i \rangle$ is the singlet entropy and $I_{2,ij} = k_B \left\langle ln\left(\frac{p_{ij}}{p_i p_j}\right)\right\rangle$ is the mutual information of the system.

Using statistical mechanics arguments given by Callen (Callen, 1985), each subsystem may be treated as a canonical ensemble that exchanges energy with its surroundings, represented by the cartoon in Figure 1. The surroundings of Subsystem 1 for example is the protein which contains Subsystem 2 also. We may choose the subsystems arbitrarily, an atom, an amino acid, or a secondary structure such as a helix, beta strand, loop or a tail. The subsystem may also be in contact with the surroundings of the protein. Mutual information is zero if the fluctuations of *i* are independent of the fluctuations of *j*. Otherwise, mutual information is always greater than zero. This leads to the conclusion that correlations always decrease the sum of the individual entropies in a system.

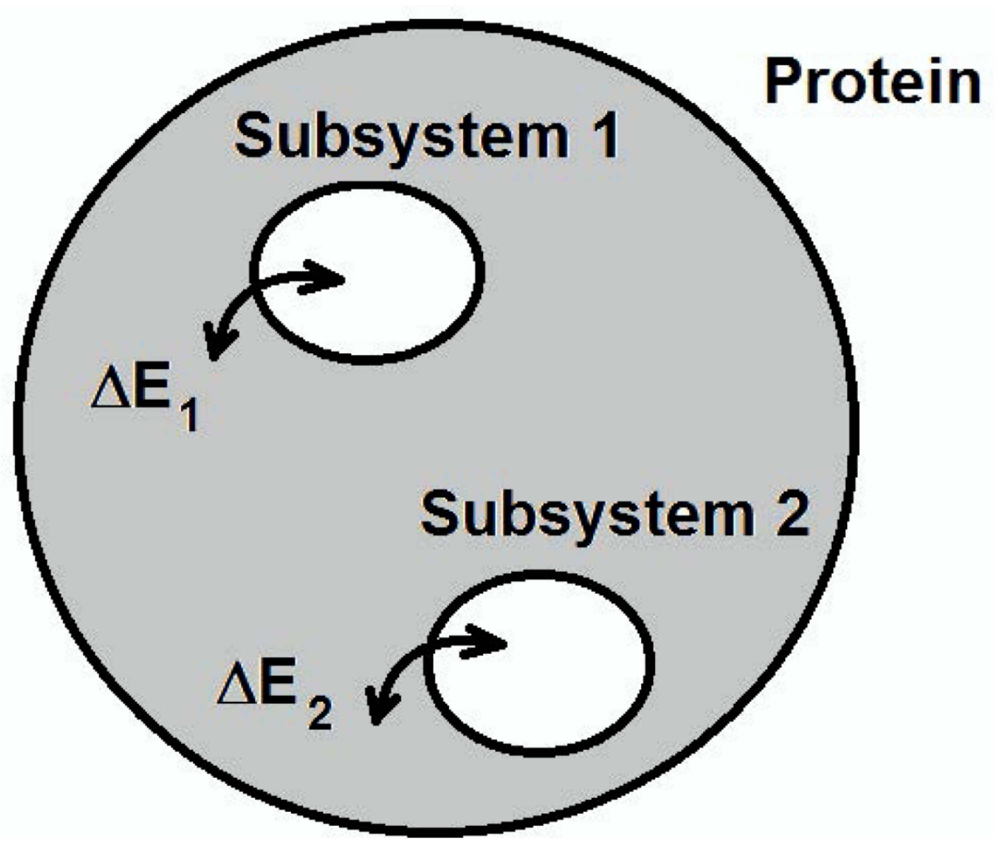


FIGURE 1. ENERGY EXCHANGE BETWEEN PROTEIN AND ITS SUBSYSTEMS.

2.2.2.6 CONDITIONAL ENTROPY

We consider two trajectories, $\Delta R_i(t)$ and $\Delta R_j(t)$. We now consider two events separated in time by τ, with the condition that ΔR_i coming before ΔR_j. The conditional entropy for these two events is defined by

$$\begin{aligned} S\left(\Delta R_j(t+\tau)|\Delta R_i(t)\right) &= -\sum p\left(\Delta R_i(t), \Delta R_j(t+\tau)\right) lnp\left(\Delta R_j(t+\tau)|\Delta R_i(t)\right) \\ &= \langle lnp\left(\Delta R_j(t+\tau)|\Delta R_i(t)\right)\rangle \\ &= \langle ln\frac{p\left(\Delta R_i(t), \Delta R_j(t+\tau)\right)}{p(\Delta R_i)}\rangle \\ &= \langle lnp\left(\Delta R_i(t), \Delta R_j(t+\tau)\right)\rangle - \langle lnp\left(\Delta R_i(t)\right)\rangle \end{aligned} \quad \text{EQUATION 14}$$

where, the summation is over all states for *i* and *j*, and the condition of stationarity is used in the last equation.

2.2.2.7 TRANSFER ENTROPY

Following Schreiber's work (Schreiber, 2000b), we write the transfer entropy $T_{i\to j}(\tau)$ from trajectory *i* to *j* at time τ as

$$T_{i\to j}(\tau) = S\left(\Delta R_j\big((t+\tau)\big)|\Delta R_j(t)\right) - S\left(\Delta R_j\big((t+\tau)\big)|\Delta R_i(t), \Delta R_j(t)\right) \qquad \text{EQUATION 15}$$

Using the last of Eq 14, this may be written as

$$T_{i\to j}(\tau) = \; -\langle \mathrm{LN}\, p\left(\Delta R_j(t), \Delta R_j(t+\tau)\right)\rangle + \langle \mathrm{LN}\, p\left(\Delta R_i(t), \Delta R_j(t), \Delta R_j(t+\tau)\right)\rangle + \\ \langle \mathrm{LN}\, p\Delta R_j(t)\rangle - \langle \mathrm{LN}\, p\left(\Delta R_i(t), \Delta R_j(t)\right)\rangle \qquad \text{EQUATION 16}$$

Through the term p*(ΔR$_i$(t), ΔR$_j$(t), ΔR$_j$(t+τ))*, Eq 16 requires the evaluation of triple probabilities. If trajectories *i* and *j* are independent, then $T_{i\to j}(\tau) = 0$ entropy transfer from *i* to *j* will be zero. In general, $T_{i\to j}(\tau) \neq T_{j\to i}(\tau)$ and this will determine the net transfer of entropy from one event to another separated in time by *τ*. Different values of *τ* shows how entropy transfer depends on prior interactions. In this study, we will take *τ*= 5 ns as the representative correlation time of cross correlations.

2.2.2.8 NET ENTROPY TRANSFER FROM AN ATOM

Eq 16 gives the entropy transferred, $T_{i\to j}(\tau)$, from atom *i* to *j*. The net entropy transferred from to all other atoms is obtained by summing over all *j* as

$$Net\ transfer\ from\ residue\ i = \sum_{j=1}^{N}\left(T_{i\to j}(\tau) - T_{j\to i}(\tau)\right) \qquad \text{EQUATION 17}$$

2.2.2.9 ENTROPY CALCULATIONS

Amplitude of fluctuations were calculated for each atom from Cartesian coordinates, $R_i(t_k)$, of the trajectory and the mean amplitude of fluctuations, $\overline{R_\iota}$ was subtracted from each $R_i(t_k)$ and the $\Delta R_{i..N}(t_k)$ matrix was generated with (t_k,N) dimensions, where N is the number of atoms and t_k is the number of frames selected for calculations. All of the calculations in this study will be based on the alpha carbon of each residue unless otherwise stated. Initial data up to equilibration was excluded from the calculations as equilibration. A binning approach was used to calculate configurational entropy, for individual and pairwise dependent atoms. Calculations were performed using MATLAB R2015b. Histogram function of MATLAB was used to cluster data into 8 bins with specified widths and partitioning of data is adaptive according to the maximum and minimum of data. Calculations were performed using 8 discrete bins. Number of bins were selected according to the Sturges' rule. The optimum number, n_{opt}, of bins is calculated from the Sturges' rule according to(Sturges, 1926)

$$n_{opt} = mean\ fluctuation\ \cdot (1 +\ log_2 N) \qquad \text{EQUATION 18}$$

Here, the mean of fluctuations, i.e., the average fluctuation of the N alpha carbons divided by maximum fluctuation is calculated from the trajectories and is equal to 0.4. For a trajectory of 600 ns, the optimum number of bins is obtained as 8 which is used throughout the calculations. After partitioning the fluctuation of each atom into 8 discrete bins, the probabilities were calculated from the frequency of occurrences and entropy was expressed in individual and pairwise mutual information terms. For comparison with benchmark calculations, the change of configurational entropy was calculated for all heavy atoms by subtracting pairwise mutual information term from individual entropy term as given by Eq 12. By using Eq 16, transfer entropy from atom *i* to *j* was calculated with a delay value of 5 ns for alpha carbons by subtracting the triple conditional entropies from pairwise conditional entropies. Result of configurational entropy calculations were compared with the benchmark data and transfer entropy results were used to study changes in entropy transfer patterns when Ubiquitin forms a complex.

The amount of mutual entropy depends on the distribution of the individual entropies and it is bounded by individual entropy terms.

$$I_2(i,j) \leq min\{S(i), S(j)\} \qquad \text{EQUATION 19}$$

Estimated entropy from a finite sample may be affected by some systematic errors and a correction term is required to get rid of this error (Scarabelli, Morra, & Colombo, 2010). Corrections were applied according to the previous studies (Paninski, June 2003) Where $S_1^{estimated}$ is the raw entropy, M is the number of histogram bins with non-zero probability. Since mutual entropy is the sum of entropies, this formula can also be used to correct $I_2(i,j)$ terms.

$$I_2^{true} \approx I_2^{estimated} + \frac{M_{ij} - M_i - M_j + 1}{2N} \qquad \text{EQUATION 20}$$

here, M_{ij}, M_i and M_j represent the numbers of the corresponding histogram bins with non-zero probabilities.

2.2.2.10 BENCHMARK FOR ENTROPY CALCULATIONS

Our method of configurational entropy calculations which we need for calculating transfer entropy are based on a histogram method using Sturges' rule. We compare the results of our configurational entropy calculations with those of MIST(Mutual Information Spanning Trees) method of PARENT(Fleck et al., 2016) . The mean entropy change result of Ubiquitin upon complex formation with human polymerase iota, 2KTF.PDB was obtained by our method as -47.64 (standard deviation of 12.33) calculated from 5 different portions from a simulation of 1200 ns. For the same system, MIST gave a mean of -40.59 (standard deviation of 28.99) for 5 different MD simulation sets for Ubiquitin and its complex. The error between the means of our result and of MIST is 17.4%, which is within 1 standard deviation of PARENT results.

2.2.3 RESULTS

2.2.3.1 STRUCTURE OF UBIQUITIN

Ubiquitin is a protein with 76 amino acids as shown in Figure 2. It consists of 8 distinct secondary structures that actively take part in its interactions with a large number of proteins.

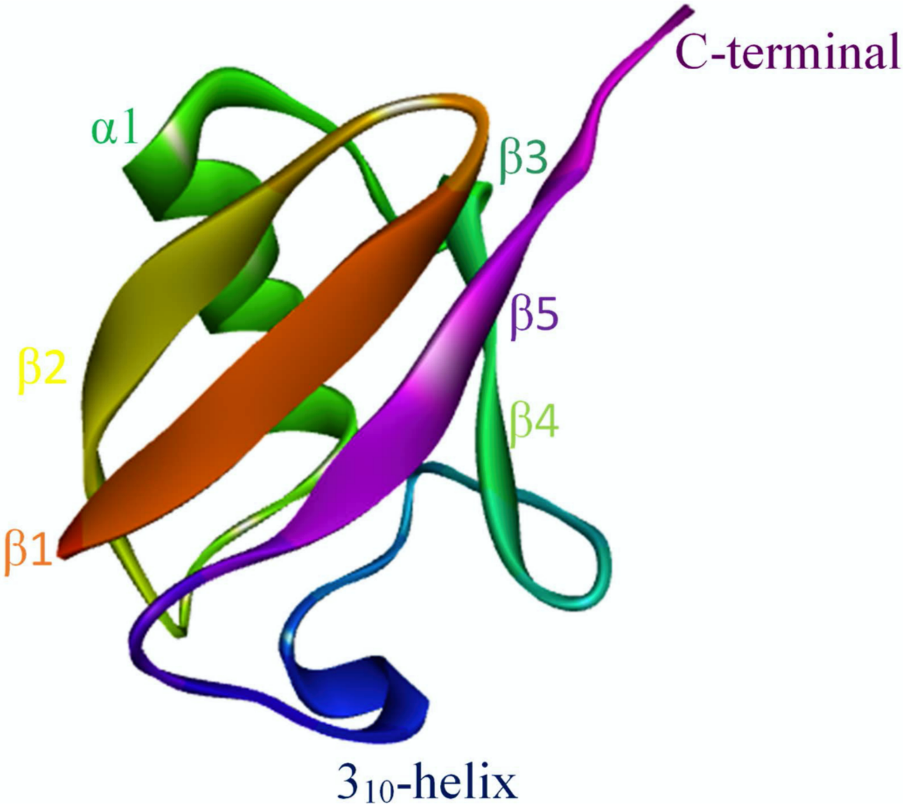


FIGURE 2. STRUCTURE OF UBIQUITIN (1UBQ)

The interactions of the secondary structures are strictly coordinated by the correlations in the protein. In Figure 3 we present the results of Pearson correlations of fluctuations, where the negative and positive correlations are shown in the left and right panels, respectively. The correlations with amplitudes (-1,-0.25) and (0.25, 1.0) are shown in the figure. The strongest negative correlation is between LEU73 and the three residues PRO37, PRO38 and ASP39. The next level of negative correlations is among distant pairs that are situated approximately diagonally opposite in the structure (examples are correlations between pairs LEU8-GLN31, ILE30-LYS63, GLU18-LYS33). The negative correlations between these pairs are expected to confer a breathing type of motion to the protein, which was indeed observed experimentally(Smith et al., 2016). We elaborate on this point in detail in the Discussion Section. Figure 3B shows that positive correlations are mostly along the diagonal, indicating that neighboring residues along the primary sequence are positively correlated. However, there are off-diagonal regions in Figure 3B showing positive correlations among residues that are not close along the primary structure. The strongest off-diagonal positive correlation in Figure 3B is between GLU24 and GLY53.

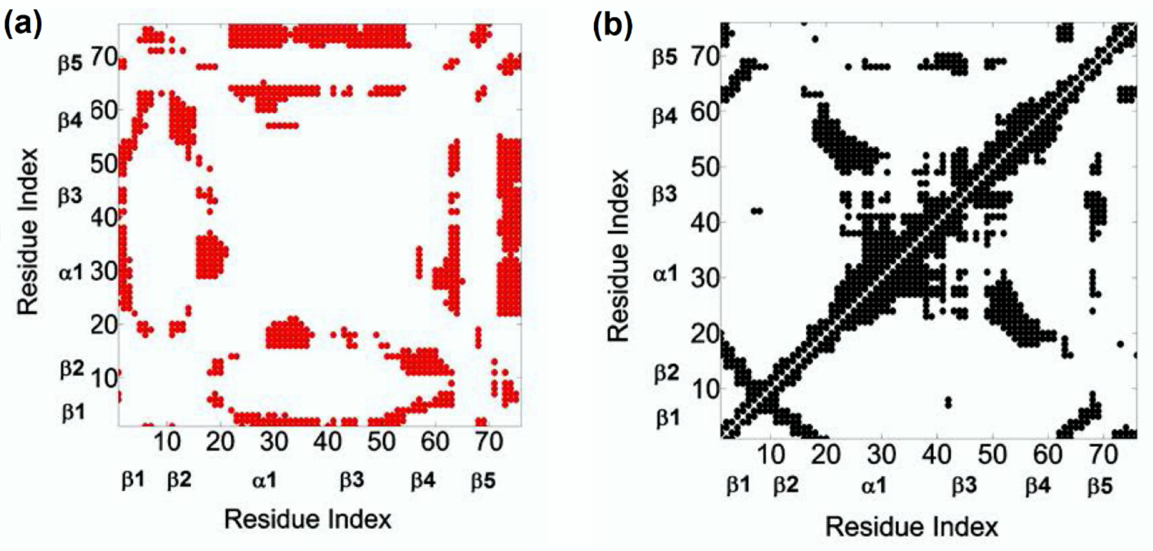


FIGURE 3. A) THE LEFT PANEL SHOWS THE NEGATIVE PEARSON CORRELATIONS IN THE RANGE (-1, -0.25), B) THE RIGHT PANEL SHOWS POSITIVE PEARSON CORRELATIONS IN THE RANGE (0.25, 1.0). PEARSON CORRELATIONS ARE CALCULATED FROM EQ 9.

2.2.3.2 TRANSFER ENTROPY IN UBIQUITIN

We present the results of entropy transfer between all residue pairs of Ubiquitin. We consider only the alpha carbons, and the values given are divided by the Boltzmann constant. Results presented below are based on a trajectory of 600 ns. Entropy transfer values calculated from a duplicate trajectory of 600 ns gave similar results. Results of entropy transfer calculations for Ubiquitin and its complex presented below showed that convergence is established after 400 ns.

Using Eq 16 we evaluated the values of entropy transfer from alpha carbon i to j, $T_{i \to j}(\tau)$, for all pairs of i and j for τ=5 ns. Calculations averaged over several time stations between 0 and 5 ns gave approximately the same values for entropy transfer. In the remaining parts we present τ=5 ns results only. The characteristic decay time of correlations of fluctuations of alpha carbons, which will be discussed in the following section, is on the average between 1 to 10 ns. The entropy transfer function $T_{i\rightarrow j}(\tau)$ that we obtain from fluctuation trajectories of alpha carbons depends on the correlation of two events that are T ns apart in time. If τ is taken very small, i. e., around zero, then the difference between $T_{i\rightarrow j}(\tau)$ and $T_{j\rightarrow i}(\tau)$will be very small because $T_{i\rightarrow j}(0) \neq T_{j\rightarrow i}(0)$. If τ is taken much larger than the characteristic decay time, then the correlations will have decayed to small values and the differences will be vanishingly small. In agreement with this reasoning, we took τ=5 ns and calculated entropy transfer at this time. The results are shown in Figure 4. The abscissa, named as entropy donor, denotes the indices of residues that act like entropy reservoirs to other residues. The ordinate, named as entropy acceptor, denotes the indices of residues that act like entropy sinks that absorb entropy from the system.

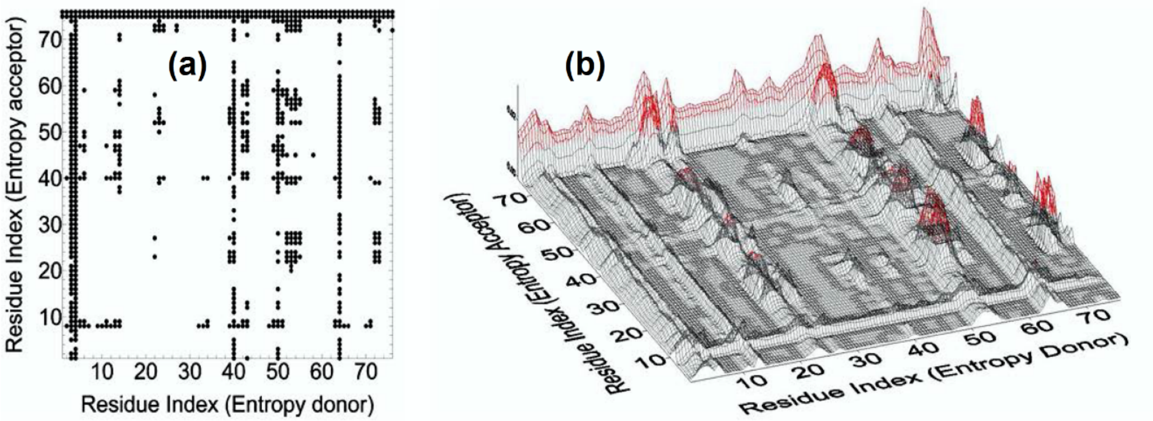


FIGURE 4. A) ENTROPY TRANSFER FROM RESIDUE I TO RESIDUE J. ABSCISSA REPRESENTS RESIDUES WHICH PROVIDE ENTROPY TO RESIDUES SHOWN ALONG THE ORDINATE. ENTROPY TRANSFERRED FROM RESIDUE I TO RESIDUE J IS OBTAINED FROM EQ 16. VALUES BETWEEN 0.0035–0.35 ARE RECORDED. VALUES BELOW 0.0035 ARE NOT SHOWN IN ORDER NOT TO CROWD THE FIGURE. $T_{I\rightarrow J}$(T) VALUES ARE CALCULATED FROM EQ 16 WITH T = 5 NS., B) THREE-DIMENSIONAL DESCRIPTION OF ENTROPY TRANSFER IN UBIQUITIN. RED REGIONS DENOTE THE RESIDUES WITH LARGE CONTRIBUTIONS TO TRANSFER ENTROPY. THE FIGURE IS A 3D VERSION OF FIG 4A.

The columns of black points in the figure show that specific residues, such as ILE3 and PHE4, ILE13, ILE23, LYS27, GLY53, GLU64, ARG72 provide entropy to several residues of the protein. The rows of black circles indicate residues such as LEU8, THR9, GLY75 and 76, that absorb entropy from several residues of the protein. Residues ILE3 and PHE4, ILE13 and GLU64 form a spatial cluster. Also, the residues ILE23, LYS27 and GLY53 form a spatial cluster. If the allosteric path description is adopted, then we can say that these two spatial clusters lie on the allosteric path.

In order to have an idea on the mechanism of communication in the system, one needs to know the transfer of entropy among specific pairs of residues. From the data of Fig 4, we can find with which residues a given amino acid interacts entropywise. Figs 5 and 6 summarize the net entropy exchange, $T_{I\rightarrow J}(5)$-$T_{J\rightarrow I}(5)$, between the labeled residue in each panel and the j^{th} residue of the protein. Figure 5 shows some examples with mostly positive entropy transfer from the labeled residue to others. The top left panel in Fig 5 shows the entropy transfer from ILE3 to other residues. Specifically, it transfers the largest entropy to LEU8 and GLY75 and GLY76. Both ILE3 and LEU8 are at the opposite extremities of β_1. ILE3 is a spatial neighbor of GLU64. GLU64 is hydrogen bonded to GLN2, and the entropy of GLU64 is transferred to ILE3 via the stated hydrogen bond. ILE3 contributes entropy to several other residues of the protein as may be seen from the figure. Entropic interactions of residues PHE4, ILE13, ILE23 and LYS27 are very similar to those of ILE3 and are not given as a separate figure. The top right panel in Figure 5 shows the interactions of GLY53 with the rest of the protein. GLY53 is situated on the long loop between β_3 and β_4, and is hydrogen bonded to the main alpha carbon of GLU24 which is at the end of α_1. Figure 5 shows that GLY53 contributes to the entropy of the segment between VAL17 and LYS29. It also transfers entropy along the chain to LEU56. GLU64 contributes entropy to several residues, in a way similar to that of ILE3. ARG72 has a unique pattern of contribution, specifically to ASP39 which is its spatial neighbor, to the loop between α_1 and β_2, to PHE45 and LEU56, both of which are spatially distant from ARG72. It also contributes to the mobility of the C-terminal. Figure 6 gives two examples for mostly negative values of $T_{I\rightarrow J}(5)$-$T_{J\rightarrow I}(5)$. The left panel in Figure 6 shows that LEU8 and GLY76 absorb entropy from most of the residues of the protein.

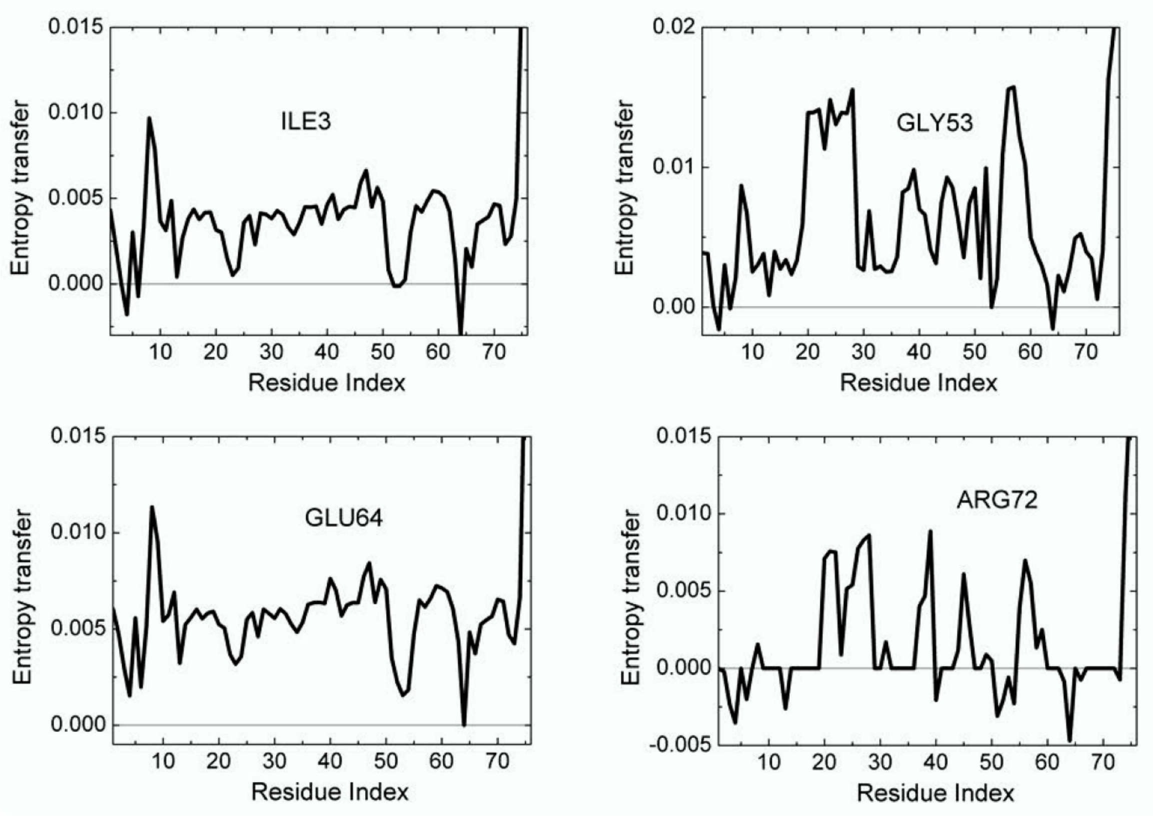


FIGURE 5. ENTROPY TRANSFER FROM A GIVEN RESIDUE TO OTHER RESIDUES OF THE PROTEIN. THE RESIDUE FROM WHICH ENTROPY IS TRANSFERRED IS MARKED IN EACH PANEL. CALCULATIONS ARE BASED ON THE RELATION $T_{I\rightarrow J}(5)$-$T_{J\rightarrow I}(5)$.

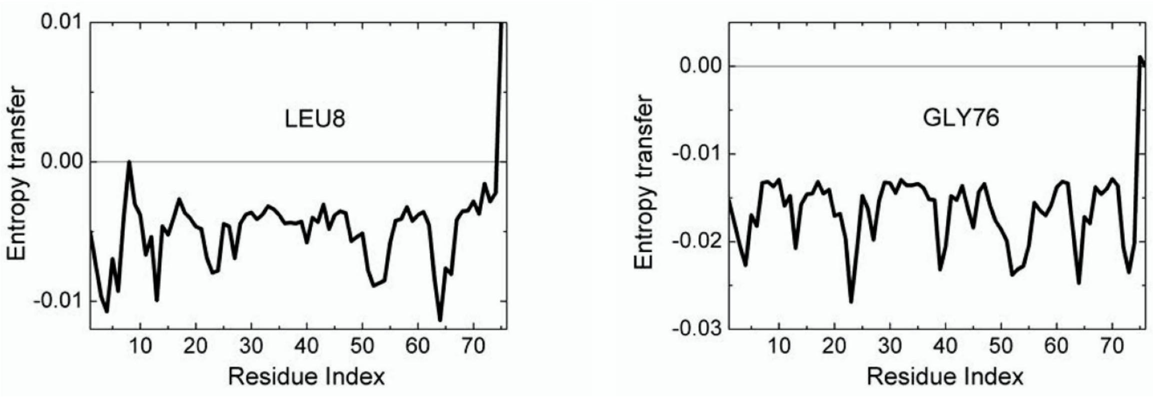


FIGURE 6. ENTROPY TRANSFER FROM RESIDUES OF THE PROTEIN INTO LEU8 AND GLY76. THE RESIDUE FROM WHICH ENTROPY IS TRANSFERRED IS MARKED IN EACH PANEL. CALCULATIONS ARE BASED ON THE RELATION TI→J(5)-TJ→I(5).

The net transfer of entropy from residue I, defined by $Net\ transfer\ from\ residue\ i = \sum_{j=1}^{N}\left(T_{i\to j}(\tau) - T_{j\to i}(\tau)\right)$ Equation 17 is presented in Figure 7. Positive values denote net entropy transfer out from a residue, and negative values denote net entropy into a residue. Similar to the pattern observed in Figure 4, we see that certain residues behave as entropy sources for the rest of the protein and some behave as entropy sinks.

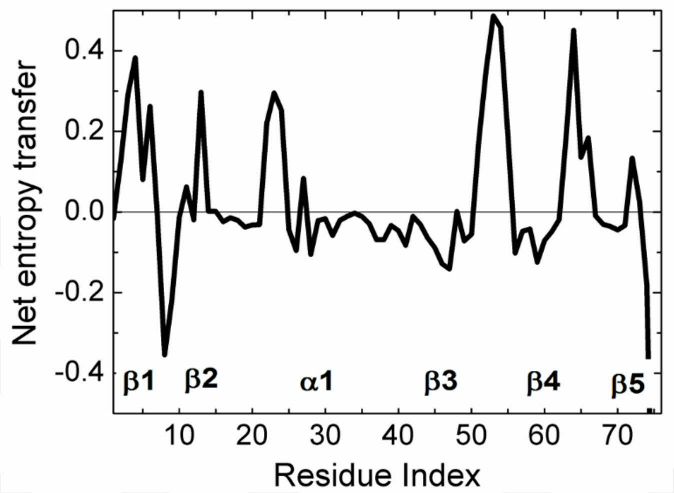


FIGURE 7. NET ENTROPY TRANSFER FROM ONE RESIDUE TO THE REST OF THE PROTEIN, CALCULATED BY EQ 17. A RESIDUE WITH A POSITIVE (NEGATIVE) VALUE OF NET ENTROPY TRANSFER IS AN ENTROPY SOURCE (SINK).

We see that β_1 and β_2 act as an entropy source as well as part of the helix α_1. The largest amount of entropy is provided by the loops between $\beta_3\beta_4$ and $\beta_4\beta_5$. The two major entropy sinks are the loop between β_1 and β_2 and the last two residues of the C-terminal. Entropy sources are located mostly at secondary structures or at their extremities. The three residues PHE4, THR14, GLU64 are spatial neighbors. Similarly, LEU43, LEU50, are spatial neighbors. The entropy source and sink residues are shown in three dimensions in Figure 8.

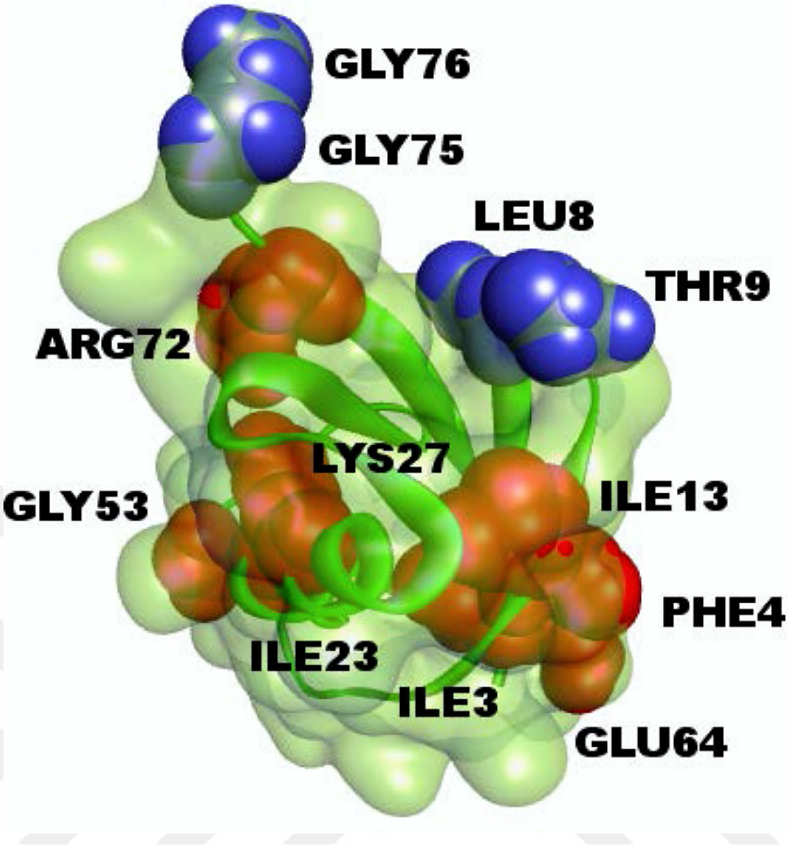


FIGURE 8. STRUCTURE OF UBIQUITIN, RESIDUES THAT ARE COLORED IN RED ARE ENTROPY ACCEPTORS AND RESIDUES THAT ARE COLORED IN BLUE ARE THE ENTROPY DONORS.

2.2.3.3 TIME DELAYED CORRELATIONS OF UBIQUITIN

Fluctuations of amino acids in Ubiquitin display characteristic decay times that are in the order of 1 to 10 ns as may be observed from the decay of the curves to 1/e of their original values. Differences arise from the unique conformational features of the amino acid and its environment. In Figure 9, we show the autocorrelations of THR7 and LEU71.

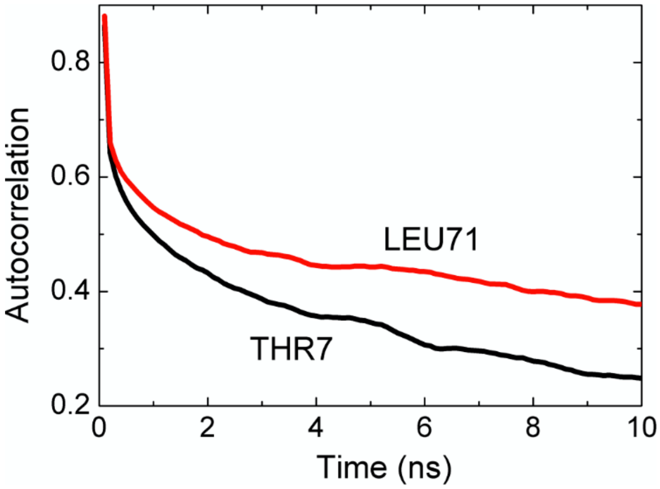


FIGURE 9. AUTOCORRELATIONS FUNCTIONS FOR THR7 AND LEU71 CALCULATED FROM EQ 7 FOR I = J. THE ABSCISSA DENOTES THE TIME DELAY PARAMETER, I.E., THE TIME BETWEEN TWO OBSERVATIONS, ONE AT TIME ZERO THE OTHER AT THE INDICATED TIME ON THE AXIS.

The autocorrelation function for THR7, i.e., the time required to decay to 1/e of the original value is 5 ns. LEU71 decays slightly slower with a decay time of 10 ns.

The time delayed cross correlations of the fluctuations of two amino acids are of interest because they yield information on the causality of events. The static correlations presented in Figure 3 are symmetric, i.e., $C_{IJ}(\tau) = C_{JI}(\tau)$. However, time delayed cross correlations of fluctuations of two amino acids show asymmetries which we discuss in this section.

In Figure 10 and 11, we present two cases that show significant causality. The strongest asymmetry is between LEU7 and THR71, shown in Figure 10.

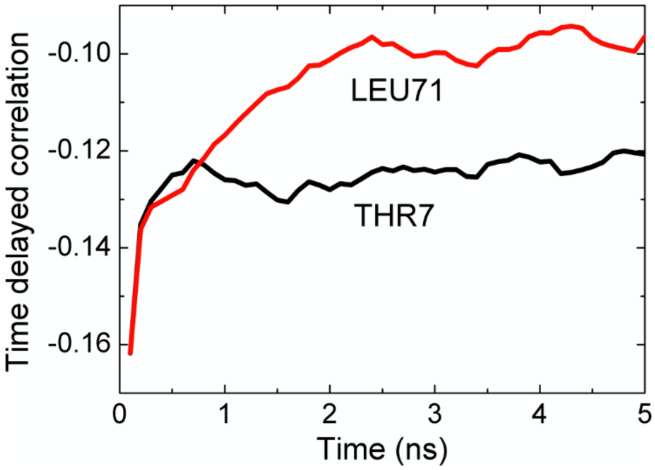


FIGURE 10. CROSS CORRELATION OF FLUCTUATIONS OF THR7 AND LEU71. BLACK LINE IS FOR CORRELATIONS WHERE THR7 PRECEDES LEU71. THE RED LINE IS FOR CORRELATIONS WHERE LEU71 PRECEDES THR7. THE ABSCISSA DENOTES THE TIME DELAY PARAMETER, I.E., THE TIME BETWEEN TWO OBSERVATIONS, ONE AT TIME ZERO THE OTHER AT THE INDICATED TIME ON THE AXIS. THE CURVES ARE CALCULATED FROM EQ 7.

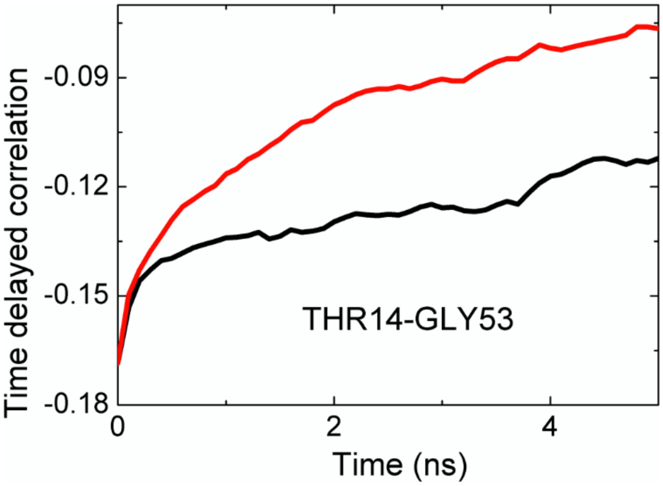


FIGURE 11. CROSS CORRELATION OF FLUCTUATIONS OF THR14 AND GLY53. BLACK LINE IS FOR CORRELATIONS WHERE THR14 PRECEDES GLY53. THE RED LINE IS FOR CORRELATIONS WHEREGLY53 PRECEDES THR14. THE ABSCISSA DENOTES THE TIME DELAY PARAMETER, I.E., THE TIME BETWEEN TWO OBSERVATIONS, ONE AT TIME ZERO THE OTHER AT THE INDICATED TIME ON THE AXIS. THE CURVES ARE CALCULATED FROM EQ 7.

In this figure, the black curve is for correlation of THR7 at time t and LEU71 at $t+\tau$. The red curve is for LEU71 at t and THR7 at $t+\tau$. The black curve decays significantly slower than the red curve, indicating that the effect of the fluctuations of THR7 on later fluctuations of LEU71 persists for longer times whereas the converse is not true. We therefore say that the motions of THR7 drive the motions of, LEU71 i.e., THR7 is the driver and is LEU71 driven. Since LEU71 is located on the C-terminal segment, and THR7 is at the end of β_1, we can say that the β_1 strand controls the fluctuations of the C-terminal. We see that the black curve remains approximately constant after a rapid initial decay. This shows that the driver action of THR7 on LEU71 persists for longer times.

In Figure 11, the black curve is for the correlation of the fluctuations of THR14 with later fluctuations of GLY53. The red curve is for the reverse case, fluctuations of GLY53 affecting later fluctuations of THR14. This figure shows that THR14 is the driver and GLY53 is driven. THR14 is on the β_2 strand and GLY53 is on the long loop connecting the $\alpha 3_{10}$ helix to β_4.

2.2.4 DISCUSSION

The entropy transfer model that we developed for understanding allosteric communication in proteins measures the amount of information transfer between the trajectories of two atoms, *i* and *j*. Knowing the fluctuations of atoms *i* and *j* at time *t*, the model evaluates the amount of uncertainty reduced in the future fluctuations of atom *j*. One extreme case is where the fluctuations of *i* have no effect on the fluctuations of *j*, i.e., their trajectories are uncorrelated. In this case, $T_{i\to j(\tau)}= -\langle \ln p\left(\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle+\langle \ln p\left(\Delta R_i(t),\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle+\langle \ln p\Delta R_j(t)\rangle-\langle \ln p\left(\Delta R_i(t),\Delta R_j(t)\right)\rangle$ Equation 16 equates to zero, and no entropy will be transferred to residue *j* from *i*. The other extreme case is where the fluctuations of *i* at time *t* are perfectly locked into those of *j* at time *t* and the knowledge of the present values of *i* and *j* fluctuations will reduce the uncertainty of the future fluctuations of *j*. In this case, the second term $S(\Delta R_j(t+\tau)/\Delta R_i(t),\Delta R_j(t))$ in $T_{i\to j(\tau)=S\left(\Delta R_j((t+\tau))|\Delta R_j(t)\right)-S\left(\Delta R_j((t+\tau))|\Delta R_i(t),\Delta R_j(t)\right)}$ Equation 15 will be modified as $S(\Delta R_j(t+\tau)/\Delta R_j(t))$-$S(\Delta R_i(t))$, which states that the reduction in uncertainty is due to the locking in of the fluctuations of *i*. Substituting these into

$T_{i\to j(\tau)= -\langle \ln p\left(\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle + \langle \ln p\left(\Delta R_i(t),\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle + \langle \ln p \Delta R_j(t)\rangle - \langle \ln p\left(\Delta R_i(t),\Delta R_j(t)\right)\rangle}$ E

quation 16 leads to $T_{i\to j}(\tau) = S(\Delta R_i)$. The range of values of entropy that may be transferred from i to j will lie within the interval $0 \le T_{i\to j}(\tau) \le S(\Delta R_i)$. In the absence of symmetry, $T_{i\to j}(\tau) \ne T_{j\to i}(\tau)$., we talk of a net transfer of entropy from i to j which will lie in the interval $-S(\Delta R_j) \le T_{i\to j}(\tau) - T_{j\to i}(\tau) \le S(\Delta R_i)$. *If* $T_{i\to j}(\tau) - T_{j\to i}(\tau)$ is greater than zero, we say that the fluctuations of atom i drive those of atom j.

Based on these explanations, we now compare the predictions of the model with experimental data. Progress in NMR spectroscopy and Relaxation Dispersion measurements allows for reliable experimental determination of correlations of fluctuations of residues that may be spatially distant (Ban et al., 2011; Fenwick et al., 2011; Smith et al., 2016). Such long-range correlations are candidate mechanisms that require information transfer, and hence may be seen as suitable indicators of entropy transfer. We show that measured correlations and patterns of entropy transfer that we calculate are complementary to each other.

2.2.4.1 COMPARISON WITH EXPERIMENT

Strong correlations between two residues, ILE23 and ASN25 have been observed by NMR studies of Ubiquitin (Dittmer & Bodenhausen, 2004; Massi, Grey, & Palmer, 2005). A more detailed investigation by Ban et al.,(Ban et al., 2011) using the recently developed NMR Relaxation Dispersion technique showed a strong correlation between ILE23, ASN25 and THR55. Later work by Smith et al., (Smith et al., 2016) using Relaxation Dispersion measurements showed that the two residues GLU24 and GLY53 act as a conformational switch and their correlated fluctuations induce breathing-like motions in the overall protein which affect the substrate binding region of Ubiquitin. Our Pearson correlation analysis shows that GLU24 and GLY53 are strongly positively correlated and there is significant entropy transfer from GLY53, which is located on a loop, to GLU24, located on the helix, as may be seen from Figure 4B and the top right panel of Figure 5. According to these figures, GLY53 is a strong entropy source for GLU24. Thus, we see a strong directionality in the interactions of the two residues that form a conformational switch that controls the overall motions of the protein: GLY53 drives the fluctuations of GLU24. In turn, GLU24 drives residues LEU8, GLN40 and the C-terminal residues.

Secondly, based on the predictive features of our model, we discuss the possible consequences of ligand binding and mutation, both of which are of significant relevance to drug design.

2.2.4.2 CHANGES UPON COMPLEX FORMATION OF UBIQUITIN

Ubiquitin forms complexes with a multitude of proteins. Here we studied its complex with Human Polymerase Iota which is a small protein of 28 amino acids, 2L0G.PDB, it binds to the following residues of Ubiquitin: LEU8, THR9, GLY10, GLY47, ARG42, ILE44 and the C-terminal. Net entropy transfer in Ubiquitin in the bound and free state is compared in Figure 12. The solid and dashed curves are for the bound and free states, respectively. Entropy transfer characteristics of four residues of Ubiquitin show significant changes upon complex formation. LEU8 which was an entropy sink in the unbound Ubiquitin is no longer a sink. GLU24, GLY53 and GLU64 which were strong entropy sources in the unbound state cease to be so in the complex. A possible interpretation of this observation is that GLY53 no longer drives GLU24 which in turn does not excite the breathing like motions of Ubiquitin, and the entropy transfer characteristics of Ubiquitin is now completely changes in favor of conformations that prefer complexation.

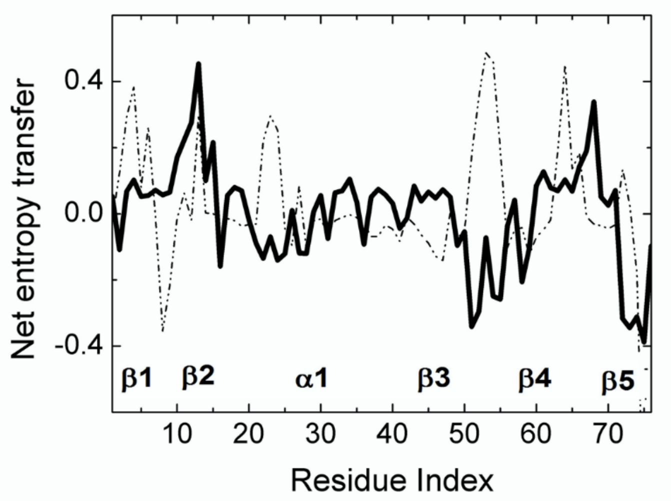


FIGURE 12. NET ENTROPY TRANSFER OF RESIDUES IN UNBOUND UBIQUITIN (DASHED CURVE) AND IN THE COMPLEX (SOLID CURVE) CURVES ARE OBTAINED BY USING EQ 17.

Mutations, although we have not performed simulations on such systems, are expected to significantly modify the entropy transfer characteristics of those residues that exhibit strong entropy transfer in the wild type. For example, mutation of GLU24 or GLY53 would abolish the driver-driven relations and change the functional dynamics of Ubiquitin significantly. In fact, Smith et al. (Smith et al., 2016)performed mutations experimentally and observed that the affinity toward ubiquitins binding partner weakened twofold in both mutants. Mutation of residues LEU8 and GLN40 should also lead to strong changes in allosteric behavior of Ubiquitin because these residues also have important role on entropy transfer, as was discussed in the preceding paragraphs.

Entropy transfer is computed only for alpha carbons in the present work. In principle, the calculations may be extended to include sidechains also since molecular dynamics trajectories are performed for all atoms of the system and the information for sidechain entropies is present in the trajectories. In the interest of brevity and clarity of presentation, only alpha carbons were treated at this preliminary analysis of entropy transfer in proteins. It is worth stating, however, that NMR measurements on backbone carbon and nitrogen were sufficient to characterize the allosteric dynamics of Ubiquitin (Massi et al., 2005; Smith et al., 2016). Sidechains and amino acid types will undoubtedly affect residue-residue communication, especially when considering differences between the bound and unbound states and should be included in a more detailed analysis. Binding of iota to Ubiquitin modifies the fluctuation patterns of residues ILE3, ILE23, GLY53 and GLU64 which are on a surface that is susceptible to the binding of a third protein, a case which may possibly affect the formation of a ternary complex (Garner et al., 2011).

In conclusion, we used Schreiber's model of entropy transfer and presented a detailed analysis of allosteric communication in Ubiquitin. Based on the analysis of time delayed events, we showed that information may be transferred between pairs of residues. The allosteric mechanisms of Ubiquitin have been understood only very recently. Our work shows that there is significant information transfer between residue pairs in this system.

From the entropy transfer point of view, all proteins may exhibit allosteric communication. This observation supports the recent hypothesis by Gunasekaran et al (Gunasekaran et al., 2004) that allostery is indeed an intrinsic property of proteins. Our work shows that the knowledge of time delayed correlations and entropy transfer is needed in order to quantify allosteric communication in proteins. Time delayed events have not been widely used in studies of protein function and allosteric communication. Recently, it was shown that causality introduced by time delayed correlations plays significant role on allosteric communications in K-Ras. In this respect, time delayed correlation functions may be viewed as a new tool for studying allosteric communication in proteins. A three-dimensional map of entropy transfer, as shown in Figure 4B may be useful for visualizing allosteric communication between pairs of residues more easily. Based on Figure 4B and the entropy transfer propensities of residue, the model serves as a suitable tool for explaining the basis of allosteric mechanisms in proteins.

Finally, it is worth noting that the present approach which maps the causality, driver-driven relations, and entropy exchange into pairs of residues, as seen in Figure 4B, should be of great significance for allosteric drug design because it tells us which residues to manipulate. In this respect, a driver residue is more critical than the driven residue and manipulating the driver will perturb the existing correlations more efficiently. The effects of mutation on allosteric communication may be quantified by calculating the changes in entropy transfer. As we showed in the UBQ-Human Polymerase Iota complex, binding may result in entropy changes in the exposed residues of the complex and change the binding propensities of the complex to other molecules such as another protein, a small molecule ligand or a DNA segment.

2.3 MEASURING ENTROPY TRANSFER WITH GAUSSIAN NETWORK MODEL

2.3.1 INTRODUCTION

Transfer of entropy from one subsystem of a protein to another is now becoming a subject of interest because of its relation to information flow and allosteric communication. Allosteric communication is the process in which action at one site of a protein is transmitted to another site at which the protein performs its activity. Protein-protein and protein-DNA interactions, drug action and all processes that depend on signal transduction involve allosteric activity for the system to carry out its normal function. Most known mutations that cause cancer lead to the disruption of normal allosteric communication. Recent findings show that allosteric activity is entropic in nature and depends on information transfer from one part of the protein to the other(Motlagh et al., 2014; Tsai & Nussinov, 2014) through coordinated fluctuations of residues. Transmission of effects through correlated fluctuations is a universal property of all proteins and not only of allosteric ones. In this sense all proteins may be regarded as intrinsically allosteric in nature (Gunasekaran et al., 2004).This new view freed the understanding of allostery from the limited picture of discrete two state transitions and opened a broader vista in terms of entropy transfer in proteins. The idea of transfer entropy, recently introduced by Schreiber,(Schreiber, 2000b) is the appropriate one for understanding information flow and communication in proteins. van der Vaart applied the Schreiber equation to determine information flow between Ets-1 transcription factor and its binding partner DNA (Kamberaj & van der Vaart, 2009b), Barr *et al.*,(Barr et al., 2011)quantified entropy transfer among several residues in a molecular dynamics analysis of mutation effects on autophosphorylation of ERK2, Corrada *et al.*,(Corrada et al., 2013) analyzed entropy transfer in antibody antigen interactions, Zhang *et al.*,(L. Q. Zhang et al., 2014) applied the method to understand changes in correlated motions of the Rho GTPase binding domain during dimerization. Presently, one of the requirements for calculating entropy transfer in proteins is to run molecular dynamics simulations in the order of microseconds. Considering the urgent need for determining information transfer in malfunctioning proteins, the molecular dynamics technique becomes a serious bottleneck and a rapid characterization is needed. The aim of this article is to provide a rapid scheme of computing entropy transfer in proteins and show that its results agree with experimental evidence and detailed molecular dynamics-based predictions. For this purpose, we formulate transfer entropy using the dynamic version of Gaussian Network Model (dGNM) which is based on harmonic interactions between contacting pairs of residues(Bahar, Atilgan, & Erman, 1997). Calculation times for determining transfer entropy between all pairs of residues of proteins, even of extremely large protein complexes can now be performed in the order of seconds on a laptop computer using dGNM.

2.3.2 METHODS

In this section, we first give a detailed discussion of why time delayed correlations are needed for explaining allosteric communication. We then discuss the meaning of entropy transfer in proteins and show how it is related to the function of the biological system. We then derive the basic equations of entropy transfer between two residues. Evaluation of transfer entropy requires the construction of third order joint probabilities for fluctuations. If molecular dynamics simulations are used, then accurate estimates of the joint probabilities require the use of long trajectories. Our recent work (Hacisuleyman & Erman, 2017b) shows that for a small protein such as Ubiquitin, the minimum trajectory lengths should be around 500 ns for convergence of entropy transfer values. This makes an approximate but fast method of estimating entropy transfer almost imperative. For this reason, we use the dGNM to obtain the time dependent probabilities. After introducing the method of Schreiber below, we give a detailed description of the dGNM, and then we use it in the Schreiber equation to arrive at the main equation of the article.

The fluctuation $\Delta R_i(t)$ of the position vector R_i of a residue from its mean position $\overline{R_\iota}$ is a rapidly changing function of time. During a long stretch of time, the fluctuations will take many different values and the occurrence of a given value of the fluctuation will be subject to a certain uncertainty. There will be a similar uncertainty in the fluctuations of the second residue. If the fluctuations of the two residues are not correlated, then, knowing the fluctuations of the first residue will not reduce the uncertainty in the trajectory of the second residue. In the other extreme, if the fluctuations of the first residue are perfectly locked into those of the second residue, then the knowledge of the first trajectory will leave no uncertainty in the second trajectory and we will have full information on the second trajectory. In this case, there will be an information transfer from the first residue to the second. If an observation is made at time $t = 0$ on the first trajectory and another observation at time $t = \tau$ on the second trajectory, then the amount of information gained on the second trajectory by the knowledge of the first may decay depending on the length of

τ and the level of correlation between the two trajectories. The time dependent correlation of two fluctuations is expressed in terms of the time delayed correlation function $\langle \Delta R_i(0)\Delta R_j(\tau)\rangle$, which shows the scalar product of the two vectors and the angular brackets denote an average over all observations, with a time delay of τ between the two. If this average is not zero, then we see that the fluctuations of *i* at one time correlate with the fluctuations of *j* at a time τ later. If the two residues are symmetrically correlated, that is, $\langle \Delta R_i(0)\Delta R_j(\tau)\rangle = \langle \Delta R_j(0)\Delta R_i(\tau)\rangle$ then there will be no net information transfer from one residue to the other, because effects propagate symmetrically between the two residues. Net information may be transferred from one residue to the other only if $\langle \Delta R_i(0)\Delta R_j(\tau)\rangle \neq \langle \Delta R_j(0)\Delta R_i(\tau)\rangle$. It is to be noted that time independent correlations, $\langle \Delta R_i(0)\Delta R_j(0)\rangle$, are always symmetric in *i* and *j*. Asymmetry is possible only in time delayed correlations.

The discussion in the preceding paragraph shows that time delayed correlations are needed to describe allosteric communication in a protein. A better description of asymmetry to describe allosteric communication is the concept of entropy transfer, which we explain in the following paragraph.

Transfer entropy $T_{i\to j}(\tau)$ from the trajectory *ΔR_i(t)* of residue i to the trajectory $\Delta R_j(t)$ of residue j is the amount of uncertainty reduced in future values of $\Delta R_j(t+\tau)$ by knowing the past values of $\Delta R_i(t)$ for the given past values of $\Delta R_j(t)$. The amount of information is measured using Shannon's entropy. Following Schreiber's work, (Schreiber, 2000b) we write the transfer entropy as Eq 15 (stated in the methods of the previous section)

$$T_{i\to j}(\tau) = S\left(\Delta R_j\big((t+\tau)\big)|\Delta R_j(t)\right) - S\left(\Delta R_j\big((t+\tau)\big)|\Delta R_i(t), \Delta R_j(t)\right)$$

EQUATION 21

Here, $S\left(\Delta R_j\big((t+\tau)\big)|\Delta R_j(t)\right)$ is the conditional entropy of residue *j* at time $t+\tau$ given the values of $\Delta R_j(t)$ at time *t*. $S\left(\Delta R_j\big((t+\tau)\big)|\Delta R_i(t), \Delta R_j(t)\right)$ is the conditional entropy of residue *j* at time $t+\tau$ given the values of $\Delta R_i(t)$ and $\Delta R_j(t)$ at time *t*. The difference shows the amount of entropy reduced in the trajectory of *j* due to a knowledge of the past values of *i*. In terms of Shannon's entropy given in Boltzmann units, that is, the Boltzmann constant taken as unity $T_{i\to j(\tau)=S\left(\Delta R_j((t+\tau))|\Delta R_j(t)\right)-S\left(\Delta R_j((t+\tau))|\Delta R_i(t),\Delta R_j(t)\right)}$ Equation 15 reads as:

$$S\left(\Delta R_j((t+\tau))|\Delta R_j(t)\right) = -\langle lnp\left(\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle + \langle lnp\left(\Delta R_j(t)\right)\rangle$$

EQUATION 22

$$S\left(\Delta R_j((t+\tau))|\Delta R_i(t),\Delta R_j(t)\right) = +\langle lnp\left(\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle - \langle lnp\left(\Delta R_i(t),\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle$$

EQUATION 23

Detailed derivations of the expressions presented in this article are provided in the Supporting Information Material. In the coarse graining approximation, we focus only on the alpha carbon of each residue.

Substituting $S\left(\Delta R_j((t+\tau))|\Delta R_j(t)\right) = -\langle lnp\left(\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle + lnp\left(\Delta R_j(t)\right)$ Equation 22 and $S\left(\Delta R_j((t+\tau))|\Delta R_i(t),\Delta R_j(t)\right) = +\langle lnp\left(\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle - \langle lnp\left(\Delta R_i(t),\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle$ Equation 23 into Eq. $T_{i\to j}(\tau) = S\left(\Delta R_j((t+\tau))|\Delta R_j(t)\right) - S\,\Delta R_j((t+\tau))|\Delta R_i(t),\Delta R_j \quad t$ Equation 21, leads to

$$T_{i\to j}(\tau) = -\langle \mathrm{LN}\,p\left(\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle + \langle \mathrm{LN}\,p\left(\Delta R_i(t),\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle + \langle \mathrm{LN}\,p\Delta R_j(t)\rangle - \langle \mathrm{LN}\,p\left(\Delta R_i(t),\Delta R_j(t)\right)\rangle$$

EQUATION 24

It is possible to determine the time dependent probabilities shown in Eq. 16 by the dGNM(Erkip & Erman, 2004; T. Haliloglu, I. Bahar, & B. Erman, 1997). The Gaussian Network Model is characterized by the spring constants matrix Γ where a spring of constant unity is assumed between residues in contact. It is defined as follows: Γ_{ij} equates to −1 if alpha carbons of residues I and j are within a cutoff distance of r_c and to zero otherwise. Each ith diagonal element Γ_{ii} is equal to the negative sum of the ith row. The time correlation of fluctuations is given by the dGNM as (T. Haliloglu et al., 1997)

$$\langle \Delta R_i(t), \Delta R_j(t+\tau)\rangle = \sum_k A_{ij}(k) exp\{-\lambda_k \tau/\tau_0\} \qquad \text{EQUATION 25}$$

Where

$$A_{ij}(k) = \lambda_k^{-1} u_i^{(k)} u_j^{(k)} \qquad \text{EQUATION 26}$$

with λ_k being the k^{th} eigenvalue and $u_i^{(k)}$being the i^{th} component of the k^{th} eigenvector of the Γmatrix. The time delayed correlation given by $S\Delta Rjt{+}\tau\,|\Delta R_i(t), \Delta R_j(t)$⍰ $=$ $+\ \langle lnp\left(\Delta R_j(t), \Delta R_j(t+\tau)\right)\rangle -$ $lnp\left(\Delta R_i(t), \Delta R_j(t), \Delta R_j(t+\tau)\right)$ Equation 23 is essentially the solution of the equation of motion $m\Delta R'' + \zeta\Delta R' + \Gamma\Delta R = F(t)$ where m is the diagonal matrix of masses of residues, ΔR is the instantaneous fluctuation vector, the primes denote differentiation with respect to time, $\zeta$is the friction coefficient for each residue, $\Gamma$is the spring constants matrix and $F(t)$ is the random force vector operating on residues. Since the masses are much smaller than friction effects, the second derivative with respect to time is negligible. With this assumption, the resulting Langevin equation may be solved leading to the solution given by $S\Delta Rjt{+}\tau\,|\Delta R_i(t), \Delta R_j(t)$⍰ $=$ $+\ \langle lnp\left(\Delta R_j(t), \Delta R_j(t+\tau)\right)\rangle -$ $lnp\left(\Delta R_i(t), \Delta R_j(t), \Delta R_j(t+\tau)\right)$ Equation 23 . Full details of the solution leading to $S\Delta Rjt{+}\tau\,|\Delta R_i(t), \Delta R_j(t)$⍰ $=$ $+\ \langle lnp\left(\Delta R_j(t), \Delta R_j(t+\tau)\right)\rangle -$ $lnp\left(\Delta R_i(t), \Delta R_j(t), \Delta R_j(t+\tau)\right)$ Equation 23 are explained in Ref. (Erkip & Erman, 2004).

The probability distribution of a Gaussian of n variables, $\Delta R= [\Delta R_1, \Delta R_2, \Delta R_3,\ldots, \Delta R_n]$ is

$$p(\Delta R) = \frac{1}{\sqrt{(2\pi) det\Gamma^{-1}}} exp\left(-\frac{1}{2}(\Delta R)^T \Gamma(\Delta R)\right) \qquad \text{EQUATION 27}$$

where, Γ^{-1} is the matrix of covariance $(\Gamma^{-1})_{ij} = \langle \Delta R_i \Delta R_j\rangle$

Taking the logarithm of $\langle \Delta R_i(t), \Delta R_j(t+\tau)\rangle = k\text{⍰⍰}A_{ij}(k)exp\ -\lambda_k\tau/\ \ \tau 0$ Equation 25 and averaging and substituting into $T_{i\to j(\tau)= -\langle \ln p\left(\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle+\langle \ln p\left(\Delta R_i(t),\Delta R_j(t),\Delta R_j(t+\tau)\right)\rangle+\langle \ln p\Delta R_j(t)\rangle-\langle \ln p\left(\Delta R_i(t),\Delta R_j(t)\right)\rangle}$ E quation 16leads to the final expression for entropy transfer in the dGNM as:

$$
\begin{aligned}
T_{i\to j} = \frac{1}{2} ln &\left(\left(\sum_k A_{jj}(k)\right)^2 - \left(\sum_k A_{jj}(k) exp\left\{-\frac{\lambda_k\tau}{\tau_0}\right\}\right)^2\right) \\
&- \frac{1}{2} ln\left[\left(\sum_k A_{ii}(k)\right)\left(\sum_k A_{jj}(k)\right)^2\right. \\
&+ 2\left(\sum_k A_{ij}(k)\right)\sum_k A_{jj}(k)\, exp\left\{-\frac{\lambda_k\tau}{\tau_0}\right\}\sum_k A_{ij}(k)\ \ exp\left\{-\frac{\lambda_k\tau}{\tau_0}\right\} \\
&- \left\{\left(\sum_k A_{ij}(k)\, exp\left\{-\frac{\lambda_k\tau}{\tau_0}\right\}\right)^2 + \left(\sum_k A_{ij}(k)\right)^2\right\}\left(\sum_k A_{jj}(k)\right) \\
&- \left(\sum_k A_{jj}(k)\, exp\left\{-\frac{\lambda_k\tau}{\tau_0}\right\}\right)^2\left(\sum_k A_{ii}(k)\right) \\
&- \frac{1}{2}\left[\left(\sum_k A_{jj}(k)\right)\right] \\
&+ \frac{1}{2} ln\left(\left(\sum_k A_{ii}(k)\right)\left(\sum_k A_{jj}(k)\right) - \left(\sum_k A_{ij}(k)\right)^2\right)
\end{aligned}
$$

EQUATION 28

The net entropy transfer from residue *i* to *j* is the difference of the entropy transfer from *i* to *j* and entropy transfer from *j* to *i*:

$$Net\ Entropy\ Transfer\ from\ i\ to\ j = T_{i\to j}(\tau) - T_{j\to i}(\tau)$$ EQUATION 29

The entropy transfer,$T_{i\to\odot}(\tau)$, out from residue *i* to the rest of the protein is obtained by the sum over *j* as

$$T_{i\to\odot}(\tau) = \sum_j T_{i\to j}\ (\tau)$$ EQUATION 30

Defined in this way, $T_{i\to\odot}(\tau)$ is a measure of entropic activity of residue i. If the residue transfers entropy to the rest of the protein, we call it an entropy source, if it accepts entropy from the rest of the protein, it is an entropy sink.

The time delay τ that appears in the equations depends on the spring constant of the harmonic interactions. Molecular dynamics simulations of proteins at physiological temperatures shows that fluctuations of two residues *i* and *j* are in the order of 5 nanoseconds (Hacisuleyman & Erman, 2017b; Vatansever et al., 2016). However, it is not possible to establish an exact quantitative correspondence between the dGNM values and real time parameters. In all the calculations below, we took τ as the time for which correlations decay to 1/*e* of their initial values.

Entropy transfer from *i* to *j* shows how the fluctuations of *i* drive the fluctuations of *j*: The entropy transfer model measures the amount of information transfer between the trajectories of two atoms, *i* and *j*. Specifically, entropy transfer is defined as follows: Knowing the fluctuations of atoms *i* and *j* at time *t*, entropy transfer is the amount of uncertainty reduced in the future fluctuations of atom *j*. One extreme case is where the fluctuations of *i* have no effect on the fluctuations of *j*, that is, their trajectories are uncorrelated. In this case, $T_{i\to j(\tau)=S\left(\Delta R_j((t+\tau))|\Delta R_j(t)\right)-S\left(\Delta R_j((t+\tau))|\Delta R_i(t),\Delta R_j(t)\right)}$ Equation 15 equates to zero, and no entropy will be transferred to residue *j* from *i*. The other extreme case is where the fluctuations of *i* at time *t* are perfectly locked into those of *j* at time *t* and the knowledge of the present values of *i* fluctuations will reduce the uncertainty in the fluctuations of *j*. In this case, the second term $S\left(\Delta R_j\left((t+\tau)\right)|\Delta R_i(t),\Delta R_j(t)\right)$ in $T_{i\to j(\tau)=S\left(\Delta R_j((t+\tau))|\Delta R_j(t)\right)-S\left(\Delta R_j((t+\tau))|\Delta R_i(t),\Delta R_j(t)\right)}$ Equation 15 will be modified as $S\left(\Delta R_j\left((t+\tau)\right)|\Delta R_j(t)\right)-S\left(\Delta R_i(t)\right)$, which states that the reduction in uncertainty is due to the locking in of the fluctuations of *i* to those of *j* at time *t*. Substituting these into $T_{i\to j(\tau)=S\left(\Delta R_j((t+\tau))|\Delta R_j(t)\right)-S\left(\Delta R_j((t+\tau))|\Delta R_i(t),\Delta R_j(t)\right)}$ Equation 15 leads to $T_{i\to j}(\tau)=S(\Delta R_i)$. The range of values of entropy that may be transferred from *i* to *j* will lie within the interval $0\le T_{i\to j}(\tau)\le S(\Delta R_i)$. In the absence of symmetry, $T_{i\to j}(t)\ne T_{j\to i}(t)$, and there will be a net transfer of entropy from *i* to *j* which will lie in the interval

$-S(\Delta R_j) \leq T_{i\rightarrow j}(\tau) - T_{j\rightarrow i}(\tau) \leq S(\Delta R_i)$. If $T_{i\rightarrow j}(\tau) - T_{j\rightarrow i}(\tau)$ is greater than zero, we say that the fluctuations of atom i reduce the uncertainty in the fluctuations of *j* because they are coupled to the fluctuations of j. In this case, we say that *i* drives j.

With this interpretation of entropy transfer, we study three allosteric proteins in the following section.

2.3.3 RESULTS

Using the dGNM version of transfer entropy developed in this article, we studied entropy transfer in the three proteins, Ubiquitin, the Pyruvate Kinase (PK) tetramer, and PDZ.

2.3.3.1 ALLOSTERIC MECHANISM OF UBIQUITIN

Ubiquitin, (PDB ID: 1UBQ), is a 76 amino acid protein. It consists of 8 distinct secondary structures that actively take part in interactions with a large number of proteins. The three-dimensional structure of Ubiquitin (1UBQ.PDB) is presented in Figure 13.

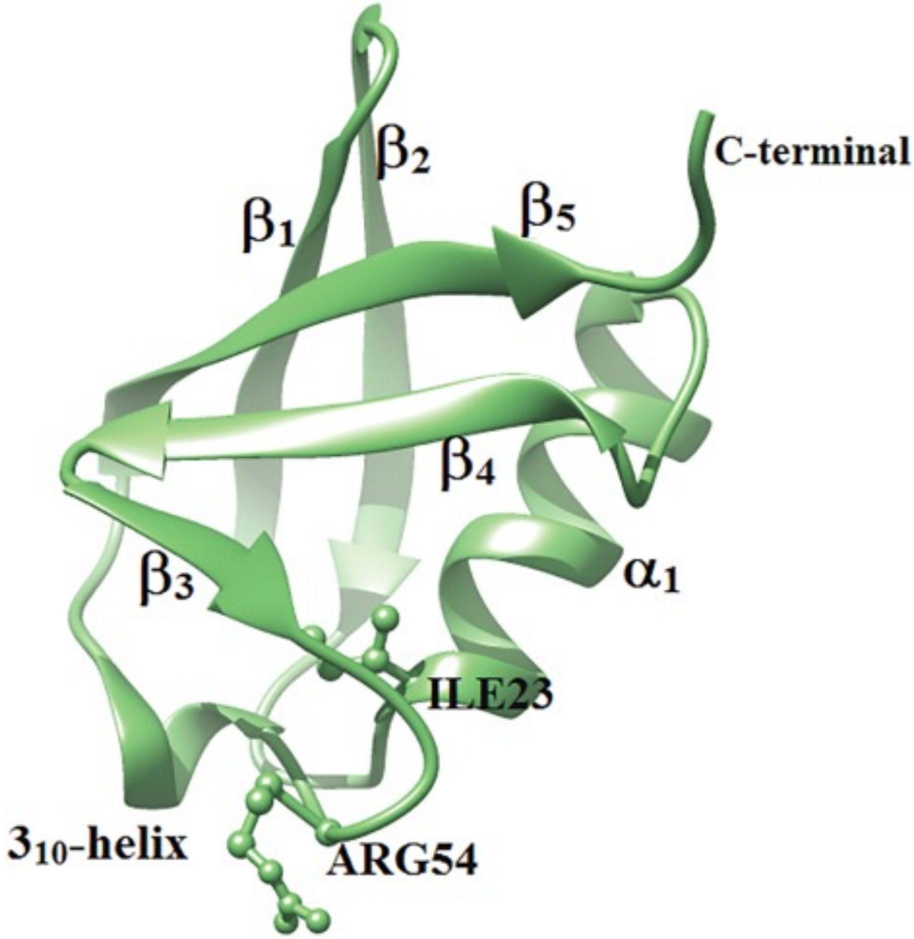


FIGURE 13. THREE-DIMENSIONAL STRUCTURE OF UBIQUITIN IN SOLID RIBBON FORMAT. THE SECONDARY STRUCTURES ARE SHOWN IN DIFFERENT COLORS. IMPORTANT RESIDUES IN ALLOSTERIC ACTIVITY ARE SHOWN IN BALL AND STICK FORMAT.

The allosteric mechanism of Ubiquitin has been understood recently by Smith *et al*.(Smith et al., 2016) Using Relaxation Dispersion measurements, they showed that two residues GLU24 and GLY53 act as a conformational switch and their correlated fluctuations induce breathing-like motions in the overall protein, which affect the substrate binding region of Ubiquitin. Earlier experiments using NMR Relaxation Dispersion measurements also showed that strong correlations exist between two residues, ILE23 and ASN25 and between three residues ILE23, ASN25, and THR55(Ban et al., 2011; Dittmer & Bodenhausen, 2004; Massi et al., 2005) .

Here, we show that allosteric activity of Ubiquitin is closely associated with entropy transfer in the system. In Figure 14, we present the net entropy transfer from a given residue to the rest of the protein obtained by using Eq. (28; lower curve) and by a 600 ns molecular dynamics trajectory (Hacisuleyman & Erman, 2017b).

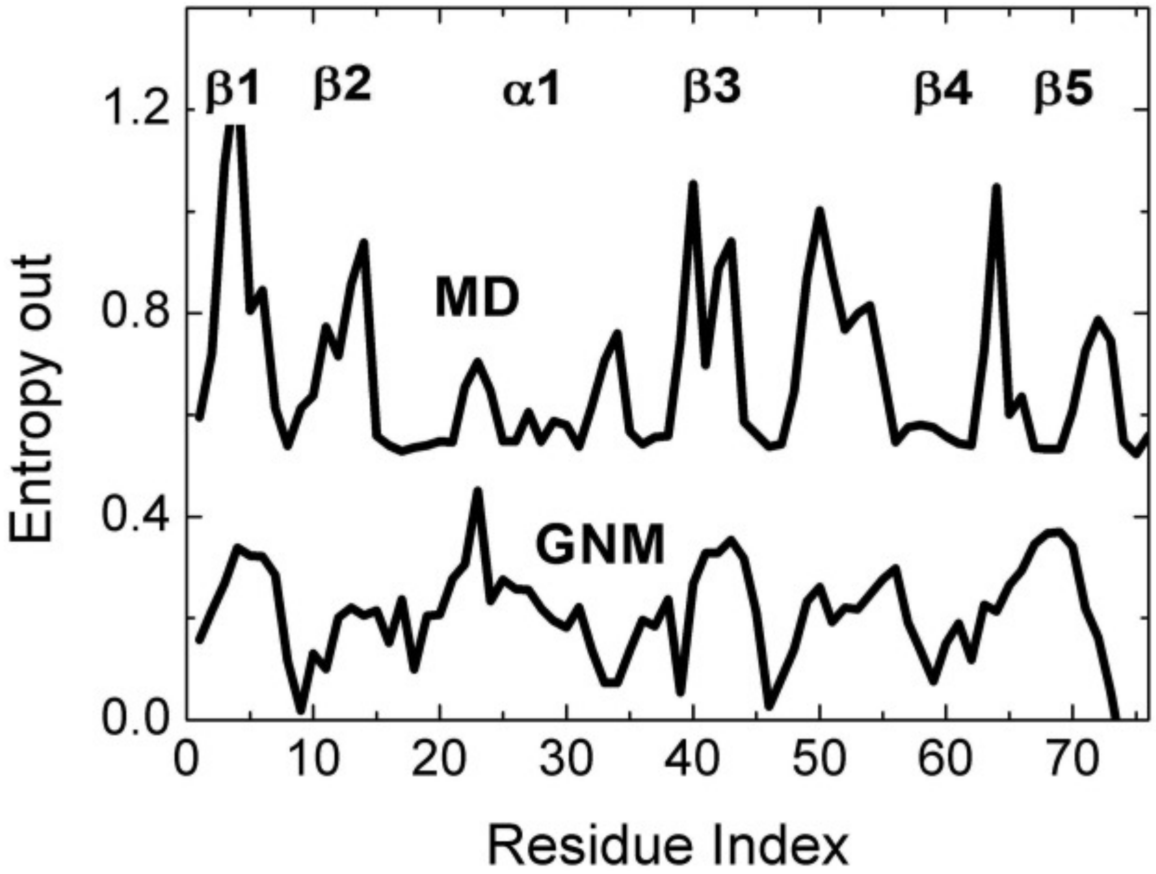


FIGURE 14. ENTROPY OUT FROM RESIDUES FOR UBIQUITIN. LOWER CURVE IS OBTAINED USING DGNM, UPPER CURVE IS FROM MOLECULAR DYNAMICS SIMULATIONS.

Comparison of total entropy transfer from a residue, obtained by dGNM and by extensive molecular dynamics simulations show good agreement as may be observed from Figure 14.

T

he values of $T_{i\to\odot}(\tau)$ for Ubiquitin from

R

E

F

_Ref22853415 \h * MERGEFORMAT

$Ti \to j$ $= \frac{1}{2} ln\left(\left(\sum_k A_{jj}(k)\right)^2 - \left(\sum_k A_{jj}(k) exp\left\{-\frac{\lambda_k \tau}{\tau_0}\right\}\right)^2\right) -$

1

2

⍰

Equation 28are shown by the lower curve indicated as GNM in Figure 14. The upper curve shows the results of MD simulations. Both curves are obtained for the entropy associated with the alpha carbons of the residues. The GNM data is scaled and translated relative to the MD data for easier comparison. The secondary structures of Ubiquitin are shown in the upper part of the figure, where α and β stand for alpha helix and beta strand, respectively. The turns between the secondary structures are not indicated in the figure. Results show that the secondary structures β_1, β_2, β_3, the N-terminal of the α_1 helix, and the loop between β_4 and β_5 are the entropically active regions of the protein.

Pearson correlation analysis $\left(\langle\Delta R_i \Delta R_j\rangle / \left(\langle\Delta R_i{}^2\rangle^{1/2}\langle\Delta R_j{}^2\rangle^{1/2}\right)\right)$ from GNM shows that LEU8, SER20, ILE23, GLU24, GLY53, LEU54, and LEU56 are the most strongly correlated residues and we focus on these residues in Figure 15 and Figure 16 below. In Figure 15, we show the entropy exchange profile of Ubiquitin. The abscissa and the ordinate show the indices of the residues that give and accept entropy, respectively. The columns of black points identify the residues that give entropy to the rest of the protein. The six residues mentioned above all indicate entropy transfer activity according to this figure. We analyze each of them individually in Figure 16. In addition to these six residues, the residues ARG72 to GLY76, that is, the C-terminal tail all act as entropy donors, as may be seen by the horizontal strip of black points in Figure 15.

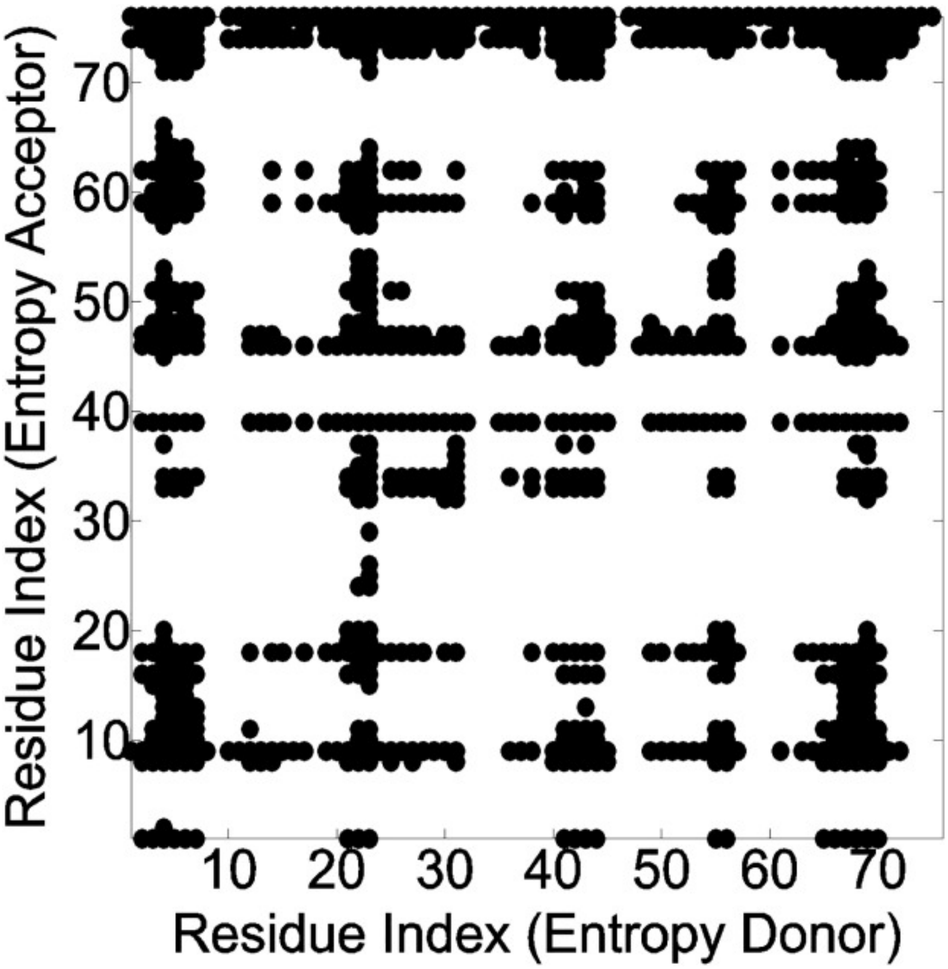


FIGURE 15. ENTROPY DONOR AND ACCEPTOR RESIDUE PAIRS OF UBIQUITIN. THE BLACK POINTS CORRESPOND TO RESIDUE PAIRS WHOSE NET ENTROPY TRANSFER VALUES ARE CALCULATED BY EQ. 27. EACH POINT SHOWS THAT THERE IS NET ENTROPY TRANSFER FROM A RESIDUE INDEXED ON THE ABSCISSA TO A RESIDUE INDEXED ON THE ORDINATE.

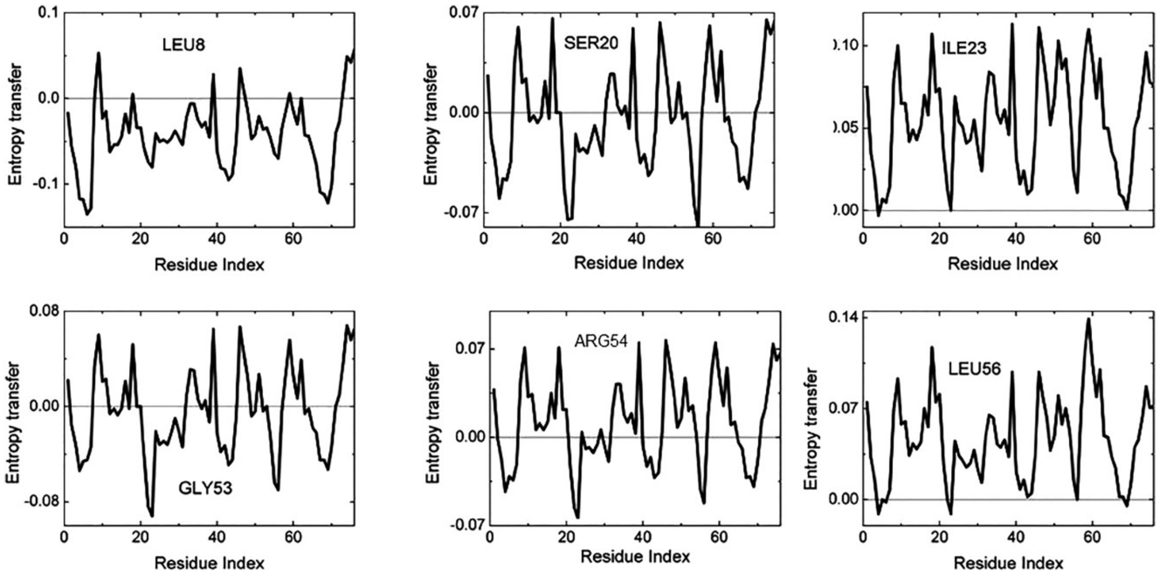


FIGURE 16. NET ENTROPY TRANSFER FROM LEU8, SER20, ILE23, GLY53, LEU54, AND LEU56 TO OTHER RESIDUES, OBTAINED FROM EQ. 27.

The ordinate of each panel in Figure 16 shows net entropy transfer from the residue identified in the panel to all others in the protein, calculated from $p\Delta R$⍰ $= \frac{1}{\sqrt{(2\pi)det\Gamma^{-1}}}exp$ $-\frac{1}{2}(\Delta R)^T\Gamma(\Delta R)$⍰ Equation 27. Positive values of the ordinate shows that net entropy is transferred from the indicated residue to the others indexed along the abscissae. Negative values show net entropy transfer to the indicated residue in each panel. ILE23 and LEU56 are strong entropy donors since they all have positive ordinate values. According to the model, the fluctuations of these residues drive the motions of others in the protein, which agrees with the experimental observation(Smith et al., 2016) that these two residues constitute the allosteric switch. LEU8 is a strong entropy acceptor and SER20, GLY53, and LEU54 are entropy donors for some residues and acceptors for others of Ubiquitin. Some observations from the panels of Figure 16. Net entropy transfer from LEU8, SER20, ILE23, GLY53, LEU54, and LEU56 to other residues, obtained from Eq. 27.that may be used to interpret the experimental data (Ban et al., 2011; Dittmer & Bodenhausen, 2004; Massi et al., 2005) on Ubiquitin are as follows: (1) ILE23 transfers entropy to GLY53 (third panel first row and first panel on second row in Figure 16). Thus, ILE23 drives the fluctuations of GLY53. This illuminates the mechanism of the allosteric switch. It is to be noted that experiments can only identify the residues but not the direction of entropy flow,

(2) ILE23 and LEU56 are essentially the two important residues that drive the motions of the rest of the protein, and (3) rapid oscillations of the curves in going from one residue to its close neighbors along the primary structure show that there is strong entropy transfer along the backbone. For example, ILE23 gives entropy to its neighbors THR22 and GLU24 as may be seen from the first panel on row 2 of Figure 16.

2.3.3.2 ALLOSTERIC MECHANISM OF PYRUVATE KINASE

PK from *Leishmania mexicana* (PDB ID:3HQQ), which is a homotetrameric enzyme having 1992 amino acids in total and 498 amino acids per protomer, each consisting of 28 secondary structures and 4 domains named as N-terminal (residues between 1 and 17), A (residues between 18 and 88 and 187 and 356), B (residues between 89 and 186) and C (residues between 357 and 498). It catalyzes the last step in glycolysis and is known to be allosterically activated by the binding of fructose biphosphate (FBP). Allosteric communication of PK takes place at two different scales, (i) intertetramer and (ii) intraprotomer communication. Details of communication at both scales have been determined experimentally by Morgan *et al*. (Morgan et al., 2010; Morgan et al., 2013)At the inter-tetramer scale, ligand free, and allosterically activated states of PK differ in rigid body rotations of the A and C subunits. This rotation results in ARG310 making hydrogen bonds with ARG262 and GLY263 on an alpha helix and this helix unwinds in the absence of FBP. ARG310 plays a crucial role because it shows the greatest conformational changes between the inactive and active states of the enzyme, but it has been proven that it is necessary but not sufficient for allosteric transition, and allosteric communication at the intraprotomer scale is required. This is realized by the presence of communication between FBP and the highly mobile B domain. FBP interacts with GLU451 and GLY487 in the C-domain and affects the motion of the B-domain by increasing the rigidity of the enzyme and consequently its activity although it binds to a site >40 Å away from the active site.(Jurica et al., 1998; Morgan et al., 2010; Naithani, Taylor, Erman, & Walkinshaw, 2015) Below, we show that the residues that are active in allosteric communication both at the inter-tetramer and intraprotomer scales may be identified by the dGNM. We used the alpha carbons of the crystal structure of the apo form of PK to determine the communication patterns using Eq. 27.

In Figure 17, we present the communication landscape for the tetramer. Experimentally significant points mentioned in the preceding paragraph coincide with the peaks of Figure 17. The communication landscape of PK may be obtained by the dGNM model in <1 min.

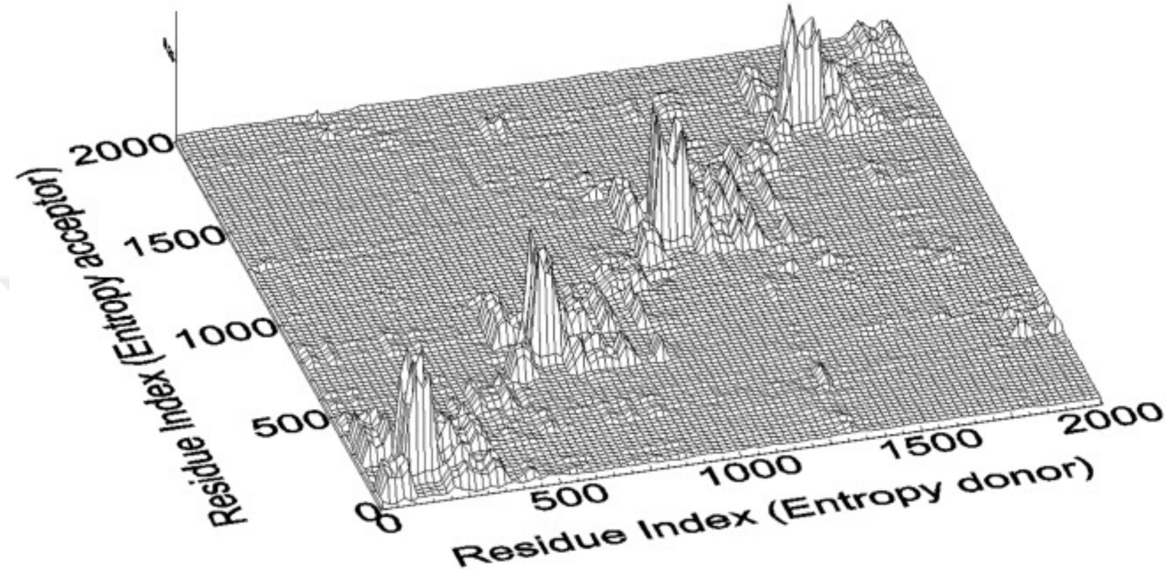


FIGURE 17. COMMUNICATION LANDSCAPE OF PK. PEAKS AT FAR OFF-DIAGONAL REGIONS REPRESENT INTERPROTOMER INTERACTIONS THAT ARE RESPONSIBLE FOR THE LARGE-SCALE ROCKING MOTIONS IN THE TETRAMER.

Entropy transfer between protomers cannot be readily seen in the resolution of Figure 17. As discussed in the preceding paragraph, ARG310 plays a significant role in establishing the allosteric communication between protomers.

In Figure 18, we show the entropy transfer from the ARG310 of one protomer (marked by the arrow on the abscissa) to other residues of the tetramer. The protomers of the tetramer are separated from the others by a thin vertical line in the figure. The location of each protomer (darkened region) in the tetramer is shown at the upper part of the figure. The peaks of points in Figure 18 identify the residues to which ARG310 of one protomer donates entropy. According to the figure, ARG310 of a given protomer provides entropy to residues centered around ARG262 and GLY263 (the large peaks in each quadrant). This is the dominant mode through which entropy is transferred from one protomer to the other in the tetramer. We also see that ARG310 provides entropy to the region between residues

3 and 70 (smaller peaks in each quadrant). It is interesting to note that there is entropy transfer from the ARG310 of a protomer to another protomer with which it is not in direct contact. Entropy transfer within one protomer is shown in Figure 19. The left panel is a dot plot that gives residue pairs indexed along the abscissa for entropy donors and along the ordinate for entropy acceptors. Here, we see that the B-domain, the strip of points indicated by residues 89–186 along the abscissa, is the dominant entropy donor. The B-domain donates entropy to almost all other residues of the protein. The right panel shows the amount of entropy given by the B-domain to other residues, indexed along the abscissa, of the protein. There are several peaks corresponding to residues: 21, 48, 78, 87, 206, 236, 259, 290, 325, 356, 379, 410, 436, 454, 480, and 492. Most of these residues or their immediate neighbors along the chain are identified as allosterically important residues in recent experimental studies (Morgan et al., 2010; Morgan et al., 2013).

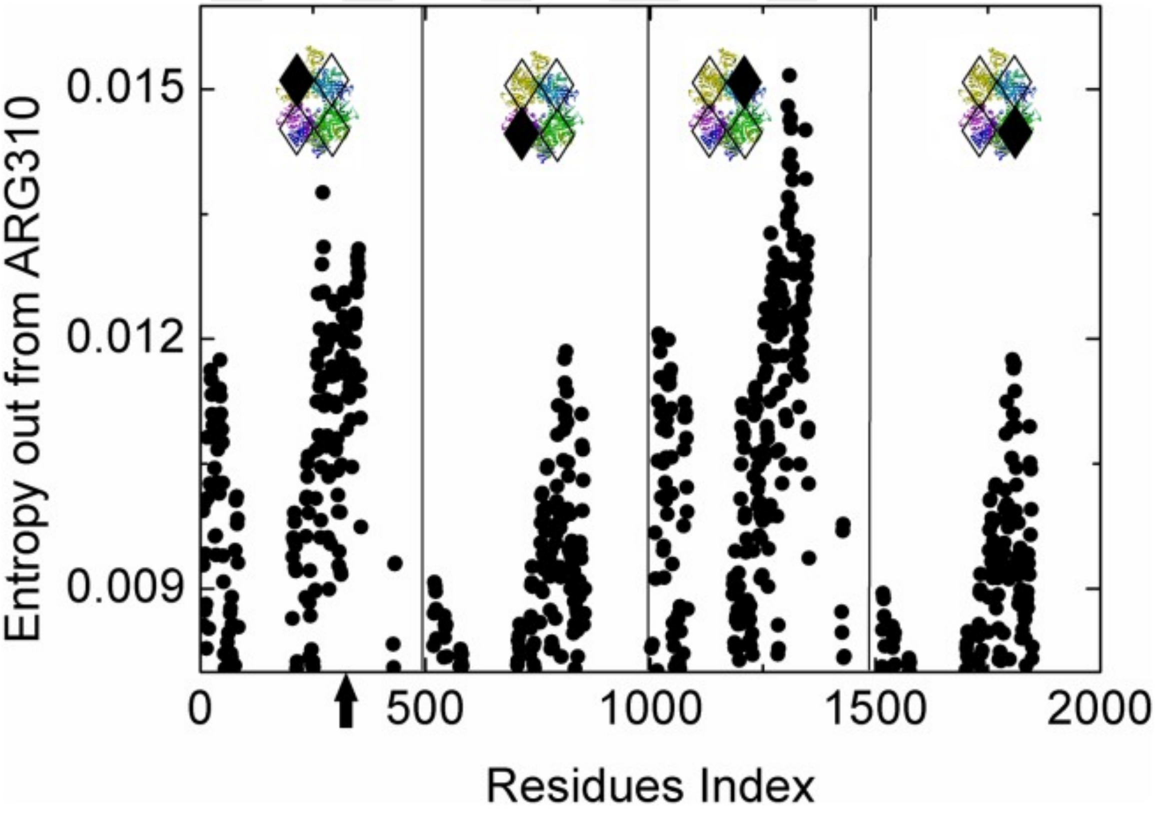


FIGURE 18. NET ENTROPY TRANSFER FROM ARG310 OF ONE PROTOMER TO OTHER PROTOMERS, OBTAINED FROM EQ. 27.

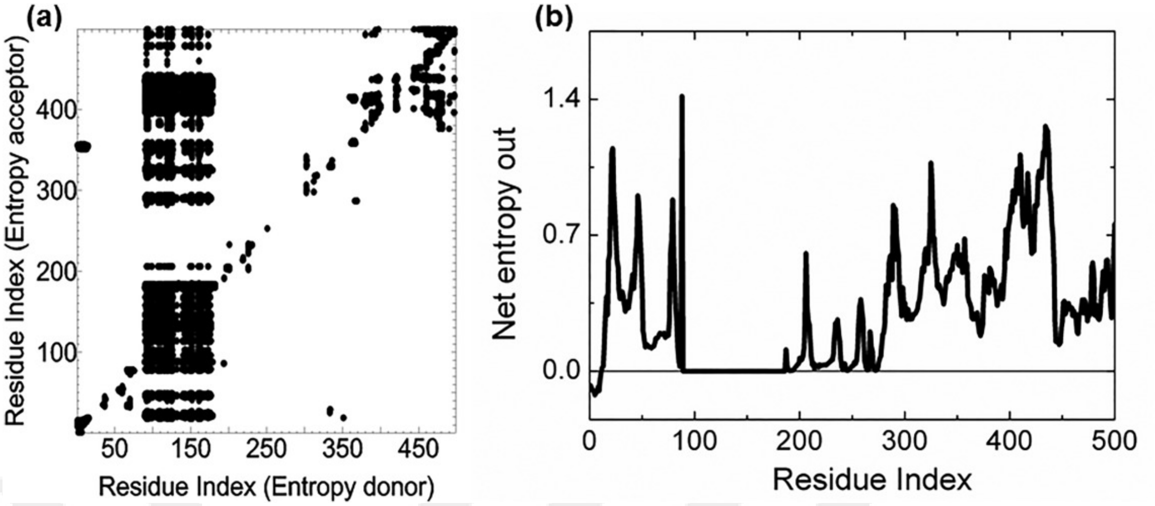


FIGURE 19. (**A**) ENTROPY DONOR AND ACCEPTOR RESIDUE PAIRS OF A SINGLE PROTOMER OF PK. BLACK POINTS CORRESPOND TO RESIDUE PAIRS WHOSE NET ENTROPY TRANSFER VALUES ARE CALCULATED BY EQ. 27. EACH POINT SHOWS THAT THERE IS NET ENTROPY TRANSFER FROM A RESIDUE INDEXED ON THE ABSCISSA TO A RESIDUE INDEXED ON THE ORDINATE. (**B**) ENTROPY TRANSFERRED FROM THE B-DOMAIN RESIDUES TO THE REMAINING RESIDUES OF THE PK PROTOMER.

2.3.3.3 ALLOSTERIC MECHANISM OF THE PDZ DOMAIN

PDZ domain proteins, which constitute a large family, function as information transmitters from one protein to another and mediate key cellular functions in the cell. They are known as allosteric proteins. Suel *et al*.(Suel, Lockless, Wall, & Ranganathan, 2003) showed the presence of evolutionarily conserved pathways in PDZ structures. Following their work, several papers were published pointing to the presence and determination of such pathways (C. N. Chi et al., 2008; Z. N. Gerek & S. B. Ozkan, 2011; C. M. Petit, J. Zhang, P. J. Sapienza, E. J. Fuentes, & A. L. Lee, 2009). Among these articles, the work of Petit *et al*.(C. M. Petit et al., 2009) showed that a member of the PDZ family, the protein PSD-95 contained a distal alpha helix, α3, the removal of which did not affect the stability of the protein but resulted in a 21-fold decrease in the binding affinity of the protein to a peptide. This effect was explained to be an allosteric activity of entropic origin. Here, we show that the distal alpha helix indeed interacts with the rest of the protein entropically. We also show that removal of this interaction results in severe changes in entropy transfer in the protein. The structure of PSD-95 (PDB ID: 1BFE) is shown in Figure 20, where the left panel is the complete protein and the right panel is the one with the missing α3.

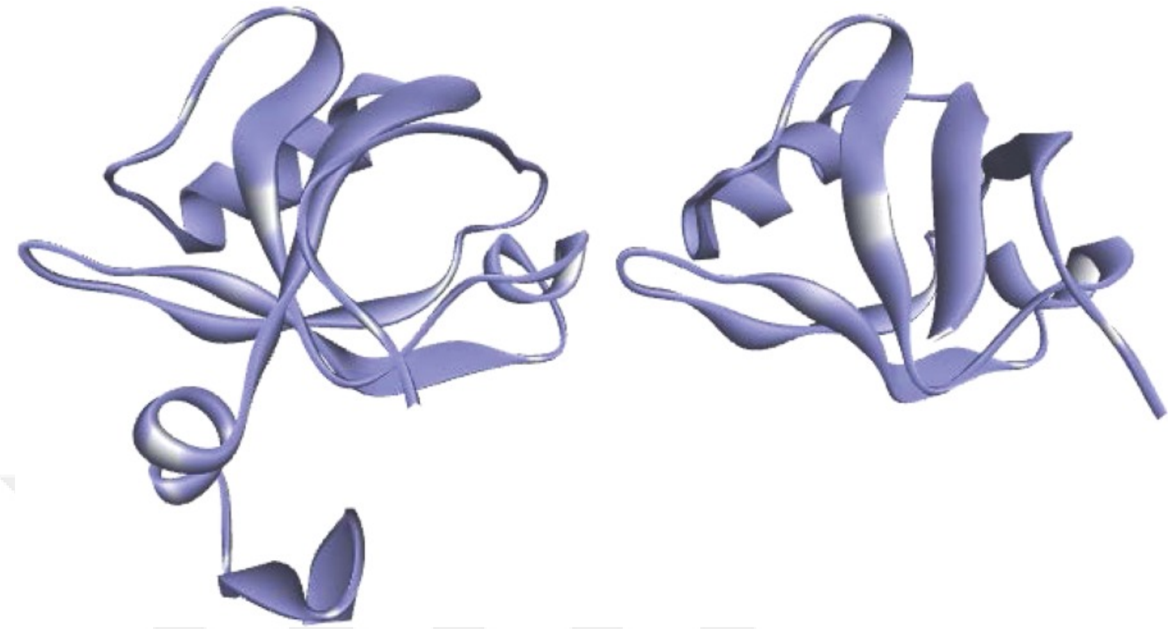


FIGURE 20. THE STRUCTURE OF PSD-95 (PDB ID: 1BFE). THE LEFT PANEL IS THE COMPLETE PROTEIN AND THE RIGHT PANEL IS THE ONE WITH THE MISSING A3.

The net entropy transfer from a residue identified along the abscissa to the one identified along the ordinate obtained using Eq. 27 is shown in Figure 21.

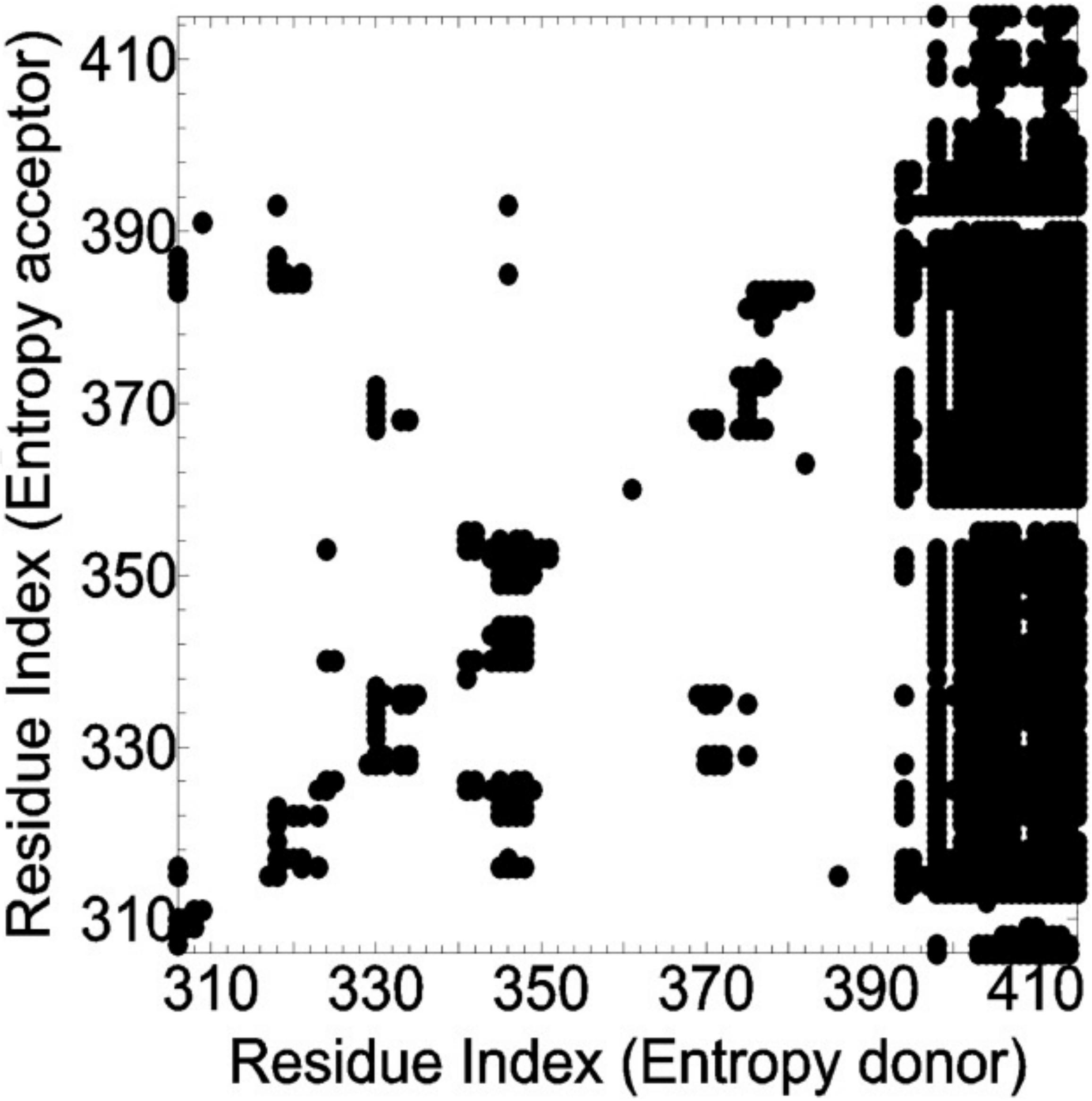


FIGURE 21. POINTS SHOWING THE PRESENCE OF NET ENTROPY TRANSFER BETWEEN PAIRS OF RESIDUES. THE ENTROPY DONOR OF THE PAIR IS IDENTIFIED ALONG THE ABSCISSA AND THE ACCEPTOR ALONG THE ORDINATE.

The strip of black dots forming a vertical region identified by the residue indices 393–415 are those of the distal helix. We see that these residues, constituting the C-terminal of the protein act as an entropy source for the remaining residues of the protein. For example, the phosphorylated TYR397 is in this region and is functionally important.(C. M. Petit et al., 2009) InFigure 22, we show how TYR397 interacts with the remaining residues of the protein. A positive value of the ordinate for a residue indicates that TYR397 drives the fluctuations of that residue and vice versa. Thus, TYR397 drives the beta strand between PRO311-HIS317, the residue GLY356, the beta-turn between ILE359-ALA370, and the beta-turn-helix cap between LYS380-ILE389. It is driven by the residues GLY344-LEU349. We also see strong entropy transfer between TYR397 and its close neighbors along the primary structure.

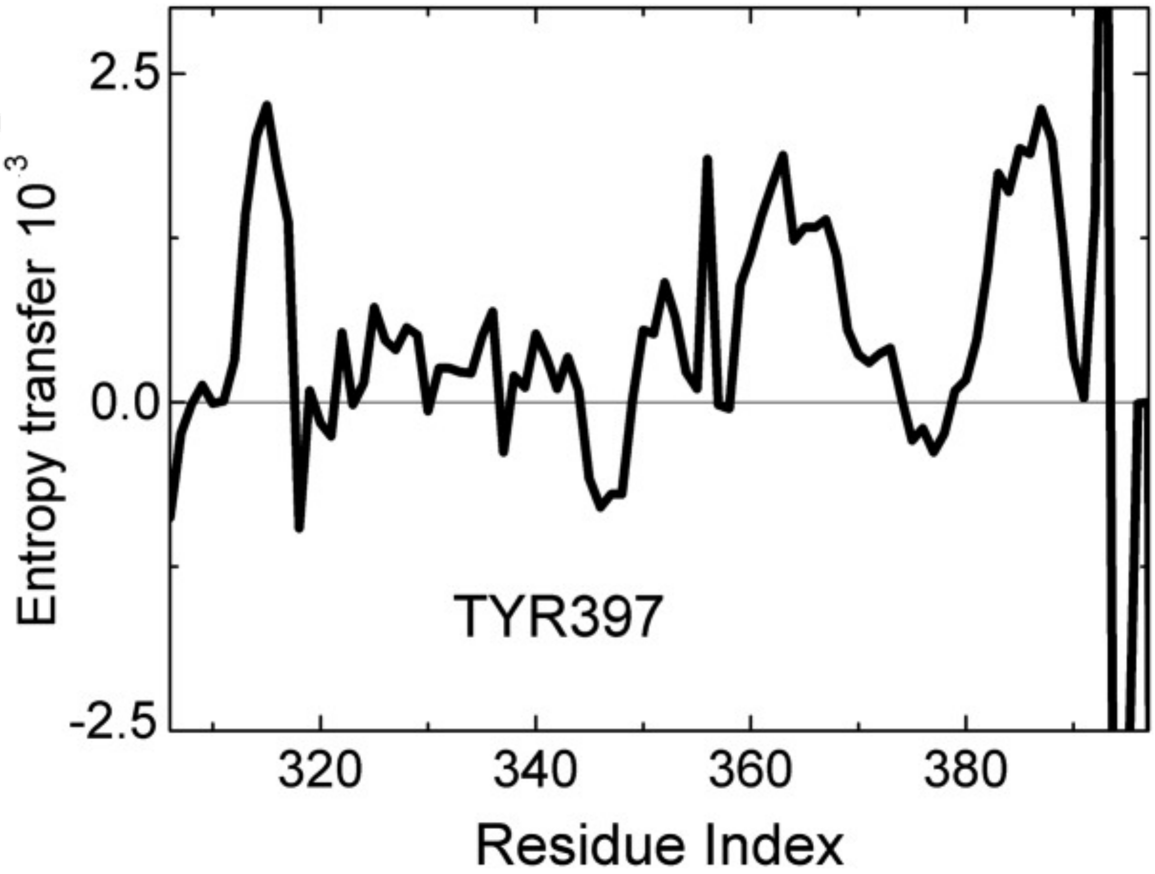


FIGURE 22. NET ENTROPY TRANSFER FROM TYR397 TO OTHER RESIDUES OF THE PROTEIN, OBTAINED BY EQ. 27. THE RESIDUE ALA347 HAS BEEN SHOWN TO LIE ON THE ALLOSTERIC PATHWAY OF THE PROTEIN.(Z. N. GEREK & S. B. OZKAN, 2011)

Results of entropy transfer calculations presented in Figure 21 show that ALA347 shows the largest entropic interactions with other residues of the protein. In Figure 23, we compare the interactions of ALA347 in the presence and absence of α3. The thin line is obtained in the presence of α3 and the thick line is in the absence of it. An examination of the three-dimensional structure shows that ALA347 makes hydrogen bonds with LEU323, LEU349, and SER350, and is in steric interaction with GLY345 and ASP348. Upon removal of α3, we see from Figure 23 that the entropic interactions of GLY345 with these five residues decrease significantly. Some of these residues are directly involved in the binding of the ligand CRIPT to PDZ. Experiments show that removal of α3 decreases the binding affinity of CRIPT to PDZ by 21-fold (C. M. Petit et al., 2009).

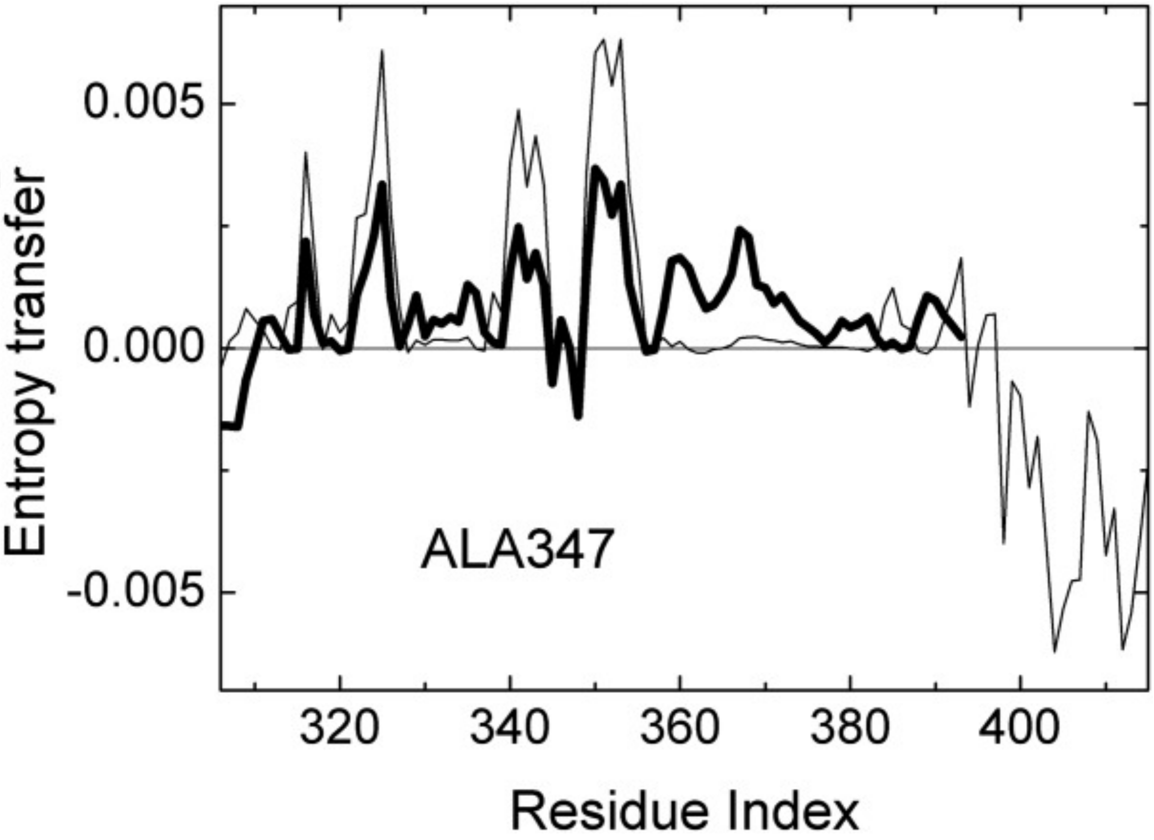


FIGURE 23. CHANGE IN THE INTERACTIONS OF ALA347 UPON REMOVAL OF THE DISTAL HELIX A3. THIN LINE OBTAINED FROM EQ. 27 IN THE PRESENCE OF A3, THICK LINE IN THE ABSENCE.

2.3.4 DISCUSSION

In this article, we present a fast method of determining allosteric communication landscapes in proteins using the dynamic version of the Gaussian Network Model, which is based on harmonic interactions between contacting residues. The model is built on the transfer entropy concept by Schreiber and shows that knowing only the energy landscape is not sufficient to predict information transfer and allosteric communication, and that time delayed correlations are necessary. The allosteric communication landscapes presented in

Figure 15, Figure 17and Figure 21 show that information transfer in proteins does not necessarily take place along a single path but involves several residues over an ensemble of pathways. The model also emphasizes that knowledge of entropy only is not sufficient for determining allosteric communication and additional information based on time delayed correlations must be introduced, which leads to the presence of causality in proteins. The possibility of causality in proteins allows for identifying driver-driven relations for pairs of residues. The GNM method of entropy transfer provides a rapid tool for determining the allosteric communication landscape for proteins. Construction of a landscape for a protein as large as 2000 residues now takes less than one minute on an ordinary laptop. We performed a comparative analysis of allosteric communication between residue pairs in Ubiquitin with the present method and showed that the results are in good agreement with molecular dynamics-based predictions. Evaluating the communication map for the PK by molecular dynamics would take several months on a supercomputer. With the GNM approach, we could determine the allosteric communication landscape of the 1992 residue protein PK. The results agree, although not shown in detail in this study, with experimentally known allosteric communication features of the complex. The GNM entropy transfer model provides a simple tool, which maps the entropy sink-source relations into pairs of residues. By this approach, residues that should be manipulated to control protein activity may be determined. This should be of great importance for allosteric drug design and for understanding effects of mutations on protein function.

2.4 AMOUNT AND RATE OF INFORMATION TRANSFERRED WITHIN THE PROTEIN

2.4.1 INTRODUCTION

Proteins are dynamic systems whose atoms exhibit fluctuations about their equilibrium positions with amplitudes in the order of nanometers and characteristic times of pico to nanoseconds. When observed individually, each atom performs fluctuations as an independent random stationary process that may be described in terms of a time-amplitude trajectory. Fluctuations and randomness of motion are built-in sources of uncertainty at the nano-scale as clearly displayed in any atomic trajectory. Considering the trajectories of pairs of atoms simultaneously, however, gives important clues on how the two atoms communicate with each other. Any degree of coupling between pairs of trajectories leads to a hint for the function of the protein. Coupling of two trajectories may be analyzed most conveniently in terms of information transfer from one to the other. Information transfer from trajectory *i* to *j* is the amount of uncertainty reduced in the states of *j* at a future time $t+\tau$ due to its coupling with the trajectory *i* at time *t(Schreiber, 2000b)*. The basic concept of information transfer relies on the calculation of average number of bits needed to encode independent events by using Shannon's entropy formulation. Shannon was interested in the capacities of telecom lines to transfer information by using minimum number of bits and derived a procedure that is essentially based on the maximum entropy principle (Max Ent) (S. Pressé, K. Ghosh, J. Lee, & K. A. Dill, 2013) and the maximum caliber, Max Cal (Dixit et al., 2018). The latter is used for predicting the relative probability that a system will take a certain trajectory in going from one state to another. As will be shown in the following sections, information transfer between two residues is a special application of Max Cal where only states allowable for a pair of residues are considered. A consequence of Max Cal formalism for pairwise interactions is that the trajectory may be treated as a Markov process (Ge, Pressé, Ghosh, & Dill, 2012). The information transfer formalism adopted here is based on conditional probabilities derived from a Markov process. With this perspective, information transfer between residues lies at the root of the dynamic view of allostery in proteins: perturbation at one site, called the allosteric site, leads to changes of conformations as well as dynamics in other regions of the protein including the site at which the protein performs its function. The present paper may be regarded as a paradigm shift from static coupling of two residues to coupling of trajectories. Coupling of trajectories may be classified into three types in terms of residue pair distances along the primary chain and in space. Residue pairs that are (i) close along the primary chain and in space, which we refer to as Type 1 coupling, (ii) distant along the primary chain but close in space, Type 2 coupling, and (iii) distant along the chain and space, Type 3 coupling. The role of Type 1 and 2 coupling in allostery, evolution, and protein function in general is well documented. The role of Type 3 coupling needs more elaboration. Starting in the past decade, information transfer in allostery has been associated with evolutionary processes. First, Ranganathan's group identified evolutionarily conserved pathways in coevolved protein families (Lockless & Ranganathan, 1999; Suel et al., 2003) and quantified Type 3 coupling on such pathways according to Boltzmann statistics, which is now referred to as 'statistical coupling'. Role of Type 3 coupling in coevolution was soon challenged by Chi et al.,(C. N. Chi et al., 2008) who determined changes in free energies resulting from mutations of the proposed statistically coupled residue pairs. Chi et al., concluded that the observed coupling, which they referred to as 'energetic coupling', is in disaccord with statistical coupling but rather depends on the distance between residues, spatially closer pairs being more strongly coupled, i.e., Type 1 and 2 coupling. More recent work based on large numbers of coevolved protein families showed that coevolution is basically controlled by Type 1 and 2 coupling (Lockless & Ranganathan, 1999; Suel et al., 2003). The role of spatially distant residue pairs on coevolution notwithstanding, the coupling of spatially distant residue pairs in relation to protein function is well documented in the literature. Millisecond molecular dynamics simulations of Lindorff-Larsen et al (C. N. Chi et al., 2008), and the corresponding NMR results (Anishchenko, Ovchinnikov, Kamisetty, & Baker, 2017; Lindorff-Larsen, Maragakis, Piana, & Shaw, 2016; Marks et al., 2011)show the importance of long range correlations in Ubiquitin. Kong and Karplus (Baldwin & Kay, 2009) used a molecular dynamics (MD) simulations-based approach, referred to as 'interaction correlation analysis', to study long range correlations in the signaling pathways of the PDZ2 domain, and identified paths which are also supported by NMR experiments. The three methods cited, (i) statistical, (ii) energetic, and (iii) MD analysis, are independent approaches. In the present paper, we propose a fourth approach, an entropy-based information transfer approach, which is independent of the other three. The model is based on the time dependent transfer entropy concept of Schreiber (Dittmer & Bodenhausen, 2004) for systems in which fluctuations of one residue are correlated with fluctuations of a second residue at a later time. The model quantifies the decrease in the uncertainty in the second residue due to coupling with past values of the first. The decrease in the uncertainty in the second residue due to its coupling with the first is a transfer of entropy or information from the first to the second. Entropy is a more suitable description of problems of physical nature due to its association with free energy transduction, but we prefer the use of the term information transfer and attempt to quantify transfer in terms of bits. Essentially one is convertible to the other by suitable choice of the proper proportionality. Although information or entropy transfer is widely used in neurosciences (Lange et al., 2008), it is relatively recent in single protein physics (Fenwick et al., 2011; Gianni et al., 2006; Wibral, Vicente, & Lizier, 2014). Throughout the paper, we use both instantaneous and cumulative information transfer. The former is the amount of transfer from a trajectory at time t to another trajectory at a future time $t+\tau$. Different values of the delay time   are used in the literature (See for example (Hacisuleyman & Erman, 2017b; Schreiber, 2000b)). For the protein used in this paper, instantaneous information as a function of   starts from zero, since it is designed to ignore static correlations (Schreiber, 2000b), makes a peak around a fraction of a nanosecond and dies off in a few nanoseconds. Cumulative information transfer is the instantaneous transfer summed up over all delay times   and may be considered as a measure of channel capacity. Cumulative information transfer divided by the peak time may be viewed as an approximate information transfer rate, which amounts to gigabytes per second. Here, we use the widely studied third PDZ domain from the synaptic protein PSD-95, (Protein Data Bank code 1BE9) as our example. We use the Gaussian Network Model of information transfer, which we implemented in Reference (C. N. Chi et al., 2008). Interestingly, this simple Gaussian model can be used to determine the amount and rate of information transfer as well as causality, i.e., the difference between transfer from i to j and from j to i. The latter is important due to its role in evolution and drug design (Hacisuleyman & Erman, 2017a). The main interest of the paper is to quantify the maximum amount of information that may go from one residue to another and the corresponding rate of information transfer.

2.4.2 MATERIALS AND METHODS

Two residues are spatially close if they lie within the first coordination shell of each other, which indicates direct contact. The radius of the first coordination shell is in the range 7.0-7.4 Å (Bahar et al., 1997). The second coordination shell has a volume twice that of the first with radii in the range 8.8-9.3. Pair contacts that are outside the first but within the second coordination shell do interact and are relatively close in space, therefore we classified them as Type 2 contacts. All residue pairs at a distance larger than the second coordination shell radius may safely be assumed as spatially distant. Contacts between pairs of residues lying beyond their second coordination shells are all classified as Type 3 interactions.

2.4.2.1 THE GAUSSIAN NETWORK MODEL (GNM)

The Gaussian network Model is based on the harmonic interactions of contacting residue pairs. The nodes of the network are defined by the alpha carbon coordinates, and the springs of the network that connect the nodes are representative of the interactions between residue pairs within a specified cutoff distance. The cutoff distance is taken as 7 Å. The matrix that contains the connectivity of the protein is described by a matrix Γ whose ij^{th} element equates to -1 if residues i and j are closer than the cutoff distance. The diagonal elements are equal to the negative sum of the corresponding row. The coefficient of this matrix is the spring constant. The spring constant between residues in contact is derived from scaling the B-factors that are obtained from the inverse of the Γ matrix and experimentally measured ones. For 1BE9.PDB, the scaling constant is calculated as 75.3.

The time correlation of fluctuations of two residues are obtained from the solution of the Langevin equation (Erkip & Erman, 2004; T. Haliloglu et al., 1997).

Using the same equations, Eqs. 23 and 24, in 2.3

$$<\mathit{\Delta} R_i(0)\mathit{\Delta} R_j(\tau)> = \sum\nolimits_k A_{ij}(k) exp\left\{-\lambda_k \frac{\tau}{\tau_0}\right\} \qquad \text{EQUATION 31}$$

where

$$A_{ij}(k) = -\lambda_k^{-1} u_i^{(k)} u_j^{(k)} \qquad \text{EQUATION 32}$$

λ_k is the k^{th} eigenvalue and $u_i^{(k)}$ is the i^{th} component of the k^{th} eigenvector of the $\mathit{\Gamma}$ matrix. The probability distribution of a Gaussian instantaneous fluctuation is given as

$$p(\mathit{\Delta} R) = \frac{1}{\sqrt{(2\pi) det\mathit{\Gamma}^{-1}}} exp\left(-\frac{1}{2}(\mathit{\Delta} R)^T \mathit{\Gamma}(\mathit{\Delta} R)\right) \qquad \text{EQUATION 33}$$

where $\mathit{\Gamma}^{-1}$ is the matrix of covariance of instantaneous fluctuations, $<\mathit{\Delta} R_i \mathit{\Delta} R_j>$.

2.4.2.2 INFORMATION TRANSFER FROM ONE RESIDUE TO ANOTHER

We consider two processes X and Y identified by the trajectories of the fluctuations, $\mathit{\Delta} R_i(t)$ and $\mathit{\Delta} R_j(t)$ at time t, of residues i and j, respectively. We identify information transfer from residue i to residue j as as the amount of uncertainty reduced in future values of Y by knowing the present values of X and Y. This concept was introduced by Schreiber (Schreiber, 2000a) where he used the term 'entropy transfer' from i to j, $t_{i\rightarrow j}$, defined by

$$t_{i\rightarrow j} = S\left(\Delta R_j(t+\tau)\middle|\Delta R_j(t)\right) - S\left(\Delta R_j(t+\tau)\middle|\Delta R_i(t), \Delta R_j(t)\right) \qquad \text{EQUATION 34}$$

Here, $S\left(\Delta R_j(t+\tau)\middle|\Delta R_j(t)\right)$ is the conditional entropy of residue j at time $t+\tau$ *given* the values of $\mathit{\Delta} R_j(t)$ at time t. $S\left(\Delta R_j(t+\tau)\middle|\Delta R_i(t), \Delta R_j(t)\right)$ is the conditional entropy of residue j at time t+□given the values of $\mathit{\Delta} R_i(t)$and $\mathit{\Delta} R_j(t)$ at time t. The difference shows the amount of entropy reduced in the trajectory of j due to a knowledge of the present values of i. In terms of Shannon's entropy Eq. 32 reads as:

$$\begin{aligned} t_{i\rightarrow j}(\tau) = &-\left\langle \ln p\left(\Delta R_j(0), \Delta R_j(\tau)\right)\right\rangle + \left\langle \ln p\left(\Delta R_i(0), \Delta R_j(0), \Delta R_j(\tau)\right)\right\rangle \\ &+ \left\langle \ln p\left(\Delta R_j(0)\right)\right\rangle - \left\langle \ln p\left(\Delta R_i(0), \Delta R_j(0)\right)\right\rangle \end{aligned} \qquad \text{EQUATION 35}$$

Here, $p\left(\Delta R_i(0),\Delta R_j(0),\Delta R_j(\tau)\right)$ is the joint probability of fluctuations of *i* and *j* at time zero and the fluctuation of *j* at time τ, with similar definitions of the remaining probabilities in Eq. 33. The joint probabilities may be obtained either by extensive molecular dynamics simulations or using the dynamic Gaussian Network Model. Here, we use the latter theory, which is outlined below.

2.4.2.3 THE GAUSSIAN NETWORK MODEL OF INFORMATION TRANSFER

Substituting from Eq. 31 into Eq. 33 leads to the following final expression for entropy transfer

$$
\begin{aligned}
t_{i\to j}(\tau) = {} & \frac{1}{2}\ln\left(\left(\sum_k A_{jj}(k)\right)^2-\left(\sum_k A_{jj}(k)\exp\{-\lambda_k\tau/\tau_0\}\right)^2\right) \\
& -\frac{1}{2}\ln\Bigg[\left(\sum_k A_{ii}(k)\right)\left(\sum_k A_{jj}(k)\right)^2 \\
& +2\left(\sum_k A_{ij}(k)\right)\sum_k A_{jj}(k)\exp\{-\lambda_k\tau/\tau_0\}\sum_k A_{ij}(k)\exp\{-\lambda_k\tau/\tau_0\} \\
& -\left\{\left(\sum_k A_{ij}(k)\exp\{-\lambda_k\tau/\tau_0\}\right)^2+\left(\sum_k A_{ij}(k)\right)^2\right\}\left(\sum_k A_{jj}(k)\right) \\
& -\left(\sum_k A_{jj}(k)\exp\{-\lambda_k\tau/\tau_0\}\right)^2+\left(\sum_k A_{ii}(k)\right)\Bigg] \\
& -\frac{1}{2}\ln\left[\left(\sum_k A_{jj}(k)\right)\right] \\
& +\frac{1}{2}\ln\left(\left(\sum_k A_{ii}(k)\right)\left(\sum_k A_{jj}(k)\right)-\left(\sum_k A_{ij}(k)^2\right)\right) \qquad \text{EQUATION 36}
\end{aligned}
$$

2.4.2.4 CUMULATIVE INFORMATION TRANSFER

The transfer of information given by Eq. 34 is the information transferred instantaneously at time $\tau+t$ resulting from an effect imposed at zero time. The model contains a characteristic time τ_0. In earlier work (Ben-Avraham, 1993; Turkan Haliloglu, Ivet Bahar, & Burak Erman, 1997) it was shown that the dynamics of folded proteins may be expressed in terms of a universal characteristic time τ_0, which was estimated to be around 5-6 ps. Adopting a value of $\tau_0 = 5\,ps$, the values of information transfer may be calculated from

Eq. 34 for each τ. Some of the instantaneous information transfer curves obtained in this manner are shown in Figs. 28 and 29.

The cumulative information transfer is obtained from the instantaneous transfer according to

$$T_{i\to j} = \int t_{i\to j}\left(\frac{t}{\tau_0}\right) d\left(\frac{t}{\tau_0}\right) \qquad \text{EQUATION 37}$$

2.4.2.5 CALCULATING THE TRANSFER RATE

The peak values shown in Figure 28 correspond to times when a large fraction of the information is transferred. As a first order approximation, we assume that all of the cumulative information is transferred at the peak time. Then, the transfer rate becomes the cumulative information transfer divided by the peak value. All the transfer rates reported in the paper are obtained in this way.

2.4.2.6 MAXIMUM CALIBER

Maximum Caliber (Max Cal) predicts the probabilities of trajectories by maximizing the trajectory entropy over all possible trajectories subject to certain dynamical constraints. A trajectory is defined as a discrete time sequence *$(t_0t_1t_2\ldots t_T)$* of length *$T+1$*. We assume that the system contains N particles, and each particle may be in *M* states. A particle may be a residue of the protein and its trajectory may be represented as the trajectory of its alpha carbon. At any given instant, the system will occupy a state among a total of M^N possible state. During the trajectory of length T, M^{TN} *states* will be available to the system. At time t_k, the state of the system is denoted as or simply as i_k. The states visited during the trajectory are denoted as *$(i_0i_1i_2\ldots i_T)$*. The set of all trajectories is shown by *$\{i_0i_1i_2\ldots i_T\}$*. The probability of the trajectory is *$p(i_0i_1i_2\ldots i_T)=p$*. The path entropy is defined as. The summation is over all possible states M^{TN}. Max Cal principle maximizes the following function, the entropy, subject to certain constraints:

$$-\sum_{\Gamma} p_{\Gamma}\,\mathrm{LOG}\frac{p_{\Gamma}}{q_{\Gamma}} - \gamma\left(\sum_{\Gamma} p_{\Gamma} F(\Gamma)\right) + \alpha\left(\sum_{\Gamma} p_{\Gamma} - 1\right) \qquad \text{EQUATION 38}$$

Here, q_Γ is a reference distribution for the problem. The distribution resulting from the variation of this equation is $p_\Gamma = \frac{q_\Gamma e^{-\gamma F(\Gamma)}}{Z}$, where $Z = \sum_\Gamma q_\Gamma e^{-\gamma F(\Gamma)}$.

Of particular interest is the constraint on the pairwise statistics where the functional *F(Γ)* is now defined as

$$F_{mn} = \sum_{M^{TN}} p(i_0 i_1 i_2 \dots i_T) \sum_{k=0}^{T-1} \delta_{i_{k,m}} \delta_{i_{k+1,n}} \qquad \text{EQUATION 39}$$

where, $\delta_{i_{k,m}}$is the Kronecker delta, which equates to unity if the state i_k is the m^{th} state and zero otherwise. Ge et al., showed that constraining the problem to pairwise statistics leads to a Markov process in which the probability distribution of the path is obtained as

$$p(i_0 i_1 i_2 \dots i_T) = \prod_{k=0}^{T-1} p(i_k i_{k+1}) \qquad \text{EQUATION 40}$$

2.4.2.7 RESEMBLANCE OF MAX CAL AND OUR METHODOLOGY

In our problem, we select two particles, *i* and *j*, out of the *N* particles of the system. The trajectory for each of the particles is defined as in the general case. The statistics then reduces to pairwise statistics. There will be a total of M^{2T} states available for each of the particles throughout the trajectory for the system.

Transfer entropy is defined as a Markov process over the phase space of M^{2T} elements as

$$T_{i\to j} = -\sum_{M^{2T}} p(j_{n+1} i_n j_n) \,\mathrm{LOG}\, \frac{p(j_{n+1}|j_n i_n)}{p(j_{n+1}|j_n)} \qquad \text{EQUATION 41}$$

Expanding gives

$$T_{i\to j} = -\sum_{M^{2T}} p(j_{n+1} i_n j_n) \,\mathrm{LOG}\, p(j_{n+1}|j_n i_n) + \sum_{M^{2T}} p(j_{n+1} i_n j_n) \,\mathrm{LOG}\, p(j_{n+1}|j_n)$$

EQUATION 42

$$T_{i\to j} = -\sum_{M^{2T}} p(j_{n+1} i_n j_n) \,\mathrm{LOG}\, p(j_{n+1} i_n j_n) + \sum_{M^{2T}} p(j_{n+1} i_n j_n) \,\mathrm{LOG}\, p(i_n j_n) + \sum_{M^{2T}} p(j_{n+1} i_n j_n) \,\mathrm{LOG}\, p(j_{n+1} j_n) - \sum_{M^{2T}} p(j_{n+1} i_n j_n) \,\mathrm{LOG}\, p(j_n) \qquad \text{EQUATION 43}$$

$$T_{i\to j} = -\sum_{M^{2T}} p(j_{n+1} i_n j_n) \,\mathrm{LOG}\, p(j_{n+1} i_n j_n) + \langle \mathrm{LOG}\, p(i_n j_n)\rangle + \langle \mathrm{LOG}\, p(j_{n+1} j_n)\rangle - \langle \mathrm{LOG}\, p(j_n)\rangle$$

EQUATION 44

Letting $\log p(j_{n+1} i_n j_n) = \langle \log p(i_n j_n)\rangle + \langle \log p(j_{n+1} j_n)\rangle - \langle \log p(j_n)\rangle$

We obtain

$$T_{i\rightarrow j} = -\sum_{M^{2T}} p(j_{n+1}i_nj_n)\,\mathrm{LOG}\frac{p(j_{n+1}i_nj_n)}{q(j_{n+1}i_nj_n)} \qquad \text{EQUATION 45}$$

Here, $q(j_{n+1}i_nj_n)$ is some reference distribution over paths. This Eq. 43 is identical with Eq. 6 of Ref. (Dixit et al., 2018).

The probability term, $p(j_{n+1}i_nj_n)$, in the transfer entropy equation conforms with pairwise statistics where j_{n+1} is the state at the $n+1^{st}$ step and i_nj_n is the state at the n^{th} step. From molecular dynamics trajectories, we created the joint distribution $p(j_{n+1}i_nj_n)$ and formed the average

$F_{rs} = \sum_{M^{TN}} p(i_0i_1i_2\dots i_T)\sum_{k=0}^{T-1}\delta_{i_k,r}\delta_{i_{k+1},s}$, where r is the state i_{n+1} at time $n+1$ and s is the state i_nj_n at step n. Thus, F_{rs} defines the states averaged over all conformations of all other residues. All calculations of $T_{i\rightarrow j}$ from molecular dynamics trajectories are performed by using F_{rs} in the equation.

2.4.3 RESULTS

Based on the transfer entropy model of Schreiber (Schreiber, 2000a) and the Gaussian network model of folded proteins (Bahar et al., 1997) we analyzed the amount of information that can be transferred from one residue to another, and the rate at which this transfer takes place. We use the crystal structure of 1BE9.PDB. The stick form of the three-dimensional structure is shown in Figure 24. Chi et al. (Celestine N Chi et al., 2008) defined six residues, H372, A376, G329, V362, F340, K380 as the energetic network residues of the PDZ domain. These residues are also significant in the works of Lockless and Ranganathan (Lockless & Ranganathan, 1999) and Kong and Karplus (Kong & Karplus, 2009). We focus on the same set of residues in this paper and refer to them as "network residues". The alpha carbons of the network residues are used in all calculations and are shown as large black spheres in Figure 24.Distances between pairs of network residues and their interaction types based on pair distances are shown in the second and third columns of Table 1. Pairs with short-range interactions are shown in bold fonts. The fourth column shows the amount of maximum information that may be transferred between each pair in bits. The fifth column shows the peak delay time at which maximum instantaneous information is transferred from one residue to the other. The last column shows the information transfer rate in gigabits per second.

Analysis of the table shows that information transfer between residues in contact, Type 1 and 2 interactions, and the corresponding transfer rates are the largest. However, transfers between residues with long-range contacts are not negligible, maximum transfer of Type 3 being between H372-K380. Transfer rate between these two residues is also large. The next highest information transfer of Type 3 is between K380 and V362 with an information transfer of 2 bits and an information transfer rate of 8.7 GB/s. Information transfer between other network residues with long range coupling also exhibit non-negligible values. These findings support the hypothesis of statistical coupling of Lockless and Ranganathan and the experiments of Suel et. al.(Lockless & Ranganathan, 1999; Süel, Lockless, Wall, & Ranganathan, 2003)

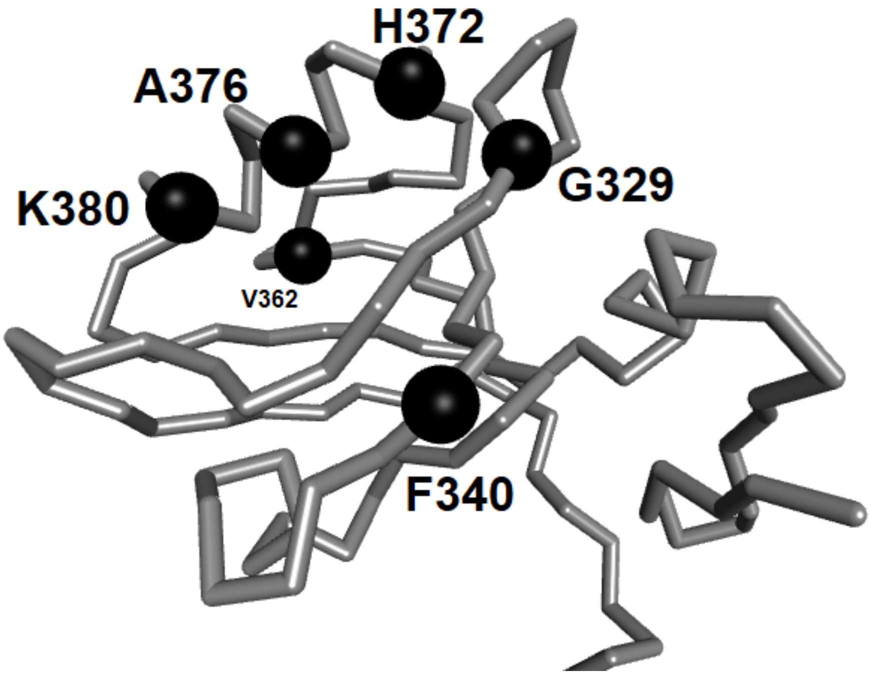


FIGURE 24. THREE-DIMENSIONAL STRUCTURE OF 1BE9. SIX RESIDUES RESPONSIBLE IN ALLOSTERIC INFORMATION TRANSFER, IDENTIFIED BOTH IN REFERENCES [84] AND [78] ARE SHOWN AS BLACK SPHERES. DISTANCES BETWEEN RESIDUES ARE GIVEN IN TABLE 1.

TABLE 1. VARIOUS METRICS OF INFORMATION TRANSFER IN 1BE9.

Network Residue Pairs	Distance (Å)*	Interaction type**	Information transfer (Bits)***	Peak Time ns	Transfer rate (GBits/s) ****
H372-K380	12.3	3	4.2	0.2	19.1
H372-F340	15.4	3	0.1	0.06/1.6	1.7/0.1
H372-G329	**6.1**	**2**	**3.3**	**0.1**	**33.0**
H372-V362	13.9	3	1.8	0.5	3.9
H372-A376	**6.3**	**1**	**4.0**	**0.1**	**40.0**
K380-F340	15.9	3	0.1	0.11/1.7	0.9/0.1
K380-G329	16.2	3	1.5	0.5	3.2
K380-V362	10.2	3	2.0	0.2	8.7
K380-A376	**6.0**	**1**	**4.1**	**0.1**	**41.0**
F340-G329	12.8	3	0.3	0.2	0.6
F340-V362	17.5	3	0.2	0.05/1.2	4/0.2
F340-A376	14.1	3	0.1	0.07/1.6	1.4/0.1
G329-V362	15.6	3	0.9	0.6	1.5
G329-A376	10.5	3	1.8	0.3	6.0
V362-A376	10.5	3	1.6	0.4	4.2

*The distance between two residues is measured as the distance between their alpha carbons

** Types 1, 2 or 3.

***Cumulative information transfer (See Eq. 7, Materials and Methods Section)

****Transfer rate is defined as Cumulative information transfer divided by peak time.

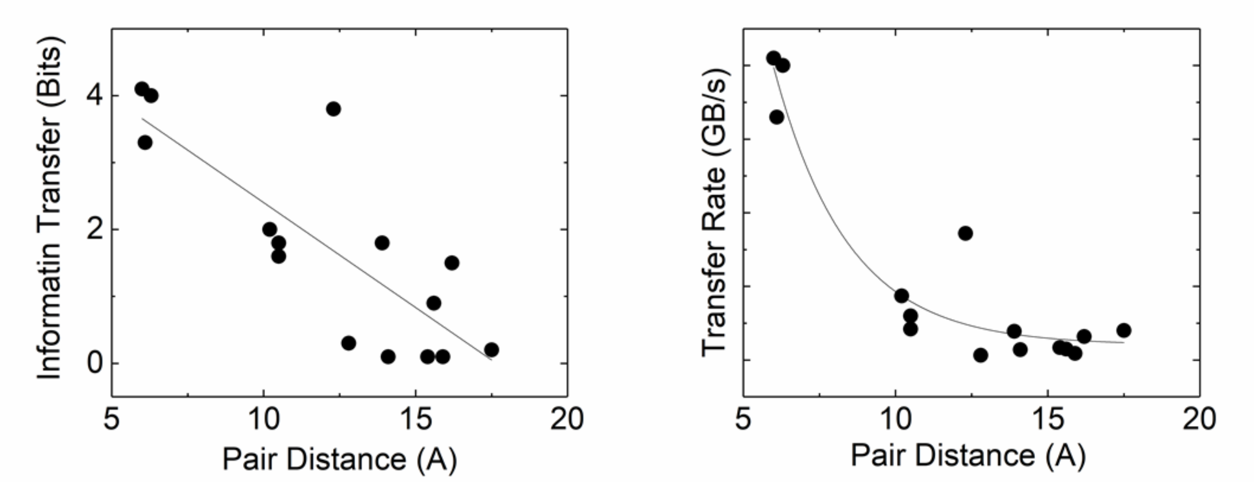


FIGURE 25. DEPENDENCE OF CUMULATIVE INTER-RESIDUE INFORMATION TRANSFER AND TRANSFER RATE ON DISTANCE BETWEEN RESIDUES. SOLID LINES ON THE LEFT AND RIGHT PANELS ARE THE BEST FITTING STRAIGHT LINE AND EXPONENTIAL DECAY, RESPECTIVELY.

The maximum amount of cumulative information that can be transferred from one residue to another, the channel capacity, is the integral of the instantaneous information transfer between the two residues. It decreases approximately linearly with the distance between the pairs, as shown in the left panel of **Error! Reference source not found.**. On the right panel, the distance dependence of transfer rate between residue pairs is shown, which decays approximately exponentially with distance.

Information transfer from network residues to others that are within a radius of 9.3 Å, i.e., the radius of the second coordination shell, is shown in Figure 26. The residue from which information is transferred is identified in each panel. In the upper left panel of Figure 26, we see a Type 2 transfer from H372 to residues I328-E331. The upper middle panel shows that K380 exhibits a weak Type 2 transfer to hinge residues between S320-L323. Upper right panel shows that the none-network residue F340 shows a Type 2 transfer to L323-I327 and L353-G356. Residue G353 is known as the active site and only F340 can transfer a Type 2 information to it. The lower left panel shows a Type 2 transfer from G329 to an alpha helix between residues H372-380 and a weaker Type 2 transfer to an alpha helix between residues P394-F400. The lower middle panel shows a Type 2 interaction from V362 to A375, A378, A382, and to the range of residues T385-A390. The lower left panel shows that V362 can exhibit a Type 1 transfer to residues D357 to R368. The lower right panel shows a weak Type 2 transfer from A376 to I327 and I336.

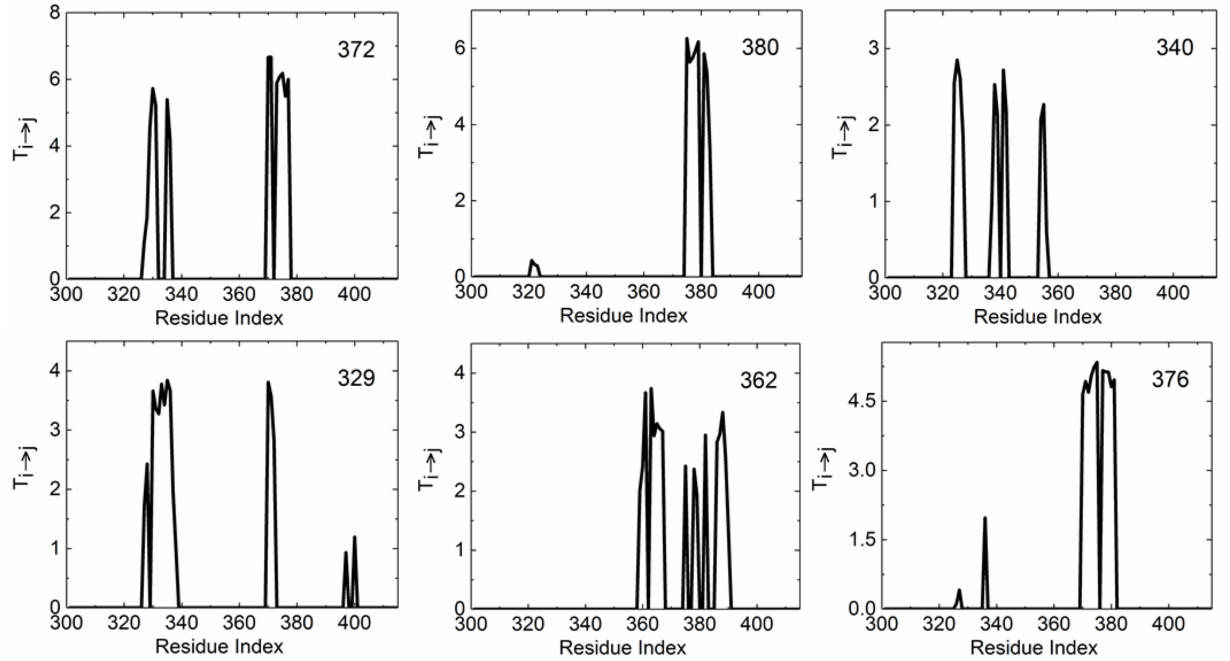


FIGURE 26. INFORMATION TRANSFER FROM THE RESIDUE INDICATED IN EACH PANEL TO ALL OTHER RESIDUES. ONLY THE INTERACTIONS OF RESIDUES THAT ARE CLOSE IN SPACE ARE CONSIDERED. THESE ARE EITHER TYPE 1 OR 2 TYPES OF TRANSFER.

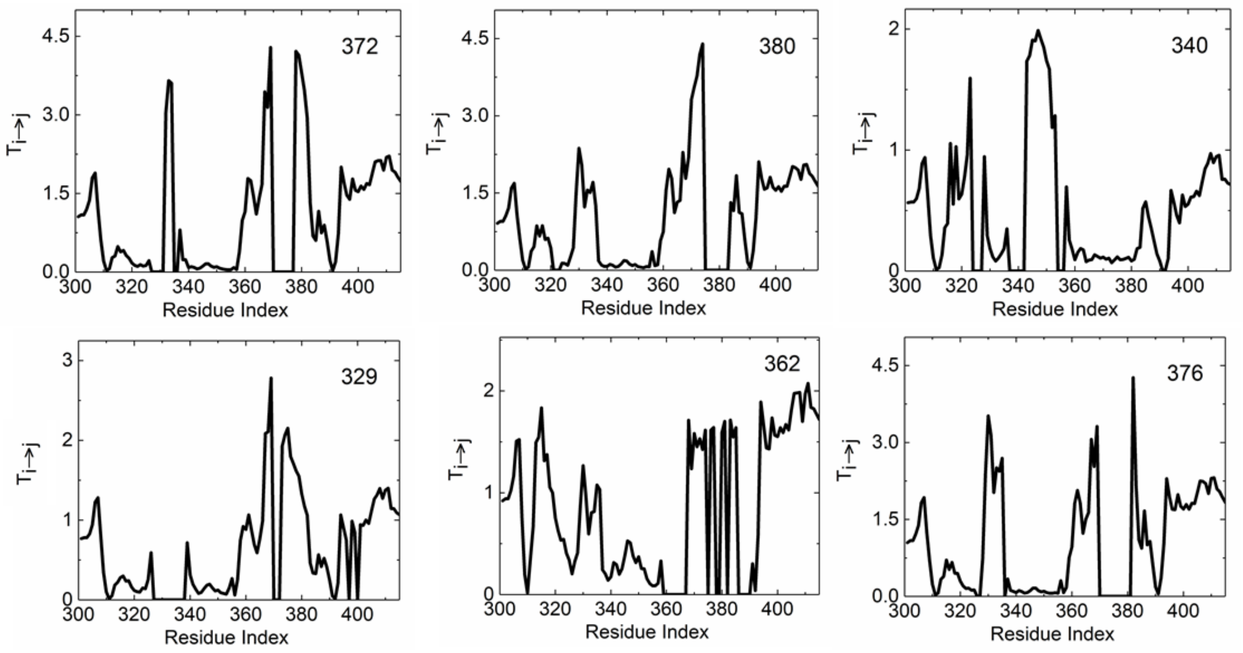


FIGURE 27. INFORMATION TRANSFER FROM THE RESIDUE INDICATED IN EACH PANEL TO ALL OTHER RESIDUES. ONLY PAIRS WITH LONG-RANGE INTERACTIONS. I.E., TYPE 3, ARE SHOWN.

Information transfer between spatially distant pairs of residues, i.e., Type 3 transfer is presented in Figure 27. Ordinate values of the six panels show that Type 3 transfer is of the same order of magnitude as those of contacting residues. It is interesting that significant long-range information transfer takes place also between network residues and non-network residues. In the upper left panel of Figure 27, information transfer of 4.2 Bits is possible from H372 to K380 with a transfer rate of 19.1 GB/s. The magnitude and rate of transfer between this pair are slightly lower than those of the contacting network residues. The two residues H372 and K380 are at the extremities of the long helix of the protein with a distance of 12.3Å between their alpha carbons. This coupling is an indicator of transfer along helices, which we will discuss below. Transfer from H372 to N369 at a distance of 9.47 Å is even more dramatic, with an amplitude of 4.3 Bits and a rate of 25.3 GB/s. H372 and N369 lie at the extremities of an elongated coil structure. H372 and G333, 10.1 Å apart exhibit a coupling through which a total of 3.65 Bits may be transferred with a rate of 21.5 GB/s. The upper middle panel of Figure 27 shows Type 3 transfer from residue K380 to the rest of the system. Maximum amount of information is transferred to Q374. Residues K380 and Q374 are 9.8 Å apart. Residue K380 also shows a Type 3 interaction with hinge residues between G330 and G335. The upper right panel shows that F340 exhibits Type 3 transfer to an alpha helix between residues P346-S350. The lower left and middle panels in Figure 27 shows that both G329 and V362 are coupled with the helix between H372-K380 and A376, a residue in this helix exhibits Type 3 transfer to the hinge regions of the protein.

All of the six panels of Figure 27 show that the network residues interact with the N and C-terminals of the protein. The C-terminal which has been the focus of earlier work (Z Nevin Gerek & S Banu Ozkan, 2011) consists of a helix-turn and two beta strands. The strongest interaction of the C-terminal is with residue V362. The distance between the centroid of the C-terminal and V362 is 24 Å, the average information transfer from the centroid of the C-terminal and V362 is around 1.3 Bits and the information transfer rate is 6.8 GB/s.

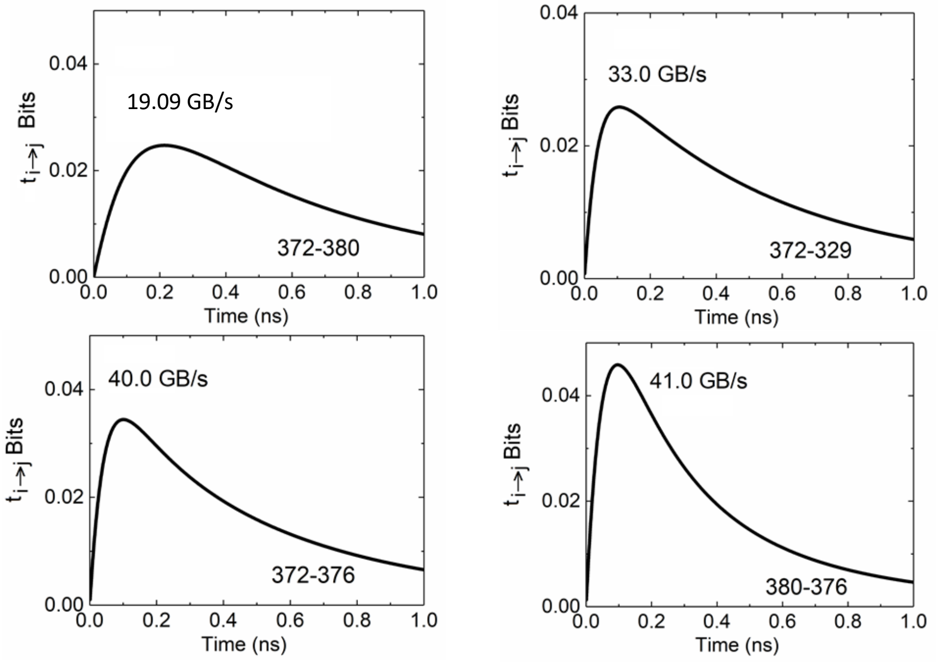


FIGURE 28. INSTANTANEOUS INFORMATION TRANSFER OBTAINED FROM EQ. 34 BETWEEN NETWORK RESIDUE PAIRS AS A FUNCTION OF DELAY TIME. THE RATES OF TRANSFER CALCULATED ACCORDING TO EQ. 35 ARE INDICATED IN THE FIGURES.

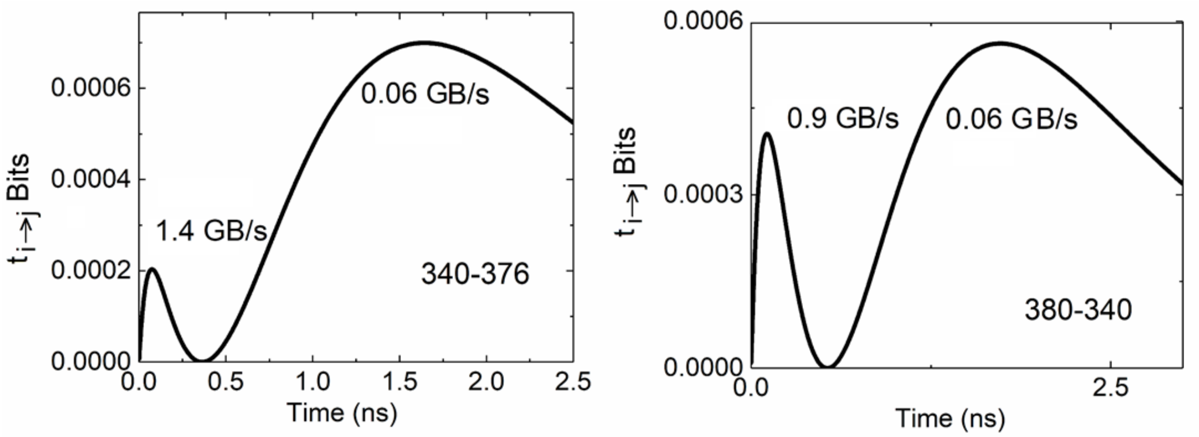


FIGURE 29. INSTANTANEOUS INFORMATION TRANSFER BETWEEN NETWORK RESIDUE PAIRS F340-A376 AND K380-F340 EXHIBIT DOUBLE PEAKS. THE TRANSFER RATE IS INDICATED FOR EACH PEAK.

2.4.3.1 INFORMATION TRANSFER ALONG SECONDARY STRUCTURES

In Figure 30, left panel shows the information transfer from residue A376, which is the central residue of the main helix to the neighboring residues along the helix structure. The cumulative transfer is significant, where approximately the same amount of information is transferred without decay as one move along the helix. The right panel of Figure 30 shows that the transfer rates along the helix are high irrespective of the distance between residue pairs along the helix.

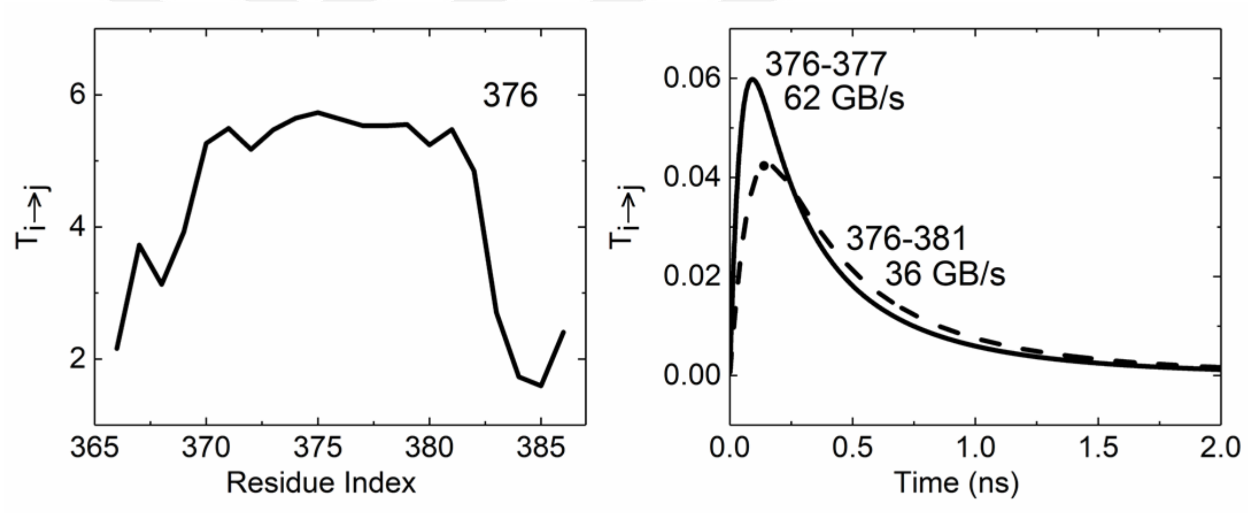


FIGURE 30. THE LEFT PANEL SHOWS THE INFORMATION TRANSFER BETWEEN RESIDUE 376 AND ITS NEIGHBORS ALONG THE PRIMARY CHAIN. THE RIGHT PANEL SHOWS THE INSTANTANEOUS INFORMATION TRANSFER AS A FUNCTION OF DELAY TIME BETWEEN NEAR NEIGHBORS (A376-I377) AND NON-NEAR NEIGHBORS (A376-N381) ALONG THE SECONDARY STRUCTURE.

2.4.3.2 CAUSALITY

The amount of information going from a residue *i* to *j* may be different than information going from *j* to *i*. This feature is referred to as causality and is implicit in the Schreiber theory (Schreiber, 2000a). The directionality can be detected either from instantaneous information transfer, where $t_{i\to j}(\tau) \neq t_{j\to i}(\tau)$, or from cumulative information transfer, $T_{i\to j} \neq T_{j\to i}$. Typical plots of $t_{i\to j}(\tau)$ between network residue pairs are shown in Figure 31.

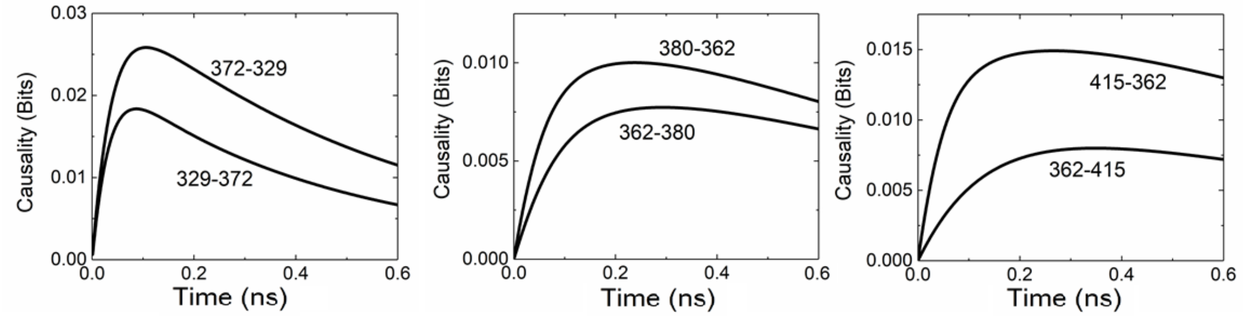


FIGURE 31. DIRECTIONALITY OF INFORMATION FLOW. ORDINATE VALUES REPRESENT THE INSTANTANEOUS INFORMATION TRANSFER.

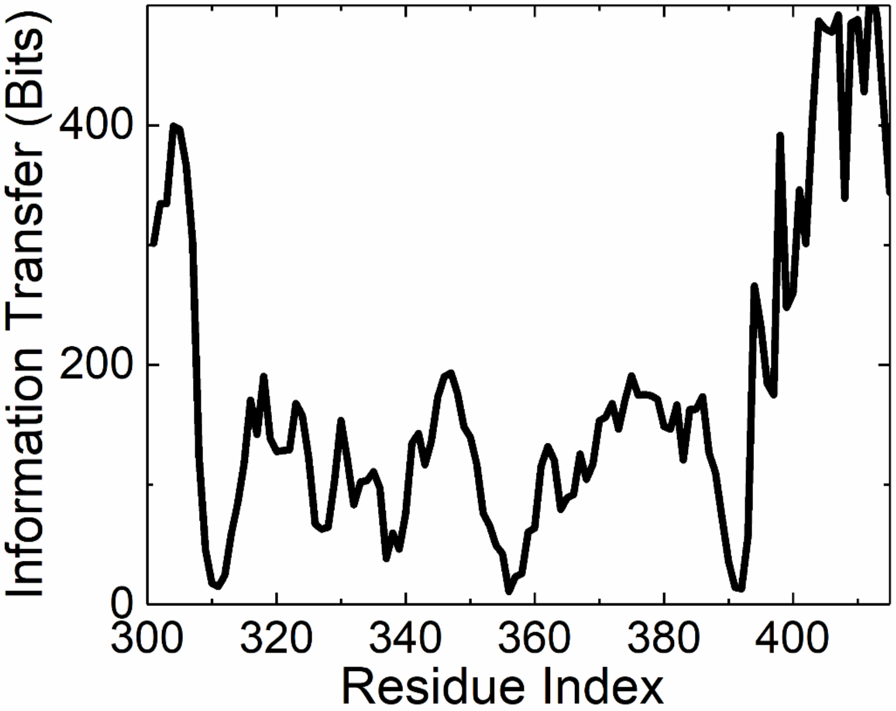


FIGURE 32. CUMULATIVE INFORMATION TRANSFER FROM RESIDUE *I* INDEXED ALONG THE ABSCISSA TO THE REST OF THE PROTEIN.

According to Rios et al, (De Los Rios et al., 2005a), the hinge region between R318-G324 and the alpha helix between H372-K380 undergoes the largest deformation upon binding. Our results show that these regions that undergo the largest deformation are the ones that show highest coupling with the remaining regions of the protein through Type 1, 2 and 3 transfers (De Los Rios et al., 2005a).

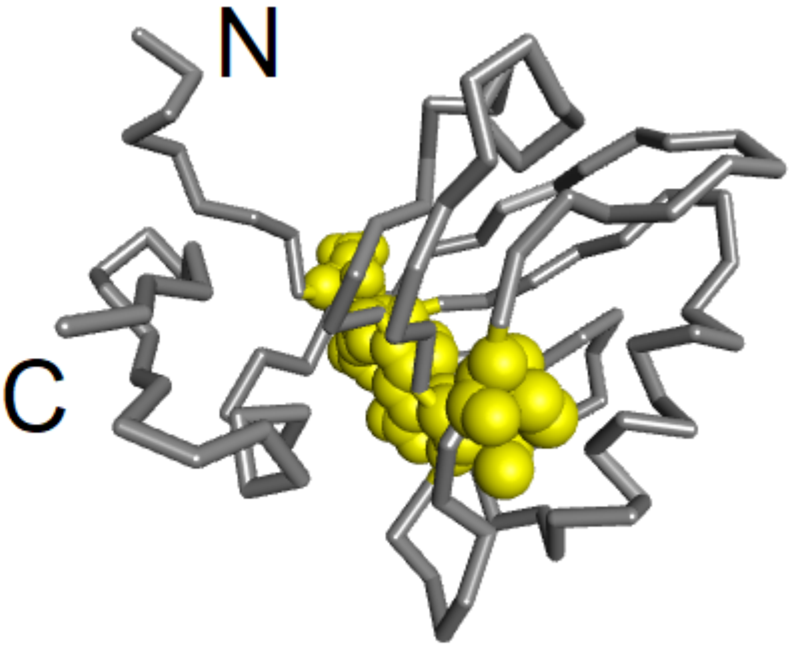


FIGURE 33. RESIDUES OF MINIMAL INFORMATION TRANSFER, HIGHLIGHTED IN YELLOW, SEPARATE THE TAIL OF THE PROTEIN FROM THE REST.

The four strong minima seen in Figure 32 correspond to minimal information transfer residues that lie approximately along a straight line that separates the amino and carboxyl tails from the rest of the protein. All of the pathway residues lie on the part that does not contain the two tails. Any information exchange that involves the tails of the protein and the rest take place through Type 3 contacts.

2.4.4 DISCUSSION

Mutation experiments of Chi et al., showed no coupling between residues with long-range correlations, including the H372-K380 pair. On the contrary, Lockless and Ranganathan observed that these two sites are statistically coupled. Kong and Karplus (Kong & Karplus, 2009) determined coupling between distant residues in PDZ domain proteins and showed that this coupling has been imprinted into the structure during evolution. In this paper, we quantified long-range coupling in terms of information transfer and showed that strong coupling is present among spatially distant residues of the PDZ domain. Whether this long-range coupling is the major factor in coevolution or not cannot be answered by information transfer, but a definite and strong long-range coupling is present among the network residues of 1BE9. There are several studies about the single domain allostery concept, which proved that PDZ domain proteins connect signals within the system and exhibit allosteric behavior (Z Nevin Gerek & S Banu Ozkan, 2011; B. K. Ho & Agard, 2010; Lee, 2015; Lockless & Ranganathan, 1999; Chad M Petit, Jun Zhang, Paul J Sapienza, Ernesto J Fuentes, & Andrew L Lee, 2009). Here, we utilize the GNM approach and detect the allosteric information transport features in 1BE9. In an experimental study, it has been confirmed that removal of the third helix, located at a distant site from the binding pocket between residues H372-K380 in 1BE9, affects the dynamics of the system and thus reduces the binding affinity (Z. N. Gerek & S. B. Ozkan, 2011). A possible allosteric pathway is constructed by a perturbation response scanning analysis (Z Nevin Gerek & S Banu Ozkan, 2011), and the residues involved in the pathway is detected by our method. The residues that are pointed out in reference (Z Nevin Gerek & S Banu Ozkan, 2011) which are involved in information transfer are I314, I327, I338, A347, L353, V362, L367, H372, K380, V386 and E396. The peaks in Figure 32 correspond to the significant information transmitting residues listed in previous studies(De Los Rios et al., 2005b; Z. N. Gerek & S. B. Ozkan, 2011; Hacisuleyman & Erman, 2017a; Kong, Karplus, & Bioinformatics, 2009; S. Pressé, K. Ghosh, J. Lee, & K. A. J. R. o. M. P. Dill, 2013)The direction of the transfer can be detected from plotting the pairwise instantaneous information transfer from *i* to *j*, $t_{i\rightarrow j}(\tau)$, and from *j* to *i*, $t_{j\rightarrow i}(\tau)$. The directionality, causality relationship, for several residue pairs is shown on Figure 31, which shows that information going from *i* to *j* may differ from the transfer from *j* to *i*. Determining the driver-driven relationship among a residue pair is a crucial step in terms of drug design and the directionality plots in Figure 31 help reveal the underlying mechanism of information transfer process.

2.5 CONCLUSION

Biological systems communicate within themselves and this communication requires the transmission of signals in terms of fluctuations. The direction and the information of physical mechanisms or interactions of proteins lie in their intramolecular dynamics. Analyzing their fluctuation relationships helps us to better understand how they behave and communicate upon different situations. Many researchers investigated mutual information to account for the shared information content between two entities but in order to get a driver-driven relationship, transfer entropy concept must be incorporated. Mutations, post-translational modifications and ligand binding change the direction of the information flow. Transfer entropy is useful to further understand, perturb the effect of mutations and detect ligand binding sites. Molecular dynamics or dGNM transfer entropy method, can be further used to design specific inhibitors that binds to important entropy source or sink residues in the protein, or residues within the transfer entropy path. It is also useful in understanding the protein-protein interactions.

CHAPTER 3.
IN SILICO NANOBODY DESIGN

3.1 OVERVIEW

Antibodies, also called immunoglobulins, are Y shaped proteins that are key elements in the adaptive immune system. They recognize unique parts of foreign targets, antigens, in the blood or mucosa and inactivate them. They activate complementary systems to destroy bacterial cells and facilitate phagocytosis of foreign substances.(Goldman & Prabhakar, 1996) They are composed of two identical copies of heavy (approximately 50 kDa and two identical copies of light chains (approximately 25 kDa). Heavy chains are connected to each other by a disulfide bond and each heavy chain is connected to a light chain by another disulfide bond. Each antibody producing B-cell produces a unique type of antibody. There are 5 antibody isotypes: IgG, IgM, IgD, IgE and IgA all differ in their heavy chains. Variance in their heavy chains result in them to bind different antigens. Antibody derived biologics have produced impressive therapeutic results. (Oliver & Jamur, 2010)Advances in antibody engineering provided improved innovative molecules with impressive achievements in treatments of several hematological malignancies and tumors. The advantages of them are their enhanced effector function with reduced immunogenicity with prolonged half-life with reduced side effects, but their large size and hydrophobic binding surfaces pose obstacles in their stability and access to the antigen. Due to these limiting factors, new antibody-formatted treatment options with similar binding specificity and higher stability are required.

Nanobodies, special derivatives of antibodies, that consist of only heavy chains produced by members of Camelidae family (*Llama glama*, *Vicugna pacos*, *Camelus dromedaries, Camelidae, Camelus bactrianus*), nurse sharks, *Ginglymostoma cirratum,* wobbegongs, *Orectolobus* and spotted ratfish, *Hydrolagus colliei*. (Serge Muyldermans, 2013)They are highly specific and exert high affinity towards their target. They can be easily expressed in microorganisms. Their toxicity is low, and their tissue penetration is not limited due their small size. The regions on nanobodies that are responsible for binding and recognition are called complementarity-determining regions (CDR).(Tiller & Tessier, 2015) The antigen binding region, paratope, diversity of antibodies are derived from their heavy and light chain combinations. Nanobodies are only composed of heavy chains compensate this diversification by taking many more loop architectures without any structural restrictions in their CDR's. Due to their biochemical functionality and economic benefits, interest in nanobodies has grown in biotechnology and medicine. In order to use nanobodies as a therapeutic reagent, their key properties such as their stability, binding specificity and affinity to the target antigen must be optimized.

The target specificity of a nanobody is provided by its CDRs. Nanobodies comprise of three CDR loops; CDR1, CDR2 and CDR3. CDR3 is the dominating contributor in antigen recognition while the impact of the first and the second CDR loops are limited. (Wong, Leem, & Deane, 2019)Residue numbers of the CDR amino acids can vary but a CDR loop can be recognized by the presence of certain amino acid repeats. These repeats are composed of evolutionarily fixed and variable regions. The template of the fixed and variable residues differs within species. There is a statistical study conducted on known, stable llama derived nanobodies, showing which CDR residues are conserved and which CDR residues are hypervariable.(McMahon et al., 2018) The increased frequency of some residues in the hypervariable positions supports convergent evolution in those regions.

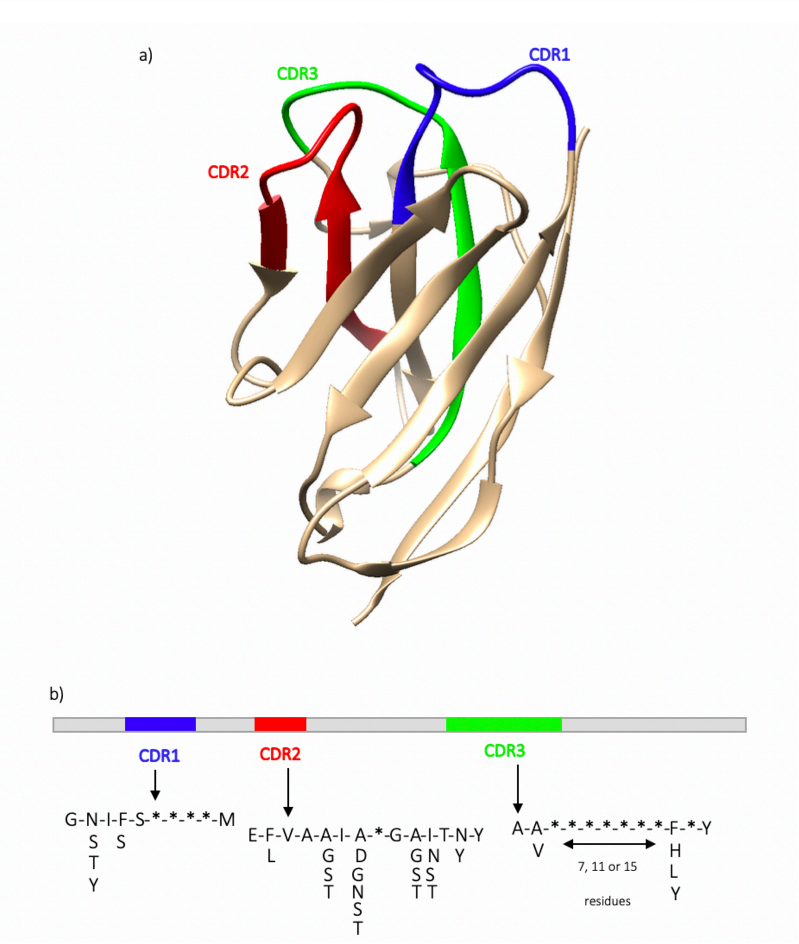


FIGURE 34. A) CARTOON REPRESENTATION OF A NANOBODY. CDR1, 2 AND 3 ARE COLORED IN BLUE, RED AND GREEN RESPECTIVELY. B) FRAMEWORK OF FIXED, VARIABLE AND HIGHLY VARIABLE AMINO ACID MOTIFS OF EACH CDR ARE

INDICATED, AMINO ACIDS INDICATED WITH * ARE THE HIGHLY VARIABLE ONES AND THE RESIDUE POSITION WITH MORE THAN ONE TYPE OF AMINO ACID ARE THE VARIABLE REGIONS THROUGHOUT THE LLAMA GLAMA NANOBODIES.

The frequency scheme of the highly variable regions are given as; 14%-Y, 12%-G, 10%-S, 9%-D, 7%-T, 6%-R, 6%-A, 5%-L, 5%-V, 4%-N, 3%-F, 3%-E, 3%-I, 3%-W, 2%-Q, 2%-K, 2%-H, 0%-C and 0%-M. It can be seen from *Figure 34* that CDRs vary in length. CDR3 of different nanobodies can also vary in length but in general, CDR3 is the longest sequence among all. In nature, antibodies are mutated naturally to increase the binding affinity and the specificity towards their target antigen. The hypervariable residues in CDRs determine the specificity of a nanobody. Similar alignment studies can be repeated for other organisms and fixed-variable residue templates for their nanobodies can be obtained.

Computational screening methods contributed greatly to the optimization and design process. There are several approaches available. The common protocol is to start the screening procedure from a known potential binder nanobody retrieved from libraries by phage display or by other selection protocols; bacterial display, yeast display, ribosome display or intracellular 2 hybrid selection, and gradually enhancing the properties by mutating CDR or non-CDR residues iteratively and measure the antigen-nanobody affinity values. (Peltomaa, Benito-Peña, Barderas, & Moreno-Bondi, 2019)

To start the mutation process, the CDR, non-CDR loops and the locations on the CDR loops which allow insertions or deletions must be identified. This can be done by aligning nanobody sequences to known structures belonging to the same species. There are publicly available CDR numbering tools that can be applied to a certain sequence.(Lefranc, 2011) After the identification of the orientation or the sequence of the nanobody, amino acids can be selected for mutation to increase the binding affinity for the antigen.

Current computational techniques are based on finding the CDR and non-CDR residues that are in contact with the antigen and performing point mutations to enhance the interaction.(Bannas, Hambach, & Koch-Nolte, 2017; Barderas, Desmet, Timmerman, Meloen, & Casal, 2008; T. Li, Pantazes, & Maranas, 2014; Mahajan et al., 2018; S. Muyldermans, 2013) The optimization strategy proposed in this study is based on finding the hypervariable positions in CDRs by an alignment process and increasing the binding affinity by mutating them iteratively and selecting the residues which are highly frequent for a particular species. This method is tested on several different proteins: human MDM4, aldolase isoforms A, B and C from human tissue, Fructose-bisphosphate aldolase from *Trypanosoma congolense* and human lysozyme. PDB codes are 2VYR, 4ALD, 1QO5, 1XFB, 5O0W and 4I0C respectively.

3.2 COMPUTATIONAL METHOD

3.2.1 ALIGNMENT

Sequences and structures of synthetic and natural nanobodies can be found in databases.(Wilton, Opyr, Kailasam, Kothe, & Wieden, 2018; Zuo et al., 2017) Nanobody sequences for specific organisms can be downloaded from the databases and aligned by using several multiple alignment tools available. (Katoh & Standley, 2013; Lassmann, Frings, & Sonnhammer, 2008; Notredame, Higgins, & Heringa, 2000; Sievers et al., 2011) The alignments can be visualized by specific softwares (Okonechnikov, Golosova, Fursov, & Team, 2012; Waterhouse, Procter, Martin, Clamp, & Barton, 2009), the consensus sequence and the conserved or hypervariable positions can be determined and a template can be obtained similar to the llama-derived nanobody framework proposed in a previously.(McMahon et al., 2018)

Nanobody sequences for *Vicugna pacos* and *Camelus dromedaries* are downloaded from single domain Antibody database(sdAb) (Wilton et al., 2018), and aligned separately by using ClustalW, a multiple sequence alignment tool.(Sievers et al., 2011) Alignments are visualized in Unipro U-gene and hypervariable residues, their evolutionary amino acid frequencies and consensus sequences for CDRs are determined.(Okonechnikov et al., 2012) Residue positions with high gap penalties are not considered, the occurrences of the amino acids in highly variable positions are counted and the frequency of each amino acid is saved to be used further in the energy correction step. A consensus sequence is obtained for each alignment and hypervariable positions are labeled by asterisks (*) in the templates. Residues corresponding to these positions will be selected for mutation.

3.2.2 OPTIMIZATION STRATEGY

The strategy that we developed in our study is to start with a bound protein and nanobody complex, determine the hypervariable CDR residues according to the given templates, then perform 20 independent point mutations on variable positions, starting from Alanine up to Valine, by using visual molecular dynamics (VMD).(Humphrey, Dalke, & Schulten, 1996) Each mutant structure will be subjected to a minimization, annealing, conventional MD run and another minimization cycle by using NAMD.(Phillips et al., 2005) In this study CHARMM36 force field is used. The interaction energy of the nanobody and the protein for each mutation case is calculated and further corrected according to the evolutionary amino acid frequency scheme obtained from the alignment. Energies are corrected by using an iterative Monte Carlo algorithm which favors higher evolutionary preferences. The end result provides us with the distribution of possible optimum mutation types on hypervariable positions of CDRs. A broad library of optimal nanobodies can be derived and tested for its affinity and selectivity. The affinity of the new nanobodies can be measured by using Steered Molecular Dynamics (SMD)(Isralewitz, Gao, & Schulten, 2001), by pulling the nanobody away from the protein and measuring the work required to abolish all of the interactions. These SMD simulations provide the dissociation constant, K_D values which are useful constants in comparing the affinity of the optimized and the wild type nanobodies.

3.2.3 MOLECULAR DYNAMICS MIN-RUN CYCLE

All of the mutant nanobodies in complex with the protein are solvated in a TIP3P water box and counter ions are added to neutralize the system. CHARMM36 force field is used to parametrize the atoms. Time step of simulations are kept as 2 fs. Initially the energy of the system is minimized for 8000 steps and the system is gradually heated up to 310K. The annealing step is followed by an equilibration period at constant temperature of 310 K and constant pressure of 1 bar for 10 000 steps. Finally, the energy of the equilibrated system is minimized for another 5000 steps.

3.2.4 INTERACTION ENERGY CORRECTION

The interaction energy includes electrostatic, VdW and non-bonded terms. Majority of the force fields favors the charged residues over the others. According to the alignment studies, stable nanobodies have highly frequent residue types in their hypervariable positions. The interaction energies calculated from the min-run cycle, are corrected according to the probability distributions, obtained from the alignment, by using the following algorithm. We assume that the system obeys Boltzmann statistics, and the probabilities are functions of the energy and the temperature of the system. Therefore, the probabilities can be written in the following form;

$$p(i_1,j) = \frac{exp[-E(i_1,j)/E_0]}{\sum_j exp[-E(i_1,j)/E_0]} \qquad \text{EQUATION 46}$$

where $E(i_i,j)$ is the interaction energy of each mutant hypervariable position and the antigen obtained from MD min-run cycles, i_l is the hypervariable residue number and j is the mutated residue type (Ala, Arg, Asn, … , Val) and E_0 scales the temperature. Small values of E_0 favor low energy residues for each individual. These priori probabilities are further corrected according to the evolutionary distributions obtained from the alignments. The choice of residues according to the a priori probabilities, Eq. 1, is made as follows: For each position, i_1, a random number is generated between 0 and 1, and the j^{th} amino acid mutation for i^{th} residue is accepted if the random number is between $p(i_1, j)$ and $p(i_1, j+1)$. The weights of evolutionary frequencies on the a priori probabilities are introduced as follows: We consider the case where a total of n residues in the nanobody are selected as hypervariable positions. For a given mutated sequence i_1, where i_1 goes from 1 to n, we represent the calculated distribution from the MD interaction energies by *C(i)*. The evolutionary distribution for the mutated sequence is *K(i)* where i goes from 1 to n. The divergence , D_{CD}, of the distribution *C* from the evolutionary distribution *K* is calculated according to the divergence equation,(Kullback & Leibler, 1951)

$$D_{ED}(C \parallel D) = -\sum_{1\le j\le 20} D(j)\,\mathrm{LN}\left(\frac{C(i,j)}{D(j)}\right) \qquad \text{EQUATION 47}$$

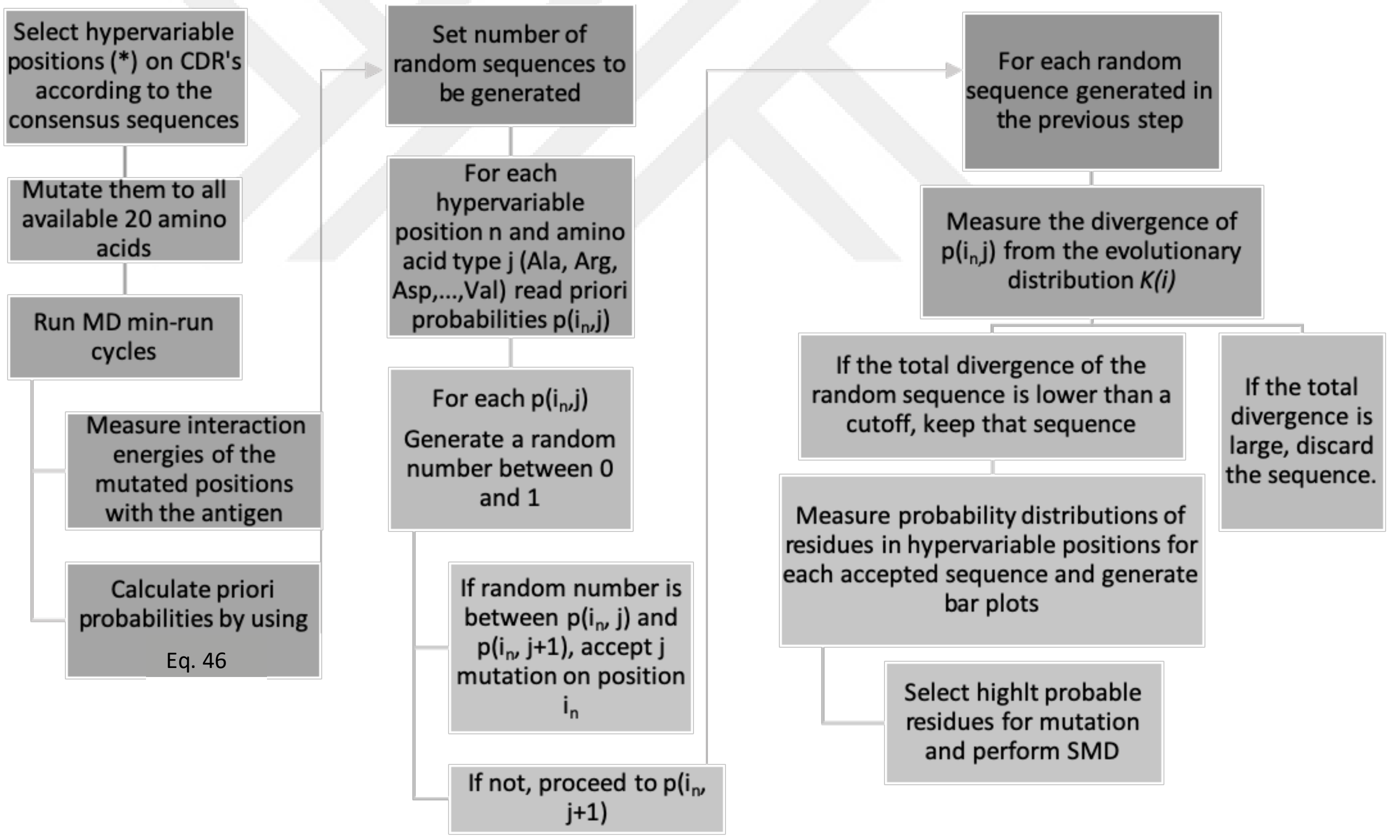

Select hypervariable positions (*) on CDR's according to the consensus sequences
Mutate them to all available 20 amino acids
Run MD min-run cycles
Measure interaction energies of the mutated positions with the antigen
Calculate priori probabilities by using Eq. 46
Set number of random sequences to be generated
For each hypervariable position n and amino acid type j (Ala, Arg, Asp,...,Val) read priori probabilities $p(i_n,j)$
For each $p(i_n,j)$ Generate a random number between 0 and 1
If random number is between $p(i_n, j)$ and $p(i_n, j+1)$, accept j mutation on position i_n
If not, proceed to $p(i_n, j+1)$
For each random sequence generated in the previous step
Measure the divergence of $p(i_n,j)$ from the evolutionary distribution *K(i)*
If the total divergence of the random sequence is lower than a cutoff, keep that sequence
If the total divergence is large, discard the sequence.
Measure probability distributions of residues in hypervariable positions for each accepted sequence and generate bar plots
Select highlt probable residues for mutation and perform SMD

Figure 35. *The flowchart of the optimization method.*

A set of random sequences are generated and the divergences of each random sequence from the evolutionary distribution is measured according to the flowchart shown in Figure 35. The sequences that have lower divergence from the evolutionary distribution are kept and the probability of residues at each hypervariable position is measured from the residual random sequences. A mutation library is generated with the proposed probability distributions for each hypervariable position. Mutations are selected from the proposed probability distributions and the affinities are calculated by the steered molecular dynamics method.

3.2.5 STEERED MOLECULAR DYNAMICS

In SMD, the binding affinity is correlated with the rupture force, where the ligand is totally detached from the protein. It has been shown that nonequilibrium pulling work also correlates well with the experimental results. (Truong & Li, 2018; Vuong, Nguyen, & Li, 2015) The proposed mutations are applied to nanobodies and the mutant nanobody is pulled away from the protein by SMD simulations at 310 K.(Isralewitz et al., 2001)For the SMD simulations the protein is anchored from a residue close to its center of mass but away from the nanobody binding site, the nanobody is pulled away with constant velocity from a residue away from the binding site and close to the free-end, where it is not bound to the protein. The force vectors required for the pulling is obtained from the simulation log file. The force vectors at each timestep are normalized with the pulling direction and the normalized force is plotted against the pulling distance. The cumulative force is calculated by summing up the area under the force vs distance curve. The distance of the SMD atom at which all the interactions between the nanobody and the protein diminished, rupture distance, is detected from the simulations, and the K_D value is calculated from the cumulative work value corresponding to that distance with the following formula:

$$K_D = exp\,(-\frac{Cumulative\ Work}{kT}) \qquad \text{EQUATION 48}$$

Where the unit of the cumulative work is kj/mol and *kT*, also written as k_BT, is the product of the Boltzmann constant, *k* and the temperature, *T*. *kT* is 2.58 kJ/mol at 310 K.

10 replicate SMD simulations were performed for each complex and the mean cumulative work values are considered in K_D calculations.

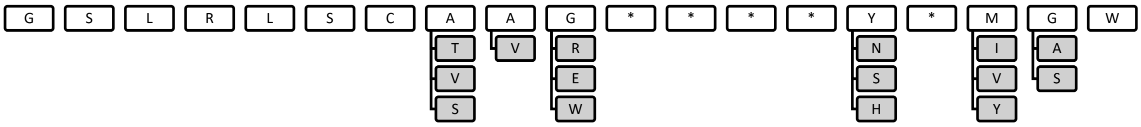
3.3 RESULTS

3.3.1 ALIGNMENT RESULTS

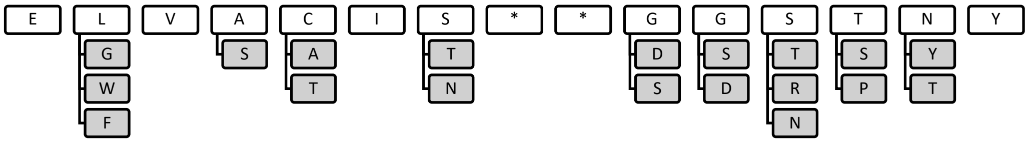
3.3.1.1 *VICUGNA PACOS*

The template for *Vicugna pacos* derived nanobodies are obtained, the consensus sequence for each CDR and the frequency of amino acids at each hypervariable position, shown by asterisks (*), is determined.

The consensus sequence and the template for CDR1



The consensus sequence and the template for CDR2



The consensus sequence and the template for CDR3

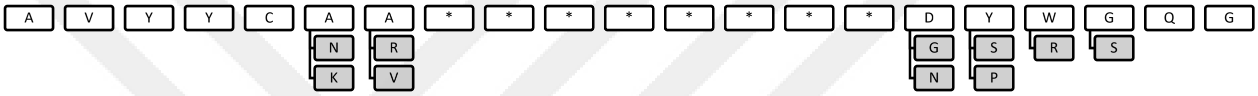


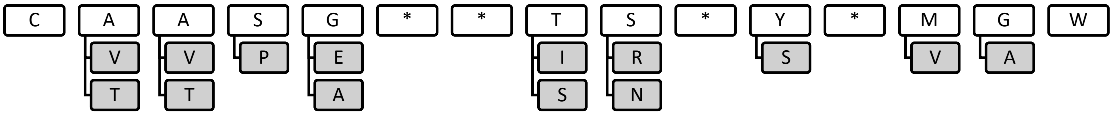
Table 2. Amino acid frequenciess for hypervariable residue positions at CDRs for *Vicugna pacos.*

Amino acid letter code	A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
Frequency	0.09	0.054	0.043	0.04	0	0.081	0	0.049	0	0.062	0.062	0	0	0.098	0.028	0.10	0.08	0.049	0.11	0.04

3.3.1.2 *CAMELUS DROMEDARIES*

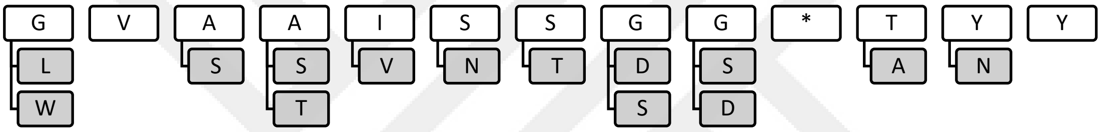
The template for *Camelus dromedaries* derived nanobodies are obtained by the same approach. The consensus sequence for each CDR and the frequency of amino acids at each hypervariable position is determined.

The consensus sequence and the template for CDR1



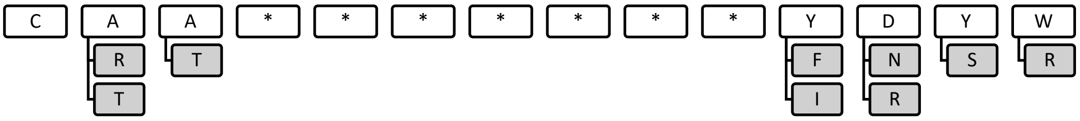
C	A	A	S	G	*	*	T	S	*	Y	*	M	G	W
	V	V	P	E			I	R		S		V	A	
	T	T		A			S	N						

The consensus sequence and the template for CDR2



G	V	A	A	I	S	S	G	G	*	T	Y	Y
L		S	S	V	N	T	D	S		A	N	
W			T				S	D				

The consensus sequence and the template for CDR3



C	A	A	*	*	*	*	*	*	*	Y	D	Y	W
	R	T								F	N	S	R
	T									I	R		

TABLE 3. AMINO ACID FREQUENCIES FOR HYPERVARIABLE RESIDUE POSITIONS AT CDRS FOR *CAMELUS DROMEDARIES*.

Amino acid letter code	A	R	D	N	C	E	Q	G	H	I	L	K	M	F	P	S	T	W	Y	V
Frequency	0.043	0.061	0.095	0.071	0	0.045	0	0.110	0	0.072	0	0	0	0.097	0.058	0.085	0.053	0	0.155	0.054

3.3.2 OPTIMIZATION RESULTS

Human MDM4, aldolase isoforms A, B and C from human tissue, Fructose-bisphosphate aldolase from *Trypanosoma congolense* and human lysozyme. PDB codes are 2VYR, 4ALD, 1QO5, 1XFB, 5O0W and 4I0C respectively are used in this study.

3.3.2.1 MDM4

Human MDM4 is a 490 amino acid, key regulatory protein of tumor suppressor p53. It consists of 3 conserved domains. An N terminal for binding p53, a Zinc-finger domain and a C-terminal RING domain. Over expression of them are associated with tumor production and it could be an important regulator in anticancer strategies.(Orit Karni-Schmidt, Maria Lokshin, & Carol Prives, 2016) MDM4 binds p53 and decreases its activity and stability against degradation. In most of the cancer studies, p53 is inactivated and reactivating p53 is thus an attractive strategy for the treatment. As the result of the recent studies, MDM proteins, negative regulators of p53, seem to be the druggable and controllable oncogenic targets. Several strategies have been proposed to reactivate p53 by suppressing MDM4. These strategies are based on disrupting the p53-MDM4 interaction by either using a small molecule or a peptidic compound. The small molecule antagonists that disrupts this interaction are nutlin and WK-298 .(O. Karni-Schmidt, M. Lokshin, & C. Prives, 2016; Wade, Li, & Wahl, 2013) Although these compounds help reactivating p53, they do not elicit all the effects of MDM overexpression. Peptidic antagonists have also been developed. Although they have larger interaction surfaces, most of them are unstable in vivo. (Toledo & Wahl, 2007)Another approach is to use mini proteins, nanobodies, to antagonize the interaction between p53 and MDM4. In a study of selecting MDM4 specific single domain antibody, a successful candidate nanobody was found to stabilize and inhibit p53-MDM4 interaction with a dissociation constant of 44 nM. This synthetic construct, VH9, is known to be the best binder for MDM4. (Grace, Vaysburd, Allen, Settanni, & Fersht, 2009)

Our aim is to use our optimization strategy to detect the wild type amino acids of the hypervariable positions and propose an amino acid probability distribution for them which will result in the similar binding affinity and the selectivity with VH9.

Crystal structure for MDM4-VH9 with the PDB id 2VYR is used. Since VH9 is a synthetic construct, thus, to determine which consensus sequence to utilize in the interaction energy update step, VH9 sequence is aligned with all consensus sequences for *Llama glama, Vicugna pacos* and *Camelus* dromedaries and the sequence similarity is measured to understand the homology of VH9 nanobody with other species. Percent identity score, which refers to a quantitative homology measure, calculated by multiplying the number of matching amino acids by 100 and dividing it by the length of the aligned sequence. Closely related sequences are expected to yield a higher percent similarity. The percent identity matrix of consensus sequences from other species and VH9 is shown on *Table 3*. VH9 and *Llama glama* alignment gives a higher percent similarity score, therefore VH9 is more homologous to *Llama glama* and its consensus sequence can be used to update MD interaction energies of VH9. The template shown on *Fig1.b* for *Llama glama,* obtained from a previous study, is used to correct the MD energies.(McMahon et al., 2018) Identity matrix of consensus sequences from other organisms and VH9 is shown in *Table 3*. VH9 showed slightly higher level of identity with *Llama glama* nanobody sequence. The hypervariable positions were detected according to the *Llama glama* template and the optimization procedure is applied. According to the template, highly variable residue positions for each CDR are residues number 30, 31, 32 and 33 for CDR1, 53 for CDR2 and 94, 95, 96, 97, 98, 99, 100, 101, 102 and 104 for CDR3.

TABLE 4. PERCENT IDENTITY MATRIX IF NANOBODY CONSENSUS SEQUENCES AND VH9, CREATED BY CLUSTAL12.1

	VH9	*Vicugna pacos*	*Llama glama*	*Camelus dromaderies*
VH9	100.00	62.31	**69.35**	66.94
Vicugna pacos	62.31	100.00	70.87	66.41
Llama glama	68.29	70.87	100.00	69.29
Camelus dromaderies	66.94	66.41	69.29	100.0

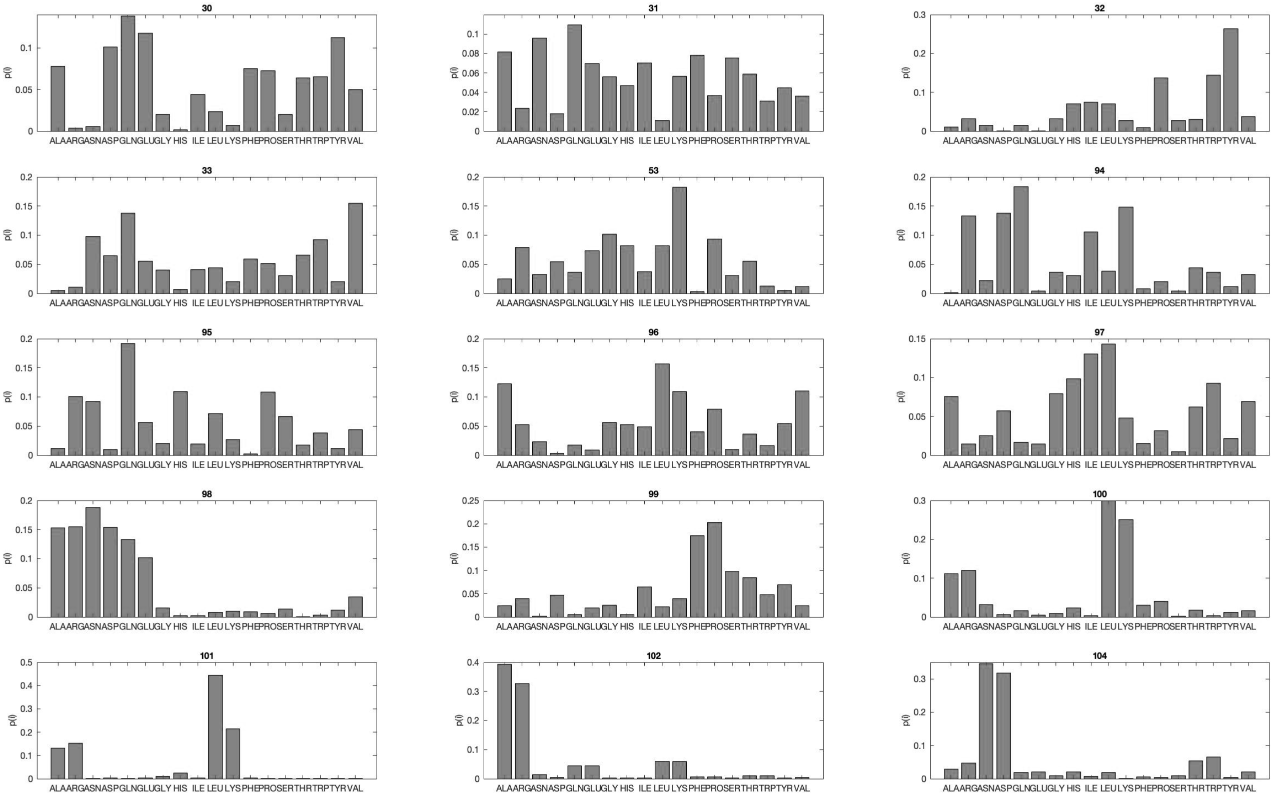


Figure 36. The probability distribution of the optimum residue mutation types of selected hypervariable positions of VH9 in complex with MDM4.

Hypervariable position mutations were selected according to the histograms and an optimized nanobody is generated. The affinity of the optimized nanobody and the wild type VH9 is compared with SMD. A constant pulling speed of 30 Å/ns is applied.

TABLE 5. RESIDUE TYPES AT THE HYPERVARIABLE POSITIONS FOR WT VH9 AND MUTANT VH9 OPTIMIZED FOR MDM4

	30	31	32	33	53	94	95	96	97	98	99	100	101	102	104
WT	GLU	GLU	TYR	ALA	ALA	TYR	TYR	CYS	ALA	LYS	PRO	TRP	TYR	PRO	MET
Mutant	GLU	TYR	THR	ARG	HIS	TYR	SER	GLY	GLY	TYR	ARG	TYR	ARG	GLY	ASP

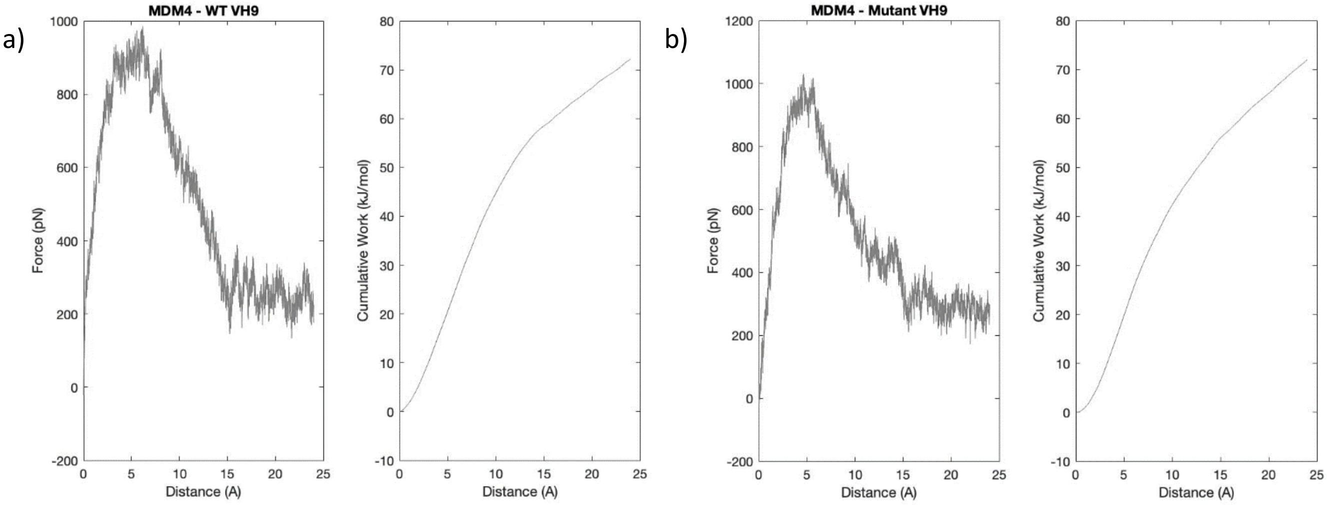


Figure 37. Mean **f**orce(pN) vs distance(Å) and mean cumulative work(kJ/mol) vs distance(Å) plots for a) MDM4-WT VH9 and b) MDM4- mutant VH9

Affinities are calculated from SMD plots given in Figure 37. The rupture distance is 10.34 Å for WT VH9 and 11.83 Å for mutant VH9. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D values are calculated from the cumulative work values corresponding to rupture distances. Calculated K_D values are 18.12 nM and 9.24 nM for WT and mutant VH9 respectively.

3.3.2.2 ALDOLASE ISOFORMS A, B AND C

Aldolase isoforms are abundant in the human body and they play crucial roles in glycolysis, fructolysis, and the synthesis of ATP and glyceraldehyde. Aldolase family members are prognostic and diagnostic markers of human cancers and diseases. The three members which are involved in metabolism and glycolysis are Aldolase isoform A, B and C. Nanobodies against the human isoforms will provide new tools to study and regulate glycolysis and potentially open the way to new forms of therapy against diabetes and cancer. Nb474 is used as the starting nanobody structure for all isoforms.(Y.-C. Chang, Yang, Tien, Yang, & Hsiao, 2018)

Fructose-bisphosphate aldolase, also known as Aldolase isoform A, is a glycolytic protein encoded by ALDOA gene. (Du et al., 2014)This enzyme catalyzes fructose-1,6-bisphosphate to glyceraldehyde 3-phosphate (G3P) and dihydroxyacetone phosphate (DHAP). This isoform is expressed in muscle and its deficiency is linked to hemolytic anemia and myopathy. The crystal structure with PDB id 4ALD is used. (A. Dalby, Dauter, & Littlechild, 1999)

Aldolase B is the isoform which is expressed in liver, kidney and enterocytes involved in both glycolysis and gluconeogenesis. (Lemaigre & Rousseau, 1994)The crystal structure of fructose 1,6-bisphosphate aldolase B from Human Liver Tissue deposited in the databank with PDB id 1QO5 is used.(A. R. Dalby, Tolan, & Littlechild, 2001)

Aldolase C is the isoform which is expressed in the brain. Aldolase C related pathways are glucose metabolism and Innate Immune System. (C.-S. Zhang et al., 2017)The crystal structure of Human Brain Fructose 1,6-bisphosphate Aldolase C deposited in the databank with PDB id 1XFB is used. (Arakaki et al., 2004)

Nb474 with PDB id: 5O0W is used as the starting structure. It is docked into all isoform. The hypervariable positions on Nb474 CDRs is selected according to the *Vicugna pacos* template. The residue numbers to be mutated are 26, 27, 28, 29, 30, 31, 32, 101, 102, 103, 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116, 117, 118, 119, 120, 121, 122, 123, 124, 125 and 126. Same residues are mutated on Nb474 for all three docked complexes.

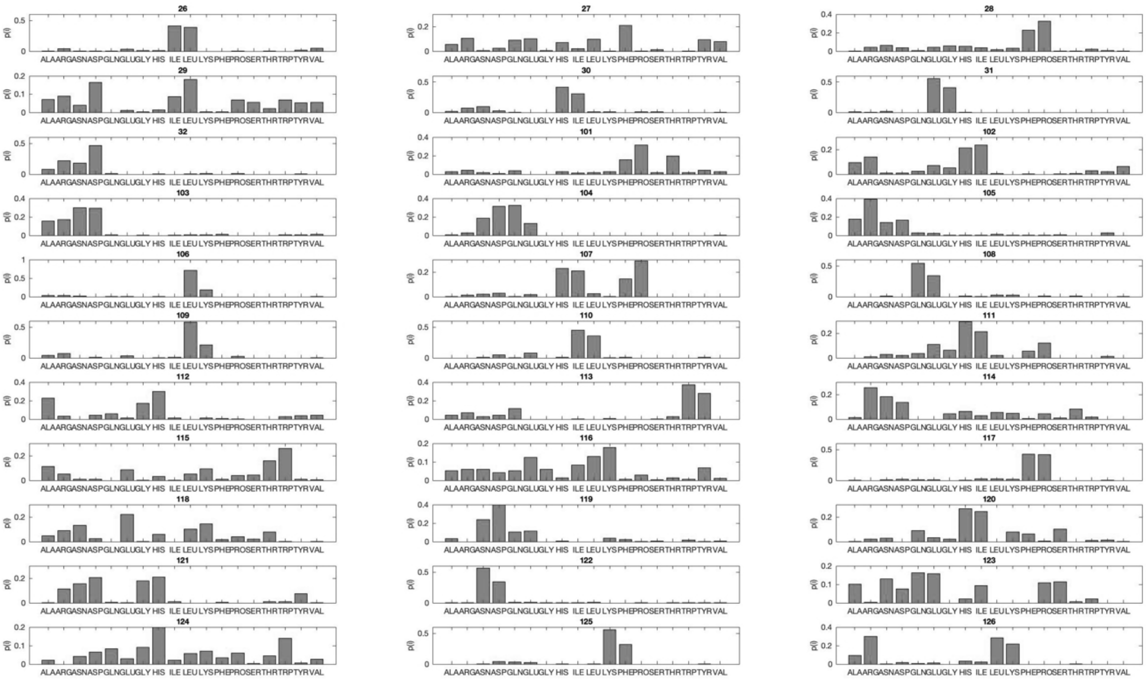


FIGURE 38. THE PROBABILITY DISTRIBUTION OF THE OPTIMUM RESIDUE MUTATION TYPES OF SELECTED HYPERVARIABLE POSITIONS OF NB474 IN COMPLEX WITH ALDOLASE ISOFORM A.

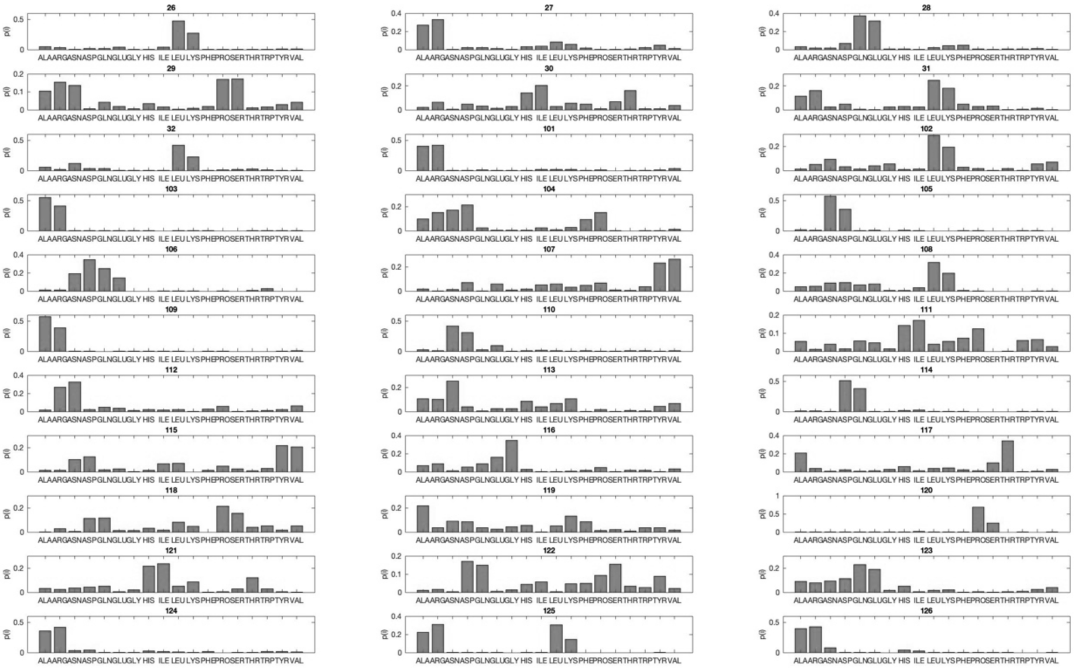


Figure 39. The probability distribution of the optimum residue mutation types of selected hypervariable positions of Nb474 in complex with Aldolase isoform B.

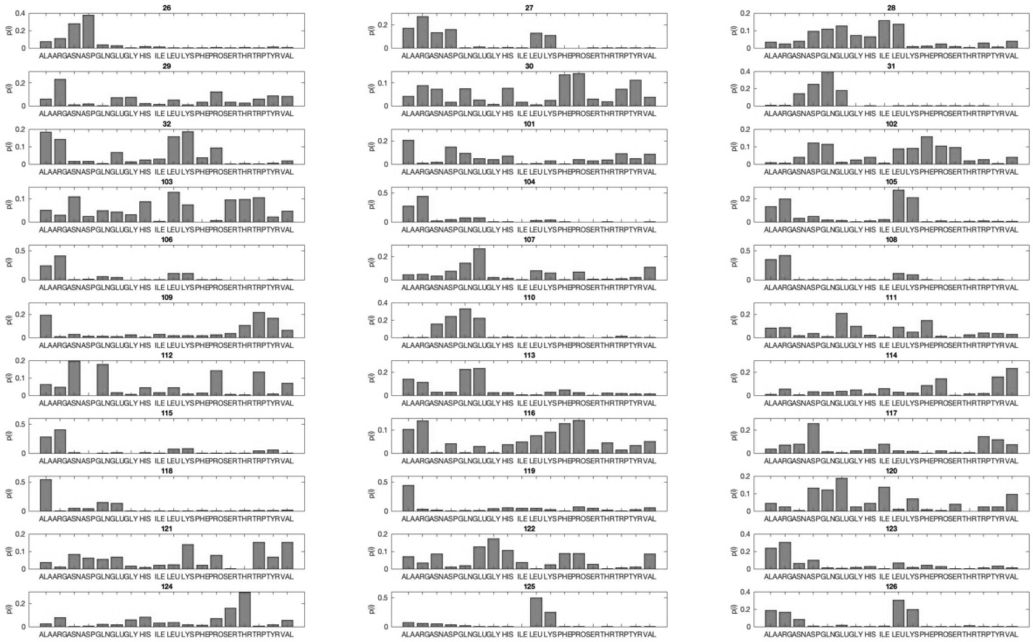
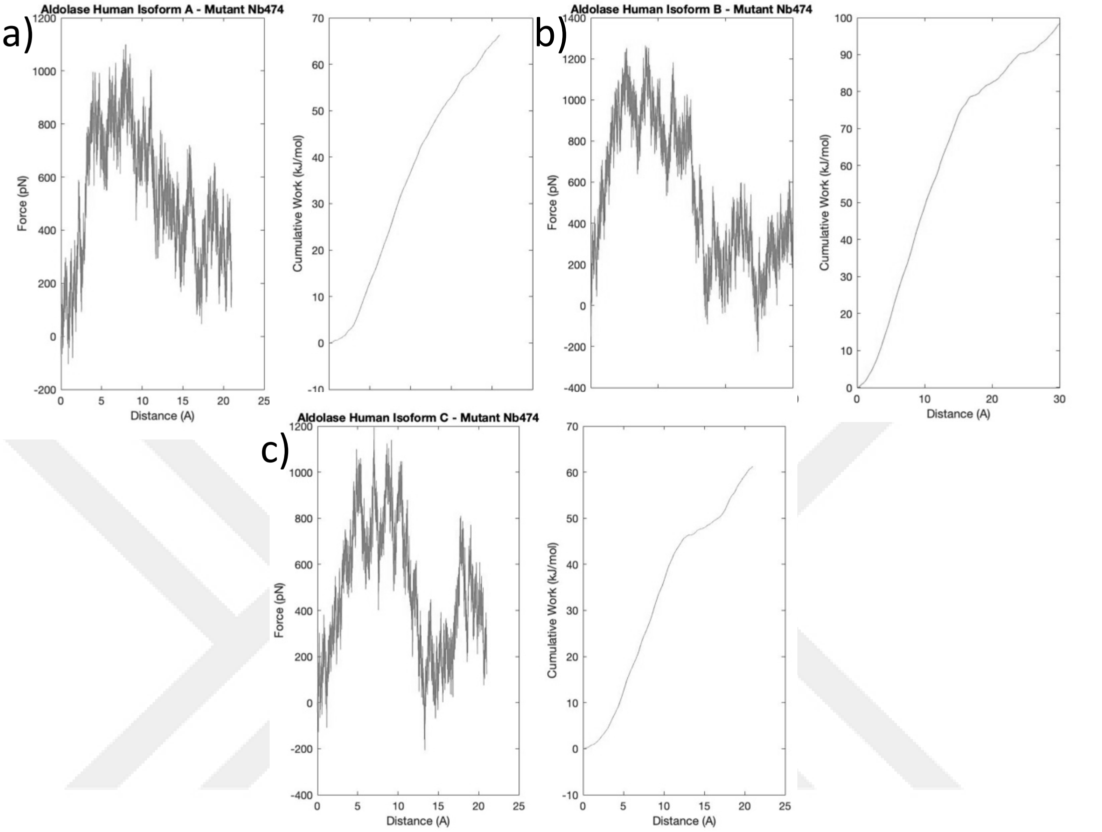


Figure 40. The probability distribution of the optimum residue mutation types of selected hypervariable positions of Nb474 in complex with Aldolase isoform C.

Nanobody sequences for isoforms A, B and C are randomly selected from the distributions and SMD simulations were performed with a constant pulling speed of 30 Å/ns.

Figure 41. Mean force(pN) vs distance(Å) and mean cumulative work(kJ/mol) vs distance(Å)



plots for a) Isoform A- optimized Nb474 b) Isoform B- optimized Nb474 and C) Isoform C- optimized Nb474

The rupture distance and measured K_D values for isoform A, B and C are: 11.64 Å- 66.42 nM, 11.69 Å- 0.26 nM and 10.89 Å – 138.73 nM respectively.

3.3.2.3 FRUCTOSE-BISPHOSPHATE ALDOLASE FROM TRYPANOSOMA CONGOLENSE

Trypanosoma genus belongs to a diverse group of parasites, which cause disease in humans and livestock. Infections of this genus leads to Human African Trypanosomosis (HAT) and Animal African Trypanosomosis (AAT). Specifically, *Trypanosoma congolense* is responsible for the infections and it also leads to major economic losses. A proper diagnostic tool is required in order to detect the infection in animals and proceed with the selective treatment. (Pinto et al., 2017) Recent studies show that nanobodies can be used as a diagnostic tool to recognize *Trypanosoma congolense* fructose-1,6-bisphosphate aldolase (TcoALD). It is a well conserved glycolytic enzyme among Trypanosoma genus. The affinity of the nanobody, Nb474, determined for enzyme detection has a dissociation affinity constant of 73.83 pM.(Pinto et al., 2017)

The aim is to use this optimization strategy to propose similar amino acids of the WT Nb474 hypervariable positions which will yield in the similar binding affinity and the selectivity with Nb474. Crystal structure for TcoALD-Nb474 with the PDB ID 5O0W is used. The template for *Vicugna pacos* is used to correct the MD energies, since it is an alpaca derived nanobody.

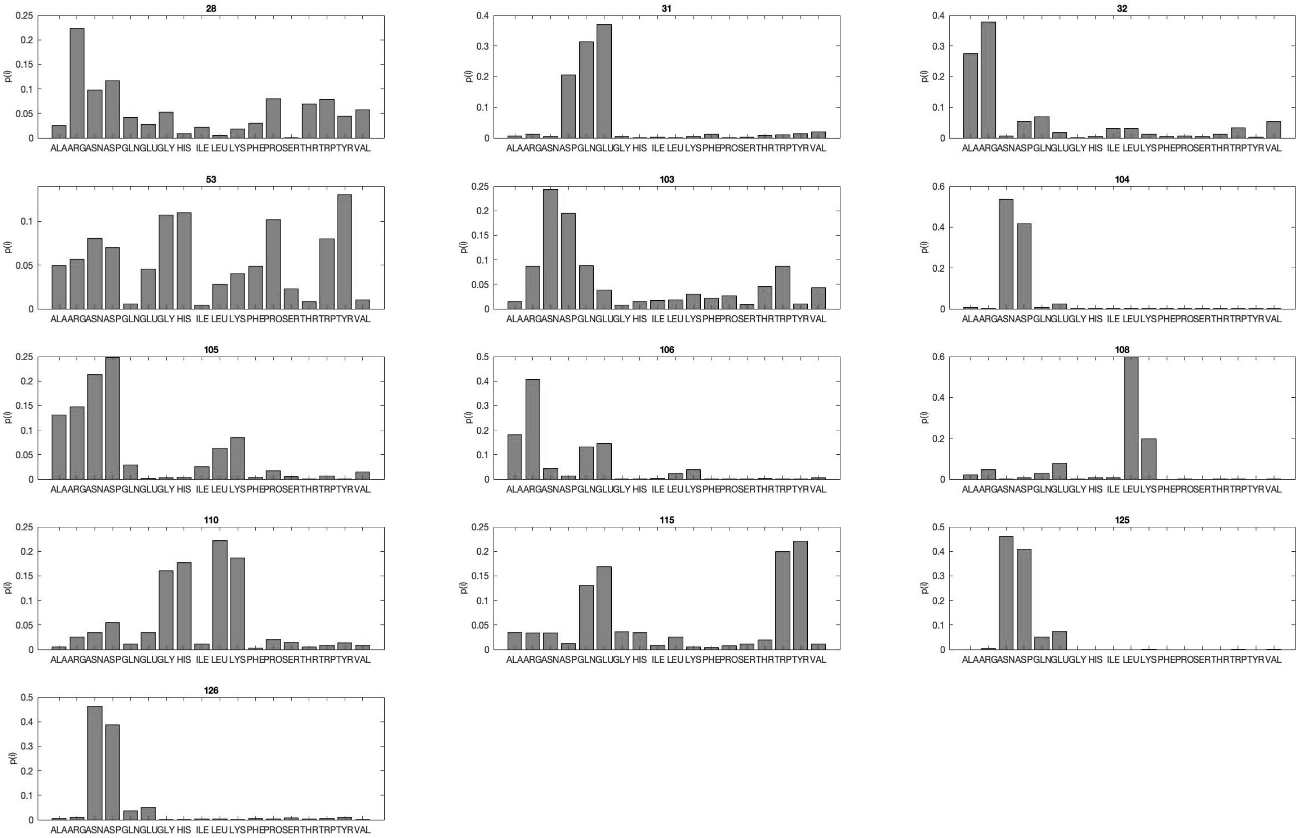


Figure 42. The probability distribution of the optimum residue mutation types of selected hypervariable positions of Nb474 in complex with TcoALD.

The hypervariable positions were detected according to the template and the optimization procedure is applied. According to the template, highly variable residue positions for each CDR are residues number 28, 31, 32 for CDR1, 53 for CDR2 and 103, 104, 105, 106, 108, 110, 115, 125 and 126 for CDR3.

Optimum mutations for hypervariable residues are selected from the histogram plot and an optimized nanobody is generated. The affinity of the optimized nanobody and the wild type Nb474 is compared with SMD. A constant pulling speed of 30 Å/ns is applied.

TABLE 6. RESIDUE TYPES AT THE HYPERVARIABLE POSITIONS FOR WT NB474 AND MUTANT NB474 OPTIMIZED FOR TCOALD

	28	31	32	53	103	104	105	106	108	110	115	125	126
WT	ALA	TYR	TYR	ARG	ASP	THR	THR	ASP	TYR	SER	TYR	ASP	TYR
Mutant	ARG	GLU	ALA	HIS	ASN	ASP	LEU	GLN	LEU	LYS	TYR	ASP	GLN

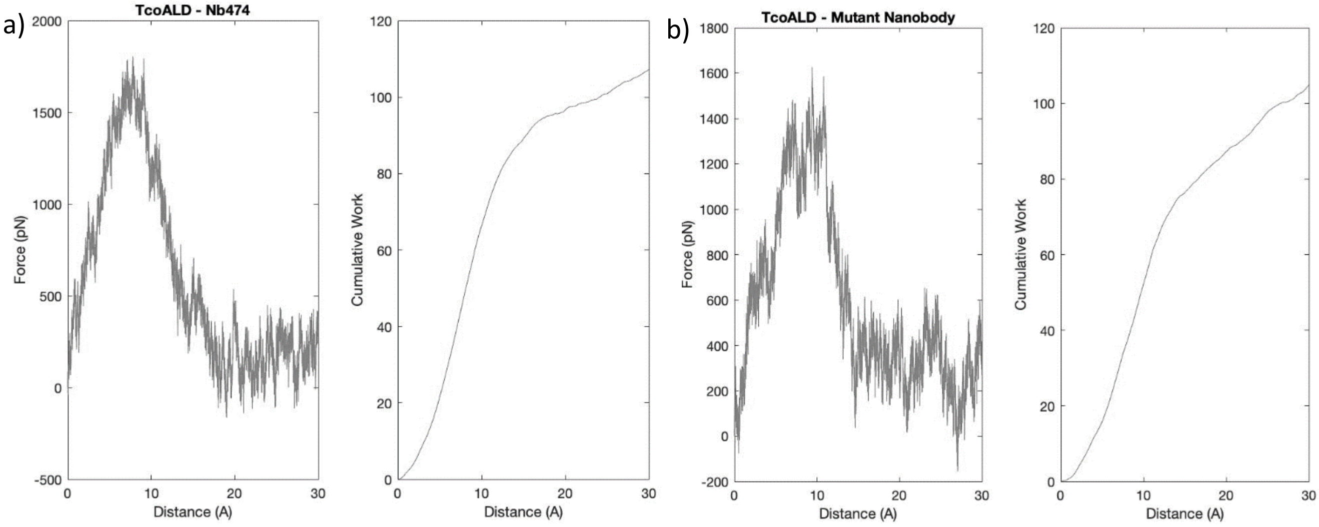


Figure 43. Mean force(pN) vs distance(Å) and mean cumulative work(kJ/mol) vs distance(Å) plots for a) WT Nb474 and b) mutant Nb474

The rupture distance is 10.26 Å for WT Nb474 and 12.07 Å for mutant Nb474. Calculated K_D values are 3.5 pM and 6.45 pM fot WT and mutant Nb474 respectively.

3.3.2.3 HUMAN LYSOZYME

Human lysozyme (HuL) belongs to the c-type class lysozymes. It is a 130 amino acid protein, capable of both hydrolysis and trans-glycosylation, plays a vital role in host defense. Several variants of this protein were found to be related with a familial systemic non-neuropathic amyloidosis. It has been revealed that several nanobodies are available that prevent the formation of pathogenic aggregates and inhibit fibril formation by the amyloidogenic variants. These nanobodies, cAbHuL5 and cAbHuL5G, can be used to treat protein misfolding diseases. cAbHuL5 and cAbHuL5G displays an affinity of 460 nM and 310 nM for wild type HuL.(Erwin De Genst et al., 2013)

The aim is to use this method to propose similar amino acids of the WT cAbHuL5 hypervariable positions which will yield in the similar binding affinity and the selectivity with cAbHuL5. Crystal structure for HuL-cAbHuL5 with the PDB ID 4I0C is used. The template for *Camelus dromedaries* is used to correct the MD energies. The hypervariable positions were detected according to the template and the optimization procedure is applied. According to the template, highly variable residue positions for each CDR are residues number 27, 28, 29, 30, 31 for CDR1, 55, 56 for CDR2 and 97, 98, 101, 110, 111, 112, 113 and 115 for CDR3.

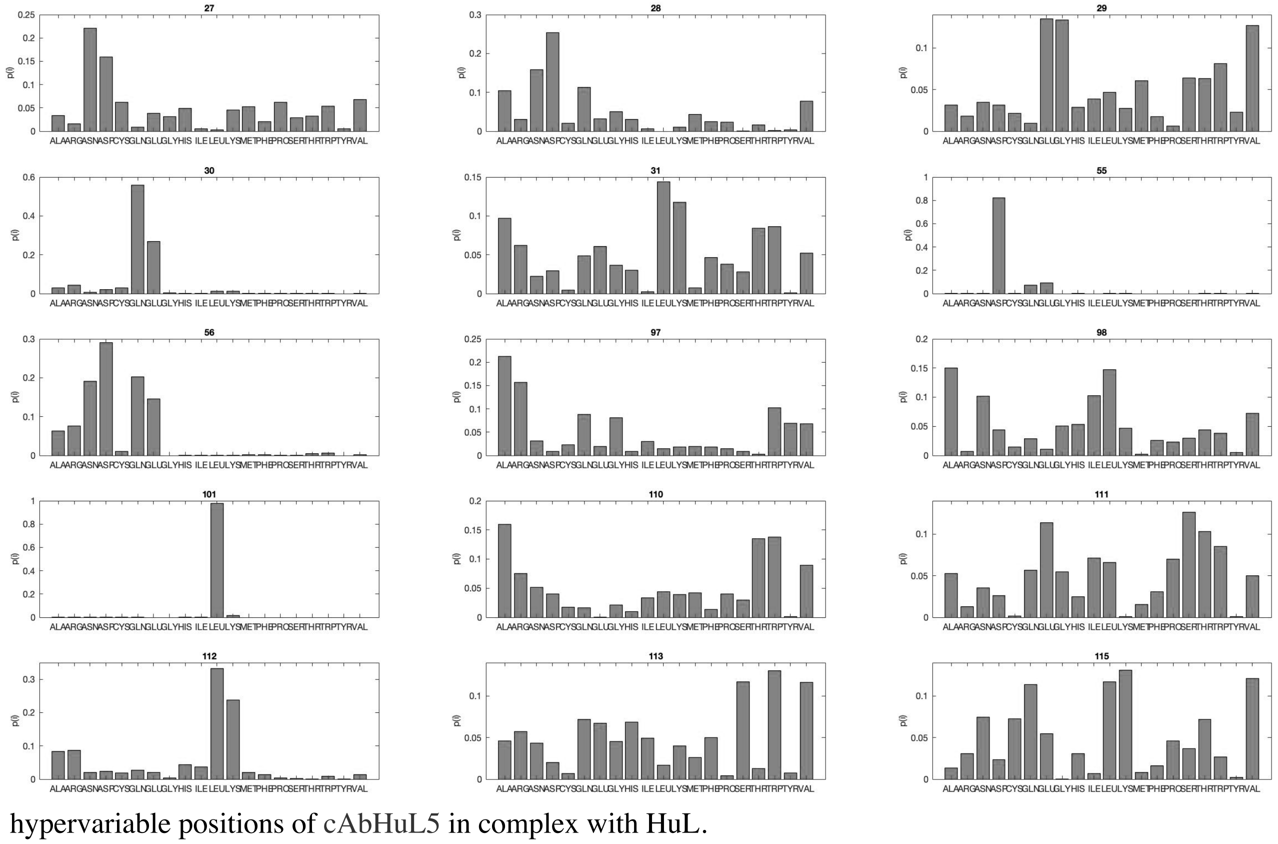
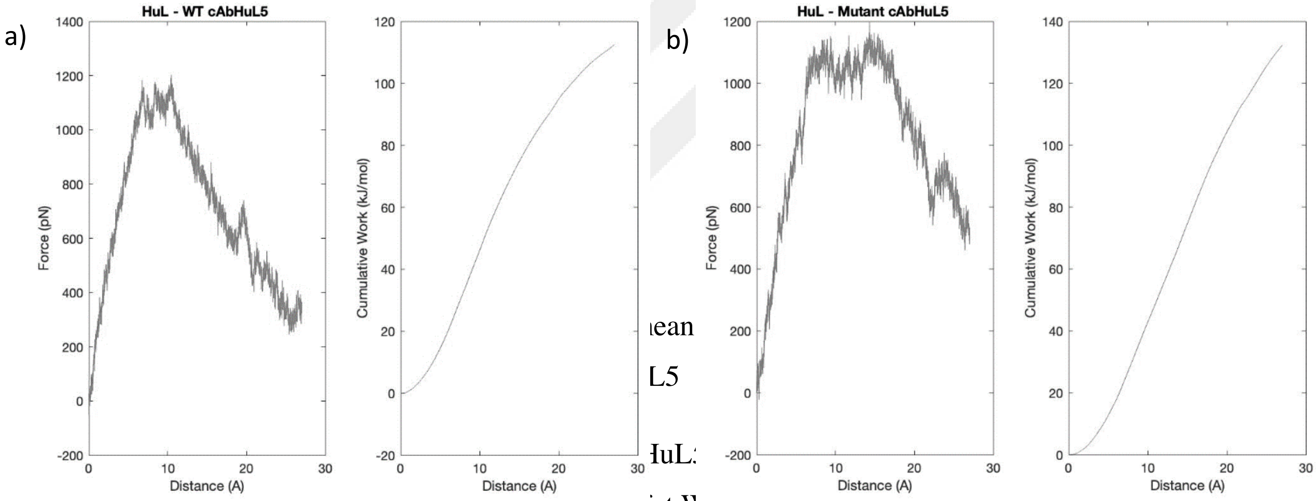


Figure 44. The probability distribution of the optimum residue mutation types of selected hypervariable positions of cAbHuL5 in complex with HuL.

TABLE 7. RESIDUE TYPES AT THE HYPERVARIABLE POSITIONS FOR WT

	27	28	29	30	31	55	56	97	98	101	110	111	112	113	115
WT	LEU	SER	THR	THR	VAL	PHE	PRO	LYS	THR	PHE	SER	ARG	ALA	TYR	HIS
Mutant	LEU	ASN	ASP	GLU	GLN	ASP	GLN	ALA	LEU	LEU	TRP	SER	LYS	TRP	VAL

CABHUL5 AND MUTANT CABHUL5 OPTIMIZED FOR HUL



ıean

L5

HuL

Calculated K_D values are [illegible] nM and [illegible] pM for WT and mutant cAbHuL5 respectively.

3.4 DISCUSSION

In this article we present a method to find better binding nanobodies for protein targets. Binding of known protein-nanobody complexes are used as benchmark. Evolutionarily variant residues on nanobodies are detected and nanobodies with similar affinities to the best binders are generated. The K_D values for wild type nanobodies are replicated with 10 replicates of Steered Molecular dynamics simulations and the affinity of mutated nanobodies are compared with the results of experimentally detected values.

3.4.1 MDM4-VH9

VH9 is the best binding nanobody to human MDM4 N-terminal domain synthesized so far. The experimentally determined dissociation constant of VH9 bound to human MDM4 N-terminal domain is 44 nM. The dissociation constant we detected in our SMD simulations is 18.12 nM, which is in the same nM range as the experimentally determined value. Hypervariable amino acids in the wild type VH9 falls into the highly probable amino acid scheme that is proposed by this optimization method, shown in Figure 36. The hypervariable amino acids of VH9 are mutated according to the probability distribution scheme and the dissociation constant of the mutant VH9 was measured as 9.24 nM.

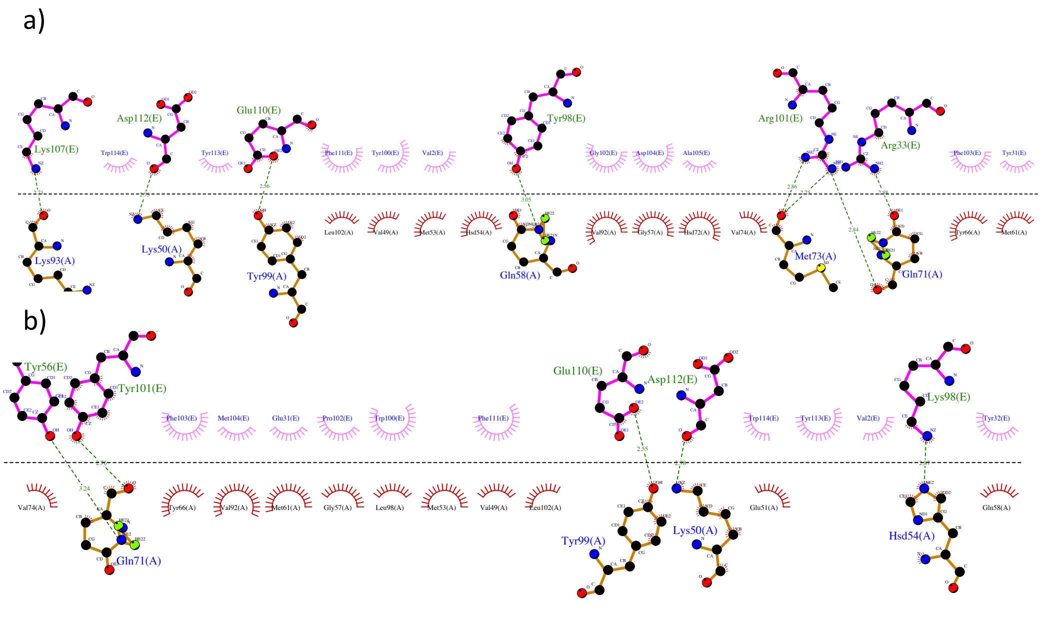


Figure 46. Residue interaction plots of a) MDM4 – WT VH9 and b) MDM4 – Mutant VH9 is plotted using Ligplus, where chain A denotes the MDM4 and chain E denotes the nanobody in both figures.

The plots in Figure 46 show that this optimization strategy helps us finding nanobodies that preserve and increase the wild type interactions. One of the CDR2 residues, residue number 56, makes a hydrogen bond in mutant VH9. The contact points shown in a and b panels of Figure 46 show that this optimization strategy helps us finding nanobodies that preserve the wild type interactions and according to the SMD results compared with the WT and presented in Figure 37. this strategy detects nanobodies with improve binding properties to the target antigen. The interaction of non-CDR residues and MDM4 in mutant VH9 is improved and the binding surface area, calculated by using Cocomaps, of the nanobody with MDM4 is increased from 792.2 Å^2 to 893.55 Å^2 in mutant structure. The affinity of both nanobodies for MDM4 is in nM range

3.4.2 ALDOLASE HUMAN ISOFORMS A, B AND C

Human isoforms of aldolase are studied. Nb474 nanobody, designed for TcAldolase was used as the starting structure. The aim was to enhance the affinity of Nb474 towards human isoforms and decrease towards TcAldolase. There were no experimental results available for human isoforms, but according to the SMD simulations nanobodies with 66.42 nM, 0.26 nM and 138.73 nM were designed for isoform A, B and C respectively. These results look promising as a further experimental analysis starting point.

3.4.3 TCOALD-NB474

A diagnostic assay, employing the smallest antigen binding Nb474, was prepared to detect *T. congolense* infections in animals. This nanobody specifically recognizes *T. congolense* fructose-1,6-bisphosphate aldolase. The affinity of this specific interaction is experimentally found as 73.83 pM, meaning a very tight binding, which we determined as 31.22 pM in our SMD simulations. The mutant Nb474 generated by mutating hypervariable residues according to the probability distribution scheme proposed for TcoAld - Nb474 template, shown in Figure 42. The affinity of the mutant Nb474 for TcoAld is measured as 634.5 pM. Both WT and the mutant Nb474 have affinities in the same magnitude of order.

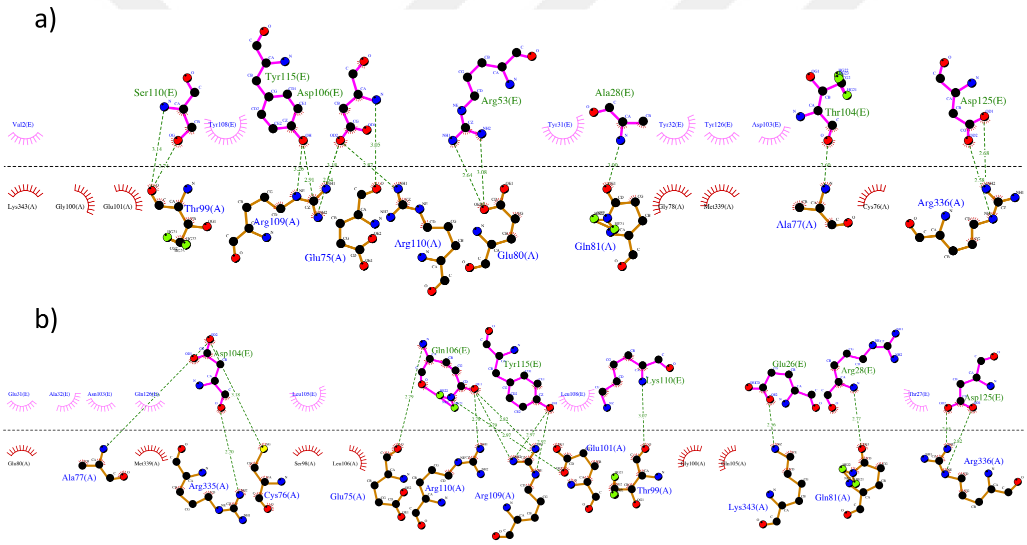


Figure 47. Residue interaction plots of a) TcoAld – WT Nb474 and b) TcoAld – Mutant Nb474 is plotted using Ligplus, where chain A denotes the TcoAld and chain E denotes the nanobody in both figures.

The interaction plots in Figure 47 and SMD plots in Figure 43 show that; similar binding patterns and affinities can be achieved by using this optimization strategy with experimentally detected nanobodies. The TcAldolase-nanobody interface area is calculated by using Cocomaps, an online tool to analyze protein interaction interface, as 820.75 Å^2 and 915.55 Å^2 for WT and mutant cases respectively. Both affinities of mutant and WT nanobodies are in pM range and interactions in WT Nb474 are preserved in the mutant case and the binding surface area is increased.

3.4.3 HUL- CABHUL5

The reported nanobody inhibits fibril formation of Human Lysozymess by preventing the unfolding and structural reorganization of the α-domain. The studies provide evidence that nanobodies can be used as therapeutic reagents for protein misfolding diseases. The suggested nanobody has an affinity of 460 nM towards HuL, which detected as 0.15 nM in SMD simulations in this study. (E. De Genst et al., 2013) The proposed mutant according to the optimization method shows an affinity in low pM range, 1.48 pM.

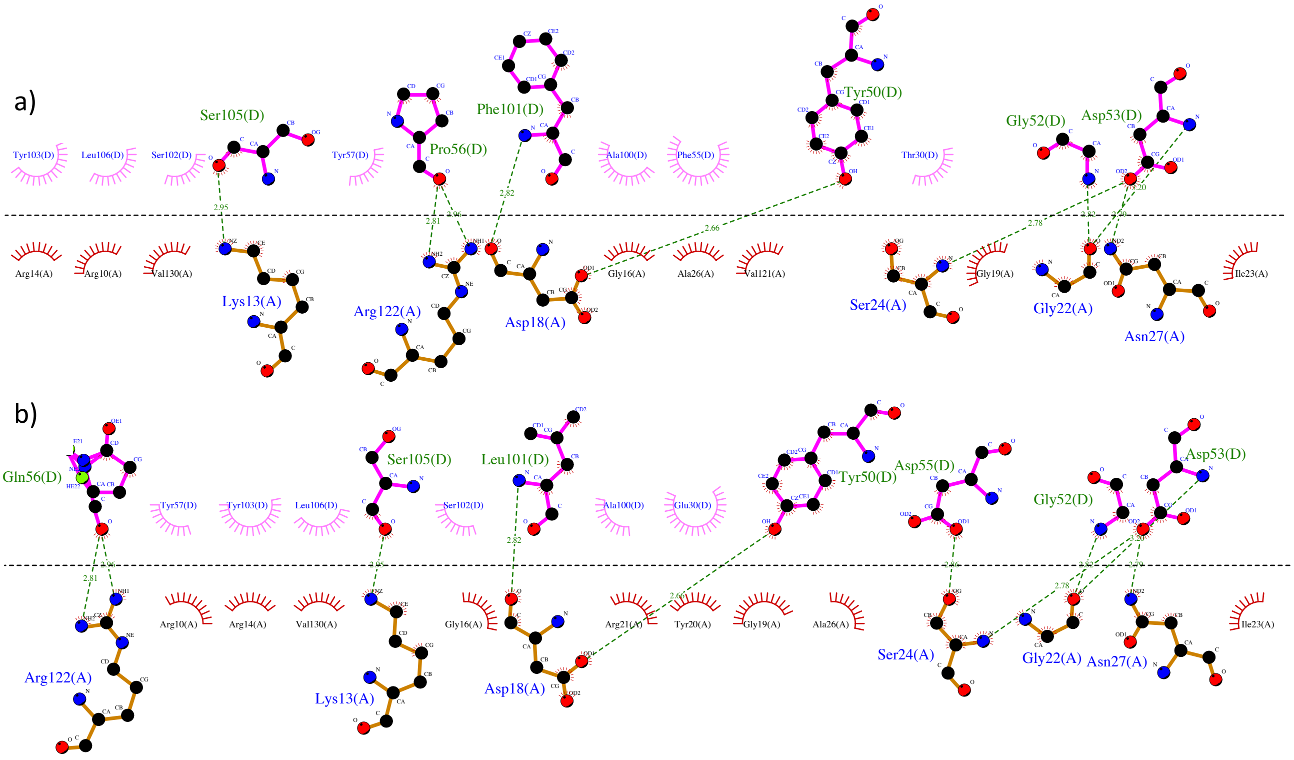


FIGURE 48. INTERACTING RESIDUES OF A) HUL – WT CABHUL5 AND B) HUL – MUTANT CABHUL5 IS SHOWN USING LIGPLUS, WHERE CHAIN A DENOTES THE HUL AND CHAIN D DENOTES THE NANOBODY IN BOTH FIGURES.

The nanobody generated by mutating its hypervariable CDR residues have higher affinity towards HuL and CDR3 interactions with HuL increased in mutant cAbHuL5. The binding surface are increased from 662.95 Å^2 to 691.6 Å^2 upon mutating the hypervariable residues on cAbHuL5.We can conclude that tighter binding nanobodies, correlated with the experimental results, can be proposed by this optimization method.

3.5 CONCLUSION

Therapeutic nanobodies are currently emerging, powerful biological drugs used for treatment and diagnostic purposes. Their small size, increased stability and solubility and large binding regions make them great therapeutic reagent candidates. Experimentally screening the best binder by phage display based methods is time-consuming and computational tools, such as molecular dynamics and steered molecular dynamics can be used to design a nanobody for a specific target and characterize the binding properties.

In this study, an optimization method is proposed by selecting the evolutionarily variable residues on nanobodies to design conformationally selective nanobodies and finding amino acid types for hypervariable residues that improve nanobody-protein interaction by applying point mutations on them and determining their contribution in binding energy and binding. Affinities of the mutant and WT nanobodies can be compared by using steered molecular dynamics. The unbinding direction can be optimized by comparing the SMD results with the actual experimental results.

This study reveals that mutation on hypervariable CDR residues improve the binding affinity of nanobodies and can detect experimentally proven best binders. Currently proposed method can be extended for wider applications and many tighter binding nanobodies can be designed for specific targets within nM-pM affinity range. We conclude that computationally optimizing binding properties in nanobody-antigen complexes will generate a reliable source of nanobody candidates for specific targets and we suggest that this method can complement existing immune and synthetic library-based methods, without the need of experiments or large libraries.

CHAPTER 4.

MOLECULAR DYNAMICS SIMULATION TRAJECTORY PROJECTION

4.1 OVERVIEW

Molecular dynamics simulations are functional tools for understanding the conformational patterns and properties of protein behavior, receptor-ligand interactions and the conformational changes that a biomolecule may undergo under various conditions. (Gane & Dean, 2000) Proteins are highly dynamic entities, their internal organization and the conformational space they span are key to their biological function. In vivo study of protein dynamics is very complicated and time consuming, therefore making use of simulations at different levels of detail and integrating them into the life sciences is a useful and powerful approach.(Hospital, Goñi, Orozco, & Gelpí, 2015; Vlachakis, Bencurova, Papangelopoulos, & Kossida, 2014) A single MD simulation for a study is not enough to get a reliable result and several duplicate simulations of the same system must be conducted to replicate the same interactions and motions observed in a single MD run. In this study we used Latent semantic indexing method, which is an information retrieval technique useful for unraveling hidden concepts in data. Singular value decomposition is used to compute similarities between different molecular dynamics runs of the same system based on the inner product in the latent semantic space. The goal is to map similar motion patterns in similar location in low dimensional space and reduce randomness in simulations by dimension reduction.

4.2 INTRODUCTION

Molecular dynamics explores dynamics of a biomolecule through a series of time frames by solving Newton's equations of motions. In principle, MD is a method that samples the response of a biomolecule under a particular set of conditions, including random motions caused by the thermal fluctuations. (Haile, 1992) A simulation trajectory represents a sample of proteins’ behavior in its dynamics under a particular condition. The motions of biomolecules are observed to be reproducible, dominant patterns in conformational dynamics are frequently visited. Progression of motions in one MD simulation is random, duplicate trajectories of the same system may not yield in the same results. Many computational studies are based on comparing the averaged structures from two or more simulations. It is easy to compare averaged structures but the phase space changes considerably depending on the direction. (Galindo-Murillo, Roe, & Cheatham III, 2015) A common approach for comparing the trajectories of the same system is the cluster analysis, but setting up a suitable distance function, the routines for delimiting the clusters might cause problems in assessment of reliability. Several simulations may converge at different time steps due to the randomness of starting velocities and motions. (Farmer, Kanwal, Nikulsin, Tsilimigras, & Jacobs, 2017)This randomness causes a noise in the generated data, but dominant collective modes of the trajectories resulted from different MD runs will be similar since the statistical significance of local dynamics of the system stays the same. Therefore, quantifying and comparing the similarity of local dynamics in a mode-based approach will be more precise to determine the quality and the accuracy of the simulation results, and will allow determining the most significant dynamic patterns of the system. This approach will also allow comparing simulations with different time lengths.

Main patterns of the data can be determined by reducing the dimension to the few, largest principal components.(Berry, Dumais, & O'Brien, 1995) By applying an orthogonal linear transformation and transforming it to a new coordinate system, biologically meaningful signals can be extracted from the data. In order to be able to compare different trajectories, they need to be brought to a common ground by using a technique called Latent Semantic Indexing (LSI). (Chen, Martin, Daimon, & Maudsley, 2013) It is an information retrieval method that uses singular value decomposition (SVD) to identify hidden patterns in the relationships between the concepts contained in a collection of terms by calculating the best rank-l approximation of the term-document matrix and reduces the dimensional space to eigen-term and eigen-document. (Yamazaki, Tomov, & Dongarra, 2017) In this context, the terms are the x, y, z components of carbon alpha atoms of each residue in the system throughout the simulation time, the concept is the conformation of the system. The document is a t x 3N matrix, T. where t denotes the time step of the simulation and 3N is the x, y, z coordinates of each alpha carbon, N in number, in the system.

This method will allow us to detect document similarity, project different trajectories onto each other without perturbing the patterns of the system, determine the common dominant motions and filter out outlier motions caused by the randomness of protein motions.

4.3 METHOD

In order to implement LSI to molecular dynamics trajectory data, a matrix of the fluctuation of x, y, z coordinates of alpha carbon of each residue is constructed. A t_1 x 3N matrix T_1, where t_1 denotes the time of simulation and 3N denotes the fluctuation of x, y, z coordinates of the alpha carbon of residues in the protein. Rows of the matrix, each fluctuation of x, y and z coordinate of alpha carbon will be denoted as terms and the columns of the matrix, time steps of simulations, will be denoted as documents according to LSI concept. This matrix is factored into 3 matrices using SVD. Then latent semantic structure model is derived from the left and right singular vectors, and the diagonal singular value matrix, U, V and Σ respectively.

$$T_1 = U\Sigma V^T \text{ EQUATION 49}$$

Each entry of the U matrix represents the terms relation with a hidden concept. Term is denoted as the mean fluctuations of all amino acids at a certain time step. Each entry of V^T represents the documents relation with a concept. Document is denoted as the mean fluctuations of all amino acids throughout the simulation. By using LSI, we'll group each alpha carbons the fluctuation in a document, along the simulation time, into concepts. Then we'll perform a dimensionality reduction by truncating each matrix and taking top k singular values, which capture the most variance from the original matrix. The truncated matrices can be further used as a measure metric for all fluctuations along the simulation time.

Another t_2x 3N matrix T_2 for the duplicate trajectory, which will be projected onto the first, is also constructed, where t_2 denotes the simulation time and 3N again denotes the fluctuation of x, y, z coordinates of the alpha carbon of residues in the protein. The second matrix is projected onto the first one according to:

$$\hat{T}_{21} = T_2 V_k \Sigma_k^{-1} \text{ EQUATION 50}$$

Where, k is the span of the time vectors of the first trajectory, $\hat{T}_{21}$is the second trajectory projected onto the first one with size t_2 x 3N.

Information about the motions of the protein lies in its largest singular values and the dominant motions of the system are covered in the largest singular values. The outlier motions of the second trajectory, which are not covered in the first trajectory are obtained as

$$\hat{T}_{21,k} = T_2 V_{(t_1-k)} \Sigma^{-1}_{(t_1-k)} \qquad \text{EQUATION 51}$$

To convert the updated trajectory back to real space, $\hat{T}_{21,k}$ is multiplied with the left singular vectors and k largest singular values of the first trajectory, to eliminate the variance between two trajectories derived from the randomness. After the conversion, the residue distance distributions of projected trajectories are compared and analyzed.

The simulation trajectories of the protein MEK1 resulting from three separate runs are used as an example.

The structures of MEK1 in the PDBdatabase had missing residues. We initially performed homology modelling by using Modeller tool to model missing parts. Structures with PDB ID: 1S9J, 3EQD, 4MNE and 3WIG were used as a template for MEK1 in complex with ATP. The structure was prepared using VMD. The CHARMM27 force field was used in all three MD simulations. Minimization was performed with explicit solvent using TIP3P water model and the energy of the system was minimized for 10 ns, followed by an annealing and 40 ps of equilibration runs. After the system is equilibrated, three independent 1 μs MD runs under constant temperature at 310 K and constant pressure of 1 bar was performed. The trajectories are translated and rotated to a reference frame, initial frame, and the coordinates of alpha carbons were saved. The mean position vector of each alpha carbon is extracted and the fluctuation vector from the mean is obtained. SVD is performed on the reference mean fluctuation matrices and 6 largest singular values were considered. Then the trajectories to be projected are multiplied with left singular vectors and 6, 4 and the largest singular values of the reference trajectory and converted to real space values. RMSF values and inter-residue distances in projected and un-projected trajectories are compared to evaluate the fitness of the indexing.

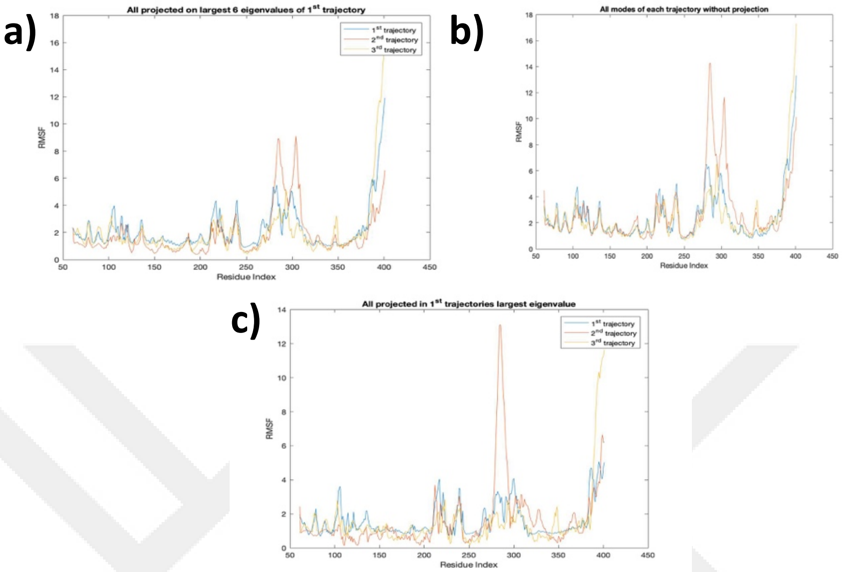


FIGURE 49. RMSF VALUES OF ALPHA CARBONS CALCULATED FROM A) TRAJECTORIES PROJECTED ON LARGEST 6 EIGENVALUE OF TRAJECTORY 1, B) RMSF PLOTS FOR ALL MODES OF EACH TRAJECTORY WITHOUT PROJECTION, C) TRAJECTORIES PROJECTED ON THE LARGEST EIGENVALUE OF TRAJECTORY 1

Second and third trajectories are projected to first 6 modes of the first trajectory and the root mean square fluctuation of residues were compared Figure 49a and RMSF values for all modes of each trajectory are plotted in Figure 49b., and all trajectories were projected on the first largest eigenvalue of the first trajectory on Figure 49c. The RMSF values become more similar when projected on the largest eigenvalue, which contributes the most on the important motions of the protein. When panel a and c are compared with panel b on Figure 49, it can be seen that RMSF values become similar when the noise in the trajectories are removed and the largest modes are considered.

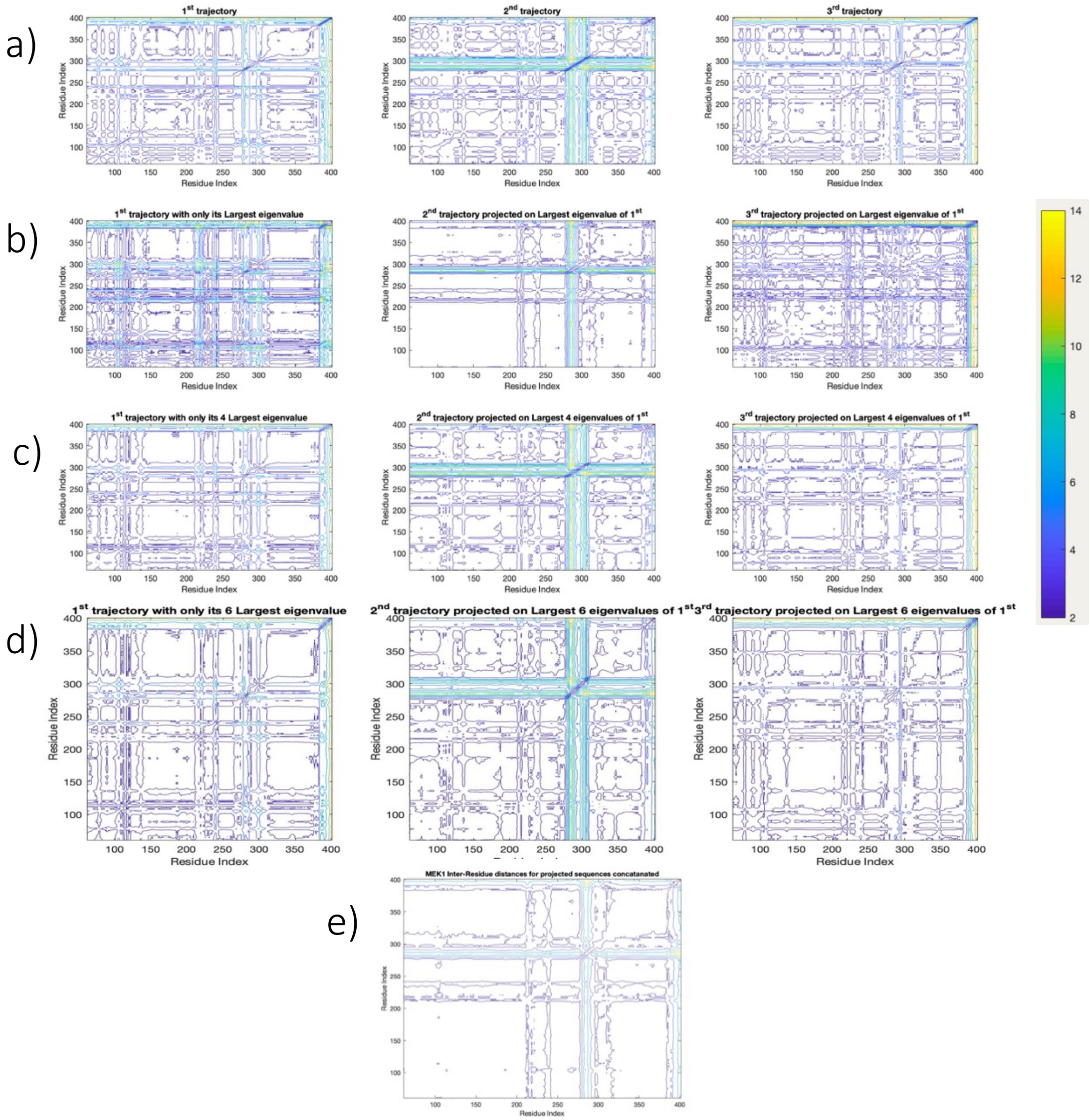


FIGURE 50. INTER RESIDUE DISTANCES OF A) UN-PROJECTED FIRST, SECOND AND THIRD TRAJECTORY, B) EACH TRAJECTORY IS PROJECTED ON THE LARGEST EIGENVALUE OF FIRST TRAJECTORY, C) EACH TRAJECTORY IS PROJECTED ON 4 LARGEST EIGEN-VALUE OF FIRST TRAJECTORY, D) EACH TRAJECTORY IS PROJECTED ON THE 6 LARGEST EIGEN-VALUE OF THE FIRST TRAJECTORY AND E) INTER-RESIDUE DISTANCES IN THE CONCATENATED TRAJECTORIES THAT ARE PROJECTED ON THE LARGEST EIGENVALUE OF THE FIRST TRAJECTORY.

When we compare the un-projected, Figure 50a., inter-residue distance plots of individuals with the plots in Figure 50b-d. we can see that the results that do not fit for all three cases, denoted as the noise in the simulations, are stripped and the patterns in residue distance plots became similar to one another. We can get the most important events in the simulations when the projected trajectories are concatenated, and inter-residue distances are calculated.

4.5. DISCUSSION

Molecular dynamics is a virtual molecular microscope which gives us an atomistic detail of the dynamical properties of atomistic systems. The MD results converge to those of experimental ones with the improved computational software and techniques. However, there are several limitations for this computational method; 1) Long scale simulations are required to sample conformations sufficiently and 2) the physical and chemical forces defined in force fields are not sufficient and they may yield noisy, biologically meaningless results. A great amount of information can be gathered by MD simulations for native state dynamics but for processes that require slow dynamics such as protein folding or non-folding, conventional MD simulations are insufficient.

The noise and accuracy problems are derived from the approximations that are built into force fields used for parameterization of molecules. The force fields used to prepare the simulations are empirical and therefore another optimization approach for MD methods are to improve the quality of force field parameters.

The methods such as principal component analysis or methods that rely on eigen value decomposition are utilized to reveal the most important properties of a system. In this study we performed three independent MD simulations for MEK1 protein under same environmental conditions and by employing the latent semantic analysis method we determined the essential dynamics in each trajectory to get rid of the noise derived from the empirical force fields and the accuracy problems derived from the rounding significant figures from the numerical integration during the velocity and force calculation steps in MD.

As it can be seen from Figure 49Figure 50, when major contributing modes are considered in calculation, the physical measures in three independent simulation trajectories become similar. This method can be useful for elongating short MD trajectories and enhance conformation space by running multiple short MD simulations.

4.6 CONCLUSION

The most important motions and properties of the simulations are extracted by only analyzing their major contributing modes and the simulation time is elongated by projecting three independent simulations onto same modal space ground and converting them back to real space with the same eigenvalue matrix. Our results show that when the noise in the simulations are removed, physically and biologically meaningful motions can be detected, and simulations can be accurately compared. Some dynamic differences may be small in magnitude but all noise in longer MD simulations will yield more extensive dynamic changes that are not meaningful and accurate. Thus, by employing this method, MD simulations can be analyzed more accurately, and the results of shorter simulations will converge to experimental results.

CHAPTER 5.

UNDERSTANDING THE ACTIVATION MECHANISM OF MEK1 AND THE EFFECTS OF LIGAND BINDING, MUTATION AND PHOSPHORYLATION ON ITS CONFORMATIONAL DYNAMICS

5.1 OVERVIEW

Ras/Raf/MEK/Erk cascade delivers signals from cell surface to transcription factors and play a role in gene expression. This signal transduction may induce or prevent apoptosis or result in cell cycle progression. This pathway is fairly complex because additional pathways may interact with kinases involved in Ras/Raf/MEK/Erk cascade and alter their targets' downstream phosphorylation status. Overactivation of this pathway is commonly seen in most of the cancer types.(F. Chang et al., 2003) There are inhibitors developed to inhibit kinases in this pathway. Ras inhibitors are challenging to identify, more attention is being focused on the study of MEK inhibitors since their solely downstream target is ERK.(Roskoski Jr, 2012) There are several MEK inhibitors have been developed; Selumetinib (AZD6244) is an allosteric, oral, potent, ATP-independent and highly selective MEK 1/2 inhibitor which has shown activity against several advanced adult cancers. (Galanina et al., 2015; Narita et al., 2014) The activation of MEK1/2 is implicated in driving cancer growth and progression. The use of concurrent inhibitors improved the responses compared with monotherapy approach. However, patients still develop resistance to medication. P124L MEK1 mutation was determined in a patient treated with Selumetinib who undergone relapse. These cells were more resistant to Selumetinib and stable disease progression was observed upon further Selumetinib treatment. The activation of MEK proteins occur upon phosphorylation by all three family members of Raf proteins (B-Raf, Raf-1 and A-Raf). Abnormal levels of MEK1 expression is observed on many types of human cancer. Phosphorylation induces a conformational change, a population shift of fluctuations and it plays a pivotal role in regulation of protein activity, but the biophysical characterization of phosphorylation is still not clearly understood. It is observed from the 1 microsecond molecular dynamics trajectory that dual phosphorylation of MEK1 by Raf induces a shift on fluctuations of residues and locks the tail of the protein in a position which facilitates Erk binding, and for the second part the activation loop is not constrained by Selumetinib when there is mutation in the system and phosphorylation route by Raf is not disrupted, the communication between important regions increases and the ligand fails to inhibit MEK1 activity when P124L mutation is present. . The activation mechanism of MEK is not clearly understood. In this study the structural and the dynamic changes upon phosphorylation, ligand binding and mutation of MEK will be unraveled by using all atom molecular dynamics simulations and the change of communication patterns upon these events will be understood by using the entropy transfer concept introduced in the previous chapters.

5.2 INTRODUCTION

The Ras pathway is one of the most widely studied pathways in humans due to its role in various cancers. The sequence of proteins along the mainstream of this path is Ras, Raf, MEK and Erk. Several cell-surface molecules activate Ras, then Ras activates Raf. MEK is the dominant substrate of Raf. MEK is a mitogen activated protein kinase kinase, which facilitates phosphoryl transfer from ATP to Erk and leads Erk activation. Activated Erk initiates other cellular processes in the cell; cell cycle progression, proliferation, differentiation, cell migration, metabolism and proliferation.(Roskoski Jr, 2012) Protein activation and phosphorylation is a process that consists of a predetermined sequence of fast events. This set of events often starts with the interaction of the protein with upstream protein and proceeds with a succession of correlated conformational rearrangements. The repertoire of events that prepare the protein for phosphorylation is extremely simple, resulting essentially from two types, both mechanical in nature: (i) changes in distance between residue pairs, i.e., relative translation, and (ii) correlated reorganization of fluctuations of residue pair distances. The sequence of events is built into the correlated structure of the protein and is robust, i.e., the same set of events is realized, except for mutations along the pathway that may disrupt the correlations. In this study, based on detailed molecular dynamics simulations, we present an atomistic view of the sequence of events before, during and after the phosphorylation of MEK1. Understanding the phosphorylation of MEK is particularly important because it is central to the

Ras/Raf/MEK/Erk pathway. (Roskoski Jr, 2012) The importance of this pathway lies in the fact that several mutations in Ras renders it active, sending information continuously along the pathway. The significance of MEK in this pathway is essentially for the interest of drug design because it has only a single downstream partner Erk while the others along the pathway interact with a multitude of proteins. MEK proteins are rarely mutated and they have a pocket structure which can be directly targeted by preventing side effects of inhibiting other proteins and hence they are suitable candidates for inhibition.

MEK1 and MEK2 are related dual-specificity protein kinases that participate in Ras/Raf/MEK/Erk pathway which mediate the phosphorylation of tyrosine and then threonine in Erk1/Erk2. (L. Li et al., 2016) They have a narrow substrate specificity. Like all protein kinases, they have a small N-terminal lobe and a large C-terminal lobe. Raf family proteins activate MEK1 and MEK2 by mediating the phosphorylation on serines 218 and 222 for MEK1 and 222 and 226 for MEK2. (Yan & Templeton, 1994; Zheng & Guan, 1994) MEK1 and 2 are %80 identical and %90 similar, the differences between them are mainly in the N terminal region. These differences may contribute to the specificity and variations of the interactions. It was reported that several Ras-Raf1 and Ras-BRaf complexes prefer to bind MEK1, not MEK2 and despite their high degree of homology MEK2 knockout mouse develop normally, whereas MEK1 knockout mouse died due to placental defects in embryonic phase. (Giroux et al., 1999)

MEK1 is composed of a small N and a large C lobe, the topological map of which are shown in Figure 51.The N and C lobes are presented in the left and right panels of Figure 51, respectively. The N lobe, residues between 1 and 144 contains the Erk binding region, residues from 1-32, and an α-C helix between Lys104 and Glu120 which plays a crucial role in activation. (Ordan et al., 2018) The large C lobe is between residues 145 and 380, contains a proline-rich sequence, residues between 262–326, required for Raf binding, the activation segment, residues between Asp208 and Glu233, where two serine residues (Ser218 and Ser222) get phosphorylated, a catalytic loop which participates in phosphoryl transfer from ATP to Erk and DFG (Asp208, Phe209, Gly210) motif which also plays a crucial role in protein activation with the conformation of the aspartate residue present in the motif.(Zheng & Guan, 1994)

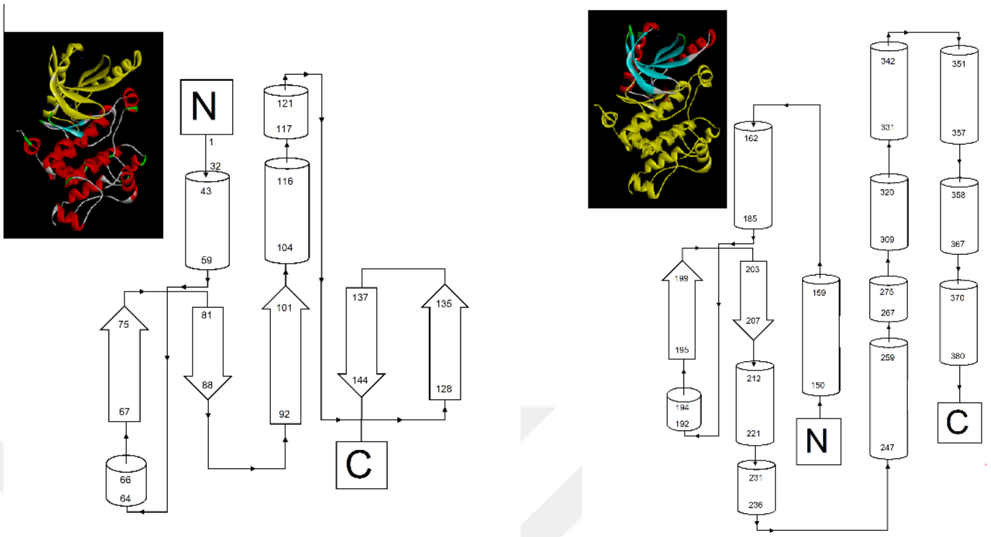


FIGURE 51. TOPOLOGY MAP OF MEK1 A) N LOBE AND B) C LOBE ARE SHOWN

There are two important sub-structures reported for kinases so far in the literature. These are the α-C helix and the DFG motif. (Roskoski Jr, 2012; Vijayan et al., 2014) For MEK1, a conserved glutamic acid, 114th residue, in the helix forms a salt bridge with a conserved lysine, 97th residue, when the protein is active. This interaction leads to the stabilization of the system and changes the interactions within the protein as follows: The Aspartic acid in the DFG motif flips towards the core of the protein upon activation and gets closer to the ATP binding region, small glycine-rich segment (P-loop) located between residues 74-82. Dual phosphorylation at Ser218 and Ser222 is required for MEK1 activation. The biophysical importance of these events, the changes they cause and why they occur have not been reported in the literature. The aim of this study is to investigate the causes and effects of these changes by integrating molecular dynamics simulations to gain insight into the atomistic details of these events.

Several MEK inhibitors have been developed; Selumetinib is an allosteric, oral, potent, ATP-independent and highly selective MEK 1/2 inhibitor which has shown activity against several advanced adult cancers.(Cheng & Tian, 2017) The activation of MEK1/2 is implicated in driving cancer growth and progression. The use of concurrent inhibitors improved the responses compared with monotherapy approach. (Zhao & Adjei, 2014) However, patients still develop resistance to medication. P124L MEK1 mutation was determined in a patient treated with Selumetinib who undergone relapse, these cells were more resistant to Selumetinib and stable disease progression was observed upon further Selumetinib treatment.(Kim & Patel, 2014; Savoia, Fava, Casoni, & Cremona, 2019; Spreafico et al., 2013)

Allostery is a property of proteins that result from long-range correlations where information is carried from one part to another asymmetrically by entropy transfer. To understand and investigate the underlying allosteric interaction mechanism in terms of entropy transfer of wild type and mutant MEK1 with Selumetinib, several Molecular Dynamics (MD) simulations were conducted for wild type (WT) MEK1 with Selumetinib and P124L mutant MEK1 with Selumetinib and information transfer networks were compared by using a fast and approximate method which combines Schreiber's entropy transfer concept with the Gaussian Network Model of proteins. Using this method, residues that are manipulated by point mutation are determined and the change of communication patterns is shown and the effect of mutation on Selumetinib inhibition is studied. The change of information transfer between two residues is related to the causality of the relationship between their fluctuations, and this should be of great importance for allosteric drug design. The structural changes on MEK1 upon ligand binding, phosphorylation and mutation are detected by observing all atom MD simulations for each scenario.

5.3 METHODOLOGY FOLLOWED IN UNDERSTANDING THE STRUCTURAL AND DYNAMIC CHANGES OF MEK1 UPON PHOSPHORYLATION, LIGAND BINDING AND MUTATION

The aim of this study is to gain insight into atomistic details of mutation, phosphorylation and ligand binding events on MEK1 by integrating all atom molecular dynamics simulations. MEK1 activation is initiated by Raf binding and further phosphorylation. To understand the whole sequence of events from Raf binding transition state to activation by phosphate transfer from ATP to MEK1, Raf- MEK1, MEK1 and dual phosphorylated (S218 and S222) MEK1 trajectories are generated. NAMD computer software is used to conduct molecular dynamics simulations. The force field parameters used in the simulations were taken from CHARMM 27 parameter files. SP2 patch for phosphorylated serines was used to construct phosphorylated MEK1. None of the MEK1 structures deposited in the Protein Data Bank contained the active MEK1 structure and several residues were missing, therefore we applied the following strategy to build the active structure. To fill missing residues, homology modelling was performed by using Modeller tool.(Eswar et al., 2006) Structures with PDB ID: 1S9J, 3EQD, 4MNE and 3WIG were used as a template for MEK1. (Fischmann et al., 2009; J. R. Haling et al., 2014; Lito et al., 2014; Ohren et al., 2004) Mg-ATP bound inactive MEK1 structure is generated, but to generate phosphorylated MEK1, 218 and 222 serine residues were phosphorylated by using SP2 patch of CHARMM 27. With the prior knowledge on required active MEK1 structural features, DFG-in and αC helix-in conformations, we generated the active MEK1. Inactive structures had DFG-in conformation but they lacked αC helix-in conformation and an important salt bridge between Lys97 and Glu114 formed upon rotation of αC helix. To be able to detect this salt bridge with molecular dynamics, energy must be provided to the system to naturally rotate the helix. Since energy can't be provided during a standard MD run, SMD was performed to rotate this helix and constraints were applied to keep the salt bridge intact. We manually rotated αC helix by performing a short steered molecular dynamics simulation. After the helix is rotated around its own axis, Lys97 and Glu114 formed a salt bridge and stabilized the structure. Activated structure is minimized further for removing the constraints and steric hinderances on the system.

For MEK1-Selumetinib simulations, structure of Selumetinib is downloaded from ZINC database and docked near the ATP binding site of MEK1 according to the preleminary information on Selumetinib-MEK1 interactions avaliable. (Caroline M Emery et al., 2009)Ligand is parameterized by using SwissParam. (Zoete, Cuendet, Grosdidier, & Michielin, 2011)

The mutated MEK1 structure is mutated by using the mutator plugin of VMD. (Humphrey et al., 1996)

5.3.1 MOLECULAR DYNAMICS SIMULATIONS

All atom MD simulations were performed using NAMD 2.10 version. CHARMM 27 force field parameters were used for protein and MgATP. Proteins are immersed in a TIP3P water box, counter ions, Na+ and Cl-, were added to neutralize the system. 50 ps of energy minimization and 500 ps of equilibration were applied to both systems. A standard protocol was performed for 1 ms MD simulation for each system at 310 K and 1 atm. Time step of simulations were 2 fs and periodic boundary conditions were applied in an isobaric-isothermal NPT ensemble. Temperature and pressure were controlled by Langevin thermostat and Langevin piston barostat respectively. 1-4 scaling is applied to van der Waals interactions with 12 Å cutoff distance.

5.4 RAF BINDING AND MEK1 ACTIVATION

5.4.1 RESULTS

5.4.1.1 RAF-MEK1 BINDING

The αC-helix marks the beginning site of MEK1 activation. Binding of Raf to MEK1 leads, due to Raf's interaction with Lys104, to the rotation and distortion of the αC-helix. A putative structure for the Raf-MEK1 complex, obtained from Protein Data Bank and missing residues for both proteins were completed by using Modeller homology modelling. Specifically, the i, i+4 helical hydrogen bond between Asn109 and Arg113 of MEK1, which is observed during %96 of the trajectory for the native state, is replaced by a non-helical hydrogen bond between Asn109 and Pro105, which is observed during %97 of the trajectory for the Raf bound MEK1 in transition state from inactive to active. The distortion of the helix brings Lys97 and Glu114 into close proximity. This closeness results in the formation of a salt bridge between Lys97 and Glu114. This salt bridge stabilizes the αC-helix in a new conformation that is suitable for the subsequent phosphorylation (see below). The presence of the salt bridge is a universal feature of kinase activation (its mechanical consequences will be explained below) and is known to stabilize the structure of most of the kinases, for example EphA4 protein tyrosine kinase. (Xu et al., 2013)

5.4.1.2 EVENTS FOLLOWING RAF-MEK1 BINDING

Rotation and distortion of the αC helix and formation of the Lys97-Glu114 salt bridge results in the breaking of several hydrogen bonds between the αC helix and the activation segment, including Leu118-Val211, Val117-Val211, Glu114-Gln214, Glu114-Gly213, Glu114-Ser212, and subsequent moving away of the activation segment from the αC helix.

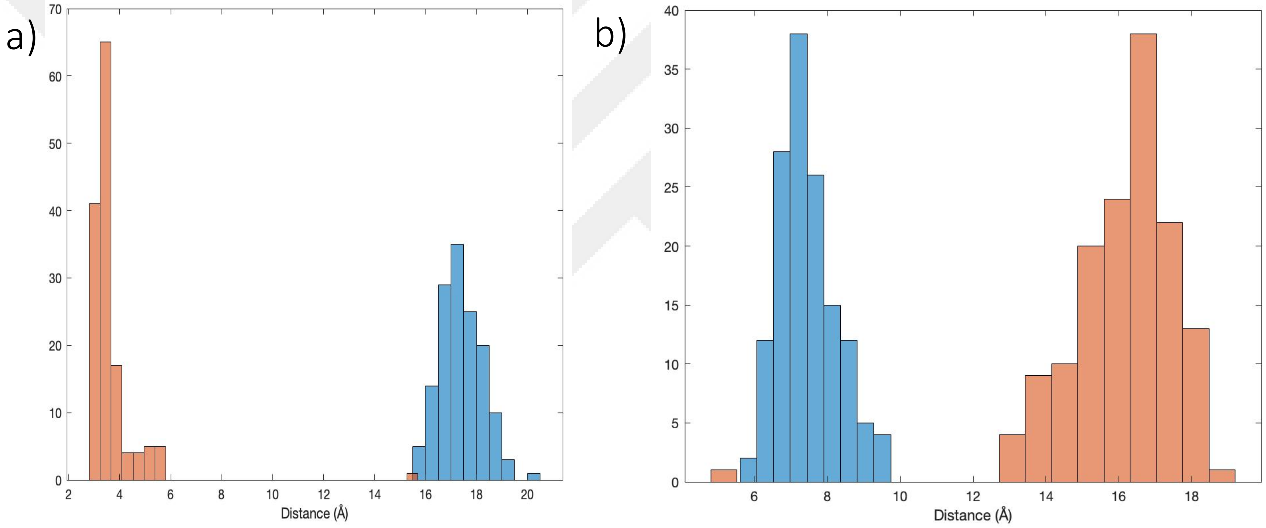


Figure 52. The histogram plot of the distance between a) Lys97-N and Glu114-O, which participate in the salt bridge and b) the histogram plot of the distance between residue 78 in the P-loop and residue 223 in the activation segment in Mek1 (blue) and Mek1 in transition state (orange).

A glycine-rich (GxGxxG) ATP-phosphate binding loop, also called the P-loop, consists of residues 74 to 82. (Taylor & Kornev, 2011) Asn78 in the P-loop and Lys 97 are connected with a hydrogen bond. Formation of a salt bridge between Lys97 and Glu114 influences the conformation of the P-loop. Conformational shift of the P-loop is required for phosphoryl transfer to Ser218 and Ser222 and it is a known issue in kinases. (Wu & Park, 2015) Translation of the activation loop upon Raf binding directly affects the distance between itself and the P-loop. This separation is clearly seen in Figure 52b. where the distance between Asn78 located on the P-loop and Phe223 located on the activation segment in the inactive state is compared with the corresponding distance in the phosphorylated state. The separation of the P-loop and the activation segment creates an opening between them, which allows Mg-ATP intake into the cavity.

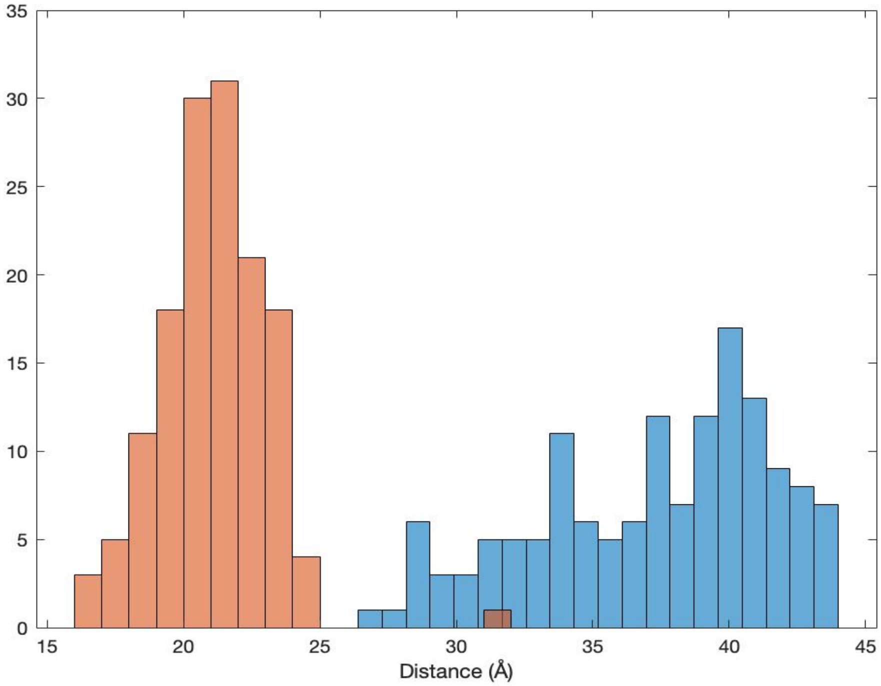


Figure 53. The histogram plot of the distance between Thr386 on tail and Pro266 on proline rich insert in Mek1 (blue) and Mek1 in transition state (orange).

The tail of MEK1, which extends between residues Gly380-Val393 exhibits large conformational changes upon activation. It forms new hydrogen bonds with residues 262-326, i.e., the proline-rich insert. This event causes the tail to stabilize at a new conformation. Mean fluctuations of the tail structure is 7.1 in inactive and 6.2 in phosphorylated MEK1. It is seen in the simulations that the tail sweeps a larger area and blocks the ERK binding region which is found in N lobe. The Erk binding region which is called the docking domain (D-domain) is located in the N lobe of MEK1. The tail blocks the entrance of the D-domain in the inactive form of MEK1 but due to the Raf induced interactions of the tail, it moves away from N-lobe, forms hydrogen bonds with residues in glycine-rich insert and enables Erk binding.

5.4.1.3 PHOSPHORYLATION OF MEK1 FOLLOWING ITS TRANSITION STATE

Upon the increase in the distance between Asn78-Ser223, ATP binding cavity opens up, the volume increases from 429.05 to 580.5 Å^3, allowing ATP intake. Phosphate binding affinity of the P-loop residues increase upon Raf binding, this enhances ATP activity and localization and results in phosphorylation of Ser218 and Ser222. The role of the phosphorylated residues in the active MEK1 is to sustain the changes in motions and conformation induced by Raf. Raf binding induces a scissor-like motion in MEK and negatively charged phosphate groups on Ser218 and Ser222 helps retaining this motion by repelling the negative and attracting the positively charged neighboring residues. Ser218 and Ser222 repel the negatively charged residues, Glu233 and Glu312 in the proline-rich insert and Glu102 on the αC-Helix. As a consequence, fluctuations of proline rich insert increase in the phosphorylated MEK and proximity of Glu114 and Lys97 is reinforced by repelling Asp102 which is found on the same αC-Helix as Glu114. This repulsion sustains the tail-proline rich insert interaction. Phosphorylated residues keep the reversed motions of MEK and helps Erk binding domain to be freely accessible by Erk.

ROLE OF ASP208 IN PHOSPHORYLATION: In the unphosphorylated state, Asp208 is deprotonated where its OD1 atom coordinates with Mg$^{2.}$ In this state Asp208, Mg^{2+} coordinates with only α and β-PO_4. Following phosphorylation, the conformation of Asp208 changes. Asp208 interacts with the catalytic loop and Lys97 in inactive state and the interaction of Asp190, in the catalytic loop, with Asp208 creates a negative polar pocket for magnesium ion, but upon conformational change it starts to make hydrogen bonds with Lys97 and Glu114 and its neighboring residue Phe209 moves out, as a result the hydrogen bond between Val211 and Phe209 breaks. When Erk binds Mek it interacts with Asp208 and changes its protonation state. It can't coordinate with Mg^{2+} in the protonated state due to the electrostatic repulsion of positive ends. The position of the OD1 atom of Asp208, which is involved in Mg^{+} coordination. In the active state, Mg^{2+} coordinates with all three phosphate groups of ATP. The activation of the nucleophile substrate hydroxyl group of threonine or tyrosine residues of Erk occurs via deprotonation by Asp208, rendering it more nucleophilic.

5.4.1.4 PHOSPHORYL TRANSFER TO ERK AND ERK BINDING TO PHOSPHORYLATED MEK1

Reversed motions enable easier Erk binding. With the help of the newly formed hydrogen bond between Glu114 and Asp208, Asp208 is located closer to the Mg-ATP binding region. The protonation state of Asp208 in the ATP binding region plays a crucial role in phosphoryl transfer. When ERK1 binds to MEK, Asp208 deprotonates Thr202 and Tyr204 residues of ERK1. This causes the negatively charged Asp208 to get protonated. Asp208 is responsible of coupling with Mg^{2+} ion and lead coordination of Mg^{2+} with phosphate groups of ATP. Asp208 can't coordinate with Mg^{2+} in the protonated state due to the electrostatic repulsion of positive ends on protonated Asp and positively charged Mg ion. Phosphate group coordination of Mg^{2+} varies with varying protonation states of Asp208. Adenine ring of ATP is anchored with the help of P-loop interactions. Sulphur atom of Met143 in hinge region forms a proton addition complex with C8 of ATP through a water molecule between them. This proton, H^{+}, is transferred from C8 to α-PO_4 directly and from α-PO_4 to β-PO_4 with the help of Lys192 in the catalytic loop. Protonation of β-PO_4 ensures that γ-PO_4 is a better leaving group. Under Asp208 control, Mg^{2+} coordinates with only α and β-PO_4, but when Asp208 is protonated Mg^{2+} coordinates with all three phosphate groups of ATP. The change of protonation state of Asp208 causes migration of Mg^{2+} to be α and β-PO_4 coordinated state. At the end of phosphoryl transfer, H^{+} transferred to β-PO_4 subsequently returns to C8 and electrons of adenine ring returns to its stable state. The mean distance of the side chain oxygen atom of Asp208 and Mg^{2+} is 3.02 Å in inactive MEK and 7.24 Å in active MEK. Asp208 couples with Mg^{2+} and prevents Mg^{2+}-γ-PO_4

coordination in inactive state. RMSD values of ATP backbone atoms show that fluctuations of γ-PO_4 group is higher in active state, which is now a better leaving group, it tends to fluctuate more in active state.

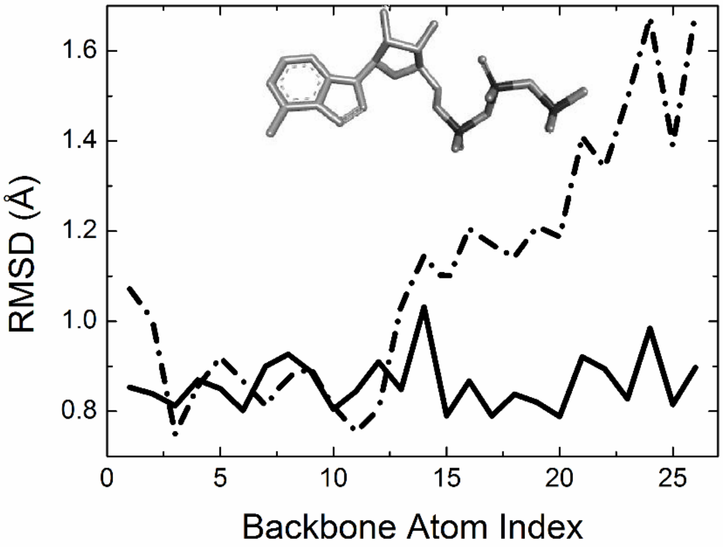


Figure 54. RMSD values of ATP backbone atoms in inactive(solid) and active(dashed) MEK1

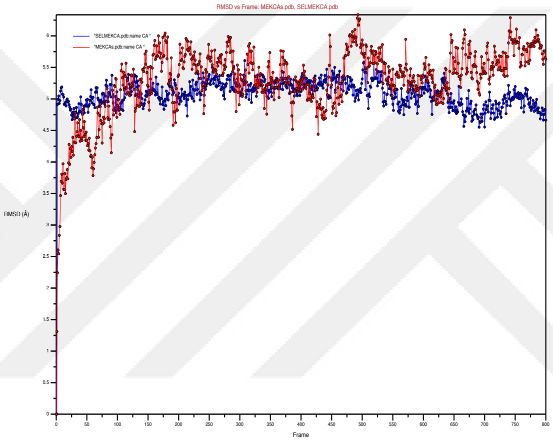
5.5 SELUMETINIB AND MEK1 INTERACTION

Selumetinib is an oral drug that is used for the treatment of several types of cancer. Selumetinib selectively blocks the activity of MEK protein activity and Erk1/2 phosphorylation, independent of the presence of the ATP molecule, to reduce the activity of Ras/Raf/MEK/Erk pathway. The narrow substrate specificity if MEK1/2 renders them as ideal therapeutic targets. MEK 1 and 2 are both consist of a kinase domain, a proline rich insert, a negative regulatory region, an Erk docking region and a nuclear export sequence. Both structures have an ATP binding site with a unique pocket adjacent to it. Since the ATP binding affinity of kinases are high, a non-competitive kinase inhibition is the best therapeutic therapy option. (Bhullar et al., 2018) There are several MEK inhibitors developed, two of them gained FDA approval, cobimetinib and trametinib. MEK inhibitors showed superior activity when combined with other therapies in treating lung cancer, melanoma and colorectal cancer. (Dombi et al., 2016)The reasons for MEK inhibitors to fail as a single agent are the signaling of Ras through other alternative effector pathways, crosstalk between multiple pathways and lack of autoregulation of negative feedback from Erk to Raf to compensate MEK inhibition. (Y. Li, Dong, & Cui, 2019) In the negative Erk-Raf feedback loop, activated Erk reduces Raf activity and downregulate the activity of its effectors. MEK inhibition also inhibits this feedback loop and frees Raf from Erk inhibition and enhances signaling through other pathways. (Saei & Eichhorn, 2019) A MEK inhibitor, AZD6244, succeeded as a single therapeutic agent in a study with BRAF V600E. BRAF V600E lacked the ability to interact with alternative effector pathways and dysregulation of Erk-Raf negative feedback loop did not caused further enhanced Raf activity.(Friday et al., 2008; Galanina et al., 2015) This study validated that concurrent usage of MEK and Raf inhibitors is the best strategy in cancer treatment. Selumetinib is another MEK inhibitor which showed promising activity in biliary cancer and recurrent, chemoresistant low-grade serous carcinoma of the ovary when used as a single agent. Also, dacarbazine combined with selumetinib prolonged progression free survival in KRas mutated non-small-cell lung cancer (NSCLC) patients, but no overall survival benefit was observed. Selumetinib is a second generation, ATP independent MEK1/2 inhibitor which binds to the allosteric binding pocket and locks MEK in an inactive conformation and disrupting the interactions required for Erk binding and catalysis. Selumetinib has an IC_{50} of 14 nM against MEK1 and this drug was found to be more effective in treating cancers involving BRaf and Ras mutations.(Ciombor & Bekaii-Saab, 2015) Aim of this study is to understand selumetinib binding mechanism and effects of this ligand to MEK1 dynamics and compare the binding mode of this ligand to inactive and active structures. Aim of this study is to understand selumetinib binding mechanism and effects of this ligand to MEK1 dynamics and compare the binding mode of this ligand to inactive and active structures.

5.5.1 RESULTS

We have observed the interaction of Selumetinib(SEL) and MEK1 from the MD simulation trajectories. The residue fluctuations of MEK1 is measured in Selumetinib bound and unbound cases and compared. According to the comparisons, Selumetinib binding decreases the fluctuations of the protein and it directly binds to the structurally and functionally important residues; Lys97, Asn78 and a residue on αC helix Val211, Asn221, near the phosphorylation site and Val401 on the tail. The radius of gyration of MEK1-MgAtp and MEK1-MgATP-Selumetinib complexes are measured and the values are 21.65 Å and 20.2 Å respectively. Selumetinib binding slowed down the motions of the protein and makes it more compact. It stabilizes the tail and physically blocks Lys97-Glu114 salt bridge formation. It also binds to Lys97 which is prone to form a salt bridge with Glu114. By making these interactions, Selumetinib prevents the transition state formation and further phosphoryl transfer from ATP to Erk.

FIGURE 55. RMSD RESULTS OF MEK1-MGATP(RED) AND MEK1-MGATP-SEL(BLUE)



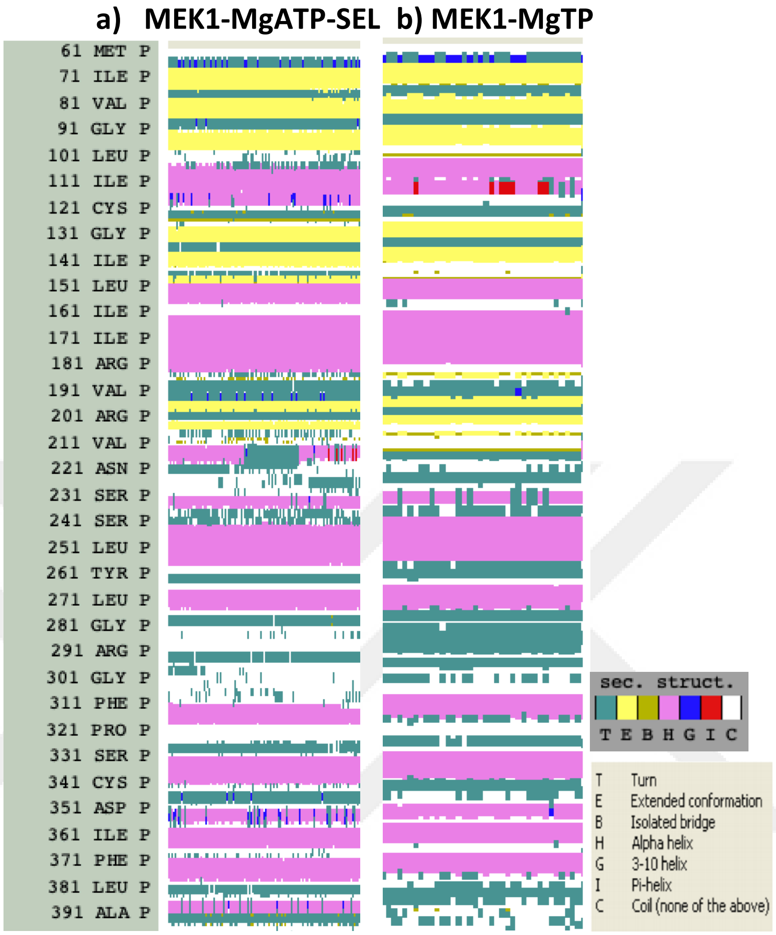

a) MEK1-MgATP-SEL b) MEK1-MgTP

Figure 56. Secondary structure of a) MEK1 with Selumetinib (SEL) and b) MEK1 without SEL, analyzed via VMD Timeline plugin from MD trajectories

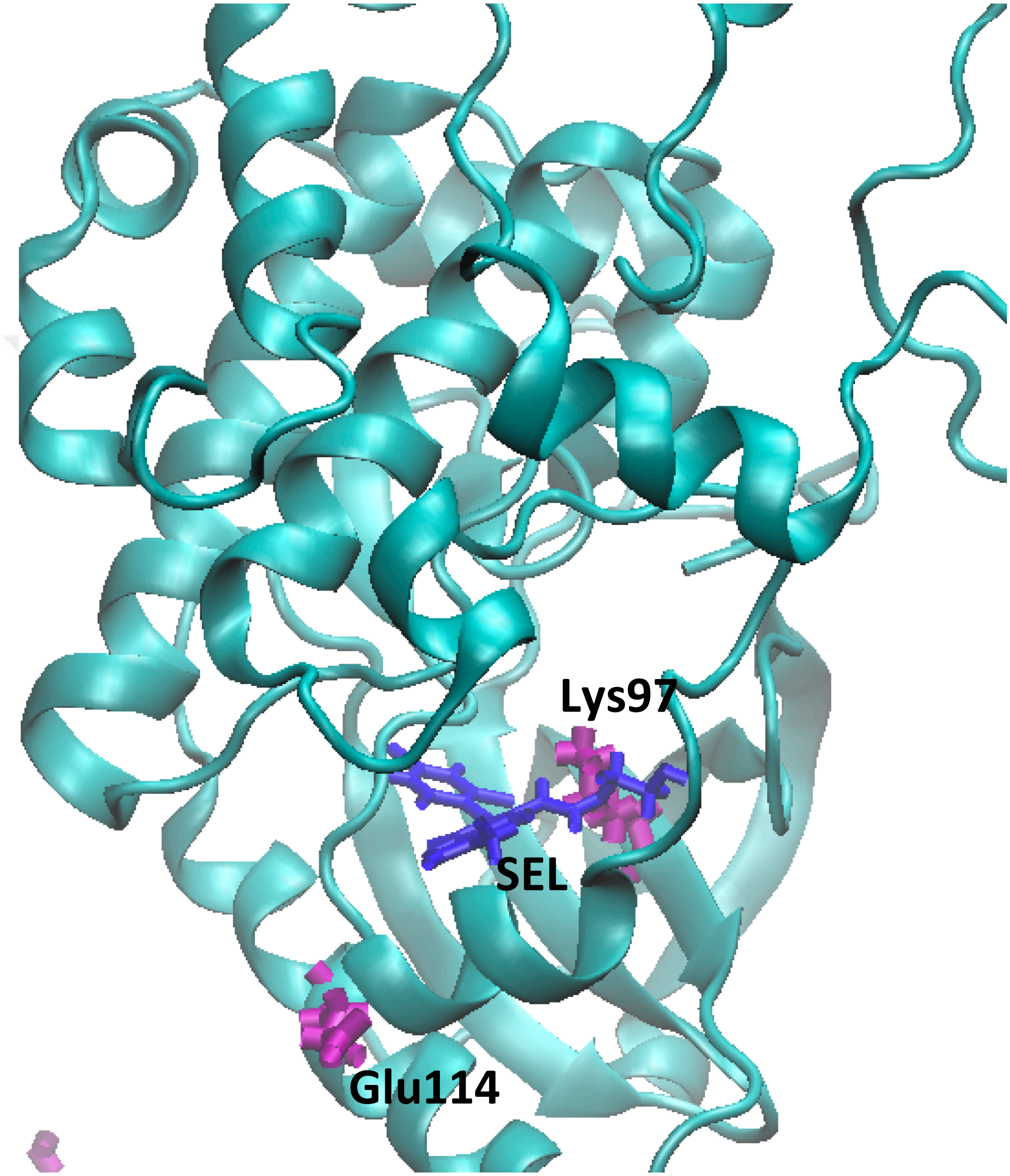


Figure 57. Cartoon representation of MEK1 in complex with SEL(blue), Lys97 and Glu114 is shown in stick representation(purple). Selumetinib prevents Lys97-Glu114 salt bridge formation by physically blocking their interaction.

5.6 EFFECT OF P124L MUTATION OF MEK1 ON MEK1-SELUMETINIB INTERACTION

The majority of the genetic alterations in the MAP kinase pathway harbor mutations in the BRAF oncogene. However, it has been recently discovered that mutations of MEK are related with chemotherapy-resistant tumors and developmental syndromes. (Bromberg-White, Andersen, & Duesbery, 2012) 75.38% of the reported cancer related MEK mutations are missense substitution and most of the mutations conferring to drug resistance are populated at the allosteric drug binding pocket of MEK. Inhibition sensitivity of RAF and MEK kinases increase in BRAF^{V600E} mutant melanomas. MEK1 P124L mutation emerged in a patient treated with Selumetinib(AZD6244). (C. M. Emery et al., 2009) MEK1 mutations are classified in 2 as the primary and the secondary mutations. Primary MEK1 resistance mutations cluster in the allosteric drug binding site and secondary mutations reside outside of the drug binding site.(C. M. Emery et al., 2009) P124 mutation is a secondary mutation on the αC-helix. MEK1^{P124L} sustained ERK phosphorylation under varying concentrations of Selumetinib, although wild type and dual phosphorylated MEK1 showed lower phosphorylated ERK concentrations at the lowest Selumetinib concentrations.

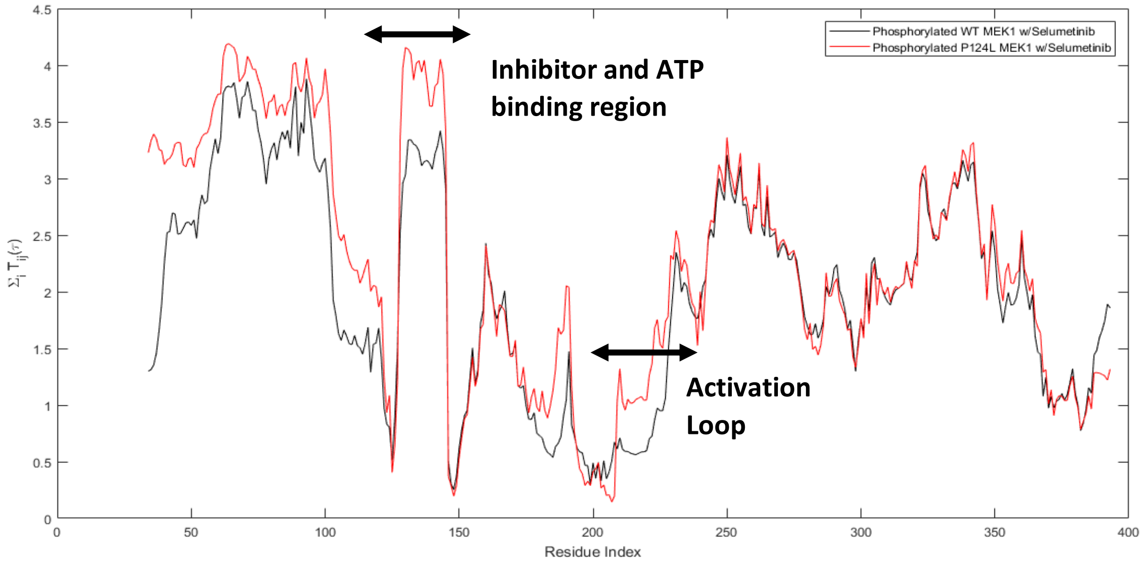
The working principal of MEK inhibitors is to lock the protein in an inactive conformation where its αC-helix becomes externally rotated. Patients with secondary MEK1 mutations showed cross-resistance to BRAF^{V600E} melanomas. Aim of this study is to determine the changes in the interaction patterns due to the mutation on 124th residue and decipher why Selumetinib fails to elicit meaningful tumor response in MEK1^{P124L} and BRAF^{V600E} melanomas.

Dual phosphorylated MEK1-MgATP-SEL and dual phosphorylated MEK1^{P124L}-MgATP-SEL MD simulations were conducted, and several snapshots were taken from the movies. The snapshots were analyzed by using dGNM method described in early chapters (Chapter 1).

5.6.1 RESULTS

Structure of selumetinib was downloaded from the ZINC database and docked into homology modeled MEK1 structure with the prior binding information given. The phosphorylation was applied on two serine residues on 218th and 222nd positions by using the SP2 patch of CHARMM27 force field. Mutation on 124th residue is applied by using VMD Mutator plugin. MD simulations were conducted with NAMD2.11 at 310 K, 1 bar by using CHARMM27 parameter files. Both systems are solvated in TIP3P water box and counter ions were added to neutralize the system. Energy of both systems were minimized, and systems were equilibrated for 50 ps and followed by 60 ns MD runs at constant temperature and pressure. Trajectories were aligned to the first frame to get rid of translational and rotational degrees of freedom. 6 snapshots were taken from the aligned system and coordinates were saved in PDB format. Information transfer within the proteins in both scenarios were calculated by using Eq. 26 in Chapter 1 and information going out from residue *i* to the rest of the protein is obtained by Eq. 28 in Chapter 1.

Figure 58. Information going out of MEK1 residues in dual phosphorylated WT MEK1-MgATP-SEL (black) and dual phosphorylated MEK1^{P124L}-MgATP-SEL(red).



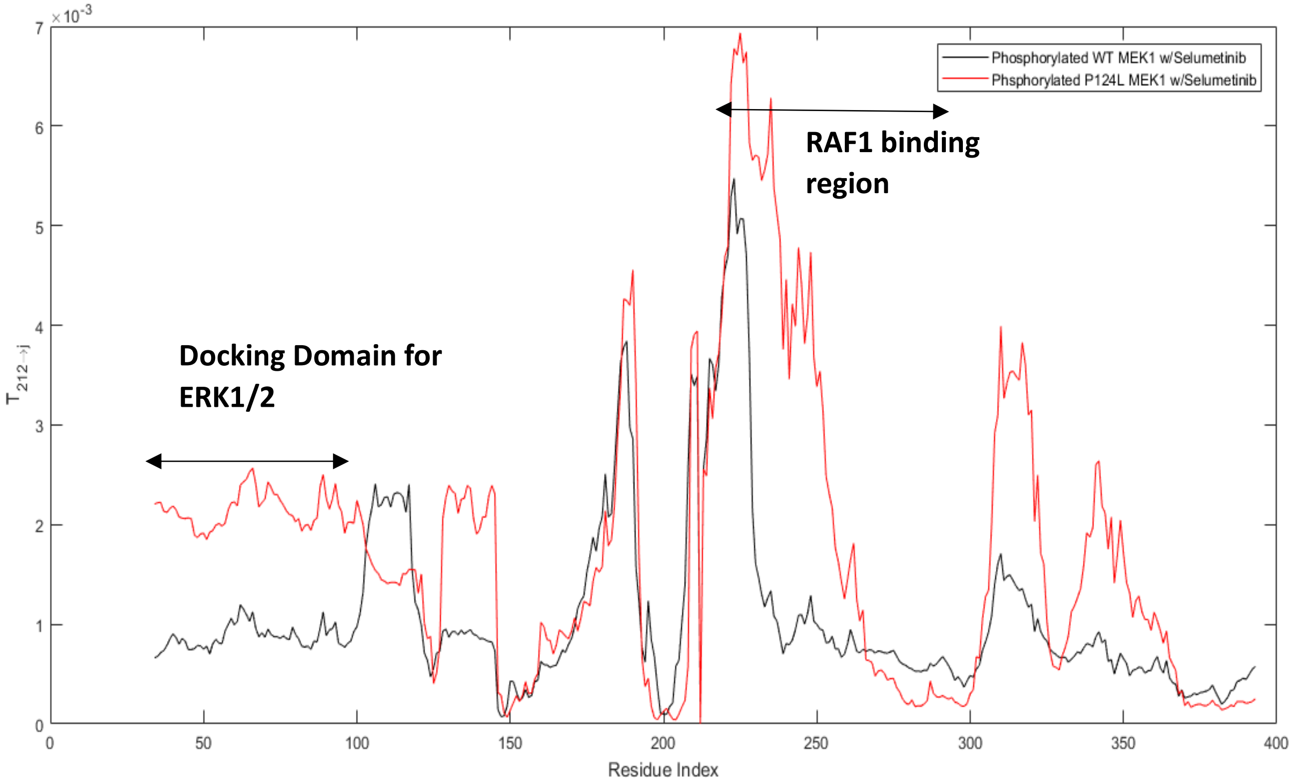


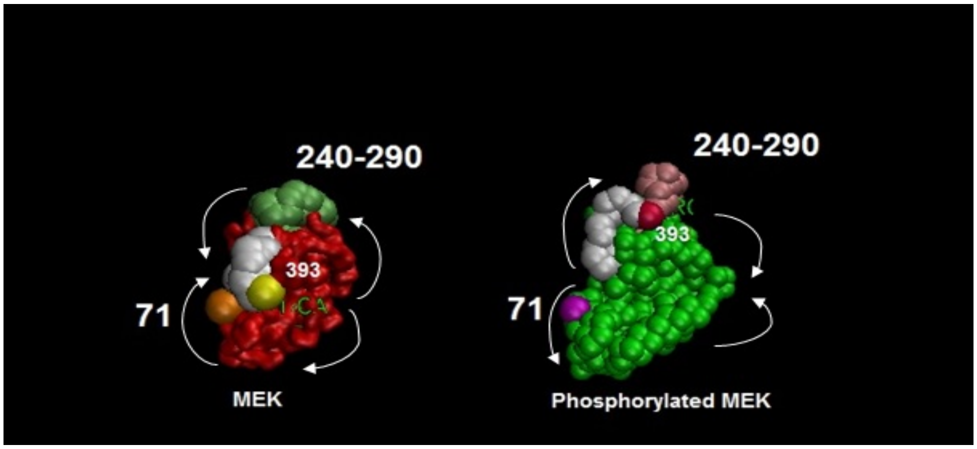
Figure 59. Information going out of residue number 212 in dual phosphorylated WT MEK1-MgATP-SEL (black) and dual phosphorylated MEK1^{P124L}-MgATP-SEL(red).

These results indicate that information transfer from residue 212 to the rest of the protein increases and pairwise communication of this residue changes upon mutation at a distant site. The information to the rest of the protein supplied from αC helix increases upon P124L mutation. In inactive MEK1 structures, this helix is found in an out-conformation and it only rotates into in-conformation in its active state. (Kooistra et al., 2015) It can be seen from Figure 59. that entropy transfer of residue number 212 with the docking domain for ERK1/2 and RAF-1 binding regions increases.

5.7 DISCUSSION

Increased ATP localization in active structure can be quantified by comparing the hydrogen bonds formed with ATP in inactive and active MEK. In active MEK, ATP forms hydrogen bonds with Asn78, Gly79 in P-loop, Lys97, Lys104, Arg189 and Lys192 in catalytic loop, Asp208 in DFG motif and Thr226 in activation segment. Mostly seen hydrogen bonds of ATP are formed with Lys97 and Lys192 which are seen in 30% and 65% of the simulation time. The adenine ring of ATP is anchored by the P-loop, Lys97 is known to couple phosphates of ATP with αC-helix. Upon increased coordination of ATP with MEK1, ATP fulfills its task and phosphorylates MEK1 from its two serine residues in the activation segment, Ser218 and Ser222. The ATP binding pocket volume increases in active structure from 429.05 to 580.5 Å^3. Binding pocket is deeper and more accessible in phosphorylated structure. Root mean square fluctuations of ATP showed an overall decreasing trend in active form although the binding pocket widened. This is related with the increased coordination and interaction between ATP and MEK1. Dual phosphorylation is required for MEK1 activation, single phosphorylated MEK1 are not considered as active, since they can't activate Erk but after the dual phosphorylation of MEK1, de-phosphorylation of its single serine residue leaves MEK1 in the active state. As a result of all simulations, the activation pattern of MEK1 is obtained and the mechanism is explained. There are several MEK1 inhibitors generically used in treatments of cancers. To prevent Erk1/2 activation by MEK1, the structural state of Erk binding region must be kept in the inactive form and important interactions in MEK1 active state must be distorted. Important feature that prevent Erk binding in inactive state is the tail, it can be kept closed on N lobe with the help of a ligand.(Jacob R Haling et al., 2014; Odendall et al., 2012)

FIGURE 60. THE MOVEMENT PATTERN OBSERVED FROM MD SIMULATIONS



FOR A) MEK1-MGATP AND B) DUAL PHOSPHORYLATED MEK1-MGATP.

Phosphorylation of MEK1 from its two serine residues on 218 and 222 is predicted to cause rotation of the N and C lobe with respect to each other and causing the catalytic cleft to close. This event causes the ATP binding residues to localize ATP more and catalyse γ phosphate of ATP onto Erk.

Currently designed MEK1 inhibitor Selumetinib binds MEK1 from Lys97, Asn78 and a residue on αC helix Val211, Asn221, near the phosphorylation site and Val401 on the tail. It decreases the overall fluctuations of the protein and prevents Erk binding by keeping the tail of MEK1 closed upon the binding site. As it can be seen on Fig 6., SEL binding causes structural changes within the protein. It physically blocks the Lys97-Glu114 interaction which is vital for the activation of MEK1. It also disturbs the positioning of ATP and changes its interactions with the DFG

motif.(Hashemzadeh, Ramezani, & Rafii-Tabar, 2019)

P124L mutation is a secondary mutation of MEK1 and the functional significance of this mutation is currently unknown. Patients with MEK1^{P124L} allele showed prolonged disease stabilization when treated with Selumetinib but it is clinically proven that P124 mutations may confer cross-resistance to BRAF inhibition.(Caroline M Emery et al., 2009) Residue 124 has an indirect influence on αC helix but directly faces helix A. Mutations on this residue causes an allosteric signal propagation in the protein and enhances the communication of ATP binding region and the activation loop. The information transfer from residue 212 to ERK and RAF binding regions increases. Ser 212 regulates the biological activity of MEK1 and MEK2, increased information transfer from S212 indicates that, S212 drives the motions of the residues which it transfers information to.(Caunt, Sale, Smith, & Cook, 2015)Increased communication between these regions result in further Raf or Erk binding and phosphorylation. P124L mutation enhances S212 control over the important regions and make MEK1 accessible for activation by RAF-1 phosphorylation. Accumulated active MEK1 further phosphorylates ERK1/2 thus inhibition of Ras/Raf/MEK/ERK pathway by Selumetinib fails.

5.8 CONCLUSION

Ras/Raf/MEK/Erk pathway is frequently represented as a linear cascade but there are other feed-back and forward components present. MEK1/2 are ERK gatekeeper proteins which process multiple inputs from other kinases but only activate Erk proteins. MEK family proteins are more therapeutically desirable. However, cells with BRAF^{V600E} mutation leaves MEK and Erk interaction intact. In this study we explained the phosphorylation, phosphoryl transfer, ligand binding effects and the effect of secondary mutations on communication network of ligand bound MEK1. Understanding the biophysical mechanism of MEK1 phosphorylation, ligand binding and mutation is essential to perceive their role in oncogenic signaling and design oncogenic therapeutics that will restrain this pathway effectively and efficiently. (Caunt et al., 2015)

CHAPTER 6.

UNDERSTANDING THE EFFECT OF DNA METHYLATION ON METHYL BINDING PROTEIN AFFINITIES

6.1 OVERVIEW

DNA methylation is a major process which changes the activity of the DNA by covalently adding a methyl group to the DNA molecule and they are not encoded by the DNA sequence. Dysregulation of DNA methylation causes diseases such as cancer(Ke Liu et al., 2018). Methylation is a key player in silencing of transcription, mostly occurs on CpG sites. Methylated CpG and methylated CH (H=A, C, T) DNAs are recognized by methyl binding proteins and Kaiso protein family. Methyl binding proteins play a vital role in coordinating crosstalk between DNA methylation, histone modification and chromatin organization(Bogdanović & Veenstra, 2009). Kaiso proteins repress transcription upon binding to methylated DNA sequences(Koh et al., 2014). Methyl binding proteins have methyl binding domains (MBD) and some of them also contains transcription repression domains (TRD). Other family members have unique domains such as a glycosylase domain (MBD4) or unmethylated-CpG-binding zinc finger domain (MBD1). Kaiso proteins contain several zinc finger domains.

There are 11 MBD proteins identified in mammalians; MeCP2, MBD1-6, SETDB1-2 and BAZ2A-B. In this study, the focus will be on two of the methyl binding proteins, MeCP2 and MBD2. They have negligible affinity towards unmethylated DNA.(Ke Liu et al., 2018)

MeCP2 is the first MBD containing protein discovered, it is involved in neuron development and mutations of this protein are linked to Rett syndrome and several neurological diseases. (Ehrhart et al., 2016; Kriaucionis & Bird, 2003) It contains a TRD domain besides its MBD domain and has 486 residues (52 kDa). It has histone de-acetylation and histone methylation binding partners and it is considered as essential in higher order or long-range chromatin remodeling and silencing and MeCP2 is reported to mediate the translation of intragenic methylation into alternative splicing. These functions are crucial for neural development.

There is a tight integration of methylation binding and transcriptional repression in MBD2 proteins because MBD and TRD domains overlap. MBD2 is a 44 kDa protein which plays a role in helper T-cell differentiation and a recent study showed that they also influence the X-chromosome inactivation (Bogdanović & Veenstra, 2009). MBD2 is not embryonically lethal but it potentially affects maternal nurturing behavior, suggesting neurological effects similar to MBD1 and MeCP2.

MeCP2 and MBD2 bind tightly to mCG and they also prefer to bind mCA over other methylations such as mCT, mCC and mCH. MBD1/2/4 and MeCP2 proteins prefer to bind mCAC motif in mCA methylated DNAs but weaker than mCG binding. MeCP2- mCAC binding is critical for cerebral gene expression. (Ke Liu et al., 2018) MeCP2 barely makes a contact other than mCG bases. It has been proposed that two conserved arginines would be the determinants in the interaction between MBD domain and methylated DNA. These are Arg111 and Arg133 for MeCP2 and Arg166 and Arg188 for MBD2. MBDs could recognize the mCA or CA via binding to their complementary TG. It has been proposed that Arg111 of MeCP2 and Arg166 for MBD2 are fixed by a salt bridge interaction with a conserved Asp residue, Asp121 for MeCP2 and Asp176 for MBD2. Second conserved arginine is more flexible and makes a hydrogen bond with 5'-mC. Mutations of both arginines reduced the mCG-MeCP2 binding affinity. (Lei, Tempel, Chen, Liu, & Min, 2019)

In this study the effect of methylation and mutations are observed by using molecular dynamics simulation program and steered molecular dynamics method is used to detect the changes in the affinity of MeCP2 and MBD2 towards DNA upon methylation. Water molecules between DNA and both methyl binding proteins are observed and their importance in binding affinity is explained. MeCP2-DNA complex used in this study is deposited in the protein data bank with 3C2I PDB ID, the PDB ID for MBD2 protein is 6c1a. (K. L. Ho et al., 2008; Ke Liu et al., 2018) MBD2 protein in 6C1A structure was superposed to MeCP2 in 3C2I to obtain the DNA-MBD2 comple. mCG, CG, mCAC and CAC simulations are conducted and mutations on important binding residues of MeCP2 are mutated to observe the change in affinity of the protein towards mCG-DNA.

6.2 INTRODUCTION

The methyl binding protein family have the ability to bind to (or tri) nucleotide sequences of genomic DNA that contain methylated cytosine (mC) (Ke Liu et al., 2018). They are highly discriminatory and bind with orders of magnitude lesser affinity to the non-methylated form. Despite a growing number of X-ray structures showing molecular details of the DNA –protein interactions, it is still not clear how the addition of one or two methyl groups can have such a profound effect on binding. A total of 11 MBD proteins have been identified in mammals; MeCP2, MBD1-6, SETDB1-2 and BAZ2A-B. In this study, we describe molecular dynamic (MD) simulations on DNA complexes of MeCP2 and MBD2 to study the effect of methylation on binding.

MeCP2 is the first MBD containing protein discovered, it is involved in neuron development and mutations of this protein are linked to Rett syndrome and several neurological diseases (Ehrhart et al., 2016; Kriaucionis & Bird, 2003). The full-length protein is 486 residues long and contains a MBD (residues 77 to 165) and a TRD domain (residues 207–307).(Heckman, Chahrour, & Zoghbi, 2014) MBD2 is a 44 kDa protein which plays a role in helper T-cell differentiation and also influences the X-chromosome inactivation (Bogdanović & Veenstra, 2009). The MBD domain of MBD2 (residues 143-220) has 43% identity with the MBD from MeCP2. It is now well established using a range of experimental methods that the methyl binding domains of MeCP2 and MBD2 show a very clear preference for binding DNA with mCG containing dinucleotides. There is also strong experimental evidence to show that MeCP2 binds DNA containing an mCAC motif and this mCAC binding plays an important role in cerebral gene expression. (Ke Liu et al., 2018) MBD2 has a lower affinity for the mCAC motif (K. Liu et al., 2018).

Though the methylated cytidine form of DNA binds MBD orders of magnitude stronger than non-methylated DNA, X-ray structural studies show that it is possible to form MBD-DNA complexes with non-methylated DNA and the protein-DNA interactions show minimal differences to the methylated complexes. Such observations beg the question of how methyl groups can have such an influence on MBD binding despite having almost no effect on the shape of the protein or any of the specific hydrogen bonded protein-DNA interactions. In this study the effect of methylation and mutations are observed by using molecular dynamics simulations. Steered molecular dynamics are also used to pull the protein away from the DNA to provide a measure of apparent affinity. The calculated energies of binding provide a reasonable correlation with experimentally determined dissociation constants for a variety of DNA-MBD constructs. Analysis if the MD simulations of different methylated and non-methylated DNA sequences allows us to explain differences in binding energies as a function of the dynamic competitive effect of water molecules and the occupancy and strength of the highly specific direct hydrogen bonds formed between MeCP2, MBD2 and the DNA bases.

6.2 PREPARING DNA-MECP2 AND DNA-MBD2 FOR MOLECULAR DYNAMICS SIMULATIONS

6.2.1 MECP2-DNA COMPLEX

DNA-MeCP2 complex with 3C2I PDB ID is used for both CG and CAC studies. QwikMD, a useful simulation plugin of VMD, is used to prepare the system. MSE in MeCP2 protein is changed to methionine. (Ribeiro et al., 2016)

The MD simulations in the present study were performed employing the NAMD molecular dynamics package (Phillips et al., 2005). The CHARMM36 force field(Best et al., 2012; MacKerell Jr et al., 1998) was used in all MD simulations.

The Minimization was performed with explicit solvent using the TIP3 water model(Jorgensen, Chandrasekhar, Madura, Impey, & Klein, 1983) in the NpT ensemble.

A distance cut-off of 12.0 Å was applied to short-range and non-bonded interactions. Long-range electrostatic interactions were treated using the particle-mesh Ewald (PME)(Darden, York, & Pedersen, 1993) method. Before the MD simulations all the systems were submitted to an energy minimization protocol for 20000 steps later on the system is gradually heated, a temperature ramp was performed consisted of 0.24 ns of simulation where the temperature was raised from 60 K to 300 K The pressure was maintained at 1 atm using Nosé-Hoover Langevin piston(Feller, Zhang, Pastor, & Brooks, 1995; Martyna, Tobias, & Klein, 1994). The time step of integration was chosen to be 2 fs for all simulations. Both ends of the piece of DNA were restrained throughout the simulation. The Equilibration was performed with explicit solvent in the NpT ensemble. The temperature was maintained at 300 K using Langevin dynamics. The equations of motion were integrated using the r-RESPA multiple time step scheme (Phillips et al., 2005) to update the short-range interactions every 1 steps and long-range electrostatics interactions every 2 steps. After the equilibration step, a short MD run is performed at 300 K for 25 ns of simulation time.

The SMD was performed with explicit solvent using in the NpT ensemble.

The SMD simulations (Izrailev, Stepaniants, Balsera, Oono, & Schulten, 1997) of constant velocity pulling protocol employing a pulling speed of 30 Å/ns and a harmonic constraint force of 7.0 kcal/mol/Å^2 was performed for 0.56 ns. In this step, SMD was employed by harmonically restraining the backbone of DNA strands and moving a second restraint on backbone of the protein, with constant velocity for all CG, mCG, CAC and mCAC SMD's.

6.2.2 MBD2-DNA COMPLEX

MBD2 protein in structure with PDB ID 6C1a is used for both CG and CAC studies. DNA in 6C1A is replaced with the DNA in 3C2I structure. QwikMD is used to prepare the system. Same steps and protocols were performed with MeCP2 preparation. (Ribeiro et al., 2016)

The MD simulations in the present study were performed employing the NAMD molecular dynamics package (Phillips et al., 2005). The CHARMM36 force field(Best et al., 2012; MacKerell Jr et al., 1998) was used in all MD simulations.

The Minimization was performed with explicit solvent using the TIP3 water model(Jorgensen et al., 1983) in the NpT ensemble.

A distance cut-off of 12.0 Å was applied to short-range and non-bonded interactions. Long-range electrostatic interactions were treated using the particle-mesh Ewald (PME)(Darden et al., 1993) method. Before the MD simulations all the systems were submitted to an energy minimization protocol for 50000 steps, longer than MeCP2 system because the DNA in this complex is synthetically docked to MBD2, later on the system is gradually heated, a temperature ramp was performed consisted of 0.24 ns of simulation where the temperature was raised from 60 K to 300 K The pressure was maintained at 1 atm using Nosé-Hoover Langevin piston(Feller et al., 1995; Martyna et al., 1994). The time step of integration was chosen to be 2 fs for all simulations. Both ends of the piece of DNA were restrained throughout the simulation. The Equilibration was performed with explicit solvent in the NpT ensemble. The temperature was maintained at 300 K using Langevin dynamics. The equations of motion were integrated using the r-RESPA multiple time step scheme (Phillips et al., 2005) to update the short-range interactions every 1 steps and long-range electrostatics interactions every 2 steps. After the equilibration step, a short MD run is performed at 300 K for 25 ns of simulation time.

The SMD was performed with explicit solvent using in the NpT ensemble.

The SMD simulations (Izrailev et al., 1997) of constant velocity pulling protocol employing a pulling speed of 30 Å/ns and a harmonic constraint force of 7.0 kcal/mol/Å^2 was performed for 0.56 ns. In this step, SMD was employed by harmonically restraining the position of DNA backbone, and moving a second restraint on protein backbone, with constant velocity for all CG, mCG, CAC and mCAC SMD's.

Methylations are applied by using 5MC2 methylation patch of CHARMM force field on 5'-C8 and 3'-C33 for mCG and 3'-C33 for mCAC simulations. (Brooks et al., 2009)

The sequence of DNA for mCG, CG, mCAC and CAC are as follows

Red – methylated C, Green – unmethylated C, Blue – mutation

For mCG:

5'-TCTGGAACGGAATTCTTCTA-3'

3'-GACCTTGCCTTAAGAAGATA-5'

For CG:

5'-TCTGGAACGGAATTCTTCTA-3'

3'-GACCTTGCCTTAAGAAGATA-5'

For mCAC:

5'-TCTGGAGTGGAATTCTTCTA-3'

3’-GACCTCACCTTAAGAAGATA-5’

For CAC:

5’-TCTGGAGTGGAATTCTTCTA-3’

3’-GACCTCACCTTAAGAAGATA-5’

*A = Adenine, G = Guanine, C = Cytosine, T = Thymine

6.3 RESULTS

The hydrogen bonds between the nucleic acid backbone and the methyl binding proteins are examined from 25 ns MD trajectories. The crystal water molecules between methyl binding site and Arg133 for MeCP2 and Arg188 for MBD2 protein are further examined to understand the role of hydration in binding. Motions of the both ends of the DNA were restrained by MD throughout the simulation about their initial positions, by applying constraints on their backbone atoms. During the course of the MD simulations the overall shape and conformation of the DNA and MBD molecules are conserved with RMS fit values for DNA of 2.52 Å and for MBD of 2.40 Å.

Previous analyses of methylated DNA X-ray structures have shown that the methyl group in the major groove counters DNA twisting and bending and leads to a widening of major groove and consequently narrowing of the minor groove. (Dantas Machado et al., 2014). The minor and major groove heights for methylated and unmethylated DNA in our simulations show a similar tendency, though the major and minor groove dimensions also vary significantly depending on base sequence (Table XX). The average heights of the minor groove where arginines bind in mCAC-DNA/ CAC are 15.58 /16.12 Å while the major groove dimensions mCAC-DNA/CAC are 20.77/20.43 Å. The comparable average dimensions for mCG-DNA/CG minor groove are 15.10/ 15.67 Å and major groove dimensions for mCG-DNA/CG are 20.71/ 20.10 Å.

For MeCP2 binding the simulations consistently show that for methylated CG sequences there is a narrowing of the minor groove by ~0.5 Å and a widening of the major groove by about the same amount.

The affinities of methyl binding properties for DNA and methylated DNA for CAC and CGcases are calculated with SMD. Pulling force is applied to all backbone atoms of the protein and the rupture distance is measured as the distance of the center of mass of the protein moved until the important bonds between the Arg111 and Arg133, for MeCP2, and Arg166 and Arg188, for MBD2, with the DNA are no longer present. The rupture distance corresponds to the first inflection point of the force vs distance plots. Cumulative work value corresponding to the measured rupture distance is considered for further affinity calculations by using $K_D = exp\ (-\frac{Cumulative\ Work}{kT})$ $\qquad Equation\ 48.$

6.3.1 MECP2

6.3.1.1 MECP2-CAC SIMULATION RESULTS

TABLE 8. HYDROGEN BONDS BETWEEN DNA^{CAC}-MECP2 CALCULATED FROM 25 NS MD SIMULATION

Donor	Acceptor	Occupancy
SegAP1-ARG111-Side	SegBN1-GUA9-Side	61.94%
SegAP1-SER116-Side	SegBN1-GUA9-Side	83.71%
SegAP1-THR158-Side	SegCN1-THY31-Side	63.99%
SegAP1-SER134-Side	SegCN1-CYT33-Side	33.34%
SegAP1-LYS135-Main	SegCN1-CYT33-Side	68.39%
SegAP1-ARG133-Side	SegBN1-GUA7-Side	71.13%
SegAP1-ARG115-Main	SegBN1-GUA9-Side	41.25%
SegAP1-LYS112-Main	SegBN1-THY8-Side	47.21%
SegAP1-LYS135-Side	SegCN1-CYT33-Side	49.81%
SegAP1-ARG162-Side	SegCN1-CYT32-Side	32.76%

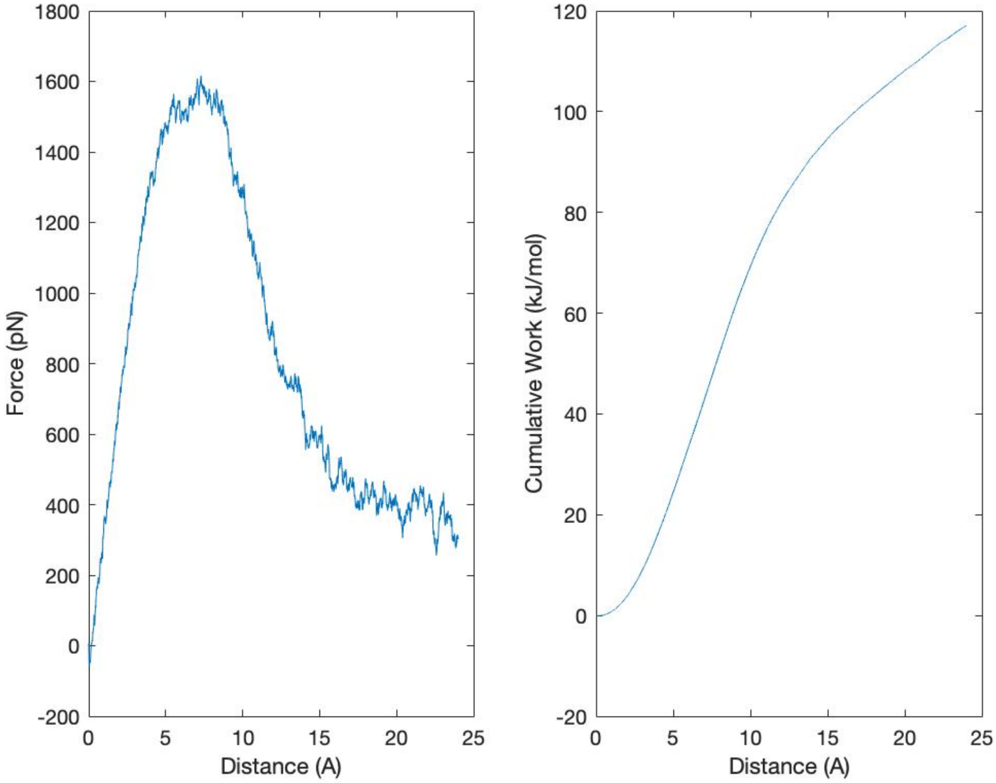


FIGURE 61. MEAN FORCE VS DISTANCE AND MEAN CUMULATIVE WORK VS DISTANCE PLOT OF 10 DNA^{CAC}-MECP2 SMD SIMULATION

Affinity is calculated from the mean SMD plot in Figure 61. The hydrogen bond rupture distance of the arginines 111 and 133, is measured as 5.8 Å. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D value is calculated from the cumulative work value corresponding to rupture distance. Calculated K_D value is 4.7 μM.

6.3.1.2 MECP2-MCAC SIMULATION RESULTS

TABLE 9. HYDROGEN BONDS BETWEEN DNA^{MCAC}-MECP2 CALCULATED FROM 25 NS MD SIMULATIONS

Donor	Acceptor	Occupancy
SegAP1-ARG111-Side	SegBN1-GUA9-Side	54.36%

SegAP1-THR160-Side	SegCN1-THY31-Side	47.00%
SegAP1-SER116-Side	SegBN1-GUA9-Side	84.92%
SegAP1-SER113-Side	SegBN1-THY8-Side	43.80%
SegAP1-THR160-Main	SegCN1-THY31-Side	42.11%
SegAP1-ARG133-Side	SegBN1-GUA7-Side	72.31%
SegAP1-TYR123-Side	SegBN1-GUA7-Side	47.81%
SegAP1-ARG115-Main	SegBN1-GUA9-Side	30.34%
SegAP1-LYS135-Side	SegCN1-CYT33-Side	40.38%
SegAP1-LYS130-Side	SegBN1-GUA5-Side	30.31%

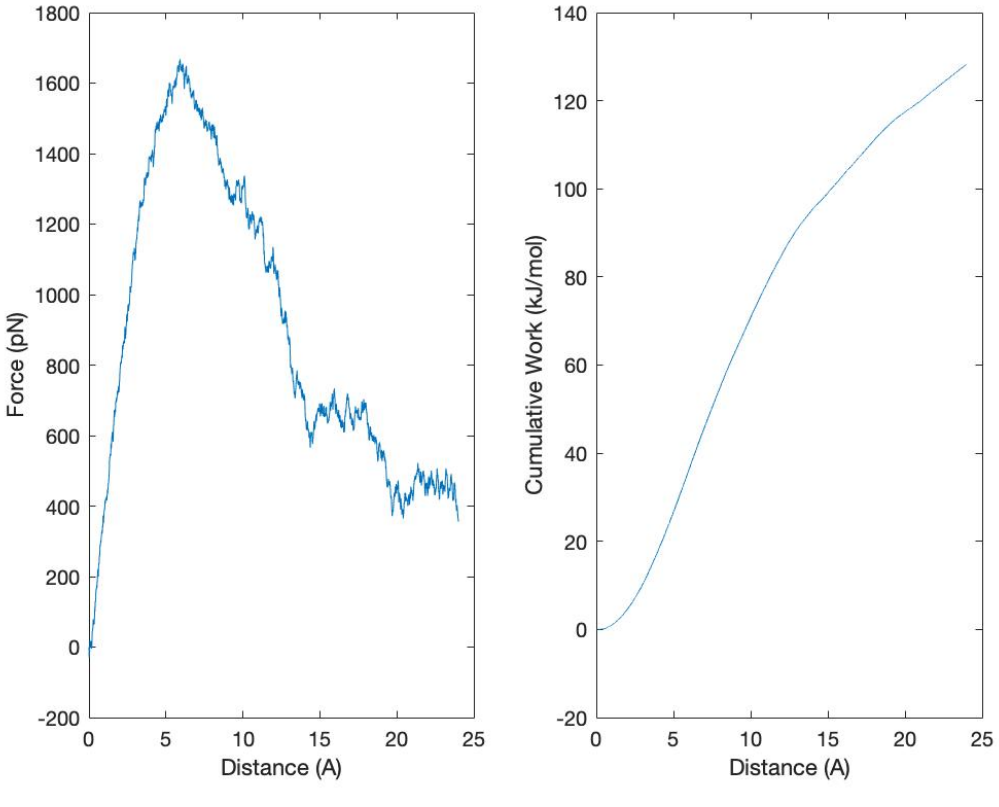


FIGURE 62. MEAN FORCE VS DISTANCE AND MEAN CUMULATIVE WORK VS DISTANCE PLOT OF 10 DNA^{MCAC}-MECP2 SMD SIMULATION

Affinity is calculated from the mean SMD plot in Figure 62. The hydrogen bond rupture distance of the arginines 111 and 133, is measured as 5.8 Å. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D value is calculated from the cumulative work value corresponding to rupture distance. Calculated K_D value is 0.94 μM.

6.3.1.3 MECP2-CG SIMULATION RESULTS

TABLE 10. HYDROGEN BONDS BETWEEN DNA^{CG}-MECP2 CALCULATED FROM 25 NS MD SIMULATIONS

Donor	Acceptor	Occupancy
SegAP1-ARG111-Side	SegBN1-GUA9-Side	67.79%
SegAP1-SER134-Side	SegCN1-CYT33-Side	18.47%
SegAP1-SER116-Side	SegBN1-GUA9-Side	82.90%
SegAP1-THR160-Main	SegCN1-THY31-Side	56.95%

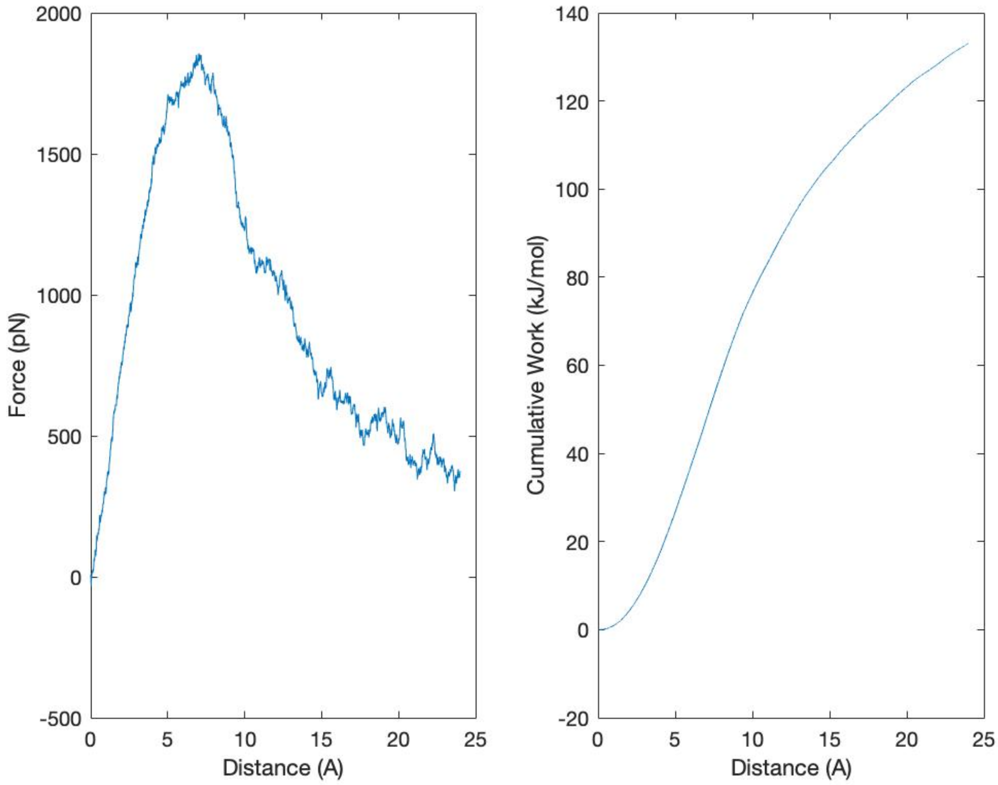


FIGURE 63. MEAN FORCE VS DISTANCE AND MEAN CUMULATIVE WORK VS DISTANCE PLOT OF 10 DNA^{CG}-MECP2 SMD SIMULATION

Affinity is calculated from the mean SMD plot in Figure 63. The hydrogen bond rupture distance of the arginines 111 and 133, is measured as 6.37 Å. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D value is calculated from the cumulative work value corresponding to rupture distance. Calculated K_D value is 133.46 nM.

6.3.1.4 MECP2-MCG SIMULATION RESULTS

TABLE 11. HYDROGEN BONDS BETWEEN DNA^{MCG}-MECP2 CALCULATED FROM 25 NS MD SIMULATIONS

Donor	Acceptor	Occupancy
SegAP1-ARG111-Side	SegBN1-GUA9-Side	62.54%
SegAP1-ARG133-Side	SegCN1-GUA34-Side	25.80%
SegBN1-CYT8-Side	SegAP1-ASP121-Side	20.40%
SegAP1-ARG133-Side	SegBN1-ADE6-Side	35.88%
SegAP1-SER134-Side	SegCN1-CYT33-Side	56.46%
SegAP1-SER116-Side	SegBN1-GUA9-Side	85.86%
SegAP1-THR160-Side	SegCN1-THY31-Side	61.99%

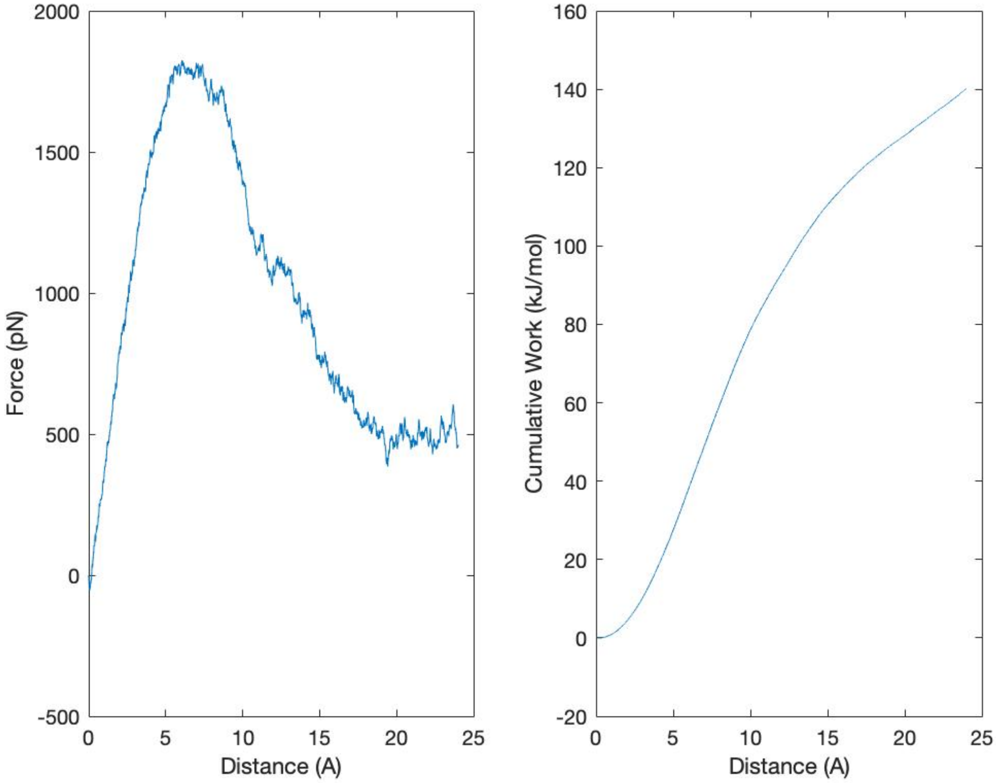


FIGURE 64. MEAN FORCE VS DISTANCE AND MEAN CUMULATIVE WORK VS DISTANCE PLOT OF 10 DNA^{MCG}-MECP2 SMD SIMULATION

Affinity is calculated from the mean SMD plot in Figure 64. The hydrogen bond rupture distance of the arginines 111 and 133, is measured as 6.53 Å. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D value is calculated from the cumulative work value corresponding to rupture distance. Calculated K_D value is 36.99 nM.

6.3.2 MBD2

6.3.2.1 MBD2-CAC SIMULATIONS

TABLE 12. HYDROGEN BONDS BETWEEN DNA^{CAC}-MBD2 CALCULATED FROM 25 NS MD SIMULATIONS

Donor	Acceptor	Occupancy
SegEP1-LYS190-Main	SegCN1-CYT32-Side	50.85%
SegEP1-ARG188-Side	SegBN1-GUA7-Side	59.92%
SegEP1-ARG166-Side	SegBN1-GUA9-Side	35.48%
SegEP1-ARG188-Side	SegBN1-ADE6-Side	61.75%
SegEP1-LYS186-Side	SegBN1-ADE6-Side	37.47%
SegEP1-SER168-Side	SegBN1-GUA9-Side	65.22%
SegEP1-LYS190-Side	SegCN1-CYT32-Side	45.12%

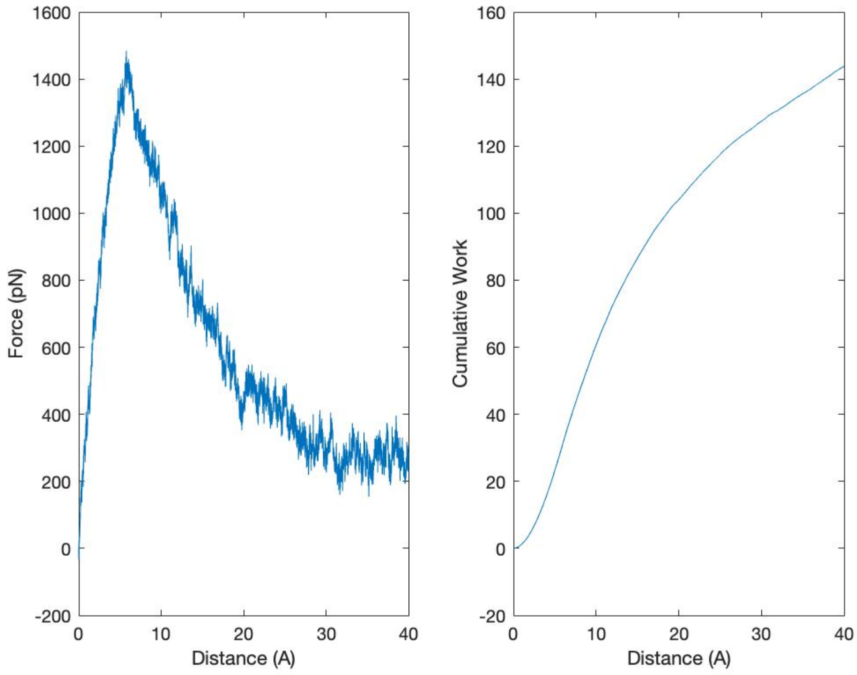


FIGURE 65. MEAN FORCE VS DISTANCE AND MEAN CUMULATIVE WORK VS DISTANCE PLOT OF 10 DNA^{CAC}-MBD2 SMD SIMULATION

Affinity is calculated from the mean SMD plot in Figure 65. The hydrogen bond rupture distance of the arginines 166 and 188, is measured as 4.33 Å. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D value is calculated from the cumulative work value corresponding to rupture distance. Calculated K_D value is 153.32 μM.

6.3.2.2 MBD2-MCAC SIMULATIONS

TABLE 13. HYDROGEN BONDS BETWEEN DNA^{MCAC}-MBD2 CALCULATED FROM 25 NS MD SIMULATIONS

Donor	Acceptor	Occupancy
SegEP1-LYS186-Side	SegBN1-ADE6-Side	50.72%
SegEP1-ARG166-Side	SegBN1-GUA9-Side	52.30%
SegEP1-SER189-Side	SegCN1-CYT33-Side	60.83%
SegEP1-LYS190-Main	SegCN1-CYT32-Side	35.27%
SegEP1-SER168-Side	SegBN1-THY8-Side	59.08%
SegEP1-SER171-Side	SegBN1-GUA9-Side	73.83%

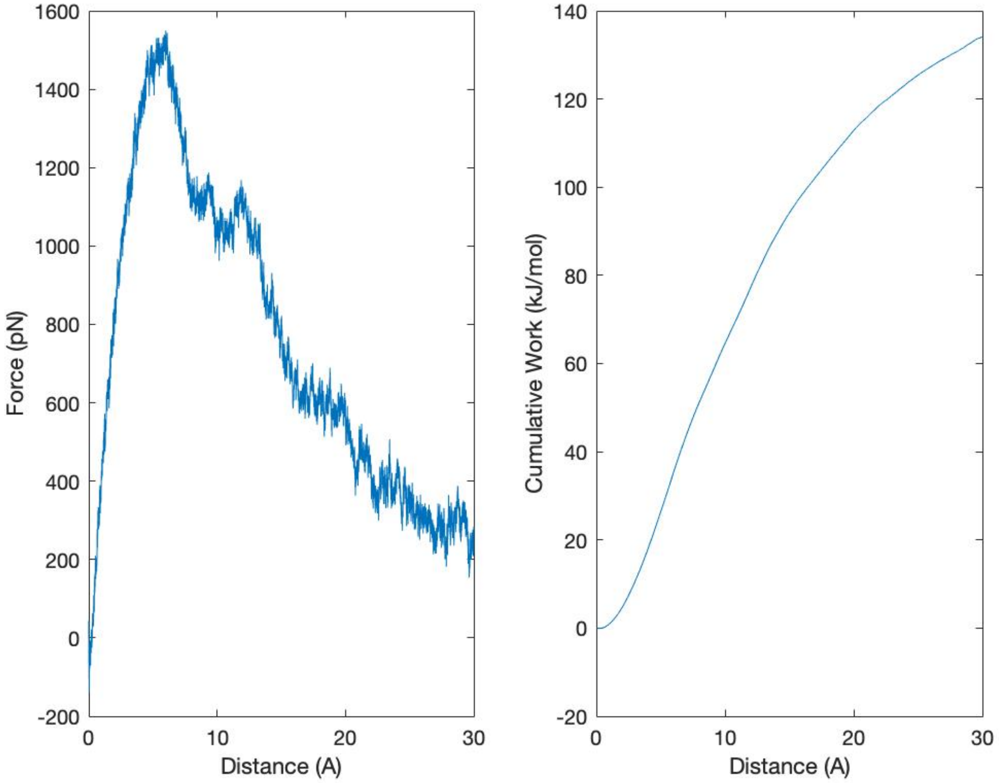


FIGURE 66. MEAN FORCE VS DISTANCE AND MEAN CUMULATIVE WORK VS DISTANCE PLOT OF 10 DNA^{MCAC}-MBD2 SMD SIMULATION

Affinity is calculated from the mean SMD plot in **Error! Reference source not found.**. The hydrogen bond rupture distance of the arginines 166 and 188, is measured as 4.71 Å. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D value is calculated from the cumulative work value corresponding to rupture distance. Calculated K_D value is 85.06 μM.

6.3.2.3 MBD2-CG SIMULATIONS

TABLE 14. HYDROGEN BONDS BETWEEN DNA^{CG}-MBD2 CALCULATED FROM 25 NS MD SIMULATIONS

Donor	Acceptor	Occupancy
SegEP1-SER171-Side	SegBN1-GUA9-Side	80.41%

SegEP1-SER168-Side	SegBN1-CYT8-Side	33.41%
SegEP1-ARG166-Side	SegBN1-GUA9-Side	74.77%
SegEP1-ARG188-Side	SegCN1-GUA34-Side	75.12%

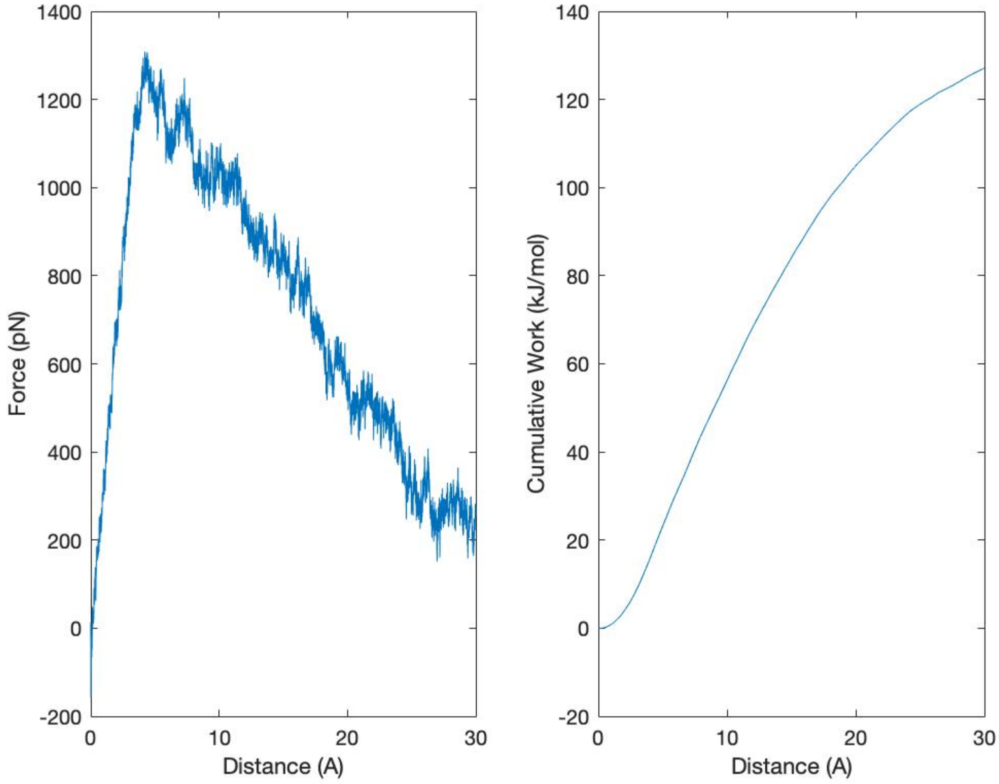
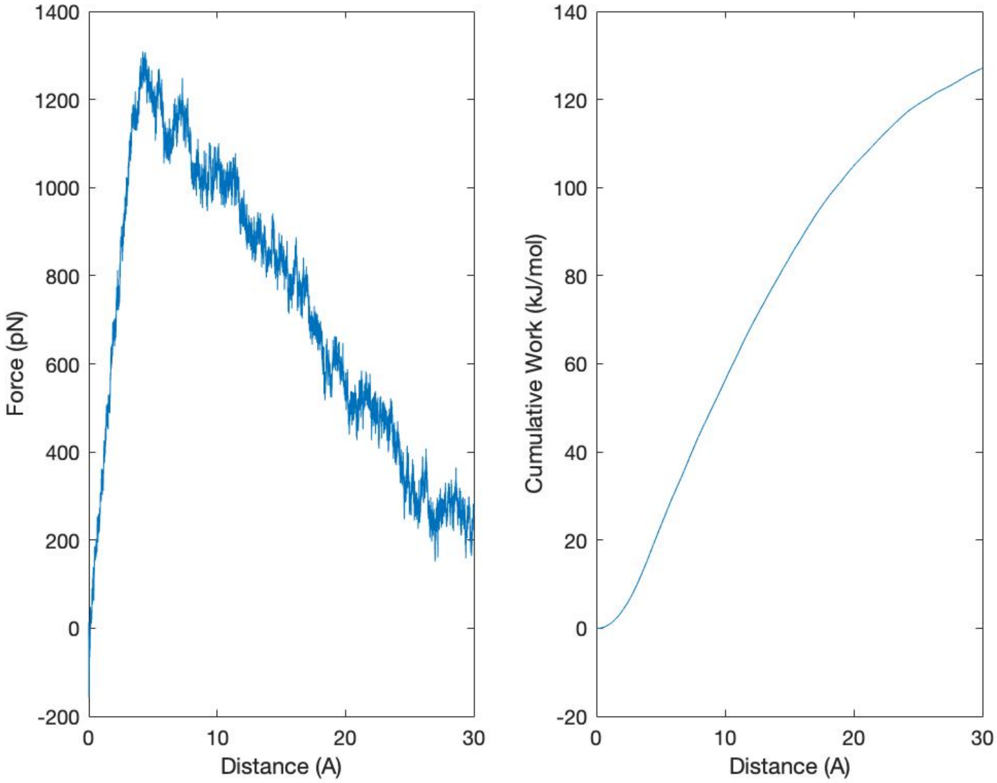


FIGURE 67. MEAN FORCE VS DISTANCE AND MEAN CUMULATIVE WORK VS DISTANCE PLOT OF 10 DNA^{CG}-MBD2 SMD SIMULATION

Affinity is calculated from the mean SMD plot in



. The hydrogen bond rupture distance of the arginines 166 and 188, is measured as 4.08 Å. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D value is calculated from the cumulative work value corresponding to rupture distance. Calculated K_D value is 2 mM.

6.3.2.4 MBD2-MCG SIMULATIONS

TABLE 15. HYDROGEN BONDS BETWEEN DNA^{MCG}-MBD2 CALCULATED FROM 25 NS MD SIMULATIONS

Donor	Acceptor	Occupancy
SegEP1-SER168-Side	SegBN1-CYT8-Side	63.38%
SegEP1-ARG166-Side	SegBN1-GUA9-Side	62.93%
SegEP1-SER189-Side	SegCN1-CYT33-Side	57.14%

SegEP1-ARG188-Side	SegCN1-GUA34-Side	41.50%
SegEP1-TYR178-Side	SegBN1-ADE7-Side	20.63%
SegCN1-CYT32-Side	SegEP1-ASP176-Side	29.73%

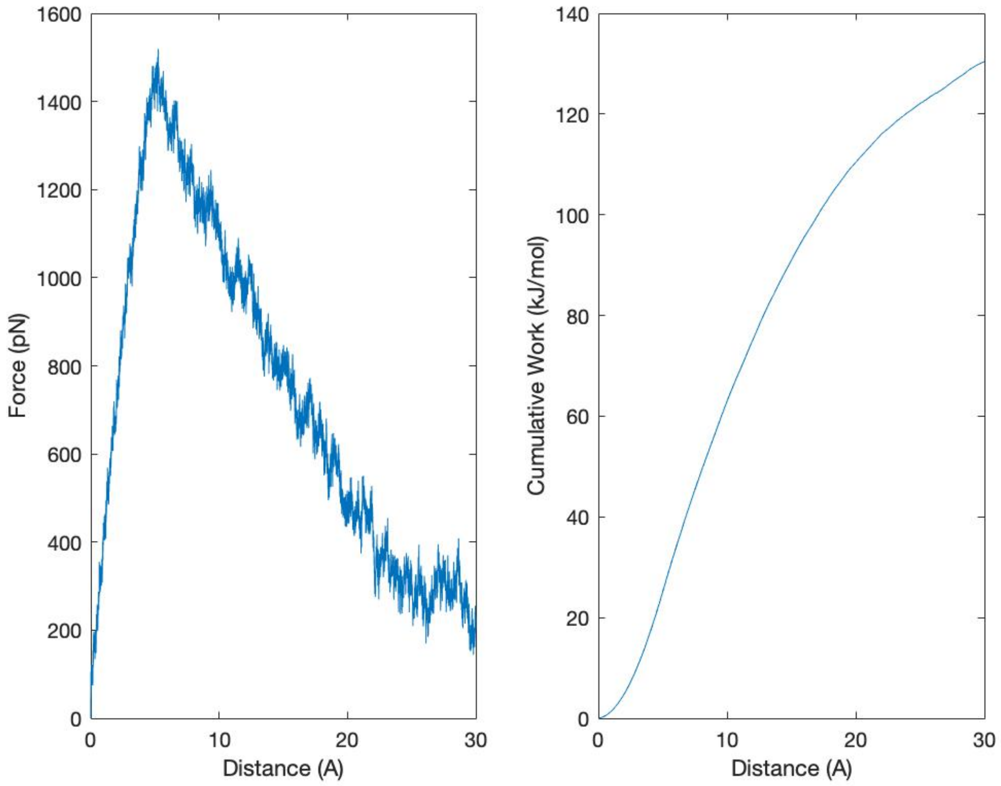


FIGURE 68. MEAN FORCE VS DISTANCE AND MEAN CUMULATIVE WORK VS DISTANCE PLOT OF 10 DNA^{MCG}-MBD2 SMD SIMULATION

Affinity is calculated from the mean SMD plot in Figure 68. The hydrogen bond rupture distance of the arginines 166 and 188, is measured as 4.79 Å. Force values are obtained from the simulation log files and the cumulative work values are measured by calculating the area under the force vs distance plots. The K_D value is calculated from the cumulative work value corresponding to rupture distance. Calculated K_D value is 0.12 mM.

TABLE 16. BINDING AFFINITIES OF MECP2 AND MBD2 TOWARDS CAC/MCAC AND CG/MCG COMPARED WITH VALUES THAT ARE CALCULATED BY SMD AND EXPERIMENTAL RESULTS

	MeCP2				MBD2			
	CAC	mCAC	CG	mCG	CAC	mCAC	CG	mCG
Affinities from SMD	4.7 μ M	0.94 μ M	133.46 nM	36.99 nM	153.32 μ M	85.06 μ M	2mM	0.12 mM
Experimental affinities	2.8 μ M	0.9 μ M	555.9 nM	50 nM	13 μ M	3.2 μ M	NB	0.7 μ M

*NB = no detectable binding.

The binding affinities calculated by SMD simulations are tabulated in **Error! Reference source not found.**. Computationally calculated values are in line with the experimental results.(Fraga et al., 2003; Ke Liu et al., 2018; Sperlazza, Bilinovich, Sinanan, Javier, & Williams Jr, 2017)

6.4 DISCUSSION

The methyl group on 5-C is present in the major groove edge of the cytosine base. The hinderance caused by the methyl group in the major groove counters DNA twisting and bending and leads to widening of major groove and consequently narrowing of the minor groove. (Dantas Machado et al., 2014) Arginines in methyl binding proteins form hydrophobic contacts with the guanine bases complement to methylated cytosines. According to the previous studies, hydrophobic contacts with methyl groups merely effect binding affinity of MBDs. Molecular dynamics simulations show that minor grooves on DNA becomes narrower upon methylation of DNA. It can be concluded from the CAC and CG simulations for methylated and unmethylated cases, contacts of methylated group do not change. The number of water molecules that can fit into the minor groove, where MBD's arginines bind to DNA, decreases upon widening of the major groove and narrowing of the minor groove.

The minor and major groove heights for CAC and mCAC-DNA for MeCP2 binding are compared. The average height of the minor groove where arginines bind in mCAC-DNA is 15.58 Å and 16.12 Å for CAC-DNA and the average height of the major groove where bulky 5-C33 face towards in mCAC-DNA is 20.77 Å and where C5 atom of 5-C33 faces is 20.43 Å for CAC-DNA.

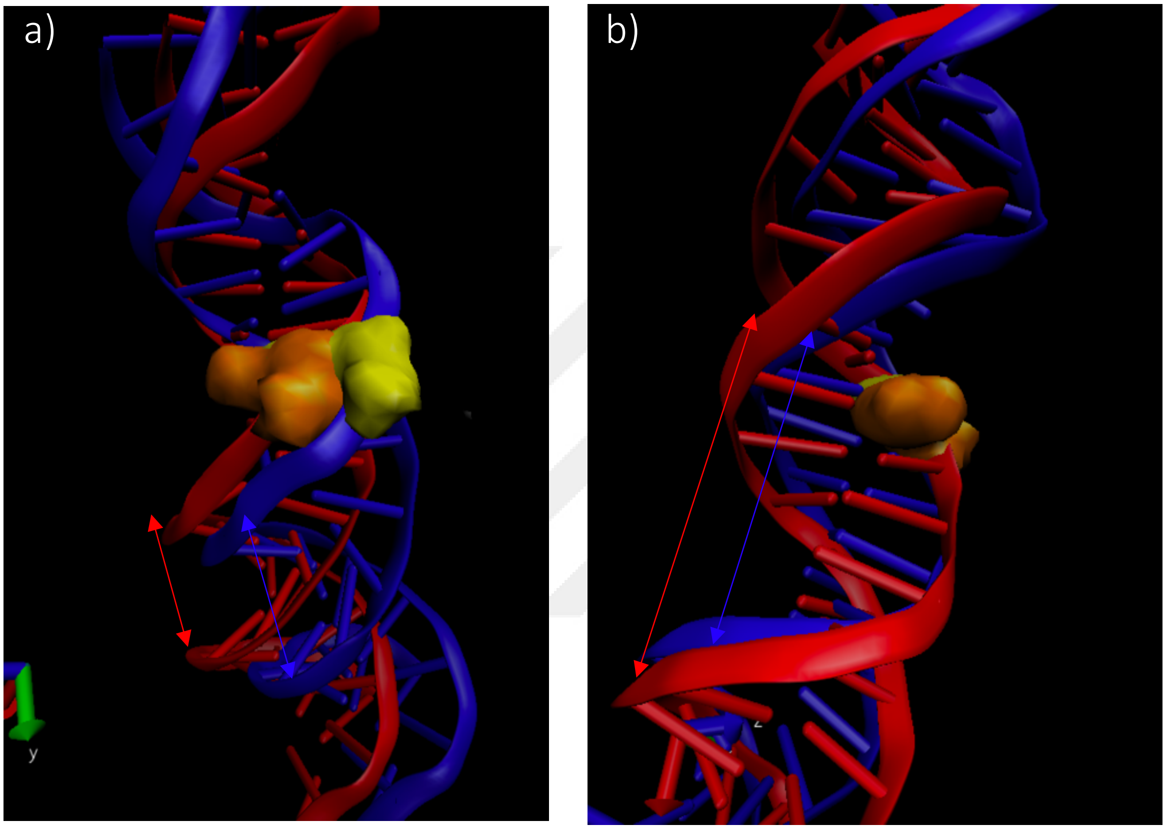


FIGURE 69. SNAPSHOT FROM DNA-CG-MECP2 MD SIMULATIONS. METHYLATED DNA IS SHOWN IN RED CARTOON FORMAT AND UN-METHYLATED DNA IS SHOWN IN BLUE CARTOON FORMAT. ORANGE AND YELLOW SURFACE REPRESENTATIONS ARE FOR METHYLATED AND UN-METHYLATED CYTOSINES RESPECTIVELY. THE MINOR GROOVE WHERE MBD'S ARGININES BIND IS SHOWN ON PANEL A AND THE MAJOR GROOVE IS SHOWN ON PANEL B.

The minor and major grooves of methylated and un-methylated DNA-CG are compared in Figure 69 for MeCP2 binding. The minor groove sizes are shown on panel and the major groove sizes are shown on panel b. The minor groove is narrower, and the major groove is wider in mCG-DNA, colored in red. The hydrophobic interaction of the arginines with guanines is much more powerful in methylated cases due to DNA accompanying smaller number of water molecules in the minor groove. The average height of the minor groove where arginines bind in mCG-DNA is 15.10 Å and 15.67 Å for CG-DNA and the average height of the major groove where bulky 5-C33 and 5-C8 face towards in mCG-DNA is 20.71 Å and where C5 atoms of 5-C33 and 5-C8 face is 20.10 Å for CG-DNA.

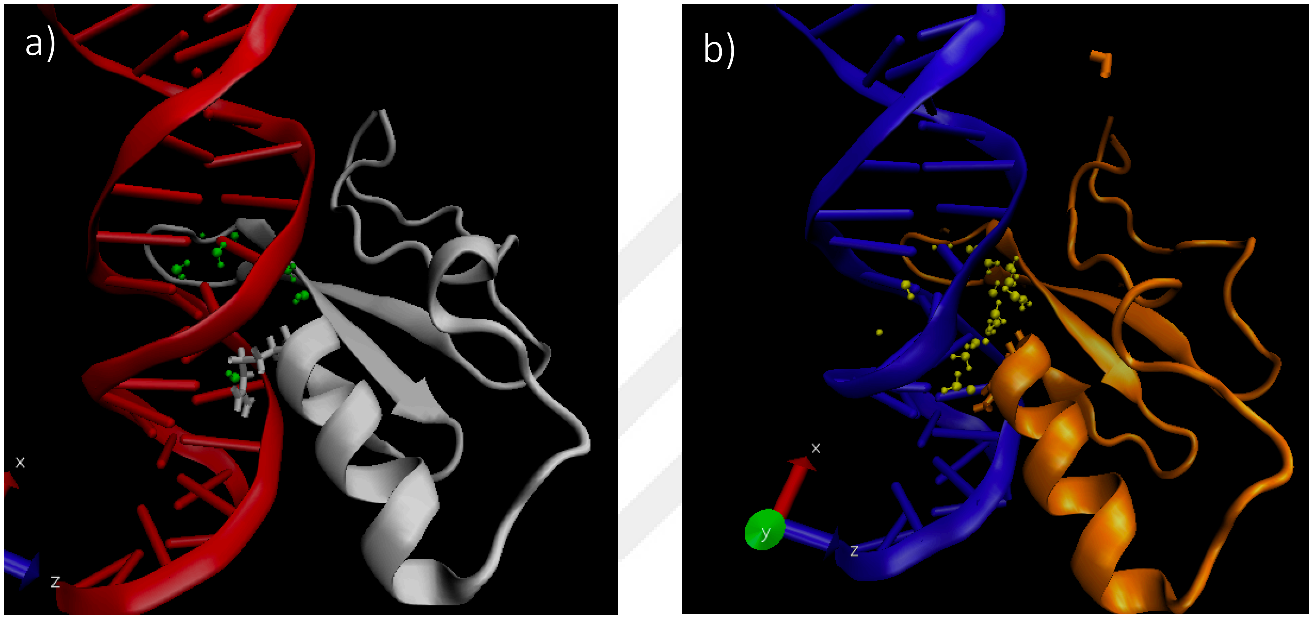


FIGURE 70. THE WATER MOLECULES NEAR 5-C33 AND 5-C8 FOR A) METHYLATED DNA-CG-MECP2 AND B) UN-METHYLATED DNA-MECP2

As it can be seen on panel a and b of Figure 70, the arginine binding region is less occupied by water molecules in methylated DNA case and water molecules form a cluster between MeCP2 and DNA in un-methylated case.

The minor and major groove heights for CAC and mCAC-DNA for MBD2 binding are also compared. The average height of the minor groove where arginines bind in mCAC-DNA is 15.75 Å and 15.47 Å for CAC-DNA and the average height of the major groove where bulky methyl group on 5-C33 face towards in mCAC-DNA is 19.49 Å and where C5 atom of 5-C33 faces is 19.52 Å for CAC-DNA.

The minor and major groove heights for CG and mCG-DNA for MBD2 binding are also compared. The average height of the minor groove where arginines bind in mCG-DNA is 15.54 Å and 15.56 Å for CG-DNA and the average height of the major groove where bulky methyl group on 5-C33 and 5-C8 face towards in mCG-DNA is 20.76 Å and where C5 atoms of 5-C33 and 5-C8 faces is 20.21 Å for CG-DNA.

TABLE 17. MINOR AND MAJOR GROOVE HEIGHTS OF MECP2 AND MBD2 BOUND CAC/MCAC AND CG/MCG DNA'S

	MeCP2				MBD2			
	CAC	mCAC	CG	mCG	CAC	mCAC	CG	mCG
Minor Groove	16.12 Å	15.58 Å	15.67 Å	15.10 Å	15.47 Å	15.75 Å	15.56 Å	15.54 Å
Major Groove	20.43 Å	20.77 Å	20.10 Å	20.71 Å	19.52 Å	19.49 Å	20.21 Å	20.76 Å

The groove heights stated in Table 17, shows that the major groove enlarges upon methylation and the amount of widening is much more noticeable in CG cases. The reason for the larger widening in CG case is a result of dual methylation of cytosine C-33 and C-8. The amount of minor groove narrowing is more noticeable in MeCP2 bound DNA than MBD2 bound DNA CAC and CG cases. This is because MeCP2 protein has more interactions with the nucleic backbone of the DNA whereas MBD2 protein has more interactions with the side chain atoms of the nucleotides.

6.4.1 MECP2 AND MBD2 CAC

When the hydrogen bonds between MeCP2, MBD2 and CAC/mCAC DNA shown on Table 8, Table 9, Table 12Table 13 are compared, it is concluded that the hydrogen bonding profile remains similar upon methylation. The calculated affinities by SMD plots in Figure 61Figure 62 as 4.7 and 0.94 μM for MeCP2 CAC and mCAC cases respectively and the affinities for MBD2 CAC and mCAC cases are calculated from SMD plots in Figure 65Figure 66 as 153.32 and 85.06 μM respectively. Two important arginines of MeCP2, R111 and R133. and MBD2, R166 and R188, form base specific interactions with the GTG (GUA9, THY8, GUA7) motif on CAC-DNA. The interactions are shown on Table 8Table 9, Table 12 Table 13are in line with the experimental results.(Lei et al., 2019) It is known that CAC-DNA is recognized by MBD proteins from their GTG motif. It is confirmed by the simulation results.

The binding of MeCP2 to mCAC-DNA is stronger than that of CAC-DNA. Despite having similar hydrogen bonding profiles, the less hydrated binding site caused the increase in binding affinity of MeCP2 in mCAC case. The widening of the major groove resulted in the narrowing of the minor groove. The binding site became less populated by water molecules which yield in a stronger hydrophobic interaction of R111 and R133 with DNA base pairs.

Similar to MeCP2 binding, MBD2-mCAC interaction is stronger than MBD2-CAC interaction. It can be seen from Table 12, MBD2 does not form an interaction with THY8 of the GTG motif in CAC-DNA. The weaker binding of MBD2 to CAC-DNA can be explained by the lack of this interaction.

The binding affinities of MeCP2 protein towards CAC and mCAC DNA are higher than MBD2. R133 in MeCP2 stands closer to GTG motif than R188 in MBD2 and the minor groove is slightly enlarged upon methylation when MBD2 is bound. This might cause more water molecules to be filled into the cavity and weaken the hydrophobic interactions of arginines in MBD2.

6.4.2 MECP2 AND MBD2 CG

When the main hydrogen bonds in MeCP2 and MBD2 for CG/mCG simulations are compared from Table 10, Table 11 Table 14Table 15, we can conclude that methylation of 5-C8 and 5-C33 enhanced the arginines' hydrophobic interactions with guanines. Interactions of the methylated cytosines are the similar in both methylated and unmethylated DNAs. The increased interactions of arginines with the guanines explain the increase in the binding affinity calculated from the SMD plots on Figure 64, Figure 65, Figure 67Figure 68. The rupture distance of the arginines are 6.37 and 6.53 Å and the calculated affinities are 133.46 nM and 36.99 nM for CG and mCG cases for MeCP2 respectively and the rupture distance of R166 and R166 are 4.08 and 4.79 Å and calculated affinities are 2mM and 0.12 mM for CG and mCG cases for MBD2 respectively. The decrease in the number of water molecules on MeCP2 arginine binding sites with the DNA, enhanced the interaction between them and since there is no considerable minor groove narrowing(see Table 17) in MBD2 bound CG and mCG, the affinity of MBD2 is not strong as MeCP2 affinity towards CG and mCG DNA.

6.5 CONCLUSION

The change of hydration and arginine interactions on CAC and CG DNA methylations when DNA is in complex with MeCP2 and MBD2. Methylation enhances the binding of both proteins but in general mCG binding is much stronger for MeCP2 and MBD2 than mCAC. 5-mC-33 and 5-mC-8 dual methylations of mCG-DNA increase the buried hydrophobic surface of MBD binding by widening the major groove and bulky methyl groups block water molecules from entering to binding site. MBD2 protein affinity for mCG and mCAC is weaker than MeCP2, the reason for this is the distance between Arg188 and guanines on DNA are higher for MBD2 protein. The residues on DNA binding site causes a hinderance and prevents this arginine, Arg188, to be located to a close proximity to DNA.

Given these complexities, of the specific interactions captured from the simulations, it is hard to detect each water molecule and the interaction from experimental results. Hence, computational approaches may be essential to gain a better insight of methylation mechanism. Methyl binding proteins are detectors of the DNA methylation. In humans, MeCP2 inactivation causes Rett syndrome. Understanding how a small methyl group can affect so many biological processes by using computational approaches will aid experimental studies and provide structural basis for disease implications in Rett syndrome.

REFERENCES

Anishchenko, I., Ovchinnikov, S., Kamisetty, H., & Baker, D. (2017). Origins of coevolution between residues distant in protein 3D structures. *Proc Natl Acad Sci U S A, 114*(34), 9122-9127. doi:10.1073/pnas.1702664114

Arakaki, T. L., Pezza, J. A., Cronin, M. A., Hopkins, C. E., Zimmer, D. B., Tolan, D. R., & Allen, K. N. (2004). Structure of human brain fructose 1, 6-(bis) phosphate aldolase: Linking isozyme structure with function. *Protein Science, 13*(12), 3077-3084.

Bahar, I., Atilgan, A. R., & Erman, B. (1997). Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. *Folding & Design, 2*(3), 173-181. doi:Doi 10.1016/S1359-0278(97)00024-2

Baldwin, A. J., & Kay, L. E. (2009). NMR spectroscopy brings invisible protein states into focus. *Nat Chem Biol, 5*(11), 808-814. doi:10.1038/nchembio.238

Ban, D., Funk, M., Gulich, R., Egger, D., Sabo, T. M., Walter, K. F., . . . Griesinger, C. (2011). Kinetics of conformational sampling in ubiquitin. *Angew Chem Int Ed Engl, 50*(48), 11437-11440. doi:10.1002/anie.201105086

Bannas, P., Hambach, J., & Koch-Nolte, F. (2017). Nanobodies and nanobody-based human heavy chain antibodies as antitumor therapeutics. *Frontiers in immunology, 8*, 1603.

Barbany, M., Meyer, T., Hospital, A., Faustino, I., D'Abramo, M., Morata, J., . . . de la Cruz, X. (2015). Molecular dynamics study of naturally existing cavity couplings in proteins. *PLoS One, 10*(3), e0119978. doi:10.1371/journal.pone.0119978

Barderas, R., Desmet, J., Timmerman, P., Meloen, R., & Casal, J. I. (2008). Affinity maturation of antibodies assisted by in silico modeling. *Proceedings of the National Academy of Sciences, 105*(26), 9029-9034.

Barr, D., Oashi, T., Burkhard, K., Lucius, S., Samadani, R., Zhang, J., . . . van der Vaart, A. (2011). Importance of Domain Closure for the Autoactivation of ERK2. *Biochemistry, 50*(37), 8038-8048. doi:10.1021/bi200503a

Ben-Avraham, D. (1993). Vibrational normal-mode spectrum of globular proteins. *Physical Review B, 47*(21), 14559.

Ben-Naim, A. (2008). *A Farewell to Entropy: Statistical Thermodynamics Based on Information: S = logW*: Wspc.

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review, 37*(4), 573-595.

Best, R. B., Zhu, X., Shim, J., Lopes, P. E., Mittal, J., Feig, M., & MacKerell Jr, A. D. (2012). Optimization of the additive CHARMM all-atom protein force field targeting improved sampling of the backbone ϕ, ψ and side-chain χ1 and χ2 dihedral angles. *Journal of Chemical Theory and Computation, 8*(9), 3257-3273.

Bettati, S., Luque, F. J., & Viappiani, C. (2011). Protein dynamics: Experimental and computational approaches Preface. *Biochimica Et Biophysica Acta-Proteins and Proteomics, 1814*(8), 913-915. doi:10.1016/j.bbapap.2011.05.003

Bhullar, K. S., Lagarón, N. O., McGowan, E. M., Parmar, I., Jha, A., Hubbard, B. P., & Rupasinghe, H. V. (2018). Kinase-targeted cancer therapies: progress, challenges and future directions. *Molecular cancer, 17*(1), 48.

Bogdanović, O., & Veenstra, G. J. C. (2009). DNA methylation and methyl-CpG binding proteins: developmental requirements and function. *Chromosoma, 118*(5), 549-565.

Bromberg-White, J. L., Andersen, N. J., & Duesbery, N. S. (2012). MEK genomics in development and disease. *Brief Funct Genomics, 11*(4), 300-310. doi:10.1093/bfgp/els022

Brooks, B. R., Brooks, C. L., 3rd, Mackerell, A. D., Jr., Nilsson, L., Petrella, R. J., Roux, B., . . . Karplus, M. (2009). CHARMM: the biomolecular simulation program. *Journal of Computational Chemistry, 30*(10), 1545-1614. doi:10.1002/jcc.21287

Callen, H. B. (1985). *Thermodynamics and an Introduction to Thermostatistics* (Vol. Second Edition): Wiley.

Capdevila, D. A., Braymer, J. J., Edmonds, K. A., Wu, H., & Giedroc, D. P. (2017). Entropy redistribution controls allostery in a metalloregulatory protein. *Proc Natl Acad Sci U S A, 114*(17), 4424-4429. doi:10.1073/pnas.1620665114

Caunt, C. J., Sale, M. J., Smith, P. D., & Cook, S. J. (2015). MEK1 and MEK2 inhibitors and cancer therapy: the long and winding road. *Nature reviews Cancer, 15*(10), 577.

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., . . . Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov, 2*(5), 401-404. doi:10.1158/2159-8290.CD-12-0095

Chang, F., Steelman, L., Lee, J., Shelton, J., Navolanic, P., Blalock, W. L., . . . McCubrey, J. (2003). Signal transduction mediated by the Ras/Raf/MEK/ERK pathway from cytokine receptors to transcription factors: potential targeting for therapeutic intervention. In: Nature Publishing Group.

Chang, Y.-C., Yang, Y.-C., Tien, C.-P., Yang, C.-J., & Hsiao, M. (2018). Roles of aldolase family genes in human cancers and diseases. *Trends in Endocrinology & Metabolism, 29*(8), 549-559.

Chen, H., Martin, B., Daimon, C. M., & Maudsley, S. (2013). Effective use of latent semantic indexing and computational linguistics in biological and biomedical applications. *Frontiers in physiology, 4*, 8.

Cheng, Y., & Tian, H. (2017). Current development status of MEK inhibitors. *Molecules, 22*(10), 1551.

Chi, C. N., Elfstrom, L., Shi, Y., Snall, T., Engstrom, A., & Jemth, P. (2008). Reassessing a sparse energetic network within a single protein domain. *Proceedings of the National Academy of Sciences of the United States of America, 105*(12), 4679-4684. doi:10.1073/pnas.0711732105

Chi, C. N., Elfström, L., Shi, Y., Snäll, T., Engström, Å., & Jemth, P. (2008). Reassessing a sparse energetic network within a single protein domain. *Proceedings of the National Academy of Sciences, 105*(12), 4679-4684.

Ciombor, K. K., & Bekaii-Saab, T. (2015). Selumetinib for the treatment of cancer. *Expert opinion on investigational drugs, 24*(1), 111-123.

Cooper, A., & Dryden, D. T. F. (1984). Allostery without Conformational Change - a Plausible Model. *European Biophysics Journal with Biophysics Letters, 11*(2), 103-109. doi:Doi 10.1007/Bf00276625

Corrada, D., Morra, G., & Colombo, G. (2013). Investigating Allostery in Molecular Recognition: Insights from a Computational Study of Multiple Antibody-Antigen

Complexes. *Journal of Physical Chemistry B, 117*(2), 535-552. doi:10.1021/jp310753z

Dalby, A., Dauter, Z., & Littlechild, J. A. (1999). Crystal structure of human muscle aldolase complexed with fructose 1, 6-bisphosphate: mechanistic implications. *Protein Science, 8*(2), 291-297.

Dalby, A. R., Tolan, D. R., & Littlechild, J. A. (2001). The structure of human liver fructose-1, 6-bisphosphate aldolase. *Acta Crystallographica Section D: Biological Crystallography, 57*(11), 1526-1533.

Dantas Machado, A. C., Zhou, T., Rao, S., Goel, P., Rastogi, C., Lazarovici, A., . . . Rohs, R. (2014). Evolving insights on how cytosine methylation affects protein–DNA binding. *Briefings in functional genomics, 14*(1), 61-73.

Darden, T., York, D., & Pedersen, L. (1993). Particle mesh Ewald: An N· log (N) method for Ewald sums in large systems. *The Journal of chemical physics, 98*(12), 10089-10092.

De Genst, E., Chan, P.-H., Pardon, E., Hsu, S.-T. D., Kumita, J. R., Christodoulou, J., . . . Muyldermans, S. (2013). A nanobody binding to non-amyloidogenic regions of the protein human lysozyme enhances partial unfolding but inhibits amyloid fibril formation. *The journal of physical chemistry B, 117*(42), 13245-13258.

De Genst, E., Chan, P. H., Pardon, E., Hsu, S. D., Kumita, J. R., Christodoulou, J., . . . Dumoulin, M. (2013). A nanobody binding to non-amyloidogenic regions of the protein human lysozyme enhances partial unfolding but inhibits amyloid fibril formation. *Journal of Physical Chemistry B, 117*(42), 13245-13258. doi:10.1021/jp403425z

De Los Rios, P., Cecconi, F., Pretre, A., Dietler, G., Michielin, O., Piazza, F., & Juanico, B. (2005a). Functional dynamics of PDZ binding domains: a normal-mode analysis. *Biophysical journal, 89*(1), 14-21.

De Los Rios, P., Cecconi, F., Pretre, A., Dietler, G., Michielin, O., Piazza, F., & Juanico, B. J. B. j. (2005b). Functional dynamics of PDZ binding domains: a normal-mode analysis. *89*(1), 14-21.

Dittmer, J., & Bodenhausen, G. (2004). Evidence for slow motion in proteins by multiple refocusing of heteronuclear nitrogen/proton multiple quantum coherences in NMR. *J Am Chem Soc, 126*(5), 1314-1315. doi:10.1021/ja0386243

Dixit, P. D., Wagoner, J., Weistuch, C., Pressé, S., Ghosh, K., & Dill, K. A. (2018). Perspective: Maximum caliber is a general variational principle for dynamical systems. *The Journal of chemical physics, 148*(1), 010901.

Dombi, E., Baldwin, A., Marcus, L. J., Fisher, M. J., Weiss, B., Kim, A., . . . Rizvi, T. A. (2016). Activity of selumetinib in neurofibromatosis type 1–related plexiform neurofibromas. *New England Journal of Medicine, 375*(26), 2550-2560.

Dror, R. O., Green, H. F., Valant, C., Borhani, D. W., Valcourt, J. R., Pan, A. C., . . . Shaw, D. E. (2013). Structural basis for modulation of a G-protein-coupled receptor by allosteric drugs. *Nature, 503*(7475), 295-+. doi:10.1038/nature12595

Du, S., Guan, Z., Hao, L., Song, Y., Wang, L., Gong, L., . . . Shao, S. (2014). Fructose-bisphosphate aldolase a is a potential metastasis-associated marker of lung squamous cell carcinoma and promotes lung cell tumorigenesis and migration. *PLoS One, 9*(1), e85804.

Ehrhart, F., Coort, S. L., Cirillo, E., Smeets, E., Evelo, C. T., & Curfs, L. M. (2016). Rett syndrome–biological pathways leading from MECP2 to disorder phenotypes. *Orphanet journal of rare diseases, 11*(1), 158.

Emery, C. M., Vijayendran, K. G., Zipser, M. C., Sawyer, A. M., Niu, L., Kim, J. J., . . . Karpova, M. B. (2009). MEK1 mutations confer resistance to MEK and B-RAF inhibition. *Proceedings of the National Academy of Sciences, 106*(48), 20411-20416.

Emery, C. M., Vijayendran, K. G., Zipser, M. C., Sawyer, A. M., Niu, L., Kim, J. J., . . . Garraway, L. A. (2009). MEK1 mutations confer resistance to MEK and B-RAF inhibition. *Proc Natl Acad Sci U S A, 106*(48), 20411-20416. doi:10.1073/pnas.0905833106

Erkip, A., & Erman, B. (2004). Dynamics of large-scale fluctuations in native proteins. Analysis based on harmonic inter-residue potentials and random external noise. *Polymer, 45*(2), 641-648.

Erman, B. (2016). Universal features of fluctuations in globular proteins. *Proteins-Structure Function and Bioinformatics, 84*(6), 721-725. doi:10.1002/prot.25032

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M., Eramian, D., Shen, M. y., . . . Sali, A. (2006). Comparative protein structure modeling using Modeller. *Current protocols in bioinformatics, 15*(1), 5.6. 1-5.6. 30.

Farmer, J., Kanwal, F., Nikulsin, N., Tsilimigras, M., & Jacobs, D. (2017). Statistical measures to quantify similarity between molecular dynamics simulation trajectories. *Entropy, 19*(12), 646.

Feller, S. E., Zhang, Y., Pastor, R. W., & Brooks, B. R. (1995). Constant pressure molecular dynamics simulation: the Langevin piston method. *The Journal of chemical physics, 103*(11), 4613-4621.

Fenwick, R. B., Esteban-Martin, S., Richter, B., Lee, D., Walter, K. F., Milovanovic, D., . . . Salvatella, X. (2011). Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition. *J Am Chem Soc, 133*(27), 10336-10339. doi:10.1021/ja200461n

Fischmann, T. O., Smith, C. K., Mayhood, T. W., Myers Jr, J. E., Reichert, P., Mannarino, A., . . . Yang, R.-S. (2009). Crystal structures of MEK1 binary and ternary complexes with nucleotides and inhibitors. *Biochemistry, 48*(12), 2661-2674.

Fleck, M., Polyansky, A. A., & Zagrovic, B. (2016). PARENT: A Parallel Software Suite for the Calculation of Configurational Entropy in Biomolecular Systems. *Journal of Chemical Theory and Computation, 12*(4), 2055-2065. doi:10.1021/acs.jctc.5b01217

Fraga, M. F., Ballestar, E., Montoya, G., Taysavang, P., Wade, P. A., & Esteller, M. (2003). The affinity of different MBD proteins for a specific methylated locus depends on their intrinsic binding properties. *Nucleic acids research, 31*(6), 1765-1774.

Friday, B. B., Yu, C., Dy, G. K., Smith, P. D., Wang, L., Thibodeau, S. N., & Adjei, A. A. (2008). BRAF V600E disrupts AZD6244-induced abrogation of negative feedback pathways between extracellular signal-regulated kinase and Raf proteins. *Cancer research, 68*(15), 6145-6153.

Galanina, N., Smith, S. M., Liao, C., Petrich, A. M., Libao, B., Gartenhaus, R., . . . Stadler, W. (2015). Selective MEK Inhibition with AZD-6244 (selumetinib) in Patients with Relapsed/Refractory Diffuse Large B-Cell Lymphoma (DLBCL): A University of Chicago Phase II Consortium Trial. In: Am Soc Hematology.

Galindo-Murillo, R., Roe, D. R., & Cheatham III, T. E. (2015). Convergence and reproducibility in molecular dynamics simulations of the DNA duplex d (GCACGAACGAACGAACGC). *Biochimica et Biophysica Acta (BBA)-General Subjects, 1850*(5), 1041-1058.

Gane, P. J., & Dean, P. M. (2000). Recent advances in structure-based rational drug design. *Current Opinion in Structural Biology, 10*(4), 401-404.

Garner, T. P., Strachan, J., Shedden, E. C., Long, J. E., Cavey, J. R., Shaw, B., . . . Searle, M. S. (2011). Independent interactions of ubiquitin-binding domains in a ubiquitin-mediated ternary complex. *Biochemistry, 50*(42), 9076-9087. doi:10.1021/bi201137e

Ge, H., Pressé, S., Ghosh, K., & Dill, K. A. (2012). Markov processes follow from the principle of maximum caliber. *The Journal of chemical physics, 136*(6), 064108.

Gerek, Z. N., & Ozkan, S. B. (2011). Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. *PLoS Comput Biol, 7*(10), e1002154.

Gerek, Z. N., & Ozkan, S. B. (2011). Change in Allosteric Network Affects Binding Affinities of PDZ Domains: Analysis through Perturbation Response Scanning. *Plos Computational Biology, 7*(10). doi:ARTN e1002154

10.1371/journal.pcbi.1002154

Gianni, S., Walma, T., Arcovito, A., Calosci, N., Bellelli, A., Engstrom, A., . . . Vuister, G. W. (2006). Demonstration of long-range interactions in a PDZ domain by NMR, kinetics, and protein engineering. *Structure, 14*(12), 1801-1809. doi:10.1016/j.str.2006.10.010

Giroux, S., Tremblay, M., Bernard, D., Cardin-Girard, J., Aubry, S., Larouche, L., . . . Jeannotte, L. (1999). Embryonic death of Mek1-deficient mice reveals a role for this kinase in angiogenesis in the labyrinthine region of the placenta. *Current Biology, 9*(7), 369-376.

Goldman, A. S., & Prabhakar, B. S. (1996). *Immunology overview*: University of Texas Medical Branch at Galveston, Galveston (TX).

Grace, W. Y., Vaysburd, M., Allen, M. D., Settanni, G., & Fersht, A. R. (2009). Structure of human MDM4 N-terminal domain bound to a single-domain antibody. *Journal of Molecular Biology, 385*(5), 1578-1589.

Grutsch, S., Bruschweiler, S., & Tollinger, M. (2016). NMR Methods to Study Dynamic Allostery. *Plos Computational Biology, 12*(3), e1004620. doi:10.1371/journal.pcbi.1004620

Gunasekaran, K., Ma, B., & Nussinov, R. (2004). Is allostery an intrinsic property of all dynamic proteins? *Proteins-Structure Function and Bioinformatics, 57*(3), 433-443. doi:10.1002/prot.20232

Hacisuleyman, A., & Erman, B. (2017a). Causality, transfer entropy, and allosteric communication landscapes in proteins with harmonic interactions. *Proteins-Structure Function and Bioinformatics, 85*(6), 1056-1064. doi:10.1002/prot.25272

Hacisuleyman, A., & Erman, B. (2017b). Entropy Transfer between Residue Pairs and Allostery in Proteins: Quantifying Allosteric Communication in Ubiquitin. *Plos Computational Biology, 13*(1). doi:ARTN e1005319

10.1371/journal.pcbi.1005319

Haile, J. M. (1992). *Molecular dynamics simulation: elementary methods* (Vol. 1): Wiley New York.

Haliloglu, T., Bahar, I., & Erman, B. (1997). Gaussian dynamics of folded proteins. *Physical Review Letters, 79*(16), 3090-3093. doi:DOI 10.1103/PhysRevLett.79.3090

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

Haling, J. R., Sudhamsu, J., Yen, I., Sideris, S., Sandoval, W., Phung, W., . . . Masselot, A. (2014). Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. *Cancer cell, 26*(3), 402-413.

Haling, J. R., Sudhamsu, J., Yen, I., Sideris, S., Sandoval, W., Phung, W., . . . Malek, S. (2014). Structure of the BRAF-MEK complex reveals a kinase activity independent role for BRAF in MAPK signaling. *Cancer Cell, 26*(3), 402-413. doi:10.1016/j.ccr.2014.07.007

Hardy, J. A., & Wells, J. A. (2004). Searching for new allosteric sites in enzymes. *Current Opinion in Structural Biology, 14*(6), 706-715. doi:10.1016/j.sbi.2004.10.009

Hashemzadeh, S., Ramezani, F., & Rafii-Tabar, H. (2019). Study of Molecular Mechanism of the Interaction Between MEK1/2 and Trametinib with Docking and Molecular Dynamic Simulation. *Interdisciplinary Sciences: Computational Life Sciences, 11*(1), 115-124.

Heckman, L. D., Chahrour, M. H., & Zoghbi, H. Y. (2014). Rett-causing mutations reveal two domains critical for MeCP2 function and for toxicity in MECP2 duplication syndrome mice. *Elife, 3*, e02676.

Henry, V. J., Bandrowski, A. E., Pepin, A. S., Gonzalez, B. J., & Desfeux, A. (2014). OMICtools: an informative directory for multi-omic data analysis. *Database (Oxford), 2014*. doi:10.1093/database/bau069

Hnizdo, V., Tan, J., Killian, B. J., & Gilson, M. K. (2008). Efficient calculation of configurational entropy from molecular simulations by combining the mutual-information expansion and nearest-neighbor methods. *Journal of Computational Chemistry, 29*(10), 1605-1614. doi:10.1002/jcc.20919

Ho, B. K., & Agard, D. A. (2010). Conserved tertiary couplings stabilize elements in the PDZ fold, leading to characteristic patterns of domain conformational flexibility. *Protein science, 19*(3), 398-411.

Ho, K. L., McNae, I. W., Schmiedeberg, L., Klose, R. J., Bird, A. P., & Walkinshaw, M. D. (2008). MeCP2 binding to DNA depends upon hydration at methyl-CpG. *Molecular cell, 29*(4), 525-531.

Hospital, A., Goñi, J. R., Orozco, M., & Gelpí, J. L. (2015). Molecular dynamics simulations: advances and applications. *Advances and applications in bioinformatics and chemistry: AABC, 8*, 37.

Huang, Z., Zhu, L., Cao, Y., Wu, G., Liu, X., Chen, Y., . . . Zhang, J. (2011). ASD: a comprehensive database of allosteric proteins and modulators. *Nucleic Acids Res, 39*(Database issue), D663-669. doi:10.1093/nar/gkq1022

Humphrey, W., Dalke, A., & Schulten, K. (1996). VMD: visual molecular dynamics. *Journal of molecular graphics, 14*(1), 33-38.

Isralewitz, B., Gao, M., & Schulten, K. (2001). Steered molecular dynamics and mechanical functions of proteins. *Current Opinion in Structural Biology, 11*(2), 224-230.

Izrailev, S., Stepaniants, S., Balsera, M., Oono, Y., & Schulten, K. (1997). Molecular dynamics study of unbinding of the avidin-biotin complex. *Biophysical Journal, 72*(4), 1568-1581.

Jha, S. K., & Udgaonkar, J. B. (2010). Free energy barriers in protein folding and unfolding reactions. *Current Science, 99*(4), 457-475. Retrieved from <Go to ISI>://WOS:000281638200023

Jo, S., Qi, Y. F., & Im, W. (2016). Preferred conformations of N-glycan core pentasaccharide in solution and in glycoproteins. *Glycobiology, 26*(1), 19-29. doi:10.1093/glycob/cwv083

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W., & Klein, M. L. (1983). Comparison of simple potential functions for simulating liquid water. *The Journal of chemical physics, 79*(2), 926-935.

Jurica, M. S., Mesecar, A., Heath, P. J., Shi, W., Nowak, T., & Stoddard, B. L. (1998). The allosteric regulation of pyruvate kinase by fructose-1,6-bisphosphate. *Structure, 6*(2), 195-210. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9519410

Kamberaj, H., & van der Vaart, A. (2009a). Extracting the causality of correlated motions from molecular dynamics simulations. *Biophysical Journal, 97*(6), 1747-1755. doi:10.1016/j.bpj.2009.07.019

Kamberaj, H., & van der Vaart, A. (2009b). Extracting the Causality of Correlated Motions from Molecular Dynamics Simulations. *Biophysical Journal, 97*(6), 1747-1755. doi:10.1016/j.bpj.2009.07.019

Karni-Schmidt, O., Lokshin, M., & Prives, C. (2016). The roles of MDM2 and MDMX in cancer. *Annual Review of Pathology: Mechanisms of Disease, 11*, 617-644.

Karni-Schmidt, O., Lokshin, M., & Prives, C. (2016). The Roles of MDM2 and MDMX in Cancer. *Annu Rev Pathol, 11*, 617-644. doi:10.1146/annurev-pathol-012414-040349

Karolak, A. (2015). *Application and Development of Computational Methods in Conformational Studies of Bio-Molecules*. University of South Florida,

Karolak, A., & van der Vaart, A. (2012). Importance of local interactions for the stability of inhibitory helix 1 in apo Ets-1. *Biophysical Chemistry, 165*, 74-78. doi:10.1016/j.bpc.2012.03.007

Karplus, M., Ichiye, T., & Pettitt, B. M. (1987). Configurational Entropy of Native Proteins. *Biophysical Journal, 52*(6), 1083-1085. doi:Doi 10.1016/S0006-3495(87)83303-9

Karplus, M., & Kushick, J. N. (1981). Method for Estimating the Configurational Entropy of Macromolecules. *Macromolecules, 14*(2), 325-332. doi:DOI 10.1021/ma50003a019

Kassem, S., Ahmed, M., El-Sheikh, S., & Barakat, K. H. (2015). Entropy in bimolecular simulations: A comprehensive review of atomic fluctuations-based methods. *J Mol Graph Model, 62*, 105-117. doi:10.1016/j.jmgm.2015.09.010

Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular biology and evolution, 30*(4), 772-780.

Kaya, C., Armutlulu, A., Ekesan, S., & Haliloglu, T. (2013). MCPath: Monte Carlo path generation approach to predict likely allosteric pathways and functional residues. *Nucleic acids research, 41*(W1), W249-W255. doi:10.1093/nar/gkt284

Kern, D., & Zuiderweg, E. R. (2003). The role of dynamics in allosteric regulation. *Curr Opin Struct Biol, 13*(6), 748-757. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/14675554

Killian, B. J., Kravitz, J. Y., & Gilson, M. K. (2007). Extraction of configurational entropy from molecular simulations via an expansion approximation. *Journal of Chemical Physics, 127*(2). doi:Artn 024107

10.1063/1.2746329

Killian, B. J., Kravitz, J. Y., Somani, S., Dasgupta, P., Pang, Y. P., & Gilson, M. K. (2009). Configurational entropy in protein-peptide binding: A computational study of the TSG101 UEV domain and an HIV-derived PTAP nonapeptide. *Abstracts of Papers of the American Chemical Society, 238*. Retrieved from <Go to ISI>://WOS:000207861909017

Kim, D. W., & Patel, S. P. (2014). Profile of selumetinib and its potential in the treatment of melanoma. *OncoTargets and therapy, 7*, 1631.

King, B. M., Silver, N. W., & Tidor, B. (2012). Efficient Calculation of Molecular Configurational Entropies Using an Information Theoretic Approximation. *Journal of Physical Chemistry B, 116*(9), 2891-2904. doi:10.1021/jp2068123

Koh, D.-I., Han, D., Ryu, H., Choi, W.-I., Jeon, B.-N., Kim, M.-K., . . . Clarke, A. R. (2014). KAISO, a critical regulator of p53-mediated transcription of CDKN1A and apoptotic genes. *Proceedings of the National Academy of Sciences, 111*(42), 15078-15083.

Kong, Y., & Karplus, M. (2009). Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. *Proteins: Structure, Function, and Bioinformatics, 74*(1), 145-154.

Kong, Y., Karplus, M. J. P. S., Function,, & Bioinformatics. (2009). Signaling pathways of PDZ2 domain: a molecular dynamics interaction correlation analysis. *74*(1), 145-154.

Kooistra, A. J., Kanev, G. K., van Linden, O. P., Leurs, R., de Esch, I. J., & de Graaf, C. (2015). KLIFS: a structural kinase-ligand interaction database. *Nucleic acids research, 44*(D1), D365-D371.

Koshland, D. E., Jr., Nemethy, G., & Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry, 5*(1), 365-385. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/5938952

Kriaucionis, S., & Bird, A. (2003). DNA methylation and Rett syndrome. *Human molecular genetics, 12*(suppl_2), R221-R227.

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *Ann. Math. Statist., 22*(1), 79-86. doi:10.1214/aoms/1177729694

Lange, O. F., Lakomek, N. A., Fares, C., Schroder, G. F., Walter, K. F., Becker, S., . . . de Groot, B. L. (2008). Recognition dynamics up to microseconds revealed from an RDC-derived ubiquitin ensemble in solution. *Science, 320*(5882), 1471-1475. doi:10.1126/science.1157092

Lassmann, T., Frings, O., & Sonnhammer, E. L. (2008). Kalign2: high-performance multiple alignment of protein and nucleotide sequences allowing external features. *Nucleic acids research, 37*(3), 858-865.

Lee, A. L. (2015). Contrasting roles of dynamics in protein allostery: NMR and structural studies of CheY and the third PDZ domain from PSD-95. *Biophysical reviews, 7*(2), 217-226.

Lefranc, M. P. (2011). IMGT unique numbering for the variable (V), constant (C), and groove (G) domains of IG, TR, MH, IgSF, and MhSF. *Cold Spring Harb Protoc, 2011*(6), 633-642. doi:10.1101/pdb.ip85

Legg, P. A., Rosin, P. L., Marshall, D., & Morgan, J. E. (2013). Improving accuracy and efficiency of mutual information for multi-modal retinal image registration using adaptive probability density estimation. *Comput Med Imaging Graph, 37*(7-8), 597-606. doi:10.1016/j.compmedimag.2013.08.004

Lei, M., Tempel, W., Chen, S., Liu, K., & Min, J. (2019). Plasticity at the DNA recognition site of the MeCP2 mCG-binding domain. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms, 1862*(9), 194409.

Lemaigre, F. P., & Rousseau, G. G. (1994). Transcriptional control of genes that regulate glycolysis and gluconeogenesis in adult liver. *Biochemical Journal, 303*(Pt 1), 1.

Li, L., Zhao, G. D., Shi, Z., Qi, L. L., Zhou, L. Y., & Fu, Z. X. (2016). The Ras/Raf/MEK/ERK signaling pathway and its role in the occurrence and development of HCC. *Oncology letters, 12*(5), 3045-3050.

Li, T., Pantazes, R. J., & Maranas, C. D. (2014). OptMAVEn–a new framework for the de novo design of antibody variable region models targeting specific antigen epitopes. *PLoS One, 9*(8), e105954.

Li, Y., Dong, Q., & Cui, Y. (2019). Synergistic inhibition of MEK and reciprocal feedback networks for targeted intervention in malignancy. *Cancer Biology & Medicine, 16*(3), 415.

Lindorff-Larsen, K., Maragakis, P., Piana, S., & Shaw, D. E. (2016). Picosecond to Millisecond Structural Dynamics in Human Ubiquitin. *J Phys Chem B, 120*(33), 8313-8320. doi:10.1021/acs.jpcb.6b02024

Lito, P., Saborowski, A., Yue, J., Solomon, M., Joseph, E., Gadal, S., . . . Ohara, K. (2014). Disruption of CRAF-mediated MEK activation is required for effective MEK inhibition in KRAS mutant tumors. *Cancer cell, 25*(5), 697-710.

Liu, J., & Nussinov, R. (2013). The role of allostery in the ubiquitin-proteasome system. *Crit Rev Biochem Mol Biol, 48*(2), 89-97. doi:10.3109/10409238.2012.742856

Liu, K., Xu, C., Lei, M., Yang, A., Loppnau, P., Hughes, T. R., & Min, J. (2018). Structural basis for the ability of MBD domains to bind methyl-CG and TG sites in DNA. *Journal of Biological Chemistry, 293*(19), 7344-7354.

Liu, K., Xu, C., Lei, M., Yang, A., Loppnau, P., Hughes, T. R., & Min, J. (2018). Structural basis for the ability of MBD domains to bind methyl-CG and TG sites in DNA. *The Journal of biological chemistry, 293*(19), 7344-7354. doi:10.1074/jbc.RA118.001785

Lockless, S. W., & Ranganathan, R. (1999). Evolutionarily conserved pathways of energetic connectivity in protein families. *Science, 286*(5438), 295-299. doi:DOI 10.1126/science.286.5438.295

Lu, S. Y., Huang, W. K., & Zhang, J. (2014). Recent computational advances in the identification of allosteric sites in proteins. *Drug Discovery Today, 19*(10), 1595-1600. doi:10.1016/j.drudis.2014.07.012

MacKay, D. J. C. (2004). *Information theory, inference, and learning algorithms*. Cambridge: Cambridge University Press.

MacKerell Jr, A. D., Bashford, D., Bellott, M., Dunbrack Jr, R. L., Evanseck, J. D., Field, M. J., . . . Ha, S. (1998). All-atom empirical potential for molecular modeling and dynamics studies of proteins. *The journal of physical chemistry B, 102*(18), 3586-3616.

Mahajan, S. P., Meksiriporn, B., Waraho-Zhmayev, D., Weyant, K. B., Kocer, I., Butler, D. C., . . . DeLisa, M. P. (2018). Computational affinity maturation of camelid single-domain intrabodies against the nonamyloid component of alpha-synuclein. *Scientific reports, 8*(1), 17611.

Marks, D. S., Colwell, L. J., Sheridan, R., Hopf, T. A., Pagnani, A., Zecchina, R., & Sander, C. (2011). Protein 3D structure computed from evolutionary sequence variation. *PLoS One, 6*(12), e28766. doi:10.1371/journal.pone.0028766

Martyna, G. J., Tobias, D. J., & Klein, M. L. (1994). Constant pressure molecular dynamics algorithms. *The Journal of chemical physics, 101*(5), 4177-4189.

Massi, F., Grey, M. J., & Palmer, A. G., 3rd. (2005). Microsecond timescale backbone conformational dynamics in ubiquitin studied with NMR R1rho relaxation experiments. *Protein Sci, 14*(3), 735-742. doi:10.1110/ps.041139505

McClendon, C. L., Friedland, G., & Jacobson, M. P. (2010). Quantifying correlations between allosteric sites in thermodynamic ensembles. *Abstracts of Papers of the American Chemical Society, 239*. Retrieved from <Go to ISI>://WOS:000208189302143

McMahon, C., Baier, A. S., Pascolutti, R., Wegrecki, M., Zheng, S., Ong, J. X., . . . Kruse, A. C. (2018). Yeast surface display platform for rapid discovery of conformationally selective nanobodies. *Nat Struct Mol Biol, 25*(3), 289-296. doi:10.1038/s41594-018-0028-6

Monod, J., Wyman, J., & Changeux, J. P. (1965). On the Nature of Allosteric Transitions: A Plausible Model. *Journal of Molecular Biology, 12*, 88-118. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/14343300

Morgan, H. P., McNae, I. W., Nowicki, M. W., Hannaert, V., Michels, P. A. M., Fothergill-Gilmore, L. A., & Walkinshaw, M. D. (2010). Allosteric Mechanism of Pyruvate Kinase from Leishmania mexicana Uses a Rock and Lock Model. *Journal of Biological Chemistry, 285*(17), 12892-12898. doi:10.1074/jbc.M109.079905

Morgan, H. P., O'Reilly, F. J., Wear, M. A., O'Neill, J. R., Fothergill-Gilmore, L. A., Hupp, T., & Walkinshaw, M. D. (2013). M2 pyruvate kinase provides a mechanism for nutrient sensing and regulation of cell proliferation. *Proc Natl Acad Sci U S A, 110*(15), 5881-5886. doi:10.1073/pnas.1217157110

Motlagh, H. N., Wrabl, J. O., Li, J., & Hilser, V. J. (2014). The ensemble nature of allostery. *Nature, 508*(7496), 331-339. doi:10.1038/nature13001

Münz, M., & University of Oxford. (2012). *Computational studies of protein dynamics and dynamic similarity*.

Muyldermans, S. (2013). Nanobodies: natural single-domain antibodies. *Annu Rev Biochem, 82*, 775-797. doi:10.1146/annurev-biochem-063011-092449

Muyldermans, S. (2013). Nanobodies: natural single-domain antibodies. *Annual review of biochemistry, 82*, 775-797.

Naithani, A., Taylor, P., Erman, B., & Walkinshaw, M. D. (2015). A Molecular Dynamics Study of Allosteric Transitions in Leishmania mexicana Pyruvate Kinase. *Biophysical Journal, 109*(6), 1149-1156. doi:10.1016/j.bpj.2015.05.040

Narita, Y., Okamoto, K., Kawada, M. I., Takase, K., Minoshima, Y., Kodama, K., . . . Sawada, K. (2014). Novel ATP-competitive MEK inhibitor E6201 is effective against vemurafenib-resistant melanoma harboring the MEK1-C121S mutation in a preclinical model. *Molecular cancer therapeutics, 13*(4), 823-832.

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of Molecular Biology, 302*(1), 205-217.

Novinec, M., Korenc, M., Caflisch, A., Ranganathan, R., Lenarcic, B., & Baici, A. (2014). A novel allosteric mechanism in the cysteine peptidase cathepsin K discovered by computational methods. *Nature Communications, 5*. doi:ARTN 3287

10.1038/ncomms4287

Numata, J., & Knapp, E. W. (2012). Balanced and Bias-Corrected Computation of Conformational Entropy Differences for Molecular Trajectories. *Journal of Chemical Theory and Computation, 8*(4), 1235-1245. doi:10.1021/ct200910z

Numata, J., Wan, M., & Knapp, E. W. (2007). Conformational entropy of biomolecules: beyond the quasi-harmonic approximation. *Genome Inform, 18*, 192-205. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/18546487

Odendall, C., Rolhion, N., Förster, A., Poh, J., Lamont, D. J., Liu, M., . . . Holden, D. W. (2012). The Salmonella kinase SteC targets the MAP kinase MEK to regulate the host actin cytoskeleton. *Cell host & microbe, 12*(5), 657-668.

Ohren, J. F., Chen, H., Pavlovsky, A., Whitehead, C., Zhang, E., Kuffa, P., . . . Banotai, C. (2004). Structures of human MAP kinase kinase 1 (MEK1) and MEK2 describe novel noncompetitive kinase inhibition. *Nature Structural & Molecular Biology, 11*(12), 1192.

Okonechnikov, K., Golosova, O., Fursov, M., & Team, U. (2012). Unipro UGENE: a unified bioinformatics toolkit. *Bioinformatics, 28*(8), 1166-1167.

Oliver, C., & Jamur, M. C. (2010). Overview of antibodies for immunochemistry. In *Immunocytochemical Methods and Protocols* (pp. 3-9): Springer.

Ordan, M., Pallara, C., Maik-Rachline, G., Hanoch, T., Gervasio, F. L., Glaser, F., . . . Seger, R. (2018). Intrinsically active MEK variants are differentially regulated by proteinases and phosphatases. *Scientific reports, 8*(1), 11830.

Paninski, L. (June 2003). Estimation of entropy and mutual information. In *Neural Computation* (Vol. 15): Massachusetts Institute of Technology.

Panjkovich, A., & Daura, X. (2014). PARS: a web server for the prediction of Protein Allosteric and Regulatory Sites. *Bioinformatics, 30*(9), 1314-1315. doi:10.1093/bioinformatics/btu002

Peltomaa, R., Benito-Peña, E., Barderas, R., & Moreno-Bondi, M. C. (2019). Phage Display in the Quest for New Selective Recognition Elements for Biosensors. *ACS omega, 4*(7), 11569-11580.

Perilla, J. R., Leahy, D. J., & Woolf, T. B. (2013). Molecular dynamics simulations of transitions for ECD epidermal growth factor receptors show key differences between human and drosophila forms of the receptors. *Proteins-Structure Function and Bioinformatics, 81*(7), 1113-1126. doi:10.1002/prot.24257

Petit, C. M., Zhang, J., Sapienza, P. J., Fuentes, E. J., & Lee, A. L. (2009). Hidden dynamic allostery in a PDZ domain. *Proceedings of the National Academy of Sciences, 106*(43), 18249-18254.

Petit, C. M., Zhang, J., Sapienza, P. J., Fuentes, E. J., & Lee, A. L. (2009). Hidden dynamic allostery in a PDZ domain. *Proceedings of the National Academy of Sciences of the United States of America, 106*(43), 18249-18254. doi:10.1073/pnas.0904492106

Phillips, J. C., Braun, R., Wang, W., Gumbart, J., Tajkhorshid, E., Villa, E., . . . Schulten, K. (2005). Scalable molecular dynamics with NAMD. *Journal of computational chemistry, 26*(16), 1781-1802.

Pinto, J., Odongo, S., Lee, F., Gaspariunaite, V., Muyldermans, S., Magez, S., & Sterckx, Y. G.-J. (2017). Structural basis for the high specificity of a Trypanosoma congolense immunoassay targeting glycosomal aldolase. *PLoS neglected tropical diseases, 11*(9), e0005932.

Pressé, S., Ghosh, K., Lee, J., & Dill, K. A. (2013). Principles of maximum entropy and maximum caliber in statistical physics. *Reviews of Modern Physics, 85*(3), 1115.

Pressé, S., Ghosh, K., Lee, J., & Dill, K. A. J. R. o. M. P. (2013). Principles of maximum entropy and maximum caliber in statistical physics. *85*(3), 1115.

Qi, Y. F., & Im, W. (2013). Quantification of Drive-Response Relationships Between Residues During Protein Folding. *Journal of Chemical Theory and Computation, 9*(8), 3799-3805. doi:10.1021/ct4002784

Ribeiro, J. V., Bernardi, R. C., Rudack, T., Stone, J. E., Phillips, J. C., Freddolino, P. L., & Schulten, K. (2016). QwikMD—integrative molecular dynamics toolkit for novices and experts. *Scientific reports, 6*, 26536.

Roskoski Jr, R. (2012). MEK1/2 dual-specificity protein kinases: structure and regulation. *Biochemical and biophysical research communications, 417*(1), 5-10.

Saei, A., & Eichhorn, P. J. A. (2019). Adaptive Responses as Mechanisms of Resistance to BRAF Inhibitors in Melanoma. *Cancers, 11*(8), 1176.

Savoia, P., Fava, P., Casoni, F., & Cremona, O. (2019). Targeting the ERK signaling pathway in melanoma. *International journal of molecular sciences, 20*(6), 1483.

Scarabelli, G., Morra, G., & Colombo, G. (2010). Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping. *Biophysical Journal, 98*(9), 1966-1975. doi:10.1016/j.bpj.2010.01.014

Schreiber, T. (2000a). Measuring information transfer. *Physical review letters, 85*(2), 461-464. doi:DOI 10.1103/PhysRevLett.85.461

Schreiber, T. (2000b). Measuring information transfer. *Phys Rev Lett, 85*(2), 461-464. doi:10.1103/PhysRevLett.85.461

Shin, Y. S., Remacle, F., Fan, R., Hwang, K., Wei, W., Ahmad, H., . . . Heath, J. R. (2011). Protein Signaling Networks from Single Cell Fluctuations and Information Theory Profiling. *Biophysical Journal, 100*(10), 2378-2386. doi:10.1016/j.bpj.2011.04.025

Shukla, D., Meng, Y. L., Roux, B., & Pande, V. S. (2014). Activation pathway of Src kinase reveals intermediate states as targets for drug design. *Nature Communications, 5*. doi:ARTN 3397

10.1038/ncomms4397

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., . . . Söding, J. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular systems biology, 7*(1).

Smith, C. A., Ban, D., Pratihar, S., Giller, K., Paulat, M., Becker, S., . . . de Groot, B. L. (2016). Allosteric switch regulates protein-protein binding through collective motion. *Proc Natl Acad Sci U S A, 113*(12), 3269-3274. doi:10.1073/pnas.1519609113

Sperlazza, M. J., Bilinovich, S. M., Sinanan, L. M., Javier, F. R., & Williams Jr, D. C. (2017). Structural basis of MeCP2 distribution on Non-CpG methylated and hydroxymethylated DNA. *Journal of Molecular Biology, 429*(10), 1581-1594.

Spreafico, A., Tentler, J. J., Pitts, T. M., Tan, A. C., Gregory, M. A., Arcaroli, J. J., . . . Kim, J. (2013). Rational combination of a MEK inhibitor, selumetinib, and the Wnt/calcium pathway modulator, cyclosporin A, in preclinical models of colorectal cancer. *Clinical Cancer Research, 19*(15), 4149-4162.

Stirzaker, C., Song, J., Ng, W., Du, Q., Armstrong, N., Locke, W., . . . Valdes-Mora, F. (2017). Methyl-CpG-binding protein MBD2 plays a key role in maintenance and spread of DNA methylation at CpG islands and shores in cancer. *Oncogene, 36*(10), 1328.

Sturges, H. A. (1926). The choice of a class interval. *Journal of the american statistical association, 21*(153), 65-66.

Suarez, E., Diaz, N., Mendez, J., & Suarez, D. (2013). CENCALC: A Computational Tool for Conformational Entropy Calculations from Molecular Simulations. *Journal of Computational Chemistry, 34*(23), 2041-2054. doi:10.1002/jcc.23350

Suarez, E., Diaz, N., & Suarez, D. (2011). Entropy Calculations of Single Molecules by Combining the Rigid-Rotor and Harmonic-Oscillator Approximations with Conformational Entropy Estimations from Molecular Dynamics Simulations. *Journal of Chemical Theory and Computation, 7*(8), 2638-2653. doi:10.1021/ct200216n

Suarez, E., & Suarez, D. (2012). Multibody local approximation: Application to conformational entropy calculations on biomolecules. *Journal of Chemical Physics, 137*(8). doi:Artn 084115

10.1063/1.4748104

Suel, G. M., Lockless, S. W., Wall, M. A., & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural Biology, 10*(1), 59-69. doi:10.1038/nsb881

Süel, G. M., Lockless, S. W., Wall, M. A., & Ranganathan, R. (2003). Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nature Structural & Molecular Biology, 10*(1), 59-69.

Tang, S., Liao, J. C., Dunn, A. R., Altman, R. B., Spudich, J. A., & Schmidt, J. P. (2007). Predicting allosteric communication in myosin via a pathway of conserved residues. *Journal of Molecular Biology, 373*(5), 1361-1373. doi:10.1016/j.jmb.2007.08.059

Taylor, S. S., & Kornev, A. P. (2011). Protein kinases: evolution of dynamic regulatory proteins. *Trends in biochemical sciences, 36*(2), 65-77.

Tiller, K. E., & Tessier, P. M. (2015). Advances in Antibody Design. *Annu Rev Biomed Eng, 17*, 191-216. doi:10.1146/annurev-bioeng-071114-040733

Toledo, F., & Wahl, G. M. (2007). MDM2 and MDM4: p53 regulators as targets in anticancer therapy. *The international journal of biochemistry & cell biology, 39*(7-8), 1476-1482.

Truong, D. T., & Li, M. S. (2018). Probing the binding affinity by Jarzynski's nonequilibrium binding free energy and rupture time. *The journal of physical chemistry B, 122*(17), 4693-4699.

Tsai, C. J., & Nussinov, R. (2014). A Unified View of "How Allostery Works". *Plos Computational Biology, 10*(2). doi:10.1371/journal.pcbi.1003394

van der Vaart, A. (2015). Coupled binding-bending-folding: The complex conformational dynamics of protein-DNA binding studied by atomistic molecular dynamics simulations. *Biochimica Et Biophysica Acta-General Subjects, 1850*(5), 1091-1098. doi:10.1016/j.bbagen.2014.08.009

Vatansever, S., Gumus, Z. H., & Erman, B. (2016). Intrinsic K-Ras dynamics: A novel molecular dynamics data analysis method shows causality between residue pair motions. *Sci Rep, 6*, 37012. doi:10.1038/srep37012

Vijayan, R., He, P., Modi, V., Duong-Ly, K. C., Ma, H., Peterson, J. R., . . . Levy, R. M. (2014). Conformational analysis of the DFG-out kinase motif and biochemical profiling of structurally validated type II inhibitors. *Journal of medicinal chemistry, 58*(1), 466-479.

Vlachakis, D., Bencurova, E., Papangelopoulos, N., & Kossida, S. (2014). Current state-of-the-art molecular dynamics methods and applications. In *Advances in protein chemistry and structural biology* (Vol. 94, pp. 269-313): Elsevier.

Vuong, Q. V., Nguyen, T. T., & Li, M. S. (2015). A new method for navigating optimal direction for pulling ligand from binding pocket: application to ranking binding affinity by steered molecular dynamics. *Journal of chemical information and modeling, 55*(12), 2731-2738.

Wade, M., Li, Y. C., & Wahl, G. M. (2013). MDM2, MDMX and p53 in oncogenesis and cancer therapy. *Nat Rev Cancer, 13*(2), 83-96. doi:10.1038/nrc3430

Waterhouse, A. M., Procter, J. B., Martin, D. M., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics, 25*(9), 1189-1191.

Wibral, M., Vicente, R., & Lizier, J. T. (2014). *Directed information measures in neuroscience*. Berlin: Springer.

Wilton, E. E., Opyr, M. P., Kailasam, S., Kothe, R. F., & Wieden, H.-J. (2018). sdAb-DB: The Single Domain antibody database. In: ACS Publications.

Wong, W. K., Leem, J., & Deane, C. M. (2019). Comparative analysis of the CDR loops of antigen receptors. *Frontiers in immunology, 10*, 2454.

Wu, P.-K., & Park, J.-I. (2015). *MEK1/2 inhibitors: molecular activity and resistance mechanisms*. Paper presented at the Seminars in oncology.

Xu, K., Tzvetkova-Robev, D., Xu, Y., Goldgur, Y., Chan, Y.-P., Himanen, J. P., & Nikolov, D. B. (2013). Insights into Eph receptor tyrosine kinase activation from crystal structures of the EphA4 ectodomain and its complex with ephrin-A5. *Proceedings of the National Academy of Sciences, 110*(36), 14634-14639.

Yamazaki, I., Tomov, S., & Dongarra, J. (2017). *Sampling algorithms to update truncated SVD*. Paper presented at the 2017 IEEE International Conference on Big Data (Big Data).

Yan, M., & Templeton, D. J. (1994). Identification of 2 serine residues of MEK-1 that are differentially phosphorylated during activation by raf and MEK kinase. *Journal of Biological Chemistry, 269*(29), 19067-19073.

Zhang, C.-S., Hawley, S. A., Zong, Y., Li, M., Wang, Z., Gray, A., . . . Zhu, M. (2017). Fructose-1, 6-bisphosphate and aldolase mediate glucose sensing by AMPK. *Nature, 548*(7665), 112.

Zhang, L. Q., Centa, T., & Buck, M. (2014). Structure and Dynamics Analysis on Plexin-B1 Rho GTPase Binding Domain as a Monomer and Dimer. *Journal of Physical Chemistry B, 118*(26), 7302-7311. doi:10.1021/jp503668k

Zhao, Y., & Adjei, A. A. (2014). The clinical development of MEK inhibitors. *Nature reviews Clinical oncology, 11*(7), 385.

Zheng, C.-F., & Guan, K.-L. (1994). Activation of MEK family kinases requires phosphorylation of two conserved Ser/Thr residues. *The EMBO Journal, 13*(5), 1123-1131.

Zidek, L., Novotny, M. V., & Stone, M. J. (1999). Increased protein backbone conformational entropy upon hydrophobic ligand binding. *Nature Structural Biology, 6*(12), 1118-1121. Retrieved from <Go to ISI>://WOS:000084022300013

Zoete, V., Cuendet, M. A., Grosdidier, A., & Michielin, O. (2011). SwissParam: a fast force field generation tool for small organic molecules. *Journal of Computational Chemistry, 32*(11), 2359-2368.

Zuo, J., Li, J., Zhang, R., Xu, L., Chen, H., Jia, X., . . . Xie, W. (2017). Institute collection and analysis of Nanobodies (iCAN): a comprehensive database and analysis platform for nanobodies. *BMC genomics, 18*(1), 797.