Deciphering Allosteric Interactions and Their Role in Protein Dynamics and Function

by

Tushar Modi

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved November 2020 by the
Graduate Supervisory Committee:

Sefika Banu Ozkan, Chair
Oliver Beckstein
Sara Vaiana
Robert Ros

ARIZONA STATE UNIVERSITY

December 2020

ABSTRACT

Traditionally, allostery is perceived as the response of a catalytic pocket to perturbations induced by binding at another distal site through the interaction network in a protein, usually associated with a conformational change responsible for functional regulation. Here, I utilize dynamics-based metrics, Dynamic Flexibility Index and Dynamic Coupling Index to provide insight into how 3D network of interactions wire communications within a protein and give rise to the long-range dynamic coupling, thus regulating key allosteric interactions. Furthermore, I investigate its role in modulating protein function through mutations in evolution. I use Thioredoxin and β-lactamase enzymes as model systems, and show that nature exploits "hinge-shift" mechanism, where the loss in rigidity of certain residue positions of a protein is compensated by reduced flexibility of other positions, for functional evolution. I also developed a novel approach based on this principle to computationally engineer new mutants of the promiscuous ancestral β-lactamase (i.e., degrading both penicillin and cephatoxime) to exhibit specificity only towards penicillin with a better catalytic efficiency through population shift in its native ensemble.

I investigate how allosteric interactions in a protein can regulate protein interactions in a cell, particularly focusing on E. coli ribosome. I describe how mutations in a ribosome can allosterically change its associating with magnesium ions, which was further shown by my collaborators to distally impact the number of biologically active Adenosine Triphosphate molecules in a cell, thereby, impacting cell growth. This allosteric modulation via magnesium ion concentrations is coined, "ionic allostery". I also describe, the role played by allosteric interactions to regulate information among proteins using a simplistic toy model of an allosteric enzyme. It shows how allostery can provide a mechanism to efficiently transmit information in a signaling pathway in a cell while up/down regulating an enzyme's activity.

The results discussed here suggest a deeper embedding of the role of allosteric interactions in a protein's function at cellular level. Therefore, bridging the molecular impact of allosteric regulation with its role in communication in cellular signaling can provide further mechanistic insights of cellular function and disease development, and allow design of novel drugs regulating cellular functions.

DEDICATION

To my parents and brother for believing in me.

ACKNOWLEDGMENTS

I would like to begin by thanking Professor S. Banu Ozkan for being a caring mentor, a dear friend and a brilliant collaborator. In my duration as a graduate student she has mentored me not only in physics, but also helped me become a better person. She has guided me through the worst and the best, always giving me a platform to try all my ideas. Her unwavering love for knowledge has been my prime motivators. I would also like to give special thanks to my friends Paul Campitelli and Dr. Sean Seyler for lending me their time and ears to discuss my thoughts, and for being my most unadulterated critics. My time at ASU would have been much drier and fruitless if it was not for my friends Ian Kenney, Jiawei Liu, Dr. Lucas Madeira, Dr. Ayush Saurabh and Can Kazan. I would also like to give special thanks to Nikita Kumari, Akanksha Singh and Pritha Bisarad for being my friends and giving me courage when I lacked it.

None of my efforts in research would have been successful without the help and contribution of my collaborators. I would like to sincerely thank Professor Kingshuk Ghosh, Professor Steve Pressè, Professor Gurol Suel, Professor Jose M. Sanchez-Ruiz and Professor Wade Van Horn for giving me guidance which helped me through my research. I am also indebted to my committee members– Professor Oliver Beckstein, Professor Sara Vaiana and Professor Robert Ros for not only guiding me through my academic career, but also providing me a helpful environment to learn and grow.

I would like to thank my family– my parents and my brother who, through all the hardships and sacrifices in their lives, have paved my way to be able to go to a different country and in a prestigious institute to chase my dreams. It would have remained a dream if not for their care, never ending love and constant support. I would be remiss to forget my friends Krishna Das, Abhishek Sharma, Deepak and Nikita Patawari and K. Ananthraman who have been my constant pillar of emotional

support.

I would like to acknowledge the computational resources provided by the Research Computing center at ASU and by Professor Oliver Beckstein. None of my work would have been possible without the computational support provided by these. Finally, I am also grateful for the financial support provided by the Gordon and Betty Moore Foundation, National Science Foundation Division of Molecular and Cellular Biosciences and Scialog Fellow Award by RCSA.

TABLE OF CONTENTS

CHAPTER Page

CHAPTER Page

CHAPTER Page

APPENDIX Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

Proteins are biological machines employed as workhorses of a cell. These were first described by the Dutch chemist Gerardus Johannes Mulder and the term Protein was coined by the Swedish chemist Jöns Jacob Berzelius in 1838 (Hartley, 1951; Leyde., 1838). The term 'protein' was derived from the Greek word *proteios* which means 'the first of rank'. These are the single most important class of macro-molecules in living organisms which perform nearly all the functions necessary for the organism's survival. This includes regulating and facilitating cell growth (Sablowski and Carnier Dornelas, 2014; Nijhout, 2003), catalyzing metabolic reactions in the cell (Zhao *et al.*, 2010), transport through cell membrane (Alexander *et al.*, 2011), providing stimuli to the ambient environment (Hernández-García and Rosenbaum, 2014; Peña *et al.*, 2005; Chockalingam *et al.*, 2007), aiding in cell division (Kim *et al.*, 2016) or apoptosis (Stewart, 1994; Reed, 2000) and many more. In order to perform such a large variety of tasks, each cell in every living organism consists of a large number of various different types of proteins. For example, the E. coli bacterium consists of approximately 4300 protein coding genes, totaling up to 3×10^6 individual protein molecules in a cell (Milo, 2013). Apart from these features, proteins are also one of the very few biological machines which possess the capability to evolve (through a genetic drift (Lynch, 2010)) and regulate their biological function with time (Soskine and Tawfik, 2010). This unique property helps the organism adapt and survive the changes in the environmental conditions like varying temperatures, pH, availability of resources, etc. Clearly, given their importance and ubiquitous presence, proteins are becoming increasingly favored targets for discovering new and novel drugs (SAL,

2010), and also for disease intervention (PEN, 2009). As a first step towards achieving these goals, it is very important to gain mechanistic insight into the functions of the proteins.

1.1 Interactions in Proteins

Each protein is composed of repeating units of different amino acids. Amino acids are simple organic compounds containing a carboxyl and an amino group (see Figure 1.1A). In addition, each amino acid also contains a side chain giving it a unique property. About 500 different naturally occurring amino acids are known to us, though only 20 more commonly appear in the genetic code (Wagner and Musso, 1983) for proteins. These can be classified based on the location of the side chain group, or on the physical properties of the side chain. For the 20 naturally occurring amino acids found in the genetic code, these are classified depending on whether the side chain is charged or uncharged and also based on the hydrophobic nature of the side chain (Vella, 1998). In a protein chain, an amino acid is connected to its neighboring amino acids via a planar (C-N) peptide bond between the carbon of one carboxyl group and the nitrogen of neighboring amide (see Figure 1.1B).

The sequence specific combination of amino-acids in this polypeptide chain leads to local and non-local interactions between amino acids give that the protein a unique 3D structure. This brings us to a "sequence–structure–function" paradigm according to which, the unique 3D structure of a protein is encoded within the 1D primary sequence of the polypeptide chain. The resultant structure of the protein facilitates its function which might involve recognition of a ligand for catalysis and/or binding of the protein with another protein, etc. This sequence-structure-function paradigm led to a significant push in development of experimental and theoretical techniques which helped categorize the interactions between amino acids, imaging structures,

and more (Halabi *et al.*, 2009; Sethi *et al.*, 2005; Socolich *et al.*, 2005; Moult, 2008; Neuwald, 2007; Harms and Thornton, 2013a; Berggård *et al.*, 2007).

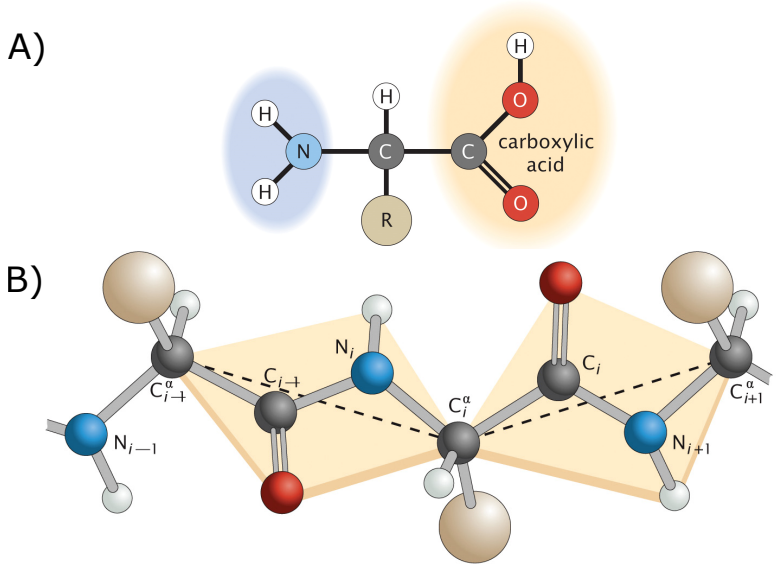

Figure 1.1: Peptide bonds connecting amino acids ($i-1$,i and $i+1$) in a protein chain. It is a bond between the carbon of one carboxyl group and the nitrogen of the neighboring amide (A). The peptide bond is a planar bond such that the bonds $C_{i-1}^{\alpha}-C_{i-1}$, $C_{i-1}-N_i$ and $N_i-C_i^{\alpha}$ are co-planar. Therefore, atoms C_{i-1}^{α}, C_{i-1}, N_i and C_i^{α} are all in one single plane such that the dotted line connecting the neighboring C^{α} atoms lies on the plane (B). (Adapted from (Urbanc, 2017))

However, with advances in structural studies and other experimental techniques, many phenomena have been observed which cannot be explained by the sequence-structure-function paradigm. For example, multiple protein families exist which have widely varying sequence and function but a highly conserved structure. Disease causing mutations is another such example where mutations (i.e., the substitutions of an amino acid with another, or deletion/insertion in the sequence) have shown to conserve the structure but deteriorate or alter the function of the protein. CATH (Knudsen and Wiuf, 2010) and SCOP (Murzin *et al.*, 1995) are two such datasets which describe a series of secondary and tertiary structures which are commonly found across a large variety of proteins found in the protein data bank (Berman *et al.*, 2000). These datasets show that proteins with very little sequence similarity

and diverging functions can adopt a very similar fold due to the spatial constraints enforced by certain type of amino acid combinations in the backbone. Osadchy and Kolodny (2011) also describes the relative diversity in the functional space and the structural space of proteins (e.g., about 1200 folds are categorized in SCOP dataset). They focus their analysis on the structures and functions for the proteins available in the protein data bank. Their study describes how proteins closer to each other in their 3D structure can be very divergent in sequences as well as function. In addition, a lack of unique 3D structure can be a strategy a protein can exploit in order to function. One such class of proteins is intrinsically disordered proteins (IDP). These proteins do not exhibit a fixed or an ordered 3D structure. The disorder in these can range from being completely unstructured or partially structured. However, despite this, these perform a variety of functions critical for organisms survival, e.g., participating in cellular signaling networks, aiding as flexible linkers between interacting domains, etc. (Oldfield and Dunker, 2014; Wright and Dyson, 2014).

These observations suggests that beyond the study of structure and sequence, another major underlying factor which plays a key role in describing the mechanism of a protein's function is an ensemble of conformation within the 3D network of interactions. These interactions range from direct interactions (i.e., between a pair of atoms) like covalent bonds, electrostatic interactions, hydrogen bonds, van der waal forces, etc. to indirect interactions which are cascaded over much larger distances via a chain of residues between two amino acids which are not in each other's coordination shell (i.e., their distance is typically much greater than 7Å). Perturbations in the structure, induced by factors such as binding of a ligand, mutations, etc. which utilize such distal indirect interactions to alter the function of the protein are called *allosteric interactions*.

1.2 Allosteric Interactions

The concept of allostery was first introduced in 1961 in a summary article for the Cold Spring Harbor Symposium on Cellular Regulatory Mechanisms by Jacques Monod and Francois Jacob (Monod and Jacob, 1961) as a phenomenon to explain the anomalous binding behaviour observed in the inhibition of the enzyme L-threonine deaminase (Changeux, 1961). Although the term was first introduced in 1961, the phenomenon was first foreshadowed about 25 years earlier (in 1935) by Pauling who proposed a long distance intramolecular control in hemoglobin to explain the positive cooperativity observed in the binding of oxygen molecules (Pauling, 1935). Monod was so much overwhelmed by the ubiquitous and rousing implications of allosteric interactions in proteins that he called it the "The second secret of Life" (Ullmann, 2011), the first being the double helical structure of the DNA.

Clearly, the study of allosteric interactions is the key factor in order to understand protein functions. Several models have been proposed which provide insights into the kinetics of allostery and resulting cooperative binding. One of the most widely accepted models is the Monod-Wyman-Changeux model (MWC model or the Concerted model) (Monod *et al.*, 1965).

The MWC model proposed in 1965, was inspired by the binding behaviour of oxygen with the hemoglobin and the role played by haem-haem interactions in it. In the study, Monod *et al.* (1965) describe allostery as– ". . . indirect interactions between distinct binding sites (allosteric effects) . . . these interactions are mediated by some kind of molecular transition (allosteric transition) which is induced or stabilized by the protein when it binds an allosteric ligand.". The model postulates that– allosteric interactions are present in symmetric oligomeric proteins, and the allosteric effect is induced by binding of a ligand to one of the subunit (a protomer) that leads to a

change in its conformation. This change has a cascading effect which trickles down to the whole oligomer in order to preserve the symmetry between subunits.

They described a model of an allosteric oligomeric protein which can exist in two states R and T (see Figure 1.2). These two states can differ in their bond networks and hence have different energies. The two states in their unbound form exist in an equilibrium where the equilibrium constant depends on the energy differences between the two states. In addition, each subunit in the oligomer is interacting with every other subunit such that its conformation is constrained by its association with the other subunits. This implies that each subunit in the state R or T of the oligomer is restricted to share a common conformation. Therefore, if one of the protomer subunits undergoes a conformation change, it forces the rest of the subunits to follow the change in their conformation such that the symmetry of the oligomer is conserved. Furthermore, the model also specifies that the binding of a ligand to a protomer is stereo specific, i.e., the binding affinity depends whether the oligomer is in state R or T. However, the affinity for the ligand to bind with a protomer, given the oligomeric state, does not depend on how many ligands are already bound to the oligomer. As a result, when the oligomer undergoes a transition from one state to another, it changes the the affinity of all the protomeric subunits towards the corresponding ligand. This model is schematically represented for an enzyme with two protomers in Figure 1.2. Monod *et al.* (1965) also generalized the model to allow the study of cooperativity in homotropic (i.e., binding of the oligomeric protein with the same ligand) as well as heterotropic (i.e., when the oligomeric protein binds with different types of ligands) effects of ligand binding.

Through the model, Monod *et al.* (1965) were able to describe the binding behaviour of several oligomeric systems like Deoxythymidine kinase from E. coli, dCMP deaminaae from ass spleen, etc. which displayed cooperativity. However, this model

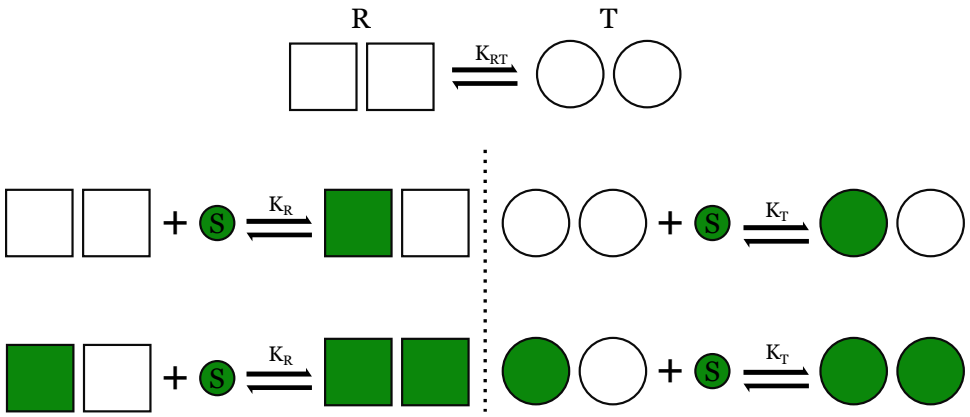

Figure 1.2: A schematic diagram of homotropic allosteric interactions in a dimeric protein following MWC model. According to the model, the enzyme can exist in two states R and T with an equilibrium constant of K_{RT} between them. The equilibrium constant for binding of ligand molecule (S) to the enzyme depends on whether it is in state R (K_R) or state T (K_T). Once the ligand has attached itself to any of the binding sites on the enzyme, the state of the enzyme is locked. Afterwards, subsequent binding events between the enzyme and the ligand have the same equilibrium constant depending on which state the enzyme is locked in.

suffers from a major disadvantage due to its historic inspiration from the allostery in the symmetric hemoglobin. The assumption of the model that ligands show similar binding affinities with each of the protomeric subunits in the oligomer does not allow for the model to exhibit negative cooperativity as shown by Kister *et al.* (1987). However, a modified version of the MWC model was also proposed (Kister *et al.*, 1987) which accounts for the change in the binding affinities of the ligand to bind with a subunit depending on how many bound subunits exist in the oligomer. This allows the intra-molecular interactions between subunits to be more efficiently included in the model and using such modifications, MWC model can be tinkered to show negative cooperativity.

Another well recognized model for allostery is the Koshland-Némethy-Filmer model (KNF model or the Sequential model) (Koshland *et al.*, 1966). This model was inspired from an earlier work of Pauling (Pauling, 1935) (the model is also known as Pauling-Koshland model). As opposed to the symmetry restrictions posed by the

MWC model, namely the interlocking of each protomeric subunit into a common conformation, here each subunit is free to change its conformation independently (between R and T, see Figure 1.3). However, similar to the previous model, the binding affinity of the ligand is stereo specific, i.e., it depends on whether the subunit is in state R or T. Further on, the interactions between the subunits in the oligomer are geometry dependent, i.e., whether the oligomer has its subunits in a tetrameric or square or some other arrangement. As a result, the equilibrium constant for transition of a subunit from R to T or vice versa depends on the relative orientation of the subunit and also on whether other neighboring subunits are already creating a complex with the ligand or not (Figure 1.3).

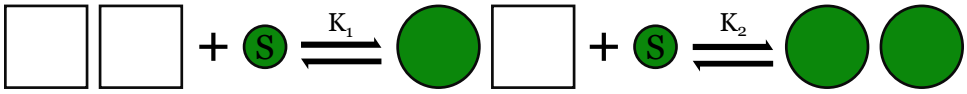

Figure 1.3: A schematic diagram of homotropic allosteric interactions in a dimeric protein following KNF model. In this model, the enzyme has two subunits which can exist in two states, R and T. However, the transition of the catalytic site in one subunit is independent of another. Also, the transition from R to T is initiated by binding of the ligand (S). The equilibrium constant for the first ligand to bind with the enzyme (K_1) is independent of the subunit to which it binds. For all subsequent binding events, the equilibrium constant depends on how many ligands are bound to the enzyme and also the relative orientation of the bound and unbound subunits.

Through this model, Koshland *et al.* (1966) were not only able to reproduce the cooperative behaviour shown by the MWC model, but were also able to exhibit negative cooperativity. In addition, they also brought forward an important result that several different arrangements of the oligomeric system can fit the same experimental data with a reasonable accuracy for the right choice of parameters (binding affinities of the ligand and the strength of interaction between the subunits).

This brings forward an interesting outlook towards these models. Both the MWC and KNF models are phenomenological in nature. These do not provide any significant mechanistic insight into how the structure facilitates allosteric interaction

between subunits (with the exception of KNF model which does show a dependence of cooperativity on the relative arrangement of subunits around each other, but without commenting on the nature of those interactions). This issue can be partially resolved by the incorporation of statistical physics into the picture. Several studies (Cooper and Dryden, 1984; Phillips *et al.*, 2008; Komorowski and Tawfik, 2019; McLeish *et al.*, 2013; Wodak *et al.*, 2019a) have been performed where allosteric interactions between two distant sites (this time not restricted to only oligomers) are given a thermodynamic and statistical treatment while not restricting themselves to a kinetic perspective.

This can be understood from a very simple toy model (adapted from Phillips *et al.* (2008) and Marzen *et al.* (2013)). Let us consider an enzyme E (not necessarily an oligomer). This enzyme can exist in two different states R and T such that this enzyme can only function while being in state R (this function can be binding with another ligand, etc.). These states differ from each other in their free energy by ϵ. As a result, the probability of finding the enzyme in state R (P_R) can be given as:

$$P_R = \frac{1}{1 + e^{-\beta\epsilon}} \tag{1.1}$$

Now, let us consider interaction of this enzyme with a ligand which acts as an allosteric regulator of the enzyme. The binding of the ligand alters (allosterically) the free energies of the two states (R and T) differently, such that, the new free energy of state R is $-l_R$ and that of state T is $\epsilon - l_T$. Due to this change, the new modified probability of finding the enzyme in state R ($P^*{}_R$) can be given as:

$$P^*{}_R = \frac{e^{\beta l_R}}{e^{\beta l_R} + e^{-\beta(\epsilon - l_T)}} \tag{1.2}$$

Dividing Eq. 1.2 with Eq. 1.1 describes the shift in the probability (hence, population)

of the enzyme's state R as follows:

$$\frac{P^*_R}{P_R} = \frac{e^{\beta l_R} + e^{-\beta(\epsilon - l_R)}}{e^{\beta l_R} + e^{-\beta(\epsilon - l_T)}} \tag{1.3}$$

$$\frac{P^*_R}{P_R} = \frac{1 + e^{-\beta\epsilon}}{1 + e^{-\beta(\epsilon - \Delta l)}} \tag{1.4}$$

where, $\Delta l = l_T - l_R$. According to this, when $l_R \neq l_T$, then the population of state R is shifted and this impacts the function of the enzyme (because as assumed earlier, only state R is functional). The impact on the function depends on the direction of the shift (i.e., whether it increases or decreases the population of state R). In this case, we observe an allosteric effect. On the other hand, when $l_R = l_T$, then the net population shift is cancelled out (from Eq. 1.4) and no allosteric effect is observed. Several different versions of this model have been proposed and used with varying degrees of success (Marzen *et al.*, 2013; Porter *et al.*, 2020). This approach has a benefit of providing a deeper mechanistic insight into the origin of allostery by describing how to link the degree of allosteric effect explicitly with the energetic differences created due to the binding of the allosteric regulator ligand as shown by Eq. 1.4.

Approaching the problem of allostery with the tools borrowed from statistical physics allows us to analyze the problem from another perspective as shown by (Cooper and Dryden, 1984). Cooper and Dryden (1984) also gave an allosteric enzyme a statistical mechanics treatment by focusing on the changes induced by ligand binding on the ensemble describing the enzyme. They showed that the net free energy difference upon multiple ligand bindings in an allosteric enzyme is a function of the partition function which describes the ensemble of states accessible to the enzyme after binding with subsequent ligands. The partition function of the enzyme can be further decomposed into its different constituents– vibrational and configurational components. For example, for an unbound enzyme E (with a partition function Q_E),

free ligand L (with a partition function Q_L) creating a bound complex EL (with a partition function Q_{EL}, the change in free energy can be expressed as:

$$\Delta G = \Delta\epsilon - kT \ln \frac{Q_E Q_L}{Q_{EL}} \tag{1.5}$$

where, $\Delta\epsilon$ is the difference in the ground state energy upon binding. Here, the partition function contains the contribution of vibrational, conformational, rotational and other forms of energy stored in the molecule. Therefore, if binding of ligand L has an impact on any of the factors contributing towards the partition function of the enzyme E, then: $Q_{EL} \neq Q_E.Q_L$, and binding will alter the ensemble describing the enzyme. This implies that allosteric effects can be induced not only by the changes in the conformational ensemble of the enzyme (as discussed in MWC or KNF model), but also by changes in the vibrational ensemble. This includes subtle changes in the normal modes describing the motion of the enzyme. These correspond to changes in the thermal fluctuations about the equilibrium position of atoms in the enzyme which are typically not easily captured through crystallographic experiments. However, these models still provide only an incomplete picture of the system. For example, these typically assume a finite number of states for the enzyme. This biases the calculations of rate constants for substrate binding and catalysis. For accurate calculations of rate kinetics, an intimate knowledge of transition pathways is primarily important which is very hard to obtain either via computational analysis or by experiments, though significant progress has been made in this direction (Doron *et al.*, 2014; Wigner, 1937; Hänggi *et al.*, 1990; Eyring, 1935; Truhlar *et al.*, 1983; Wang *et al.*, 2016).

Therefore, in order to explore a mechanistic picture of allostery, a deeper insight into the network of interactions in a protein, which includes a way to quantify their strength is needed. This also allow us to discuss the implications and the role of allostery in a more general way which is not restricted to only binding induced changes

but also highlight how these long distance indirect interactions impact the dynamics of a protein and its function. Here we will be focusing on some of these implications of allosteric interactions described below.

1.3 The Role of Allosteric Interactions

In the thesis we begin by studying the role played by allosteric interactions in protein evolution. Protein evolution is another example of a phenomenon, where a violation of "sequence-structure-function" paradigm is observed. Life is believed to have evolved on earth over a period of around 4 billion years. Several theories suggest that cells with all its constituents like membranes, proteins, etc. have evolved from a primordial soup of organic molecules of carbon, nitrogen, hydrogen and oxygen (Fuchs, 2011). The steps taken by nature to evolve life in its current form has been an open problem captivating the interests of biologists, chemists and physicists alike for centuries (Schrödinger, 1944). This branch has been contributed significantly by progress in information theory and more accurate sequencing techniques (Agozzino and Dill, 2018; Copp *et al.*, 2019; Finnigan *et al.*, 2012; Ingles-Prieto *et al.*, 2013; Risso and Sanchez-Ruiz, 2017; Wang *et al.*, 2019). With advances in computational and mathematical techniques, attempts have been made to estimate the amino acid sequences of the ancestral counterparts of proteins (Risso *et al.*, 2013; Trudeau *et al.*, 2016; Risso *et al.*, 2014). These ancestral sequences are then resurrected *in vitro* where their functions and physical properties can be evaluated. From such studies, it has been observed that, proteins have evolved through amino-acid substitutions (in some case more than 50% positions are substituted (Copp *et al.*, 2019; Wang *et al.*, 2019; Risso *et al.*, 2015)) to develop a new function or modify their activity while conserving their fold. This clearly is in contrast to the typical sequence-structure-function paradigm. Here, we study an alternate dynamics-function paradigm where, as opposed to con-

centrating on a static structure, we focus on the ensemble of structures available to the protein in its native state. By analyzing the equilibrium dynamics of a protein we aim to elucidate the relationship between a protein and its function/activity. These dynamics incorporate the effect of the previously discussed direct as well as indirect interactions between amino acids as dictated by the bio-chemistry of the type of amino acids. For this, we use molecular dynamics simulations to sample the protein dynamics (through covariance in fluctuations between different residues). From the covariance obtained, we use in-house developed tools, namely Dynamic Flexibility Index (DFI) and Dynamic Coupling Index (DCI) to quantify the residue flexibility and the strength of pairwise residue coupling in proteins respectively. The details of the tools used here are described in chapter 2.

Using these tools, in chapter 3, we first focus on the role played by native state dynamics and allosteric interactions in the evolution of Thioredoxin proteins (Modi *et al.*, 2018). For the analysis we used the molecular dynamics simulation trajectories for the reconstructed ancestral and extant Thioredoxin proteins from bacterial, eukaryota and archea branches in the phylogeny. These molecular dynamic simulations were performed by our collaborators while we used this data to obtain the dynamics and thus flexibility profiles of each protein. The results obtained showed overwhelming evidence of how substitutions during evolution have modified the flexibility profile of the proteins and also how dynamic coupling of each residue with the catalytic site has evolved. It further describes the general mechanism used by nature to fine tune the function by a hinge-shift mechanism, where enhancement in the flexibility of a rigid part in the protein is compensated by a loss in the flexibility of another region in the protein.

Afterwards, in chapter 4, we direct our attention to analyze how the dynamics and allosteric interactions between residues impact the evolution of modern proteins. For

this, we studied the emergence of antibiotic resistance in extant TEM-1 β-lactamase. β-lactamases are the bacteria's defense to β-lactam antibiotics. They confer resistance to penicillin and, third generation antibiotics, cephalosporins. We used the data compiled by Salverda *et al.* (2010) which describes antibiotic resistant variants of TEM-1 β-lactamase from clinical and laboratory isolates. First, a majority of the mutations in these variants are located at amino acid positions distal from the catalytic site. Second, through molecular dynamic simulations of these variants, we perform the DFI analysis, and observed that these distal substitutions particularly alter the dynamics of the the catalytic site. Furthermore, using the data from exhaustive mutagenesis experiments on TEM-1 β-lactamase performed by Stiffler *et al.* (2015), we also attempted to decipher how the flexibility and the allosteric coupling of a residue position with the catalytic sites can pre-determine its evolvability towards emergence of antibiotic resistance.

With the knowledge gained from the last two studies about the general principles involved in the emergence of a new function through mutations in evolution, we designed a dynamics-based computational protocol to engineer mutations in an ancestral enzyme such that it mimics the flexibility profile (and hence, the function) of a target enzyme. We use this computational approach in chapter 5 to propose mutations in the sequence of ancestral GNCA β-lactamase to alter its function from a promiscuous enzyme with moderate degrading activity towards cefotaxime and penicillin to a penicillin specific enzyme, catalyzing penicillin with a much higher activity, while showing a very low activity towards cefotaxime. Our computational approach was further biophysically characterized by our collaborators in the study.

Motivated by our understanding of the role played by the generalized view of allostery in protein evolution (or emergence of a new function), we turned our attention to study the implications of allostery in a protein at cellular level. Proteins in a

cell do not function in isolation. Typically, these are cogs in a much more elaborate machinery where communication between different proteins in a cell is established indirectly, either via ions (or a common substrate shared by several proteins) or via the product molecules generated by one which is used by the other.

In chapter 6, we study the first mechanism of allostery at a cellular level by analyzing the dynamics of E. coli ribosome. Ribosomes require a large amount of magnesium ions for the stability and functionality to neutralize the electrostatic repulsion and to spatially coordinate rRNA functional groups. We focused on the impact of a mutation in the L22 protein in E. coli ribosome (called L22* variant) on the association of magnesium ions with the ribosome. Other studies on the impact of mutation suggests that the variant is responsible for providing antibiotic resistance to the bacteria, for example, against aminoglycosides (Sharrock *et al.*, 1981; Thorbjarnardóttir *et al.*, 1978; Buckel *et al.*, 1977; Nessar *et al.*, 2011; Criswell *et al.*, 2006). However, despite the added benefit of antibiotic resistance, the L22* variant of ribosome has not established itself as the "wild type". In order to understand the reason why, we used coarse grained elastic network models to study differences in the dynamics of the two ribosomes through the DFI analysis. We observed that the mutation enhances the association of magnesium ions with the ribosome. The magnesium ions in a cell are also important for biologically activating the ATP molecules by binding with them. The altered association of magnesium ions with the L22* variant disrupts the delicate balance between the magnesium ions bound to the ribosome and to the ATP molecules in the cell, which has an impact on the cell growth. The disruption of the balance is further shown by the kinetic model proposed by our collaborator which also confirms that an increase in association of magnesium ions with the ribosome will reduce the concentration of ATP molecules bound to magnesium ions in the cell. These results are further verified by our collaborators where L22* variant impacts the

concentration of magnesium bound ATP molecules in the cell, hence impairing the growth rate of bacteria. This also provides a possible mechanism for the reason why the L22* mutation is not dominant in nature, as under low extracellular magnesium ion concentration, they exhibit a reduced growth and thereby the cell has to pay a higher physiological cost due to increased sensitivity to environmental conditions. Therefore, in this study, we observed how magnesium interactions within a ribosome are allosterically linked to the distal ATP molecules via cellular magnesium ions. This phenomenon is coined "Ionic Allostery".

We further address how allosteric interactions in a signaling pathway can modulate information among proteins using a simplistic toy model in chapter 7. In the model, an allosterically regulated enzyme generates product molecules which interact with another protein downstream in the signaling pathway. This protein acts as a receiver for the signal transmitted by product molecules. We describe how this type of a system can be reduced to a two state model under the conditions of detail balance, which can be further reduced to a hidden markov model by discretizing time. We also calculate the mutual information (MI) between the state of the enzyme (i.e., its allosteric states) and the state of the receiver protein (i.e., whether it is bound to the product molecule or not). Here MI provides a measure of the strength of regulation the enzyme has on the state of the receiver protein. Therefore, an MI value of zero would imply that the state of the enzyme do not carry any influence over the state of the receiver protein. Using this analysis, we observed that the strength of regulation (MI) depends on the time signature of the product arrivals at the receiver proteins which can be altered with the changes in the kinetics of allosteric regulation. In addition, a positive value of MI is observed when the product formation rate is up-regulated as well as down-regulated. This observation is counter-intuitive as it shows that down-regulation can be a strategy used by nature to efficiently regulate the

signaling pathway.

Finally, in Chapter 8, the results described in the thesis are summarized where we have discussed the success of the computational tools we have used, to characterize the role of allosteric interactions in protein function and how this influences its evolution. We have also addressed the key role played by these interactions in regulating the other molecules in a cell, thereby, showing the ubiquitous influence of allostery in regulation protein functions and cellular mechanisms.

Chapter 2

STUDYING PROTEIN DYNAMICS

2.1 Modeling Interactions in a Protein

As discussed in the previous chapter, in contrast to the sequence-structure-function paradigm, the function of a protein relies on its dynamics. To be more specific, the function is dictated by not only a static native 3D structure of the protein, but also by the ensemble of conformations accessible the protein while it is in its native state. A first step towards studying this ensemble perspective is to quantify the type of interactions between various amino acids in a protein sequence. A variety of *in silico* models have been proposed over time with varying level of complexities and details to model these interactions. These can be broadly classified as all-atom and coarse-grained models.

2.1.1 All-Atom Models.

All-atom models for interactions in proteins make use of semi-classical force-fields where the interaction potential between every pair of atoms is taken into account. In order to reduce the order of complexity, a typical force-field uses a semi-empirical potential where bonded interactions and interaction between bond planes and dihedral angles are represented by harmonic potentials. For showing non-bonded electrostatic and Van der Waal interactions, coulombic and Lenard-Jones potentials are used respectively (Levitt, 1983; Brooks *et al.*, 1983; Weiner *et al.*, 1984). These can be shown

as:

$$U = \sum_{bonded} K_b(b - b_o)^2 + \sum_{Angles} K_\theta(\theta - \theta_o)^2 + \sum_{dihedral} K_\phi(1 + \cos(n\phi - \delta))^2 + \sum_{non-bonded} (\frac{A}{r^{12}} - \frac{B}{r^6} + \frac{q_1 q_2}{Dr}). \quad (2.1)$$

These force-fields are usually parameterized by comparing the free parameters against some other model where experimentally obtained data, in either gas phase or solid phase, is available or through *ab initio* calculations involving quantum mechanical potentials. In the first three harmonic terms, the information of bond lengths, bond angles and dihedral angles is obtained through the crystallographic data. The force constants on the other hand are typically estimated with the help of experimental studies like infrared or Raman spectroscopy (González, M.A., 2011). The parameters for dihedral angles are usually derived through *ab initio* calculations and further refined using experimental data such as molecular geometries or vibrational spectra. The non-bonded term in Eq. 2.1 involves contribution from Van der Waal interactions and electrostatic interactions. Several models exist for representing Van der Waal interactions, e.g., the 12-6 Lennard-Jones (LJ) potential (Jones and Chapman, 1924), Buckingham potential (Buckingham and Lennard-Jones, 1938), etc. For electrostatic interactions a coulomb potential is used, which depends on the molecular electronic density which are not trivial to calculate. Therefore, partial charges are assigned to each atom which are derived from a fit using fitting methods such as RESP (Bayly *et al.*, 1993) and CHELPG (Breneman and Wiberg, 1990) or using *ab initio* calculations and then deriving them from the quantum mechanical potential. For these, a number of computational tools are available which provide data from electronic structure calculations, e.g., the Merck Molecular Force Field (Halgren, 1996), or quantum mechanical force field (Ewig *et al.*, 2001). Given the desired level of detail, these models also allow the protein to be submerged in a water box where each molecule of

water is allowed to interact with the protein and also with itself (Price and Brooks, 2004), or a field based approach can be used where an implicit field of water is applied in order to replicate some of the properties of water (e.g., the dielectric constant, etc.) (Still *et al.*, 1990). The combined system of protein submerged in water typically known as the simulation box. Using the potential energy terms obtained from force-fields, the equations of motion can be solved providing a time series of coordinates of atoms in the protein at equilibrium, or under an applied force. Generally, this involves solving the equations of motion such as Langevin equation (Van Kampen, 2007). It is a stochastic differential equation which includes the Newton's equation of motion (F=ma) along with the contribution of thermal noise, pressure, dissipation, memory effects, etc. Thermal noise and pressure accounts for the presence of an external bath which regulates the temperature and pressure of the simulation box. These are accounted for by the help of specialized algorithms called Thermostats (e.g., Nose-Hoover (Nosé, 1984; Hoover, 1985), Anderson (Andersen, 1980), Langevin, etc.) and Barostats (e.g., Anderson (Andersen, 1980), Berendsen (Berendsen *et al.*, 1984)) respectively. These are stochastic processes which perform the crucial job of coupling a bath to the simulation box, thereby, regulating the pumping in or draining out of the energy from the box. Thus making sure that the temperature and pressure of the simulation box remains constant at a fixed value (thus, coupled to an external bath). Due to the immense complexities and the size of these differential equations, specialized software packages are used for solving these, e.g., AMBER (Pearlman *et al.*, 1995; Case *et al.*, 2014), etc.

For all the analysis performed in the thesis, we have used AMBER molecular dynamics package (Pearlman *et al.*, 1995; Case *et al.*, 2014) for performing equilibrium simulations of the proteins. For each protein, the initial structure was obtained from the protein data bank. If a mutant need to be synthesized, then PyMol's mutagenesis

package (Schrödinger, LLC, 2015) was used to introduce substitutions *in silico* on the initial structure making sure that the initial rotamer with least steric hindrance was selected.

Afterwards, the H++ web server was used to predict the protonation state of the charged side chains (Anandakrishnan *et al.*, 2012; Myers *et al.*, 2006; Gordon *et al.*, 2005). The refined structure was then loaded into TLEAP using the ff14SB force field (Maier *et al.*, 2015). Protein hydrogens were then added and a 9.0 Å cubic box of TIP3P surrounding water atoms was added, followed by neutralizing ions (Jorgensen *et al.*, 1983; Neria *et al.*, 1996; Joung and Cheatham, 2008). The system was then energy-minimized using the SANDER module of AMBER 14 (Case *et al.*, 2014; Pearlman *et al.*, 1995; Salomon-Ferrer *et al.*, 2013). The first cycle of minimization reduced the energy and steric clashes of the solvent while the protein is restricted using harmonic restraints. The second cycle of minimization was then performed without the harmonic restraints, so the entire solution could adjust to the local minimum.

Heating, density equilibration and production were all then run using the GPU-accelerated PMEMD module of AMBER 14 (Salomon-Ferrer *et al.*, 2013). These simulations were performed with periodic boundary conditions and the bond lengths of all covalent hydrogen bonds were constrained using SHAKE (Pearlman *et al.*, 1995). Direct-sum, non-bonded interactions were cut off at 9.0 Å, and long-range electrostatics was calculated using the particle mesh Ewald method (Darden *et al.*, 1993; Essmann *et al.*, 1995; Hockney and Eastwood, 1988). The heating phase ran over 100 ps from 0 to 300 K. The density of the system was then allowed to equilibrate over 5 ns at constant temperature and pressure. All production simulations were run using the Langevin thermostat and Berendsen barostat (Berendsen *et al.*, 1984) to keep temperature and pressure, respectively, constant. A time step of 2 fs was used

and structural conformations were saved every 10 ps.

2.1.2 Coarse-Grained Models.

On the other side of the complexity spectrum, coarse-grained approaches can be used to model interactions in a protein, but with a significantly diminished level of detail. These are designed to capture the salient features of their motion under quasi-static equilibrium conditions. In these models, the atoms in a protein are segregated into several group where each group can be represented by a single node at their center of mass or at the cartesian coordinate of a representative atom in the group (typically the α-carbon (C_α) atom in each amino acid). These atom groups can be constructed depending on a number of factors, such as the amino acid identity, shared dynamics of the participating atoms, etc. The set of nodes obtained subsequently, comprised of the whole protein (with or without water and other atoms around it), are then connected to each other with energy potentials representing the physics of their interaction. One of the most commonly used model to represent proteins in this manner is the Elastic Network Model (ENM), where harmonic springs are used to connect these nodes in a protein. The application of this model was demonstrated by the pioneering work of Tirion (1996), who showed that in order to observe the motion under equilibrium conditions, the far too complex terms in Eq. 2.1 can be approximated by a simple Hookean potential between each pair of atoms as:

$$U = \frac{1}{2}\sum_{ij} K_{ij}(\vec{r_i} - \vec{r_j})^2 \tag{2.2}$$

where, K_{ij} is the force constant depicting the interaction between $i-j$ pair of nodes and $\vec{r_i}$ represents the coordinate vector for ith node in the protein. Through this, one can perform the normal mode analysis (Goldstein *et al.*, 2002) on the protein structure in order to calculate the eigenvectors and their corresponding eigenvalues describing

their motion. This analysis assumes that the protein lies deep inside a harmonic potential minima and that there are no net external forces acting on the system. Under these assumptions, for a system with N interacting nodes, the potential energy function in Eq. 2.2 can be rewritten for the system close to equilibrium conditions as (using Taylor's expansion):

$$U = \sum_{i=1}^{3N} U(x_i^o) + \frac{1}{2!} \sum_{i,j} \left. \frac{\delta^2 U}{\delta x_i \delta x_j} \right|_{x_i = x_i^o x_j = x_j^o} \Delta x_i^o \Delta x_j^o + ... \tag{2.3}$$

In this expression, x_i^o represents the equilibrium position of x_i. Therefore, the first expression can be ignored as it represents the ground state reference energy of the interacting nodes (thereby, taking it as zero). Ignoring higher order terms (considering that their contribution will be of the order of Δx^3 and higher, hence too small), we get:

$$U = \frac{1}{2!} \sum_{i,j} \frac{\delta^2 U}{\delta x_i \delta x_j} |_{x_i = x_i^o x_j = x_j^o} \Delta x_i^o \Delta x_j^o. \tag{2.4}$$

This can be rewritten conveniently using matrix notations as:

$$U = \Delta \vec{X} \mathbf{H} \Delta \vec{X}^T \tag{2.5}$$

where, $\vec{X}$ is the $3N \times 1$ dimensional vector representing the coordinates of the network, and $\mathbf{H}$ is the Hessian of the network of interactions which is defined as an operator of second order derivatives of their potential energy.

$$\mathbf{H} = \frac{1}{2} \sum_{i,j} \frac{\delta^2 U}{\delta x_i \delta x_j}. \tag{2.6}$$

Performing eigenvalue decomposition of Hessian reveals the normal modes of oscillations of nodes within the network of interactions. These normal modes reveal information regarding the direction and amplitude of oscillations in each node in the network (Goldstein *et al.*, 2002).

Through this simplistic model, Tirion (1996) was able to show that with the help of a single-parameter potential, it is possible to accurately reproduce the low frequency dynamics of the globular proteins otherwise obtained from a significantly more computationally expensive all-atom MD simulations. In addition, the model was also able to accurately predict the x-ray crystallographic temperature factors which are otherwise obtained through much more expensive calculations using a multi-parameter potential (Weiner *et al.*, 1984; Brooks *et al.*, 1983; Levitt, 1983).

Several different variations of ENM exist, one of the prominent models is the Anisotropic Network Model (ANM) (Atilgan *et al.*, 2001; Tama and Sanejouand, 2001). In this model, the Elastic Network Model used by Tirion is further reduced in its complexity by using only C_α atoms instead of all the atoms for representing a protein. Here a uniform spring constant is used for joining all residues which are within a cutoff distance from each other (typically 12 to 15 Å (Eyal *et al.*, 2006; Tama and Sanejouand, 2001; Atilgan *et al.*, 2001; Tama and Sanejouand, 2001)) in the native structure of the protein. In their native structure, it can be shown that most proteins can be described by using only a few degrees of freedom– usually those associated with the C_α atoms on the backbone on the protein chain (Tama and Sanejouand, 2001). This is because, using only C_α atoms provides a basis set for the conformational representation of the protein which can not only show the complete spatial extent of fluctuations in a protein but can also be used to re-normalize out the other degrees of freedom in the structure (Thorpe, 2007). This can be further shown by observing the small frequency-dependence of the force-constants for the interaction between these C_α atoms, and hence the success of ANM for predicting the global dynamics of the proteins. This enables the force-constants to be treated as invariable parameters in the low frequency regime, dominating the global motions in a protein, and can be subsequently fitted to experimental variables like Debye-Weller B factors determined

from x-ray diffraction (Thorpe, 2007; Bahar *et al.*, 1997; Tirion, 1996) with a large degree of agreement. Other versions of coarse-grained Elastic Network Models include Gaussian Network model (GNM), Born Model, etc. (Thorpe, 2007; Bahar *et al.*, 1997; Born and Huang, 1998).

Despite the huge success of the coarse-grained ENM on predicting global motions in proteins and their relatively lower computational costs, these suffer from a major drawback. These rely heavily on the information contained in the static native structure in order to build the potential energy function of the protein. It does not account for the biochemical composition of the amino acid sequence of the protein. As a result, it cannot capture the shift in the native conformational ensemble of the protein upon mutations (i.e., substitution, insertion or deletion). Therefore, we utilized MD simulations in order to observe the differences created by the impact of mutations in the amino acid sequence of a protein.

2.2 Metrics for Quantifying the Dynamics of Residues in a Protein

The previous section gave a brief description of several models available to model the potential energy of interactions between atoms in a protein. It also demonstrated how to perform normal mode analysis starting from the potential energy function in order to calculate the eigenvectors and eigenvalues describing the motions in a protein. This type of analysis can provide information regarding the type of motion involved with several domains in a protein (e.g., as shown by the normal mode analysis of a protein (Bahar *et al.*, 2010; Zhang *et al.*, 2019)). Other notable methods to quantify the dynamics of motion of residues in a protein are described below.

2.2.1 Root Mean Square Fluctuations (RMSF).

RMSF, as the name suggests, is the square root of the mean of squared fluctuations in the atomic coordinates of an atom or a group of atoms in a protein structure around their equilibrium position. For an atom or a node i representing group of atoms, the RMSF can be calculated as:

$$RMSF_i = \sqrt{\langle |\delta r_i|^2 \rangle_t}. \tag{2.7}$$

Here, $\delta r_i = (r_i - \langle r_i \rangle)$ is the displacement of node i around its equilibrium position. For the calculation of RMSF, δr_i is averaged over a time window of duration t. RMSF is one of the most commonly used metrics to depict the magnitude of motions due to thermal fluctuation. (Keskin *et al.*, 2000; Maguid *et al.*, 2006; Papaleo *et al.*, 2006). The calculation of RMSF typically restricts itself to the fluctuations in C_α atoms in the backbone of the protein chain. These can be readily calculated using any ensemble of protein conformations obtained by methods like MD simulations or Monte-Carlo (MC) simulations of protein structure. However, calculation of RMSF comes with a major limitation. As one would imagine, it depends heavily on the choice of the equilibrium coordinates. Therefore, the type of sampling and its frequency used to generate conformational ensembles can sway the results dramatically. Moreover, for the calculation of the equilibrium coordinates, the global translational and rotational motions are eliminated by fitting each structure with a reference frame (typically the first frame of the trajectory). This is typically performed by choosing a set of atoms for fitting. As a result, the outcome of fitting can also depend on the choice of the subset of atoms for coordinate fitting.

Furthermore, from the ensemble of conformations, one can also calculate the co-

variance between the fluctuations of C_α atoms in a protein as:

$$\begin{aligned} G_{ij} &= \langle (r_i - \langle r_i \rangle).(r_j - \langle r_j \rangle) \rangle, \\ G_{ij} &= \langle \delta r_i . \delta r_j \rangle. \end{aligned} \tag{2.8}$$

Here, δr_i is the fluctuation in coordinate r_i around its equilibrium mean position. Therefore, if these calculations are performed for a protein with N residues, using only their C_α atoms, then i goes from 1 to $3N$ and G would be a matrix of dimensions $3N \times 3N$. Using Eq. 2.8 one can clearly observe that the RMSF score of residue i can be calculated with the help of the diagonal elements of the covariance matrix as:

$$\begin{aligned} RMSF_i &= \sqrt{\langle |\delta r_i|^2 \rangle_t} \\ &= \sqrt{\langle |\delta r_{xi}|^2 \rangle_t + \langle |\delta r_{yi}|^2 \rangle_t + \langle |\delta r_{zi}|^2 \rangle_t} \\ &= \sqrt{G_{3i,3i} + G_{(3i+1),(3i+1)} + G_{(3i+2),(3i+2)}}. \end{aligned} \tag{2.9}$$

In addition, one can also perform the principal component analysis (Balsera *et al.*, 1996) on the ensemble of conformations sampled and calculate the contribution of specific modes to the RMSF of C_α atoms in a protein. In order to do so, the covariance matrix can be written by expressing it in terms of its principal components as:

$$G_{3N \times 3N} = U_{3N \times 3N} W_{3N \times 3N} V_{3N \times 3N} \tag{2.10}$$

where, U and V are the left and right singular vectors, and W is a diagonal matrix with singular values. Using this, one can reduce the rank of covariance matrix by incorporating the contribution of only a specific set of normal modes (principal components) as:

$${G^*}_{3N \times 3N} = {U^*}_{3N \times p} {W^*}_{p \times p} {V^*}_{p \times 3N} \tag{2.11}$$

where, p are the number of normal modes used in the reduced rank, U*, V* and W* matrices contain only the respective p normal modes and singular values. G* is the

covariance matrix with reduced rank. Finally, Eq. 2.11 can be once again used to calculate the RMSF from the reduced covariance matrix including the contribution of the selected normal modes. However, as Eq. 2.7 suggests, the accuracy of the prediction of fluctuations by this metric relies heavily on the number of conformations sampled by the protein and also on the convergence of the sampling method employed. As a result, the calculation of RMSF which can predict the dynamics of a protein with a reasonable accuracy is computationally very expensive (Tirion, 1996; Fuglebakk *et al.*, 2012; Echave and Fernández, 2010; Maguid *et al.*, 2006).

In order to avoid the computational cost, one can also utilize the slowest global normal modes obtained through static structure through coarse-grained models like ENM, ANM, etc. Using Eq. 2.6, these models can be used to calculate the Hessian describing the potential energy of interactions between various C_α atoms in residues in the protein. Furthermore, under the conditions of quasi-static equilibrium it can be shown that:

$$G_{3N\times 3N} \propto {H^{-1}}_{3N\times 3N}. \tag{2.12}$$

with the proportionality constant depending on physiological parameters like the temperature and the spring constant used for the ANM (Bahar *et al.*, 2010). Afterwards, the steps described in Eq. 2.11 can be followed to reduce the covariance matrix while retaining only the required normal modes and calculate the RMSF profile of the residues in the protein. Several studies performed on coarse-grained and all-atom models of proteins have shown the success of RMSF to predict the x-ray crystallographic B-factors with remarkable accuracy (Tirion, 1996; Echave and Fernández, 2010; Bahar *et al.*, 1998, 2010). However, since the calculation of RMSF is based on capturing the response of stochastic thermal forces on the protein, it lacks the resolution of observing the response to forces at the level of individual residues in a protein chain. For this purpose, Perturbation Response Scanning (Gerek *et al.*,

2009; Nevin Gerek *et al.*, 2013; Gerek and Ozkan, 2010; Atilgan *et al.*, 2010) is utilized which provides the required resolution in observation of response to forces as described in the next section.

2.2.2 Dynamic Flexibility Index (DFI).

Dynamic Flexibility Index (DFI) is a novel metric which is related to the vibrational entropy of residues by calculating the relative resilience each residue experiences to force perturbations in a protein (this is true for coarse-grained ENM, however, if data from MD simulation is used, then it is related to the conformational entropy instead). Computation of DFI utilizes a Perturbation Response Scanning (PRS) technique (Gerek *et al.*, 2009; Nevin Gerek *et al.*, 2013; Gerek and Ozkan, 2010; Atilgan *et al.*, 2010), using unit perturbative forces as probes to sample the local vibrational ensemble of each residue in the protein. Here, the protein is coarse-grained into a network of interacting amino acids represented by nodes at their C_α atoms using ANM. For a protein under overdamped conditions (i.e., net average acceleration is zero), the equation of motion for a node i of mass m_i in a protein under an external force F can be written as:

$$\begin{aligned} m_i \frac{d^2 x_i}{dt^2} &= -\frac{\partial U}{\partial x_i} + F = 0 \\ F &= -\frac{\partial U}{\partial x_i}. \end{aligned} \tag{2.13}$$

Here, U is the potential energy describing the interaction between different nodes of the protein. Therefore, using Eq. 2.6,

$$\begin{aligned} F &= \frac{\partial}{\partial x_i}\left[\frac{1}{2}\sum_{i,j}\frac{\partial^2 U}{\partial x_i \partial x_j}\partial x_i \partial x_j\right] \\ F &= \left[\frac{1}{2}\sum_{i,j}\frac{\partial^2 U}{\partial x_i \partial x_j}\partial x_j\right]. \end{aligned} \tag{2.14}$$

This can be rewritten using convenient matrix notations as:

$$F = H\Delta R \tag{2.15}$$

$$\Delta R_{3N\times 1} = {H^{-1}}_{3N\times 3N} F_{3N\times 1}. \tag{2.16}$$

Here, ΔR is the response vector of residues due to the external perturbative force F. Eq. 2.16 expresses the linear dependence of a response upon an external perturbation under equilibrium condition. This *back-of-the-envelope* calculation describes the basic principle involved in the Linear Response Theory (LRT) (Gerek *et al.*, 2009; Nevin Gerek *et al.*, 2013; Gerek and Ozkan, 2010; Ikeguchi *et al.*, 2005).

Using Eq. 2.16, we can calculate the response of each residue in a protein when unit random Brownian kicks are applied to each individual residue. In order to obtain an isotropic response profile of the protein, several unit random Brownian kicks, distributed uniformly in all directions in space are used, see Figure 2.1. Finally, the average response of every residue upon random perturbations at each residue position can be stored in the Perturbation Response matrix, A as:

$$A = \begin{bmatrix} |\Delta R^1|_1 & \cdots & |\Delta R^1|_N \\ \vdots & \ddots & \vdots \\ |\Delta R^N|_1 & \cdots & |\Delta R^N|_N \end{bmatrix}. \tag{2.17}$$

Here, $|\Delta R^j|_i = \sqrt{\langle (\Delta R_i^j)^2 \rangle}$ is the magnitude of fluctuation response at site i due to the perturbations at site j averaged over a large number of random unit perturbations in different directions. Therefore, the sum of each column along a row i of the perturbation matrix gives the net average displacement of the residue i from its equilibrium position when each individual residue is perturbed by an isotropic unit force one at a time. The DFI score of a residue position i is defined as the ratio of its net response as all the residues in the protein chain are perturbed one by one in a

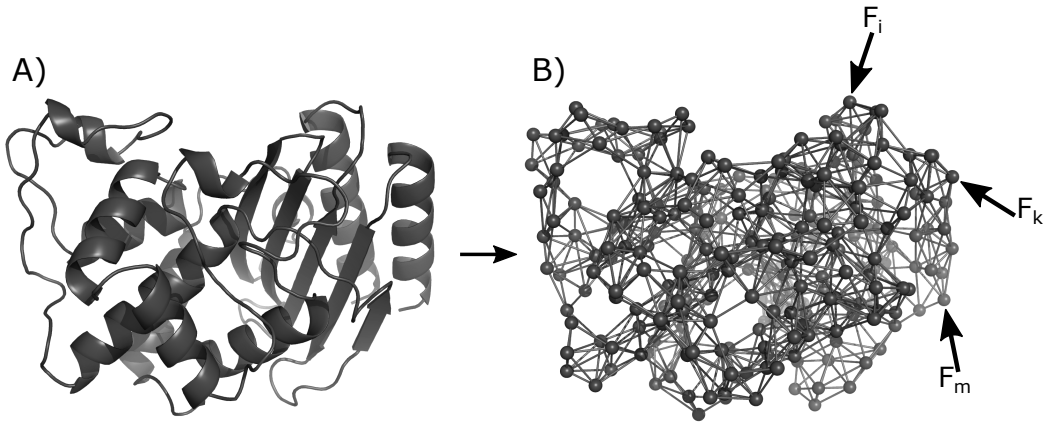

Figure 2.1: The interactions in TEM-1 β-lactamase (shown in cartoon representation on left) are being modeled using harmonic springs in the Elastic Network Model (right). Further on, random Brownian kicks are being applied on C_α atoms (i, k and m) in the network to probe its response profile using Perturbation Response Scanning.

sequential manner and the net displacement of all the residues when each residue is perturbed one by one in a sequential manner. It can be described as:

$$DFI_i = \frac{\sum_{j=1}^{N} |\Delta R^j|_i}{\sum_{i=1}^{N} \sum_{j=1}^{N} |\Delta R^j|_i}. \tag{2.18}$$

Therefore, a residue with a higher DFI score, is more susceptible to random perturbations in the protein and samples the local conformation space more freely, hence is labeled as a "flexible" residue. On the other hand, residues with a lower DFI score, are more resilient to motions in the protein, and are therefore called "rigid" residues.

As discussed earlier, ENM based coarse-grained models are able to accurately capture the global dynamics of the protein (Tirion, 1996; Thorpe, 2007; Zhang *et al.*, 2019). However, as the use of the Hessian matrix suggests, this technique is limited by the use of harmonic approximation on a static three-dimensional structure. As a result, it fails to incorporate the changes in the dynamics of the protein upon changes in the chemistry of the protein by residue substitutions. Therefore, in order to predict the dynamics of the point mutations on the wild type proteins all-atom MD simulations need to be performed. The simulations provide the pairwise correlations

in the fluctuations of alpha carbon atoms in the protein in the form of a covariance matrix (G) which using Eq. 2.12 can replace the Hessian (H) in Eq. 2.16 as:

$$\Delta R_{3N\times 1} \propto {G^{-1}}_{3N\times 3N} F_{3N\times 1}. \tag{2.19}$$

As earlier, the proportionality constant is a function of the force-constants used in the model and temperature. Therefore, it can be treated as a constant for a given network. This enable us to calculate the DFI scores using Eq. 2.18. Since DFI is a relative metric which measures the relative responses only, the proportionality constant is canceled out.

Further on, the metric DFI provides some crucial advantages over the metric discussed earlier– RMSF. As described, RMSF provides the average response of a residue to thermal fluctuations. This implies that the metric is highly susceptible to falling into the trap of insufficient sampling. On the other hand, the use of a perturbative force as a probe reduces that dependency on sampling as it will always capture a given response of perturbations as long as the normal mode corresponding to it is sampled. This provides a better resolution in DFI profile as compared to root mean square profiles. This implies that, while RMSF profile and DFI profile for a protein might share some common features, DFI profile should provide a deeper insight into the dynamics of the protein via relative flexibilities. It should also be noted that, as discussed earlier, since covariance matrices are obtained from MD simulations, these are susceptible to sampling or fitting related artifacts. Since DFI also uses covariance matrices, one must be careful while calculating DFI profiles. In order to ensure that DFI profile is sampled from a well converged energy minima without fitting related defects or sampling defects, we check for convergence of dynamics as discussed in the section below.

Earlier we described how the DFI analysis quantifies the flexibility of each po-

sition through equilibrium dynamics which is related to the conformational (in the case of all-atom MD or vibrational while using coarse-grained ENM) entropy of each residue position in a protein. It is also important to compare the DFI values of each positions with the experimentally measured equilibrium dynamics. Experimental techniques like Nuclear Magnetic Resonance (NMR) (Rabi *et al.*, 1938), Infrared (IR) Spectroscopy (Skoff and Zanni, 2013), etc. can aid in exploring the dynamics of protein motion (i.e., quantifying time scales of fluctuations in bond lengths, twisting/rolling of bonds, etc.). However, the analysis through these depends on the time scales and the type of interactions being explored. Here, we will be describing how NMR technique can be used to provide details on dynamics of protein residues.

For proteins, NMR exploits the magnetic properties of the (called NMR-active isotopes) spin 1/2 isotopes ^{1}H, ^{13}C, ^{15}N and ^{31}P which exhibit magnetic dipole moments, and the spin 1 isotope ^{2}H, which exhibits both magnetic dipole and quadrupole moments (Levitt, 2013; Cavanagh *et al.*, 1995; Rule and Hitchens, 2006). These isotopes can be introduced at specific locations in the protein and used as non-perturbative probes to explore local structure and dynamics (Ohki and Kainosho, 2008). For NMR (specifically solution NMR), a solution containing protein is put under a strong external magnetic field, B_o which aligns the magnetic moment of the NMR-active isotopes in the protein sample. Afterwards, another magnetic field is then applied orthogonally to create perturbations in the magnetic moment of the sample, which undergoes Larmor precession about B_o (Levitt, 2013). The precession rate depends on the local structure and dynamics around the NMR-active isotope. The precessing nuclei in the sample induces a time dependent current which can be detected in the receiver coil of the instrument. The Fourier transform of this current is called free induction decay (FID). The rate of decay of FID can be obtained by fitting it with mono-exponential functions to give the transverse relaxation rate (R_2) as $I(t) = \exp(-R_2 t)$. This

provides information regarding the backbone dynamic of the protein. In addition, the peak in the frequency spectrum of the signal provides information regarding the chemical shift (δ) which reports on the local structure of the protein.

Using NMR, one can also measure the nuclear Overhauser effect (NOE) which describes the degree of cross-relaxation in the structure by quantifying the transfer of nuclear spin polarization from one NMR-active isotope to another (typically ^{1}H-^{15}N along a protein backbone) (Overhauser, 1953). This provides a 2D representation of the chemical shifts of the isotopes resonating. The number of resonances observed, peak resolution and intensity heterogeneity, together, provide information regarding the dynamics of the residues in a protein. In a well-structured and rigid protein, the number of resonances is related to the number of residues (because of relaxation through the backbone). Therefore, a decrease in the number of peaks is an indicator of a more flexible protein. Similarly, a poor chemical shift peak resolution also suggests an enhancement in the flexibility of the protein (Jaudzems *et al.*, 2010). Using these approaches, several different variations of NMR exist depending of the time scale explored, type of perturbative field applied, observable detected, etc. These are– real-time NMR (time scale $>$ 1s)(Zeeb and Balbach, 2004), exchange spectroscopy (time scale$\approx$10-5000ms) (Jeener *et al.*, 1979), lineshape analysis (time scale $\approx$ 10-100ms) (McConnell, 1958), Carr-Purcess Meiboom-Gill relaxation dispersion (CPMG-RD) (time scape$\approx$0.3-10ms) (Palmer 3rd *et al.*, 2001), nuclear spin relaxation (time scale in ps-ns) (Jarymowycz and Stone, 2006), etc.

Of these, CPMG-RD in particular is a very powerful technique to quantify conformational dynamics of a protein. It help explore the dynamic processes with exchange time scales ranging from 0.3 to 10 ms, which include side chain reorientation, loop motion, secondary structure changes and hinged domain movements (Vallurupalli and Kay, 2006; Fraser *et al.*, 2009). In CPMG-RD, a series of perturbative pulses

are applied to the sample containing NMR-active isotope and the observable (FID) is measured as a function of the frequency of these pulses (ν_{CPMG}). Using this, an effective relaxation rate ($R_{EFF}(\nu_{CPMG})$), quantifying the decay of the signal, can be calculated (assuming mono-exponential decay) as a function of ν_{CPMG}. $R_{EFF}(\nu_{CPMG})$ depends (in a nonlinear fashion) on the exchange rates between the ground and excited state of the NMR-active isotope, which is influenced by local structure and dynamics (Meiboom and Gill, 1958; Carr and Purcell, 1954).

2.2.3 Checking for Convergence of Dynamics.

As described by Eq. 2.16, when a Hessian is used to calculate the response of a perturbative force, we are restricting ourselves to a harmonic potential. Therefore, as we sample data from a simulation trajectory in order to calculate the covariance matrix, we are essentially assuming that the data is sampled from a gaussian distribution (because of harmonic potential). In order to achieve appropriate sampling, two of the basic conditions discussed in the sections above have to be met– (i) All conformations sampled must belong to the same distribution. Otherwise, the potential energy well underlying the distribution is different for different configurations in our sample. (ii) The covariance matrix thus obtained, should be independent of the choice of the subset of atoms used for fitting coordinates (in order to find the equilibrium coordinates and eliminate global motions). In order to ensure that these two criteria are met following steps are taken.

1. The trajectory is divided into smaller time windows with different starting points separated by a time lag (25ns for the analysis shown in the thesis). Here different window-sizes are used, corresponding to 25ns, 50ns and 75ns. For each window, we calculate a covariance matrix by fitting all the configurations in the window with its first frame. For fitting, we use only the heavy atoms along the

backbone of the protein chain. It should be specified that, for the analysis, the first 100 to 150 ns of the trajectory is rejected to avoid relaxation artifacts.

2. For each window size separately, we obtain the average DFI profile. The average DFI profile for each window size is then compared with other to see if they exhibit common features (namely, location of low flexibility and high flexibility regions). If all the window-sizes are sampled from the same potential energy minima, then the final average DFI profile should be independent of the window-size, i.e., averaging of DFI profile from 25ns, 50ns and 75ns window size will give similar flexibility profiles. As a result, a consensus profile observed from all different window sizes suggest us consistent and converged dynamics. In addition, this will also ensure that the final resultant DFI profile is independent of the choice of subset of atoms for fitting as the initial coordinates for fitting is different for each time window.

If the condition described above is not satisfied, then we consider the trajectory as not converged and extend the simulation till above described convergence criterion is achieved (see Figure 2.2).

2.2.4 Dynamic Coupling Index (DCI).

Dynamic Coupling Index (DCI) (Butler *et al.*, 2018, 2015; Gerek and Ozkan, 2011; Kumar *et al.*, 2015b; Modi and Ozkan, 2018; Campitelli *et al.*, 2020; Modi *et al.*, 2018), as the name suggests, is a novel metric designed to quantify the strength of coupling between residues in a protein. Similar to DFI it utilizes the principles like PRS and LRT in order to probe the coupling of a residue i with another residue j with functional importance. The DCI score of a residue i with another residue j is defined as the ratio of the total displacement at i when the C_α at residue j is perturbed by

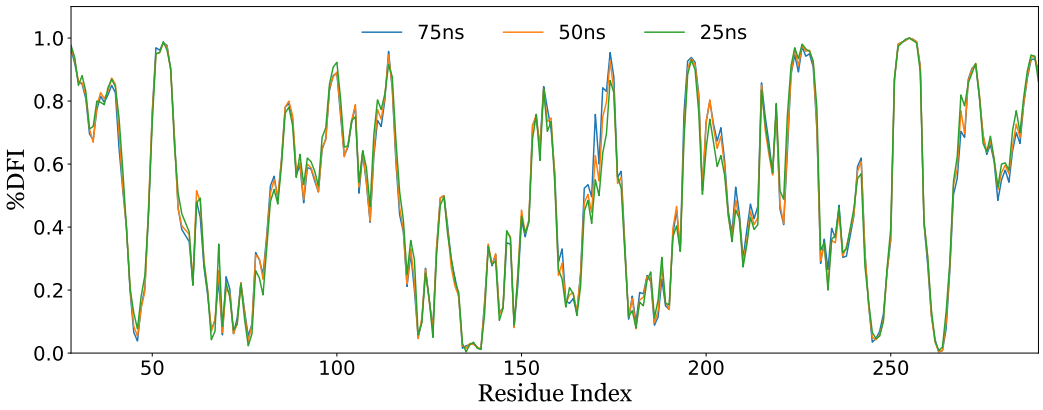

Figure 2.2: Convergence of DFI in a protein from MD simulation. Here, a 1000 ns simulation of a protein (engineered mutant of GNCA β-lactamase) is used to obtain covariance matrices of window size 25ns, 50ns and 75ns. These are obtained by analyzing the trajectory data from 100 ns to 1000 ns. Using covariance matrices of each window size, the DFI profile for the protein is obtained and averaged. These are then compared with each other. Here, we observe that consensus results are obtained for low flexibility and high flexibility regions in the profile, i.e., the results are independent of the selection of window size. Hence the trajectory is considered to be converged.

random unit Brownian kicks, to the average displacement of residue i when all the residues in the protein chain are perturbed by random unit Brownian kicks. It is expressed as:

$$DCI_i^j = \frac{|\Delta R^j|_i}{\sum_{j=1}^{N} |\Delta R^j|_i / N}. \tag{2.20}$$

As earlier, $|\Delta R^j|_i$ is the response fluctuation profile of residue i upon perturbation of residue j. According to Eq. 2.20, a higher DCI score of a residue with another functional sites would imply that perturbation at functionally important sites have a larger impact at that residue as compared to the rest of the protein indicating a higher coupling between the two. On the other hand, a lower DCI score would mean the opposite.

These techniques described above enables us to explore the dynamics-function paradigm introduced earlier. The DFI profile of the protein gives insight into the relative fluctuations along with the DCI score which can quantify the strength of

long-distance allosteric interactions in a protein giving us an edge in understanding the fundamental principles involved in the evolution. This analysis can also benefit from a convergence check to minimize sampling artifacts.

2.2.5 Clustering DFI/DCI Profiles Using Principal Component Analysis.

In order to compare the dynamical differences between three or more proteins, it is important to define a metric which can quantify the magnitude of difference between the dynamic of the proteins. For this, we compare the percentile rankings of the DFI score of each protein by performing principal component analysis. This analysis is used to represent the data in reduced dimensions. This allow us to focus on key parameters (corresponding to highest principal components) which contribute to largest differences between protein.

In this process, firstly, the proteins being compared are aligned according to their multiple sequence alignment. This is an important step if we are comparing the DFI profiles of proteins with unequal number of residues. Therefore, performing sequence alignment ensures that gaps or insertions into the sequence are accounted for and equal number of residues from each protein are considered into the comparison. Afterwards, the DFI profile from all the protein are concatenated into a data matrix X such that each row represents a different protein. Afterwards, Singular value decomposition (SVD), is used to factorize the data into the orthonormal basis, which represents the vector space containing data. The data matrix X is of dimensions $(m \times n)$. Here m is the number of different proteins being clustered together, each having n number of attributes (i.e., the number of residues after multiple sequence alignment). On performing SVD, X is decomposed as follows:

$$[X]_{m\times n} = [U]_{m\times m}[\Sigma]_{m\times n}[V]_{n\times n}. \tag{2.21}$$

Here, U and V are the unitary matrices with orthonormal columns and are called the left singular vectors and right singular vectors, respectively. Σ is a diagonal matrix with elements known as the singular values of X. The singular values of X, by convention, are arranged in a decreasing order of their magnitude. These are $\sigma = \sigma_i$ which represents the variances in the corresponding left and right singular vectors.

The set of highest singular values representing the largest variance in the orthonormal singular vectors can be interpreted to show the characteristics in the data X and the right singular vectors can be used to create the orthonormal basis which spans the vector space representing the data. The left singular vectors contain weights indicating the significance of each attribute in the dataset as:

$$w_i = \sum_{k=1}^{r} \sigma_k |u_{ik}|. \tag{2.22}$$

Here, we use the r largest singular values and corresponding left singular vectors to obtain these weights. Using these features of the decomposed singular vectors, one can create another matrix X^* with a reduced dimension r. Using only the highest r singular values ensures that the data is described by only the most prominent and fundamental features in the dataset (thereby, increasing the noise to signal ratio by eliminating less important features). This is performed as:

$$[X^*]_{m\times r} = [U^*]_{m\times r}[\Sigma^*]_{r\times r}. \tag{2.23}$$

Here, Σ^* is a diagonal matrix containing only largest r singular values and U^* contains the corresponding left singular vectors.

Initially, for a pair of proteins in the data (say, k and l), the distance between them in the original form of data (before dimensional reduction) was given by:

$$d_{kl} = \sqrt{\sum_{i=1}^{n} (X_i^k - X_i^l)^2}. \tag{2.24}$$

This equation can now be rewritten to accommodate the reduced dimensional form of data as:

$$d^*_{kl} = \sqrt{\sum_{i=1}^{r} (X_i^{*k} - X_i^{*l})^2}. \tag{2.25}$$

where, d^*_{kl} represents the distance between proteins k and l in reduced dimensions.

In addition to above, the data can be clustered hierarchically based on the pairwise distance between different proteins in the reconstructed DFI data with reduced dimensions.

Chapter 3

NATURE ALTERS THE NATIVE STATE ENSEMBLE TO EVOLVE FUNCTION

This chapter is adapted from "Modi, T., Huihui, J., Ghosh, K. & Ozkan, S. B. Ancient thioredoxins evolved to modern-day stability–function requirement by altering native state ensemble. Philos. Trans. R. Soc. B Biol. Sci. 373, 20170184 (2018)"

In chapter 1, we provided a general perspective of allosteric interactions in an enzyme which regulate its conformational dynamics to modulate substrate binding affinity and/or its catalytic activity. We reviewed how binding at a distal site, can alter catalytic activity of an enzyme through modulation of dynamics upon change in local network of interactions at the distal binding site. Here, we aim to generalize the concept of allostery by studying how mutations distal from the catalytic site in an enzyme can impact their function through the similar mechanism. Thus, we analyze how such mutations rewires the network of interactions in the enzyme which has a global effect on the dynamics of the catalytic site. In this chapter, we focus on the evolution of Thioredoxin proteins through the lens of allosteric regulations. We examine the equilibrium dynamics of several ancestral and extant Thioredoxin proteins using their flexibility profiles calculated through Dynamic Flexibility Index analysis. Particularly, we analyze how the flexibility profiles evolve with time and also how this impacts the network of interactions in the enzyme by studying the changes in the dynamic coupling of the catalytic site with the rest of the enzyme through Dynamic Coupling Index.

These analyses shed light onto the complex relationship between the native state

ensemble and the stability-function of the enzymes giving an insight into how they adapt to ambient conditions and alter their function. Furthermore, the results suggest that nature sculpts the native ensemble by substituting the residues with medium flexibilities to adapt and alter the function. Comparison of how the flexibility profile of the residues differs between ancestral and extant proteins provides a plausible molecular mechanism– hinge-shift mechanism where the increased flexibility of specific regions (e.g., $\alpha 3$ helix in LBCA Thioredoxin) is compensated by the decreased flexibility of another distal region (e.g., $\alpha 4$ helix in E. coli Thioredoxin). For the analysis, the dynamics of Thioredoxin proteins were obtained via all-atom molecular dynamic simulations performed by our collaborator.

3.1 Abstract

Thioredoxins (Thrx) — small globular proteins that reduce other proteins — are ubiquitous in all forms of life, from Archaea to mammals. Although ancestral Thrx share sequential and structural similarity with the modern-day (extant) homologues, they exhibit significantly different functional activity and stability. We investigate this puzzle by comparative studies of their (ancient and modern-day Thrxs) native state ensemble, as quantified by the dynamic flexibility index (DFI), a metric for the relative resilience of an amino acid to perturbations in the rest of the protein. Clustering proteins using DFI profiles strongly resemble an alternative classification scheme based on their activity and stability. The DFI profiles of the extant proteins are substantially different around the $\alpha 3$, $\alpha 4$ helices and catalytic regions. Likewise, allosteric coupling of the active site with the rest of the protein is different between ancient and extant Thrxs, possibly explaining the decreased catalytic activity at low pH with evolution. At a global level, we note that the population of low-flexibility (called hinges) and high-flexibility sites increases with evolution. The heterogeneity

(quantified by the variance) in DFI distribution increases with the decrease in the melting temperature typically associated with the evolution of ancient proteins to their modern-day counterparts.

3.2 Introduction

Proteins are one of the unique biological machines which have the ability to undergo evolution with time. These evolve to adapt to the ever-changing ambient conditions to ensure the survival of the organism. It involves fine-tuning their pre-existing functional activity or developing new ones altogether. Clearly, the mechanism of the underlying evolutionary process is an interesting question which begs to be answered.

From the observation of evolutionary trajectories of proteins and their functions (Zou *et al.*, 2015; Risso *et al.*, 2014, 2013; Risso and Sanchez-Ruiz, 2017; Ingles-Prieto *et al.*, 2013; Harms and Thornton, 2013b), it can be inferred that the modern proteins have evolved through a series of small changes from ancient times. Much of this information is encoded in protein classes from different species in the three kingdoms of life (Bacteria, Archaea and Eukarya). With advances in the phylogenetics and DNA-synthesis techniques, various ancient genes, including those from the last common ancestors of bacteria, bilaterian animals and vertebrates, have been resurrected in the laboratories. These studies have provided crucial insights on the environmental adaptations and the evolution of functions (Carroll *et al.*, 2008; Ortlund *et al.*, 2007; Perez-Jimenez *et al.*, 2011; Wilson *et al.*, 2015; Bar-Rogovsky *et al.*, 2013; Smith *et al.*, 2013; Boucher *et al.*, 2014; Ingles-Prieto *et al.*, 2013): (i) ancestral proteins are more robust, showing high thermal and chemical stability (Trudeau *et al.*, 2016; Akanuma *et al.*, 2013; Hart *et al.*, 2014; Zou *et al.*, 2015; Risso *et al.*, 2015) and, more interestingly, (ii) the protein structures are highly conserved, much more than the protein sequences throughout the molecular evolution (Zou *et al.*, 2015; Risso

et al., 2015; Bridgham *et al.*, 2009; Choi and Kim, 2006). Thus, the current challenge in molecular evolution is to understand the molecular mechanism of how nature alters the function and biophysical properties through amino acid substitutions, while conserving the 3D structure.

Several biophysical studies exploring the nature of interactions in a protein (Tirion, 1996; Thorpe, 2007) have revealed that all the residue positions in a protein are dynamically linked to each other, thereby creating a network of interactions, where the strength of each link varies across the protein. This network of interactions leads to intrinsic fluctuations in the network due to a combination of the stochastic nature of thermal noise and the external forces originating from the protein molecule's interactions with solvent and other solute molecules. These fluctuations are encoded in the structure of the protein as well as the sequence —that govern the protein's function (Wilson *et al.*, 2015; McLeish *et al.*, 2015; Bahar *et al.*, 2010, 2007; Boehr *et al.*, 2006; Tobi and Bahar, 2005; Mazal *et al.*, 2017; Kar *et al.*, 2010; Tsai and Nussinov, 2014; Tokuriki and Tawfik, 2009a; Henzler-Wildman *et al.*, 2007; Zheng *et al.*, 2006; Dima and Thirumalai, 2006; Liu *et al.*, 2007; Gruber and Horovitz, 2016; Buchenberg *et al.*, 2017; Sawle *et al.*, 2017). Therefore, it is pertinent to account for the contribution of these fluctuations in the protein's structure while studying its function. As a result, the obsolete view of the single native protein structure has long been replaced by 'an ensemble of sub-states' that – as a whole – accurately represents the dynamics of the native state (Tokuriki and Tawfik, 2009a). To further shed light onto the mechanism of evolution and study how evolution shapes the native ensemble, we utilize Dynamic Flexibility Index (DFI) analysis (Nevin Gerek *et al.*, 2013).

The DFI is a position-specific metric that quantifies the resilience of a given residue (amino acid) to the perturbations occurring at various parts of the protein using linear response theory (Nevin Gerek *et al.*, 2013). Thereby, it mimics the multi-

dimensional response of the protein when it's conformational space is probed upon interaction with small molecules or other cellular constituents. The DFI is related to a residue's relative contribution to the conformational entropy of the protein. Since it is a position-specific metric, it can also be used to quantify the change in flexibility per residue position throughout evolution. The DFI identifies flexible and rigid residue positions (or sites) within the 3D interaction network of the protein structure. The low DFI sites are labelled as rigid sites (i.e., hinge sites) and are robust to perturbations owing to their interactions to other residue positions within the 3D structure. They efficiently transfer the force due to perturbations to the rest of the protein chain, similar to joints in a skeleton. Thus, they play a critical role in the conformational dynamics, and usually correspond to the functionally critical and conserved residue positions in a protein (Nevin Gerek *et al.*, 2013). On the other hand, high DFI sites are flexible in nature, thus mutation on these highly flexible sites are typically more acceptable or neutral for the function. A previous study of the DFI analysis on over 100 human proteins has shown that there is a strong positive correlation between the DFI score of each residue position and their evolutionary rates (Nevin Gerek *et al.*, 2013). This demonstrates that the rigid residue positions are more conserved, while the highly evolving residue positions corresponds to sites with high flexibility (Nevin Gerek *et al.*, 2013). The DFI analysis of evolution of different protein families including GFP proteins (Kim *et al.*, 2015), β-lactamase inhibitors (Zou *et al.*, 2015) and nuclear receptors (Glembo *et al.*, 2012) has shown that manipulations in the conformational dynamics of the catalytic region in the protein either through substitutions in nearby region or through distal allosteric regulations leads to functional changes in the protein. Furthermore, we also utilized another site-specific and dynamics-based metric, Dynamic Coupling Index (DCI) which quantifies the strength of coupling between two different residue positions. It is used to analyse

the dynamically coupled residues, which form an allosteric communication network with the active sites, thereby, regulating the enzymatic function. It has been observed that evolution uses substitutions at such residues to regulate the dynamics of the active sites and/or the binding interface (Kim *et al.*, 2015; Kumar *et al.*, 2015a).

In this study, we focused on the evolution of Thioredoxins (Thrx). Thrxs are versatile, small globular protein molecules comprising about 108 amino acid residues. They belong to a class of oxidoreductase enzymes present in all living organisms from Archae-bacteria to humans. These are labelled as the ultimate moonlighting proteins with functions as ubiquitous as being a reducing agent for other proteins in biological reactions (Holmgren, 1985; Romero-Romero *et al.*, 2016).

Thrx proteins from Archaea to humans share about 27–69% sequence similarities and a common 3D fold with a central β-sheet core surrounded by four α-helices (Risso *et al.*, 2015; Romero-Romero *et al.*, 2016). The structure of the reduced and oxidized states of Thrx has been studied extensively over the last few decades. Their structures contain a highly conserved functional site comprising of two neighbouring redox-active cystines, Cys-Gly-Pro-Cys (CGPC) (Eklund *et al.*, 1991; Weichsel *et al.*, 1996). These functional cystines play a key role in Thrx function by participating in the redox reactions. In the reduced state, these two cystines exist with a "thiol" group attached to each. This reduced state is maintained using another class of reducing agents called thioredoxin reductases like nicotinamide adenine dinucleotide phosphate, flavin adenine dinucleotide (FAD), etc. (Mustacich and Powis, 2000; Eklund *et al.*, 1991; Arnér and Holmgren, 2000). The reduced state of Thrx facilitates the reduction of the target protein while turning itself into an oxidized state with the cystines reduced to create a "disulfide" bond between its two functional cystines by losing an electron.

Although, the oxidized state and the reduced state of the protein are very similar to each other in their fold (with most of the differences localized around the

disulfide active site), the chemical denaturation (Chakrabarti *et al.*, 1999) and thermal (Ladbury *et al.*, 1994; Godoy-Ruiz *et al.*, 2004) unfolding experiments suggest that Thrxs are more stable in their oxidized state. The mutagenesis analyses of Thrxs have shown that the C-terminal α-helix, $\alpha 4$, is known to play a critical role in the folding kinetics of the protein (Vazquez *et al.*, 2015). A shorter helix, $\alpha 3$, has been identified as important for the thermal stability of the fold, having several stabilizing mutations (Cabrera *et al.*, 2017). The central core with β-strand $\beta 5$ acts like a bridge for interaction between the two α-helices (Cabrera *et al.*, 2017) (see Figure 3.1).

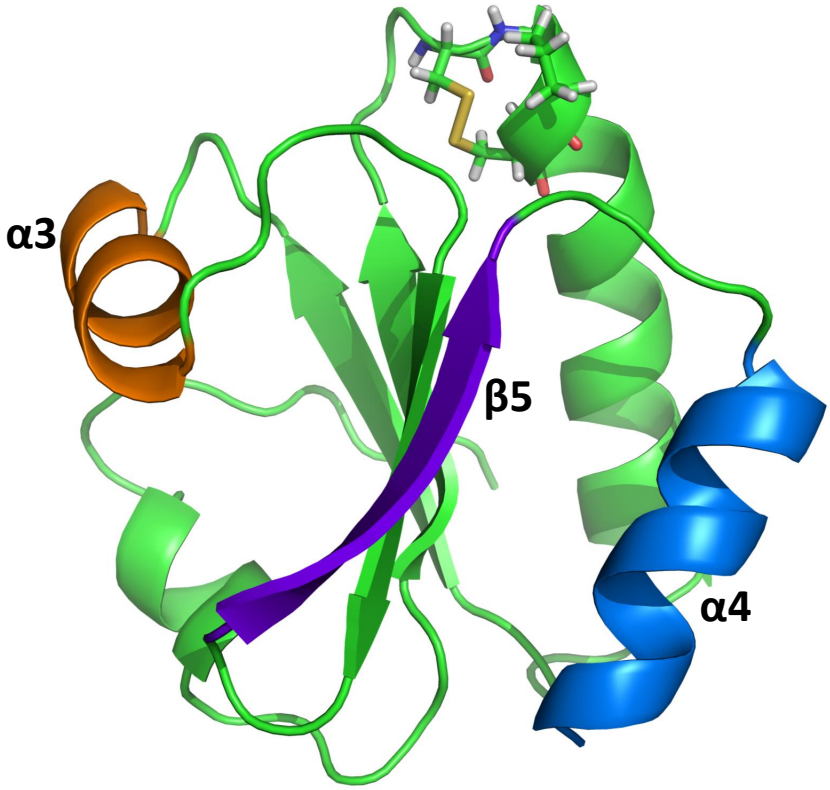

Figure 3.1: The cartoon representation of Archea Common ancestor Thrx (AECA). The active sites, CGPC, are shown in stick representation; the β-strand, $\beta 5$, is shown in purple; and the α-helices, $\alpha 3$ in orange and $\alpha 4$ in blue.

Despite a significant difference in the sequences between the ancestral and extant Thrxs, the ancestral Thrxs, which existed almost 4 billion years ago, display the canonical Thrx fold with only minimal structural changes (Romero-Romero *et al.*,

2016; Risso *et al.*, 2014; Perez-Jimenez *et al.*, 2011). However, during 4 billion years of evolution, the stability of Thrxs has decreased in order to adapt to the cooling temperatures of the Earth (in mesophilic lineage). Moreover, the catalytic rates of ancestral Thrxs at low pH are different from those of extant one (Romero-Romero *et al.*, 2016).

In summary, Thrxs achieved adaptation to a cooler and less acidic Earth by altering their stability and changing their catalytic rates while maintaining the same 3D fold. Therefore, in order to provide a deeper insight into how variations in the sequence alters the stability and the function while conserving the 3D structure, we explore how the native state ensemble of Thrx has evolved throughout evolution. To this aim, we performed 1 μs long molecular dynamics simulations for all ancestral and extant Thrxs, and using these obtained their DFI profiles. Comparison of the DFI profiles between ancestral and extant Thrxs on each branch shows a common pattern. The α3 helix, which contributes the most to stability, exhibits enhanced flexibility in modern Thrxs, correlating with the decrease in stability observed in modern enzymes. The increased flexibility of α3 was also associated with increased rigidity in the α4 helix along with conserved rigidity of the β-sheet core. This may have helped the protein to maintain the 3D fold through the evolution of Thrxs. Further on, The clustering of the DFI profiles of all nine Thrxs and mapping them onto a two-dimensional landscape of experimentally measured stability and catalytic rates suggest that the native ensemble of Thrxs holds the clue to adaptation at a cooler ambient temperature and lower acidic environment.

3.3 Methods for Modelling Interactions and Obtaining Protein Dynamics

All starting structures for Thrx proteins were taken from the Protein Data Bank using the respective accession numbers for the oxidized form of the proteins– Last

Bacterial Common Ancestor (LBCA) Thrx (PDB id: 4BA7), Archaea-Eukaryotes Common Ancestor (AECA) Thrx (PDB id: 3ZIV), Last Animal and Fungi Common Ancestor (LAFCA) Thrx (PDB id: 2YPM), Last Eukaryotes Common Ancestor (LECA) Thrx (PDB id: 2YOI), Last Archaea Common Ancestor (LACA) Thrx (PDB id: 2YNX), Last Gamma- Proteo-bacteria Common Ancestor (LGPCA) Thrx (PDB id:2YN1), Last Common Ancestor of the cyano-bacterial, deinococcus, and thermus groups (LPBCA) Thrx (PDB id: 2YJ7), Escherichia coli (E. coli) Thrx (PDB id: 2TRX) and Human Thrx (PDB id: 1ERU) (Risso *et al.*, 2013; Katti *et al.*, 1990; Capitani *et al.*, 1998). Afterwards, $1\mu s$ long MD simulations are performed following the protocol described in section 2.1.1.

Following the molecular dynamics simulations, the flexibility profiles of each protein was obtained by calculating their DFI profiles following the method described in section 2.2.2, making sure that converged dynamics are obtained in accordance to section 2.2.3. In addition, we also obtained the coupling of each residue with the catalytic sites by calculating the DCI profiles of each protein with respect to the catalytic site residues (described by residues "CXXC") following the method described in section 2.2.4. For this analysis, we used 50ns windows for covariance matrices sampled from 500ns - 1000ns time slot of the trajectory.

3.4 Results and Discussion

3.4.1 The Change in Dynamic Flexibility Index Profiles Provides Insight on Decrease in Melting Temperatures During Evolution of Thioredoxins.

Comparing the DFI profiles of the ancestral and extant Thrxs, marked differences between the flexibility profiles are observed. Particularly, in the flexibilities of helices $\alpha 3$ and $\alpha 4$. Through mutagenesis analysis, it has been shown that disruption of $\alpha 3$

impacts the overall stability of the Thrxs (Cabrera *et al.*, 2017). Interestingly, the enhancement in the flexibility of $\alpha 3$ in E. coli Thrx as compared to the flexibility profile of LBCA Thrx (see Figure 3.2) shows the flexibility–rigidity compensation between the $\alpha 3$ helix and the $\alpha 4$ helix. Specifically, the $\alpha 3$ helix in E. coli has substantially higher DFI values when compared with LBCA, while the opposite is seen in the $\alpha 4$ helix. However, the core region of the β-sheet does not show any noticeable alteration in flexibility between E. coli and LBCA apart from a slight enhancement in rigidity of the $\beta 5$ β-strand.

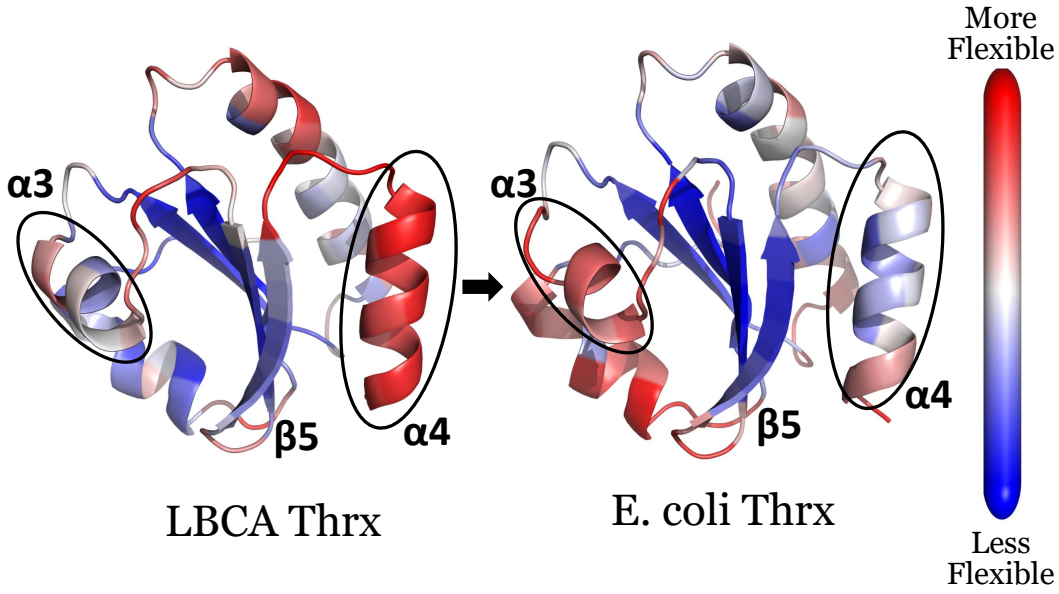

Figure 3.2: The Cartoon representations of ancestral and extant Thrxs from bacterial thioredoxins LBCA and E. coli, respectively, colour coded with their DFI profiles. Red sites are the most flexible and blue sites are the least flexible (i.e., rigid). A shift in the flexibility profiles of the helices $\alpha 3$ and $\alpha 4$ is observed through evolution owing to changes in the thermal stability. By contrast, the β-sheet core remains rigid throughout evolution, highlighting its importance in mediating the interaction between $\alpha 3$ and $\alpha 4$ helices necessary for the Thrx fold.

These observations can be further correlated with the measured differences in the stability due to differences in amino acids in these regions. After performing a sequence alignment between E. coli, LBCA and LPBCA Thrx, it is noted that the ancestral protein LPBCA Thrx (a close homologue of LBCA Thrx) has a critical

mutation, P68A (E. coli Thrx position 68 versus the LPBCA Thrx aligned position), in the α3 helix region when compared with E. coli Thrx. This substitution is typically stabilizing (Cabrera *et al.*, 2017). In the context of two other critical mutations, G74S and K90 L, seen in LPBCA Thrx, it has been hypothesized that the α3 helix may tilt towards arginine (R) 89 (another mutation in LPBCA Thrx), allowing for a stronger charge–dipole interaction (Cabrera *et al.*, 2017). However, E. coli Thrx does not allow such favourable interactions owing to the different amino acids in these positions. Particularly, 89 is threonine (T) instead of arginine (R) in E. coli Thrx. Thus, the loss of this favourable interaction may explain the lowered stability of E. coli Thrx and the enhanced flexibility of the α3 helix as evident from their DFI profiles. This hypothesis was also tested in the context of LBCA Thrx. Similar to LPBCA Thrx, we note that LBCA Thrx also has all the three critical mutations, i.e., P68A, G74S and K90L. Furthermore, after alignment, it was also observed that position 89 has a positively charged amino acid, lysine, in LBCA Thrx.

Motivated by these similarities between LBCA and LPBCA Thrx, we further analysed the relative orientation between the α3 helix dipole and the side chain of lysine (K) (at the equivalent of position 89, after considering sequence alignment) in LBCA. It was observed that the relative orientation between the α3 helix dipole and K87 is more rigid in LBCA Thrx, whereas T, at position 89 in E. coli Thrx, has a fluctuating orientation with the α3 helix (see Figure 3.3). This difference indicates that a favourable interaction between the charge and the dipole (of the α3 helix) may be operative in LBCA Thrx, similar to LPBCA Thrx and responsible for the higher stability of LBCA Thrx and also the rigidity of its α3 helix when compared with E. coli Thrx.

Turning the attention to the α4 helix, guided by mutagenesis experiments (Cabrera *et al.*, 2017), a set of key substitutions between E. coli and LPBCA Thrx were

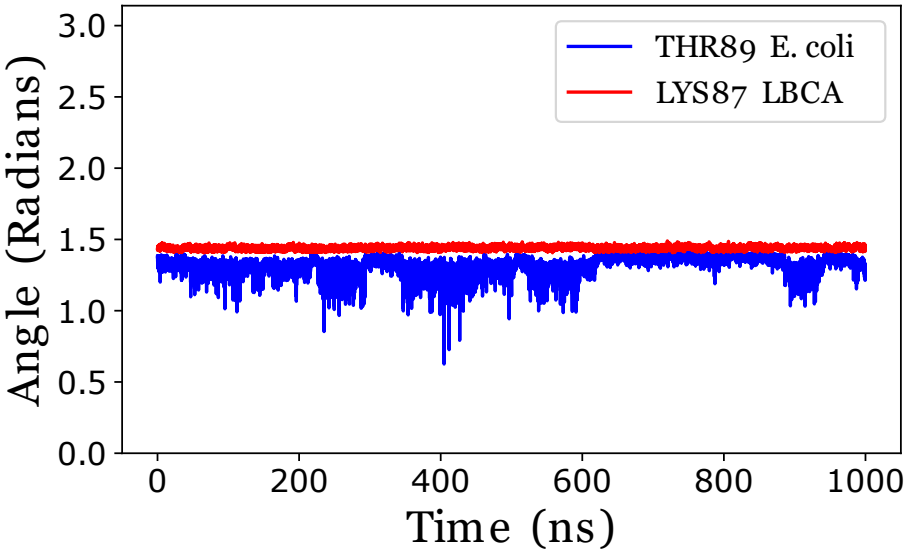

Figure 3.3: The time evolution plot of the angle (y-axis) between the specific position in β-strand $\beta 5$ and the $\alpha 3$ helix dipole over the course of the simulation (x-axis) for E. coli Thrx (in blue) and LBCA Thrx (in red). This position is K87 in LBCA Thrx which creates a stable charge-dipole interaction, leading to a stable helix for LBCA Thrx. Same position based on alignment corresponds to T89 in E. coli Thrx, thus substitution from Lysine to Threonine leads to more flexible $\alpha 3$ helix as shown by high fluctuation of the angle between dipole moment of $\alpha 3$ helix and the position T89 in E. coli Thrx.

observed which were responsible for the folding stability differences between the two proteins. For example, S95P and Q98A—both occurring near the end of the $\alpha 4$ helix in LPBCA Thrx—cause a reduction in stability. It has been hypothesized that both the serine (S) and glutamine (Q) in E. coli Thrx may be responsible for stabilizing the loop connecting the $\alpha 4$ helix and the $\beta 5$ β-strand by possibly using dipole–dipole interactions. However, upon mutating to alanine (A) (for Q98A) and proline (P) (for S95P) in LPBCA Thrx, these dipolar interactions are lost, thereby causing a decrease in their stability and consequently creating high flexibility regions. A second set of mutations, L94Q and F102R (also present in the $\alpha 4$ helix region), in LPBCA Thrx have been further implicated for lowering the melting temperature by possibly destabilizing the hydrophobic network of interactions between the $\alpha 4$ helix and the $\beta 5$

β-strand (Cabrera *et al.*, 2017). All these mutations are also present in LBCA Thrx, after taking the sequence alignment into consideration, apart from L94R, present in LBCA Thrx as opposed to L94Q in LPBCA Thrx. This suggests that the same acting physical principles, i.e., loss of charge–dipole or dipole–dipole interaction and/or disruption of the hydrophobic network, may be responsible for the higher flexibility of the α4 helix in LBCA Thrx compared with E. coli Thrx, as observed by DFI analysis (Figure 3.2).

Similar changes, particularly the compensation for the change in DFI profiles of the α3 and α4 helices, have also been observed in the evolutionary branch of human Thrxs when the DFI profiles of AECA Thrx and the extant human Thrx were compared (Cabrera *et al.*, 2017) (Figure 3.4). Overall, based on the mutagenesis analysis (Cabrera *et al.*, 2017) and DFI comparison between ancestral and extant Thrxs, a plausible mechanism of how Thrx has evolved to a lower stability is suggested. While the increased flexibility of the α3 helical region achieved a decrease in stability to adapt to cooler ambient temperatures, the rigidity conservation of the core and increased rigidity in helix α4 ensured conservation of the canonical Thrx fold throughout evolution. The proposed mechanism should be further verified with experimental analysis in the future.

3.4.2 Dynamic Coupling of the Active Site Changes Throughout Evolution.

When the catalytic rates of the last ancestral Thrx and their extant variant are compared on each branch in phylogeny, it shows that the kinetic rates of disulfide bond reduction has decreased during the evolution at pH 5 (Perez-Jimenez *et al.*, 2011). Particularly, in the human branch, there is an approximately six-fold decrease in the kinetic rates between its first ancestor, AECA Thrx, and the modern-day extant Human Thrx. Earlier work on protein evolution shows that nature uses distal sites

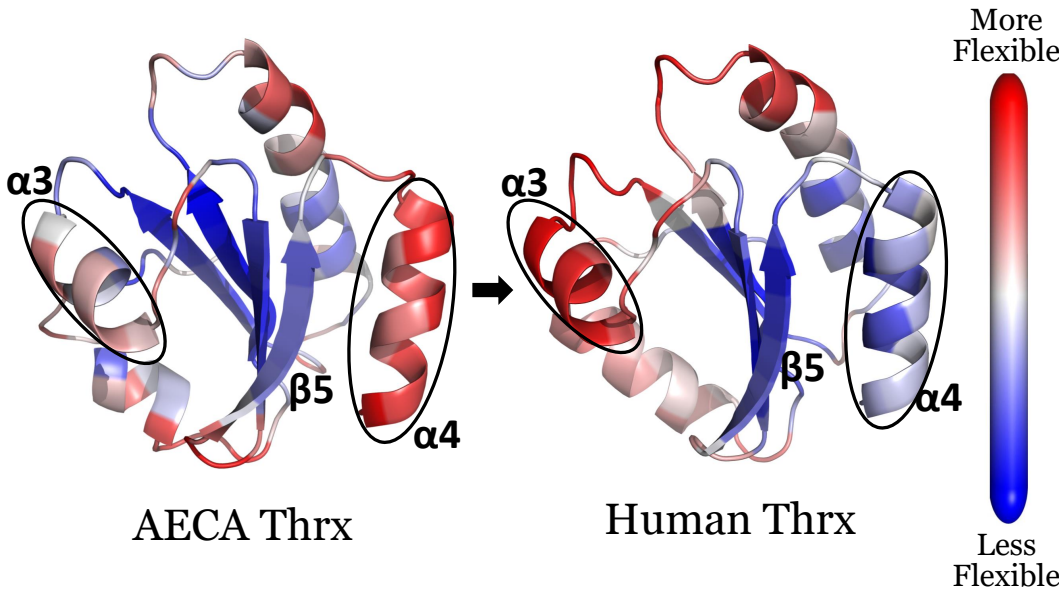

Figure 3.4: The Cartoon representations of ancestral and extant Thrxs from Human branch AECA and Human Thrx respectively, color coded with their DFI profiles, red being the most flexible sites and blue as the least flexible (i.e., rigid) sites. Through evolution, we observe a change in the flexibility profiles of α helices $\alpha 3$ and $\alpha 4$ owing to changes in thermal stability and kinetics. However, the β-sheet core, particularly $\beta 5$, remains rigid.

that are dynamically coupled with the active sites to allosterically control the active site's dynamics (Larrimore *et al.*, 2017; Kumar *et al.*, 2015b; Butler *et al.*, 2018). Therefore, the DCI analysis was used to observe how the coupling of the active site with the rest of the protein changed during Thrx evolution.

Interestingly, it was observed that the dynamic coupling of the $\alpha 3$ helix with the active site had decreased drastically between the ancestral and extant Thrxs in the human branch (Figure 3.5). A similar difference between the ancestral and extant Thrxs in the bacterial branch (Figure 3.6) was observed. Previous studies have also shown that the $\alpha 3$ helix is part of the substrate-binding region and the change in dynamics of this binding region plays a critical role in the catalytic activity of the protein (Perez-Jimenez *et al.*, 2009). Thus, the decrease in the allosteric dynamic coupling of the $\alpha 3$ helix region with the catalytic site may be linked with the lowered

activity observed in modern-day Thrxs. However, careful mutagenesis experiment is needed to conclusively prove or disprove this hypothesis.

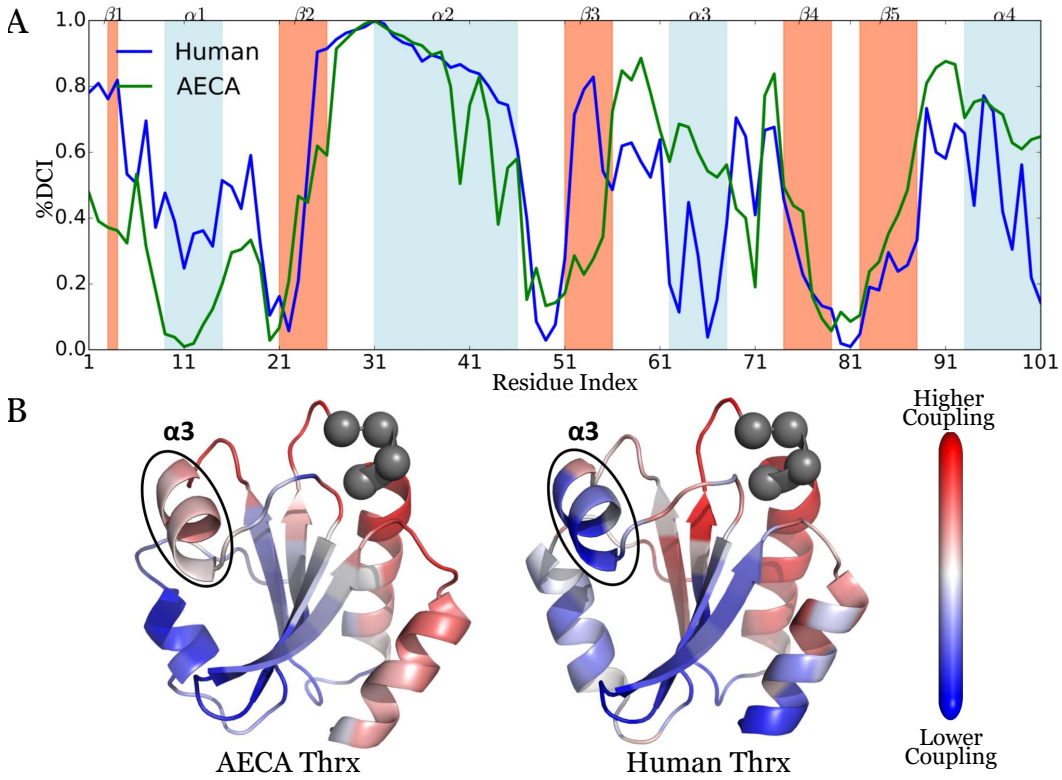

Figure 3.5: Comparison of the coupling of catalytic sites (CGPC) in ancestral (AECA) with extant Human Thrx using a percentile ranking of the dynamic coupling index (%DCI) (A). A striking difference in the couplings of the $\alpha 3$ region in the respective proteins with their corresponding active sites is observed. In KBCA Thrx, the helix $\alpha 3$ is highly coupled to the catalytic site with %DCI>0.8 and on the other hand, it is not coupled to the catalytic site in E. coli Thrx with %DCI<0.8. This suggests the role of coupling of $\alpha 3$ helix in altering the catalytic rates. This difference can also be visualized on the cartoon representations (B) of AECA and Human Thrxs where red represents sites highly coupled to the active sites, CGPC (grey spheres) and blue as sites with no significant coupling.

3.4.3 The Variance in Dynamic Flexibility Index Profile Distribution Correlates With Change in Melting Temperatures.

The DFI analysis of an ancestral versus extant Thrxs in the E. coli and human branches provides an insight on how the 'fine-tuning' of flexibility profiles of some

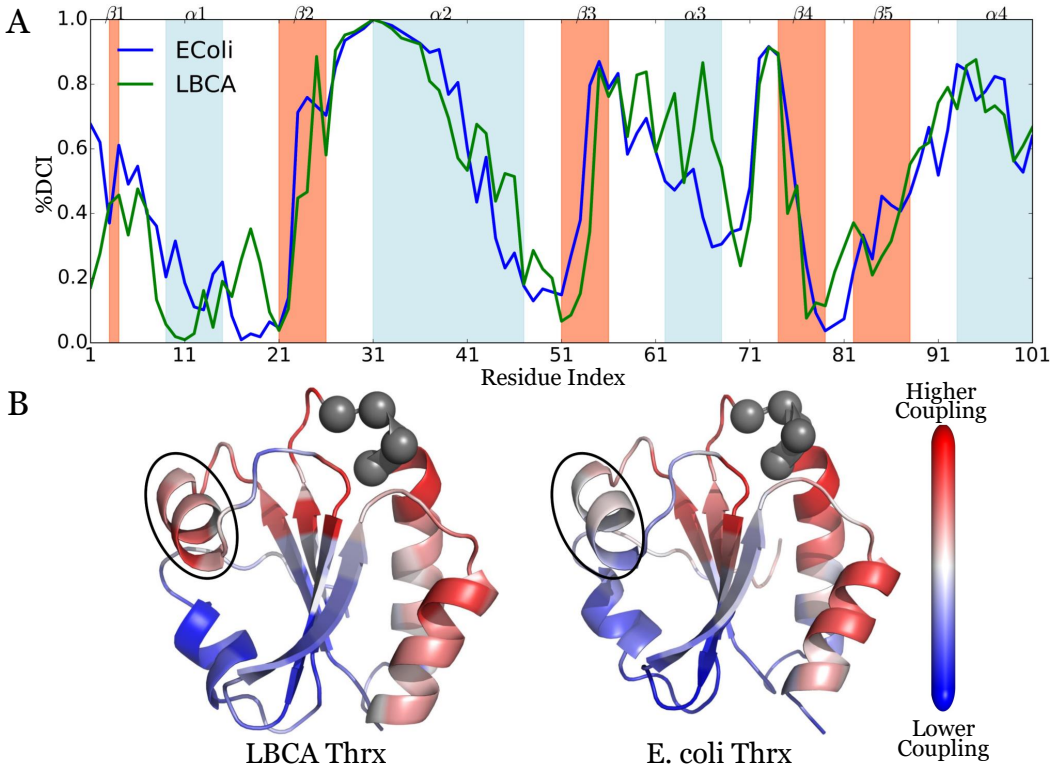

Figure 3.6: Comparison of the coupling of catalytic sites (CGPC) in ancestral (LBCA) with extant E. coli Thrx using a percentile ranking of the dynamic coupling index (%DCI) (A). A striking difference in the couplings of the $\alpha 3$ region in the respective proteins with their corresponding active sites is observed, suggesting their role in altering the catalytic rates. This difference can also be visualized on the cartoon representations (B) of LBCA and E. coli Thrxs where red represents sites highly coupled to the active sites, CGPC (grey spheres) and blue as sites with no significant coupling.

functionally important structural features is used during evolution. Comparing the distribution of flexibility of various residue positions in ancestral Thrxs with the extant ones can additionally reveal information about the change in their native landscape through evolution. The distribution of the flexibility of residues in each Thrx protein is obtained by binning the residues according to their DFI scores (Figure 3.7).

Interestingly, in the case of evolution of Thrx proteins in the bacterial branch of the phylogenic tree, it is observed that the distribution of the DFI profile for the 4-Gyr-old Thrx (from LBCA) differs from that of extant Thrx (from E. coli)

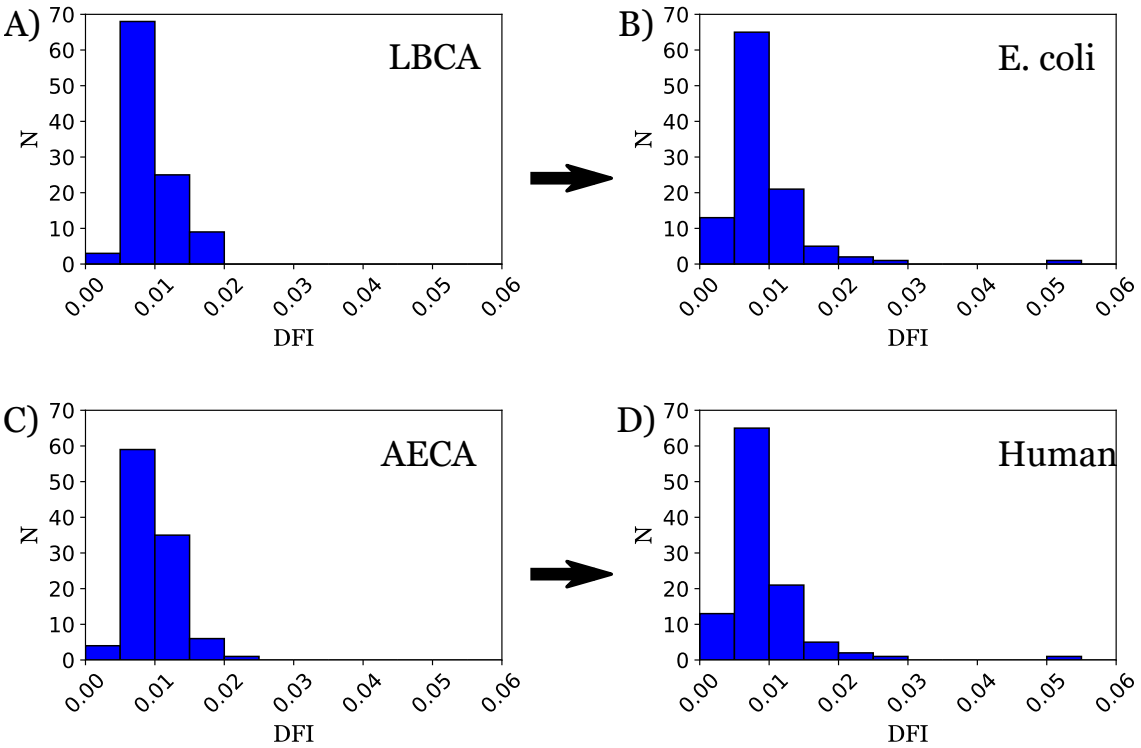

Figure 3.7: The distributions of DFI profiles in the ancestral and extant Thrx proteins belonging to the ancestor of the bacterial branch, LBCA (A), evolving to E. coli (B), and the ancestor of the human branch, AECA (C), evolving to extant human Thrx (D). It is observed that through evolution more residues populate low-flexibility and high-flexibility regions in the distribution, making it wider. The fraction of residues populating high-flexibility (DFI greater than 0.02) regions increased in E. coli from 0% to approximately 4% and in humans from 1% to 4%.

(Figure 3.7A,B). The high flexibility tail region of the distribution (i.e., high DFI) has a higher population in modern Thrxs. Likewise, the probability density of the low flexibility regions (low DFI) has also witnessed an increase in the Thrxs of modern organisms, representing a gain of high-flexibility and low-flexibility regions during evolution. This behaviour was observed not only in the bacterial but also in the human branch of Thrx's evolutionary tree (Figure 3.7C,D). The redistribution of flexible sites in Thrx structures is a mechanism which helps the protein 'fine-tune' its activity (or function) in accordance with the functional requirement.

The observation of the characteristic pattern of the increase in the 'width' of the

distributions of flexibility with evolution is further supported by the high correlation between the variance of the DFI distributions of Thrx proteins and their corresponding melting temperatures (correlation, R = -0.86 and p = 3.2×10^{-3}) (Figure 3.8A). In addition, since the decreasing melting temperature should also correlate with the time duration in evolution (because of in general decreasing ambient temperature of the earth), a significant correlation between the evolution time of Thrx proteins and their variance in DFI distributions is also observed (correlation, R = 0.77 and p = 1.6×10^{-2}) (Figure 3.8B).

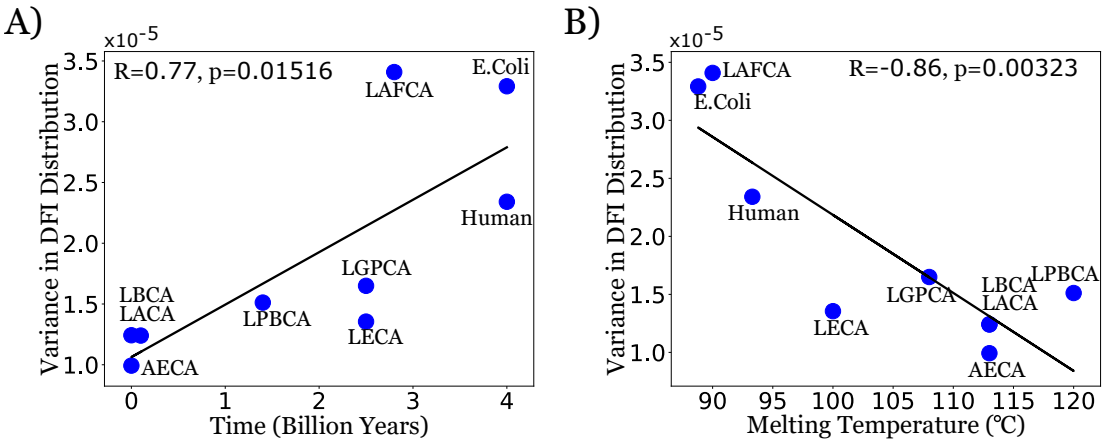

Figure 3.8: (A) The variance in distribution of the flexibility profiles in ancestral and modern-day Thrx proteins is observed to correlate strongly with their time of evolution (Romero-Romero *et al.*, 2016) (R=0.77). (B) The variance in DFI distributions is also observed to correlated with the melting temperatures the proteins (Romero-Romero *et al.*, 2016) (R=-0.86), indicating the implication of flexibility profiles in their thermo-stability.

Overall, these results are in agreement with the previously observations (Kim *et al.*, 2015; Zou *et al.*, 2015) that concludes that evolution shapes the conformational landscape of the native state. From the observations shown above, two major changes in the distribution of DFI profiles can be consistently seen in all the phylogenetic branches of Thrx. Firstly, the DFI profiles change as the sequences evolve. Second, the probability distributions at the beginning (ancestral sequences) are more compact, having a higher propensity towards medium DFI values. This type of distribution

ensures an evolvability in the sequence through mutations of various flexible positions in the protein. This is because, the flexible sites are typically more robust to mutations and generally correspond to highly evolving residues (Glembo *et al.*, 2012; Gerek *et al.*, 2009; Nevin Gerek *et al.*, 2013). On the other hand, as we get closer to the modern extant protein, the distribution widens up. There is an increase in the population density of the low DFI region along with a longer tail corresponding to higher DFI values. In other words, a well-distributed set of very rigid and very flexible sites could be an evolutionary mechanism to adapt to low temperature and/or adjust to the functional needs of the protein.

3.4.4 Dynamic Flexibility Index Captures the Functional Evolution in Thioredoxin.

In order to further test how the DFI profiles of the nine ancestral and extant Thrxs capture the change in function throughout evolution, their DFI profiles were clustered using Principal Component Analysis (PCA). This analysis is used to understand the structure of the data or to increase the signal-to-noise ratio in data by eliminating the redundant dimensions and mapping them on to a lower-dimensional space. Further on, clustering by PCA acts as an effective noise filter by isolating the highest variances among data points in the top principal vectors. Consequently, the remaining insignificant singular vectors can be omitted from the reconstruction.

For this, the percentile rankings of the DFI score of each protein was used. Firstly, the proteins were aligned according to their multiple sequence alignment and were concatenated into a data matrix X. Afterwards, Singular value decomposition (SVD), is used to reduce the dimensions of the data following the method described in section 2.2.5. Here we used the largest three largest principal components to reconstruct the data matrix.

After dimensional reduction, the data can now be clustered hierarchically based on

the pairwise distance between different proteins in the reconstructed DFI data with reduced dimensions. The pairwise distance between each protein after dimensional reduction was calculated using Eq. 2.25.

Finally, a bottom-up approach was used for hierarchically clustering the Thrx proteins, where initially each protein is assigned its own cluster and then, in successive iterations, closest clusters are merged together to create a new common cluster. In this approach, the distance between the clusters obtained was defined by the average pairwise distance between each of their components (average linkage clustering (Day and Edelsbrunner, 1984). In the end, the clusters were represented hierarchically using a dendrogram, where the vertical axis denotes the euclidean distance between various clusters and also among their sub-clusters. The horizontal axis carries no significance whatsoever, other than representing the cluster names.

The Thrxs are clustered together in accordance with the similarities in their flexibility profiles (Figure 3.9). Figure 3.9 also shows that the Thrxs in a same cluster also exhibit similar rate constants for disulfide bond reduction obtained from single-molecule experiments (Perez-Jimenez *et al.*, 2011). For example, ancestral Thrxs LACA and AECA, belonging to the Archaea branch, have very high rate constants for disulfide bond reduction. This is in consistency with the clustering based on DFI scores which arranges them in the same group. On the other hand, LBCA Thrx, from the bacterial branch, which had evolved around the same period as AECA Thrx (around 4 billion years ago), is cluster differently along with LPBCA and LGPCA Thrxs. LBCA, LPBCA and LGPCA Thrxs all share a much lower rates for disulfide bond reduction than those of LACA and AECA Thrxs. Interestingly, human, LAFCA and LECA Thrx, all grouped together in a common cluster, share similar kinetic rates of disulfide bond reduction. It is also noted that E. coli Thrx, having the lowest reduction rate of all other Thrx proteins, was clustered in a separate group

(Figure 3.9A).

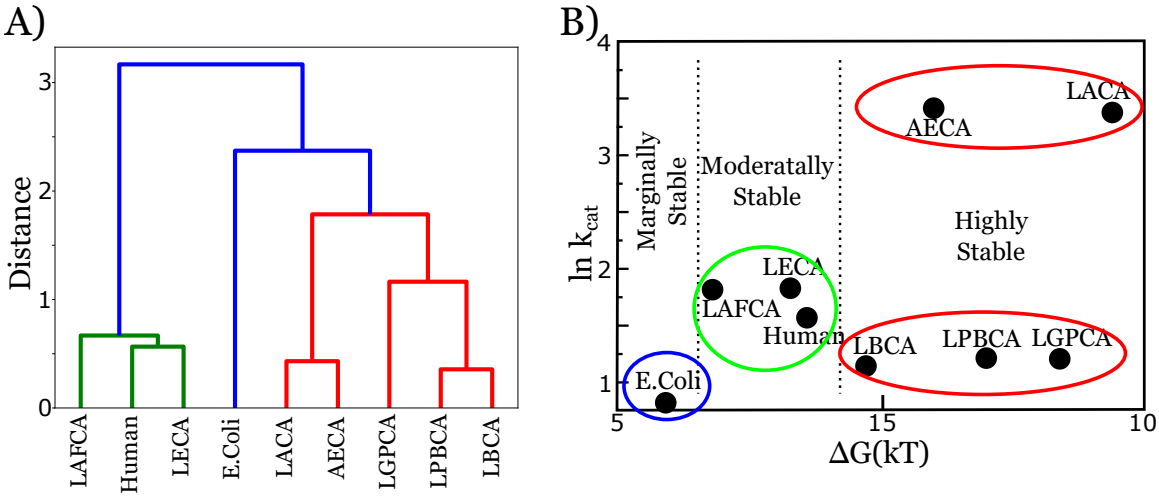

Figure 3.9: The clustering of the ancestral and extant Thrxs based on their DFI profiles (A). The nine Thrx proteins, mapped on a 2D landscape based on their experimental rate constants of disulfide bond reduction, $k_{cat}(M^{-1}s^{-1})$ (Tzul *et al.*, 2017) and the change in free energy of folding, ΔG(kT) (Tzul *et al.*, 2017) (B). The two different maps one using a clustering criteria based on computational analysis and the other using experimental characteristics give similar results as they segregate the nine Thrx proteins based on their activities and thermal stability.

While the argument given above clearly explains the segregation of Thrx proteins into sub-classes, the clustering also raises a question about merging these different sub-classes together. For example, what is the rationalization behind the grouping of the LACA–AECA Thrx dyad (red) with the triad LGPCA-LPBCA–LBCA Thrx (also red) in spite of the two subclasses having widely different catalytic rates (Figure 3.9B)? Interestingly, it is also noted that LACA, AECA, LGPCA, LPBCA and LBCA Thrxs are all highly stable when compared with the triad LAFCA, human and LECA Thrxs.

Therefore, in order to understand the DFI-based classification it is important to take into consideration both– the catalytic rates (k_{cat}) (Perez-Jimenez *et al.*, 2011) and the free energy of folding (Tzul *et al.*, 2017). It is interesting to note that although DFI solely uses the native state ensemble, it is successfully able to sculpt

the landscape in these two coordinates (Figure 3.9B) based on stability and catalytic activity. Thereby, Thrxs can be grouped into three broad categories: (i) highly stable proteins (AECA and LACA Thrxs with high activity; LBCA, LPBCA and LGPCA Thrxs with low activity), (ii) moderately stable proteins (human, LAFCA and LECA Thrxs) with moderate catalytic activity and (iii) marginally stable protein with low catalytic activity (E. coli Thrx) (Figure 3.9B).

In summary, clustering solely based on DFI profiles successfully captures the clustering based on catalytic rates and stability. This suggests that evolution has shaped the native state ensemble of Thrxs in order to adapt and function at cooler temperatures and lower acidic ambient conditions. It is also in agreement with previous studies on protein evolution (Kumar *et al.*, 2015a; Zou *et al.*, 2015), highlighting that evolution exploits native state conformational dynamics to alter function.

3.5 Conclusion

Despite the significant structural similarities between ancestral and extant Thrxs, these have evolved towards a lower stability and kinetic turnover rates. In order to gain insights into the underlying molecular mechanism of evolution, we have explored how changes in the native state ensemble might have impacted the evolution of Thrx proteins. To study this, the difference in the DFI profiles of ancestral and extant Thrx proteins are compared with each other. The enhanced flexibility of the $\alpha 3$ helix in the extant proteins, compared with their ancestral counterparts, was observed to be compensated by the decrease in flexibility of the $\alpha 4$ helix. This mechanism might be responsible for lowering the stability of Thrx proteins which was used to adapt to cooler ambient temperatures, while keeping the fold conserved (since the lower flexibility of the β-sheets was found to be preserved). It was further noted that the dynamic coupling of important positions with the catalytic site has changed

during evolution. Particularly, the decrease in the allosteric dynamic coupling of the $\alpha 3$ helical region, critical in substrate binding, with the catalytic site in extant Thrxs may be associated with the decrease in their catalytic activity at lower acidic conditions.

Furthermore, comparison of the distribution of the flexibility of residues between ancestral and extant proteins revealed that the population density of high- and low-flexibility residues increased as they evolve. These common features observed in evolution suggest a 'fine-tuning' of their native ensemble to adjust to ambient conditions in accordance with the evolution in their function. The high correlation between the variance of flexibility distribution of proteins and their melting temperature quantitatively supports this hypothesis.

In addition, clustering these proteins based on their flexibility profiles (through DFI) closely matches the grouping obtained using their kinetic rates of disulfide bond reduction and thermal stability. These observations, in agreement with previous results, highlight that nature uses native state conformational dynamics to adapt and evolve by altering the protein dynamics (hence flexibilities) and the dynamic allosteric interactions in the protein.

Chapter 4

MUTATIONS UTILIZE DYNAMIC ALLOSTERY TO CONFER RESISTANCE IN TEM-1 β-LACTAMASE

This chapter is adapted from "Modi, T. & Ozkan, S. B. Mutations Utilize Dynamic Allostery to Confer Resistance in TEM-1 β-lactamase. Int. J. Mol. Sci. 19, (2018)."

Chapter 3 presents how nature shapes the native state ensemble of Thioredoxin to modify its function and stability during evolution by comparative dynamics studies of ancestrally resurrected proteins with their modern homolog. In this chapter, we explore whether the observed modulation of dynamics through allosteric interactions is also a general mechanism for the evolution of modern proteins. Thus, we explore the role of protein dynamics in functional evolution of the extant TEM-1 β-lactamases. They are bacteria's defense to β-lactam antibiotics. They confer resistance to penicillin and, third generation antibiotics, cephalosporins. Moreover, they also exhibit functional plasticity in response to the introduction of novel drugs derived from these antibiotics. Over 170 variants of TEM-1 β-lactamase have been isolated in hospitals and clinics. Their high evolvability also made them one of the most studied enzymes through laboratory evolution experiments in an effort to understand the natural evolution of resistance and also to predict resistance driving mutations.

Here we have particularly focused on the allosteric mutations which are distal from the active site, but still play crucial role in emergence of resistances using our conformational dynamics-based approach. Our analysis revealed that these distal mutations remotely alter the flexibility of the active site to accommodate the hydrol-

ysis of newer antibiotics. In addition, we also performed the dynamical analysis of an exhaustive set of 5000 mutations in TEM-1 β-lactamase. It has shown that the sites exhibiting medium flexibility and high dynamic coupling with the active site contribute most to the evolvability towards emergence of resistance.

4.1 Abstract

β-lactamases are enzymes produced by bacteria to hydrolyze β-lactam antibiotics as a common mechanism of resistance. Evolution in such enzymes has been rendering a wide variety of antibiotics impotent, therefore posing a major threat. Clinical and *in vitro* studies of evolution in TEM-1 β-lactamase have revealed a large number of single point mutations that are responsible for driving resistance to antibiotics and/or inhibitors. The distal locations of these mutations from the active sites suggest that these allosterically modulate the antibiotic resistance. We investigated the effects of resistance driver mutations on the conformational dynamics of the enzyme to provide insights about the mechanism of their long-distance interactions. Through all-atom molecular dynamics (MD) simulations, we obtained the dynamic flexibility profiles of the variants and compared those with that of the wild type TEM-1. While the mutational sites in the variants did not have any direct van der Waals interactions with the active site position S70 and E166, we observed a change in the flexibility of these sites, which play a very critical role in hydrolysis. Such long distance dynamic interactions were further confirmed by dynamic coupling index (DCI) analysis as the sites involved in resistance driving mutations exhibited high dynamic coupling with the active sites. A more exhaustive dynamic analysis, using a selection pressure for ampicillin and cefotaxime resistance on all possible types of substitutions in the amino acid sequence of TEM-1, further demonstrated the observed mechanism. Mutational positions that play a crucial role for the emergence of resistance to new antibiotics

exhibited high dynamic coupling with the active site irrespective of their locations. These dynamically coupled positions were neither particularly rigid nor particularly flexible, making them more evolvable positions. Nature utilizes these sites to modulate the dynamics of the catalytic sites instead of mutating the highly rigid positions around the catalytic site.

4.2 Introduction

β-lactamases are enzymes produced by bacteria to hydrolyze β-lactam antibiotics as a common mechanism to provide resistance. Since the introduction of penicillin in the 1940s, β-lactam antibiotics are the most popular antibiotic agents, accounting for about 65% of all antibiotic consumption across the world (Appelbaum, 2012). β-lactam antibiotics target enzymes that synthesize the bacterial cell wall. They constitute an effective low-cost method to treat infections. As a protective measure, TEM-1 β-lactamase enzymes synthesized in bacteria cleave the β-lactam ring in β-lactam antibiotics through hydrolysis, thereby rendering the antibiotic inactive. As a result, β-lactamases constitute a major defense mechanism bacteria uses for survival against β-lactam based antibiotics, particularly for Gram-negative bacteria (Appelbaum, 2012; Risso *et al.*, 2013). It confers resistance to penicillin and early cephalosporins and has been shown to exhibit functional plasticity in response to the introduction of novel drugs derived from these antibiotics. Since its discovery in the early 1960s, over 170 variants of TEM-1 β-lactamase have been isolated in hospitals and clinics (Salverda *et al.*, 2010). Their high evolvability has also made them one of the highly studied enzymes through laboratory evolution experiments in an effort to understand the natural evolution of resistance as well as predict resistance driving mutations (Zou *et al.*, 2015; Risso *et al.*, 2013; Salverda *et al.*, 2010; Ingles-Prieto *et al.*, 2013; Weinreich *et al.*, 2006).

Upon comparison of the amino acid substitutions found in all known clinical TEM-1 β-lactamase isolates with those found in the *in vitro* or in vivo laboratory evolutionary experiments of TEM-1 revealed that, the substitutions found with a high frequency in clinical isolates overlapped with the resistance driving mutations observed in laboratory experiments (i.e., E104K, R164C, R164H, R164S, A237T, G238S and E240K) (Salverda *et al.*, 2010). However, there are a significant number of mutations that have been observed in the laboratory evolution experiments, but have not been found in clinical isolates. Moreover, the structure alone did not help explain the molecular mechanism by which these mutations contribute to antibiotic resistance. Indeed, the mutational landscape that drives β-lactamase enzymes toward more specific functions is extremely complex. It is a multi-dimensional landscape which incorporates a large number of neutral, deleterious, and beneficial mutations distributed throughout the sequence space of the enzymes (Raynes *et al.*, 2018; Weinreich *et al.*, 2013; Knies *et al.*, 2017; Tan *et al.*, 2011). While several of these mutations are found close to the active site in the 3D structure having direct interactions with the active site, many of the mutations that contribute to resistance are distally located, and thereby imposing an effect on the function allosterically. A number of studies have been performed on clinical and laboratory isolates of mutations which exhibited how allosteric mutations modulated the degradation of antibiotics and binding of inhibitors (Bowman and Geissler, 2012; Zimmerman *et al.*, 2017; Cortina and Kasson, 2016, 2018; Bowman *et al.*, 2015; Horn and Shoichet, 2004).

However, despite these studies, the general mechanism by which these mutations allosterically modulate the activity is still elusive. In particular, predicting the mutational residue positions beforehand that could be observed in clinical isolates for developing drug resistance is still posing a great challenge. It is essential to understand the highly complex relationship between mutations and the active sites in

β-lactamase enzymes to design more efficient and effective antibiotic therapeutics. In order to answer the question of whether there are generalized molecular mechanisms for the positions that contribute to resistance, we explored the role of conformational dynamics.

Through the analysis of laboratory-resurrected ancestral β-lactamases it is revealed that these were also resistant to antibiotics. In fact, these could not only degrade the first-generation antibiotics (e.g., penicillin), but also the later generation ones with the same efficiency (Ingles-Prieto *et al.*, 2013). Furthermore, a comparison of the conformational dynamics of ancestral β-lactamases with penicillin-specific extant TEM-1 β-lactamase has shown that the ancestral proteins have an increased conformational diversity to degrade a variety of antibiotics (Zou *et al.*, 2015). Indeed, in other resurrected ancestral protein studies, it has been observed that nature fine-tunes the native state ensemble in order to evolve to a new function-requirement or to adapt to a new environment (Zou *et al.*, 2015; Kim *et al.*, 2015; Modi *et al.*, 2018). In this study, we have focused on comparing the conformational dynamics and the role of dynamic allostery in resistance driving mutations that have been frequently observed in both the clinical isolates and laboratory evolution experiments (Salverda *et al.*, 2010). In addition, we have also investigated whether the similar principle of "fine-tuning" of conformational dynamics, observed in the previous chapter, dictates the positions that are substituted for resistance. To achieve these goals, we apply the Dynamic Coupling Index (DCI) and Dynamic Flexibility Index analysis (DFI) analysis to TEM-1 β-lactamase. As described earlier, the DFI is a position-specific metric which qualitatively measures the relative conformational entropy (flexibility) of a residue position with respect to the rest of the protein chain. The DCI is another position specific metric which gives a quantitative score of the strength of the long range dynamic communication between a given position and the active site. Here we

are particularly interested in the DCI profiles of the positions that are observed as being frequently mutated in the clinical isolates.

It was observed that a majority of single point mutations found in clinical isolates and laboratory evolution experiments (Salverda *et al.*, 2010) exhibit higher coupling with the active site residues (a higher DCI score). In addition, many of these sites were found to impact the dynamics of the active sites allosterically by modulating their flexibility. Moreover, in order to test the completeness of the approach, the data from an exhaustive study was also examined (Stiffler *et al.*, 2015) that provides description of the fitness landscape of TEM-1 β-lactamase enzyme. The data incorporates the information of surviving populations of bacteria under a selection pressure of different concentrations of the antibiotics ampicillin and cefotaxime upon performing all possible single point substitutions. After applying the dynamics-based metrics on this dataset, we observed a pattern where the mutations with a higher impact on the fitness also exhibit a higher DCI score with the active site. Thus, indicating that the positions contributing to resistance indeed have a long range allosteric dynamic coupling with the active sites. In addition, these positions also belong to regions in the enzyme with a medium to a higher flexibility that allow them to be more robust (i.e., accepting to the perturbative impact) to mutations. Moreover, deleterious mutations are typically located at positions with low flexibility, and high flexibility regions house neutral mutations. A similar pattern of impact of mutations at low flexibility region and high flexibility region was also observed in other proteins (Glembo *et al.*, 2012; Kim *et al.*, 2015; Nevin Gerek *et al.*, 2013; Butler *et al.*, 2015). Overall, a generic behavior was recorded in evolution (Modi *et al.*, 2018) of TEM-1 β-lactamase where mutations were more frequently observed in regions exhibiting medium flexibility and high dynamic coupling with the active site to fine-tune the dynamics of the active site, thereby modulating the function.

4.3 Methods for Modelling Interactions and Obtaining Protein Dynamics

In order to analyze the impact of single point mutations, all-atom MD simulations of the wild type and its variants were used. The starting structure of the wild type TEM-1 β-lactamase was obtained from the protein data bank (PDB id: 1BTL (Jelsch *et al.*, 1993)). Subsequent variants were created using the mutagenesis tool of PyMol (Schrödinger, LLC, 2015) by replacing the wild type amino acid with the mutant amino acid template where the initial rotameric state was selected in order to have minimum steric hinderance. Afterwards, 600ns MD simulations were performed for the wild type and mutant protein following the protocol described in section 2.1.1.

After the MD simulations, the flexibility profiles of the wild type TEM-1 β-lactamase and its variants were obtained by calculating their DFI profiles following the method described in section 2.2.2 and making sure that converged dynamics are obtained in accordance to section 2.2.3. For the analysis, 50ns window size was used for calculating the covariance matrix sampled over a time slot of 100ns to 600ns from the MD trajectories. In addition, we also obtained the coupling of each residue in wild type TEM-1 β-lactamase with the catalytic site by calculating the DCI profile with respect to the catalytic site residues– 70, 130, 132, 166 and 234 following the method described in section 2.2.4.

4.4 Results and Discussion

The single point mutations that have been most frequently observed in a large number of both clinical and laboratory isolates of evolution experiments on TEM-1 β-lactamase, compiled by Salverda *et al.* (2010) (Table 2 in (Salverda *et al.*, 2010)), were obtained. These mutations modulate the binding of inhibitor and/or β-lactam antibiotics and give rise to antibiotic resistance (Martínez *et al.*, 2007). While these

alter the function of TEM-1 β-lactamase to degrade different antibiotics, the experimental biophysical characterization of these antibiotic resistant variants has shown that these mutations do not alter the 3D structure of the enzyme and also yield similar protein expression levels (Salverda *et al.*, 2010; Ingles-Prieto *et al.*, 2013). These observations lead to the question that, how does these mutations alter the function of the TEM-1 β-lactamase while conserving its 3D fold.

4.4.1 A Majority of the Resistance Driving Mutations are Distal to the Active Site.

As a first step to study the impact of the resistance driving mutations on the dynamics and activity of the TEM-1 β-lactamase enzyme, the locations of the mutations were mapped on the 3D structure of the enzyme. Through this, a radial distribution of the location of each mutational position from the center of the catalytic region (S70, S130, N132, E166, and K234) was calculated.

Interestingly, it was observed that the majority of these mutations lie farther than 10 Åfrom the catalytic site (Figure 4.1), which is significantly beyond the first coordination shell for creating a direct contact with the active sites. This indicates that these mutations must alter the function by allosterically modulating the network of interactions in the protein chain. This suggested mechanism of antibiotic resistance driven by single point mutations is in agreement with the previous work on the functional evolution of GFP proteins (Kim *et al.*, 2015), Thioredoxins (Modi *et al.*, 2018), β-lactamases (Zou *et al.*, 2015), etc. (Nevin Gerek *et al.*, 2013; Gerek *et al.*, 2009). Thereby showing that the mutations critical for the emergence of a new function, or those which alter pre-xisting functions (or catalytic activity) are usually situated farther away from the active site. These allosterically manipulate the function by fine-tuning the native state ensemble while conserving the 3D fold of the native state.

With the contribution from other studies, it is now well established that the muta-

tions endowing an enzyme with a novel function are generally destabilizing (Tokuriki *et al.*, 2008). Therefore, it was suggested that the enzymes adapting to new novel functions should require mutations which are both functionally beneficial but are also thermodynamically destabilizing mutations while being accompanied with compensatory stabilizing mutations. This type of non-additive interaction between groups of single point mutations (i.e., when the net affect of two point mutations is different from the sum of the effect of the two mutations individually) is called Epistasis (Weinreich *et al.*, 2018, 2013; Miton and Tokuriki, 2016). This complex epistatic relationship among the resistance driving mutations allows for the emergence of novel functions. Several studies involving *in vitro* laboratory evolution (Levin-Reisman *et al.*, 2017; Barbosa *et al.*, 2017) and computational techniques employing the use of Markov state models (Bowman *et al.*, 2015; Zimmerman *et al.*, 2017), statistical models (Figliuzzi *et al.*, 2016; Weinreich *et al.*, 2006; Raynes *et al.*, 2018; Weinreich *et al.*, 2013), structure-based (Cusack *et al.*, 2007), and dynamics-based (Modi *et al.*, 2018) models have been performed where the epistatic effects of pairs of mutations in evolutionary pathways have been explored and the role of allosteric mutations has been emphasized in different protein systems including β-lactamases (Bowman *et al.*, 2015; Zimmerman *et al.*, 2017; Cortina and Kasson, 2016, 2018; Bowman *et al.*, 2015) and other enzymes (Otten *et al.*, 2018; Bershtein *et al.*, 2006; Miton *et al.*, 2018; Keedy *et al.*, 2018). In particular, it has been shown that the thermodynamic effects of beneficial mutations are uncorrelated with cefotaxime resistance in TEM-1 β-lactamase (Knies *et al.*, 2017). The molecular mechanism by which these single point mutations are modulating the resistance of antibiotic resistance of TEM-1 β-lactamase while keeping the 3D fold conserved is now explored through the analysis of conformational dynamics.

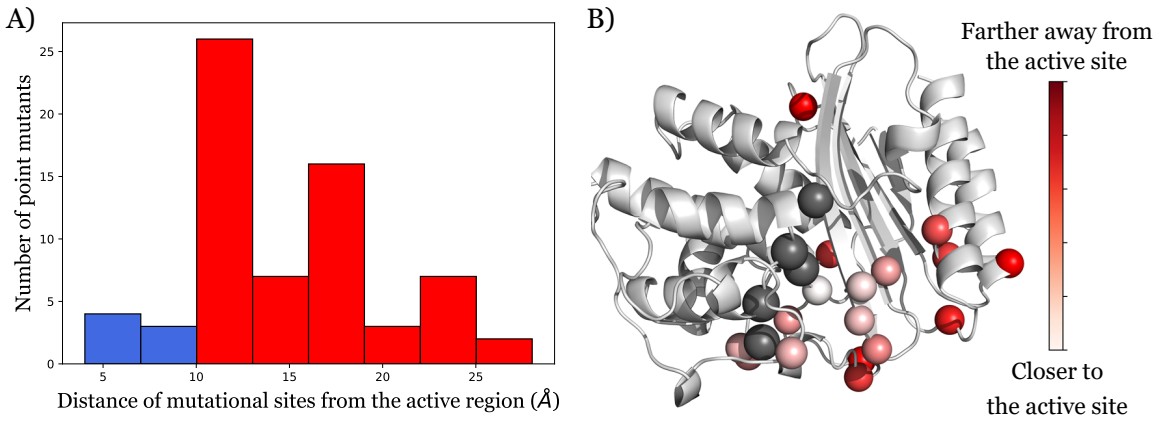

Figure 4.1: (A) Distribution of the distance of mutational sites in single point mutations observed in clinical isolates and laboratory evolution experiments in wild type TEM-1 β-lactamase around the active site. It was observed that the majority of the mutations were distal to the active site (i.e., >10 Å, shown in red) and hence impacted the function via allosteric interactions. On the other hand, a very small number of variants were observed where mutational sites were closer to the active site (in blue). (B) Cartoon representation of TEM-1 β-lactamase with active sites in black spheres and the mutational sites studied in colored spheres, where the color depends on their distance from the centroid of the catalytic region.

4.4.2 Antibiotic Resistance Driving Single Point Mutations Alter the Flexibility Profile of TEM-1 β-lactamase.

The dataset of distal single point resistance driving mutations was divided into two parts based on their modus operandi: (i) the mutations that drive inhibitor resistance, namely: N276D, R244C, R244S, R275L, R275Q, and S268G; and (ii) the mutations that drive the resistance of β-lactam antibiotics, E240K, I173V and Q39K. These mutants were synthesized *in silico* and their dynamics were obtained as described in the section 4.3.

As described earlier, the DFI is a position specific metric which gives the measure of relative resilience of an amino acid to random force perturbations, thereby, mimicking the natural environment of a protein in a crowded cell. Comparing the DFI profiles upon mutation gives an estimate of the impact of the mutation on the equilibrium dynamics of the protein. It also provides insight into the acting mechanism

of the mutation by focussing on the region with the largest change in their DFI score. Therefore, we compare the DFI profile of the wild type TEM-1 β-lactamase with the averaged DFI profile of two sets of mutants (Figure 4.2A,B). Both types of resistance driving mutations exhibited a similar mechanistic behavior. These increased the flexibility of the low DFI hinge regions (i.e., rigid sites), particularly around the catalytic sites S70 and E166, despite that fact that the mutations did not have any direct interaction with the active sites.

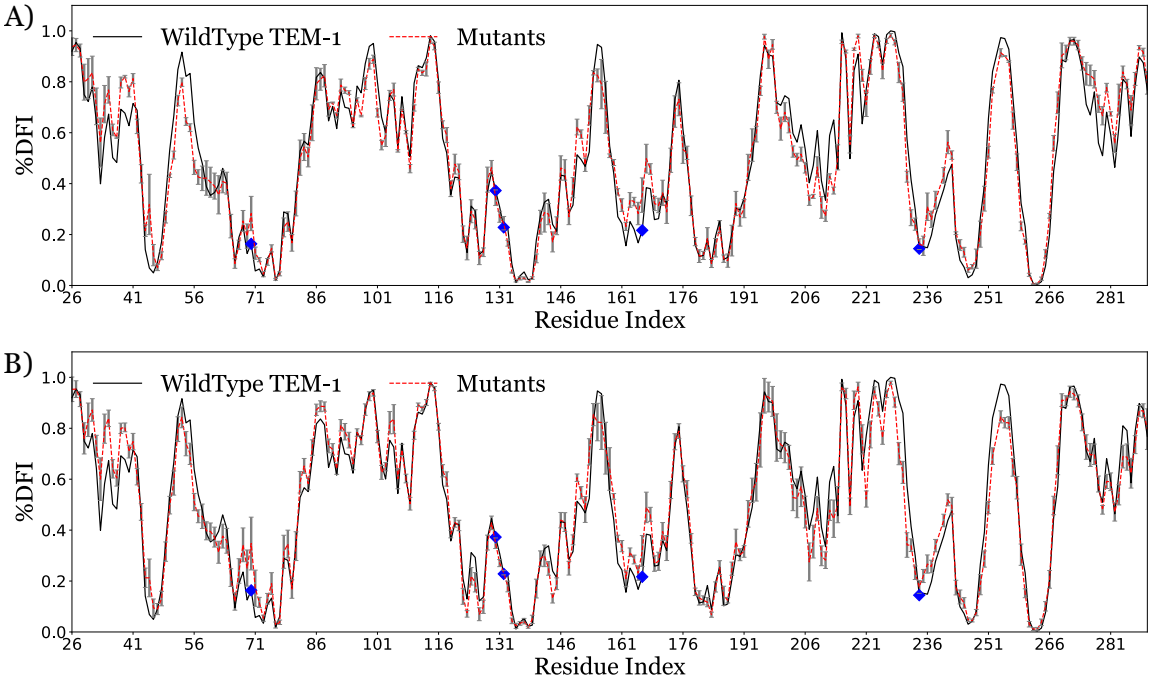

Figure 4.2: The comparison of the conformational dynamics of the wild type TEM-1 β-lactamase (black) using the average %DFI profile (y-axis) of the single point variants (red), distal from the catalytic region (blue diamonds). It was observed that the mutants, despite being distal to the active region (Figure 4.1) impacted the dynamics of the active sites, particularly around S70 and E166. This behavior was observed in the single point variants providing resistance to β-lactam antibiotics (A) as well as in the mutants impacting the resistance through inhibitors (B). Here, in each plot, the error bars in grey represent the standard error of mean obtained from comparing the DFI profile of all the mutants.

The residue positions exhibiting low DFI scores typically play a crucial role in the dynamics of the protein. These participate in the transfer of forces through the

protein's network of interactions, like hinges in a door frame (i.e., facilitating an application of torque without actually moving themselves). Therefore, such residue positions are necessary for mediating the collective motion dynamics, which are key for the protein to function. Studies on other protein systems using DFI like on the evolution of Thioredoxin (Modi *et al.*, 2018), GFP (Kim *et al.*, 2015), etc. Nevin Gerek *et al.* (2013) have revealed that, (i) the mutations that alter the flexibility of hinge regions lead to changes in the protein's dynamics and hence function, and (ii) nature fine-tunes the flexibility of hinge regions through mutations leading to the emergence of new functions. Other studies have also shown that mutations which alter the conformational dynamics of a protein play a key role in its evolution to a new function (Campbell *et al.*, 2018). In alignment with the above-mentioned studies, it was interesting to see a change in the flexibility (DFI values) of the hinge site S70. S70 plays a critical role in the catalytic mechanism by serving as the nucleophile for attack on the carbonyl carbon of the amide bond. The change in DFI profile of S70 suggested that these allosteric mutations fine-tune the flexibility of the catalytic positions by altering their conformational dynamics for the hydrolysis of new antibiotics. Likewise, the same mechanism (i.e., the change in the flexibility of active sites which play a major role in binding and catalysis) was also observed for the other set of mutations, which are known for driving the resistance through inhibitors (Figure 4.2).

These observations were further consolidated by calculating the change in the flexibility profile of TEM-1 β-lactamase post mutations. The change in DFI profile (ΔDFI) was obtained by comparing the averaged DFI profiles of the mutants with that of the wild type. In particular, a significant change was observed on the positions that exhibited a large change in DFI, i.e., more than one standard deviation of change than the average change in DFI (Figure 4.3). Interestingly, this analysis revealed that while the DFI score of most of the residue positions being mutated in the

present work (orange diamonds) did not change significantly. However, mutations at those positions triggered changes in the DFI scores of the distal positions, particularly near the active site (blue diamonds, of note S70 and E166). Furthermore, it was also discovered that some of the positions that exhibited significant change upon resistance driving mutations from clinical isolates corresponded to the mutational points that had already been observed in the laboratory evolution of TEM-1 β-lactamase (Figure 4.3A,B green diamonds). However, further mutational analysis is required to understand the underlying mechanistic details governing the antibiotic resistance emerging from these mutations.

4.4.3 Dynamic Coupling Index (DCI) Gives an Insight Into the Internal Network of Interactions in TEM-1 β-lactamase.

In the section above, it has been observed that the analyzed mutations providing antibiotic resistance to TEM-1 β-lactamase, allosterically modulated the flexibility of the active sites. Buoyed by this analysis, we measured the long-distance dynamic interactions of the mutation sites with the active sites using the DCI analysis (Kumar *et al.*, 2015b; Butler *et al.*, 2015; Larrimore *et al.*, 2017). As described earlier, the DCI score quantifies the strength of the coupling interaction between any pair of residues or between a residue and a region (e.g., active sites). This analysis helps identify the Dynamic Allosteric Residue Coupling (DARC) spots (Butler *et al.*, 2015; Larrimore *et al.*, 2017). These are the residue positions that are distal from active sites but are capable of remotely regulating the active site by modulating the flexibility of functionally critical sites. This type of allosteric coupling, regardless of the distance of separation, is also likely to contribute to the function.

Through the DCI analysis, it was observed that irrespective of the distance of separation between the mutational sites and the active sites (Figure 4.1) most of the

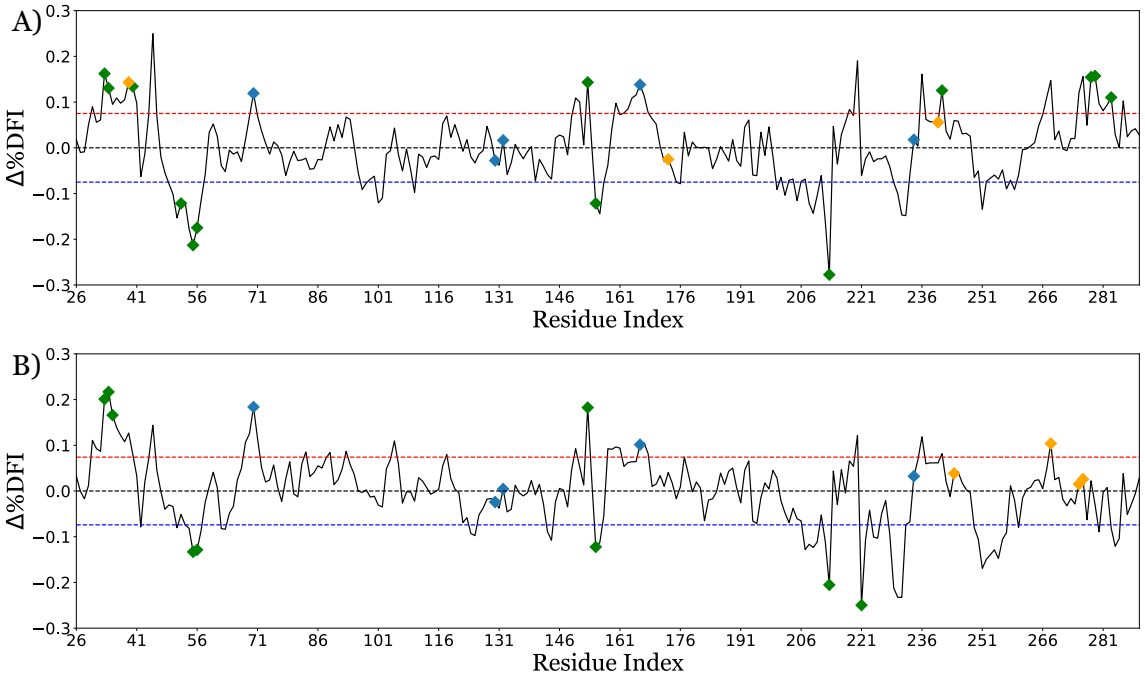

Figure 4.3: Comparison of the conformational dynamics of the wild type TEM-1 β-lactamase with the average DFI profile of the single point variants, distal from the catalytic region (blue diamonds), by calculating the difference in their %DFI profiles (Δ%DFI, y-axis) of the mutants impacting the hydrolysis of β-lactam antibiotics (A) and the mutations driving resistance through inhibitors (B). The red and blue dash line represents ± 1.0 times the standard deviation around mean (black dash line) of the differences between the profiles. It was observed that the mutations (orange diamonds) impacted the dynamics of the catalytic region particularly around S70 and E166. In addition, it was also observed that many of the single point mutations explored in laboratory evolution experiments were found in regions with significant differences in their conformational dynamics when compared to the wild type (green diamonds) (Salverda *et al.*, 2010). This hints towards their possible contribution to the dynamics of the catalytic region.

resistance driving mutations studied here were dynamically coupled with the active sites as shown by their overall higher DCI scores (Figure 4.4, color-coded where darker spots exhibit high DCI values). Apart from being highly coupled to the active sites, the mutational sites also exhibit a medium range of flexibilities (i.e., neither too flexible nor rigid, inside the gray shaded region in Figure 4.4). The medium flexibility of these sites make them more robust to the perturbative impact of mutations, while the high dynamic coupling of these residues with the actives site ensures their ability

to remotely modulate the active site flexibility. This observed mechanistic picture is also in agreement with other studies of missense variants, where the disease associated mutations on the DARC spots were found to regulate the flexibility and dynamics of functionally critical sites from a distance leading to a loss or gain in the protein function (Butler *et al.*, 2018; Kumar *et al.*, 2015a).

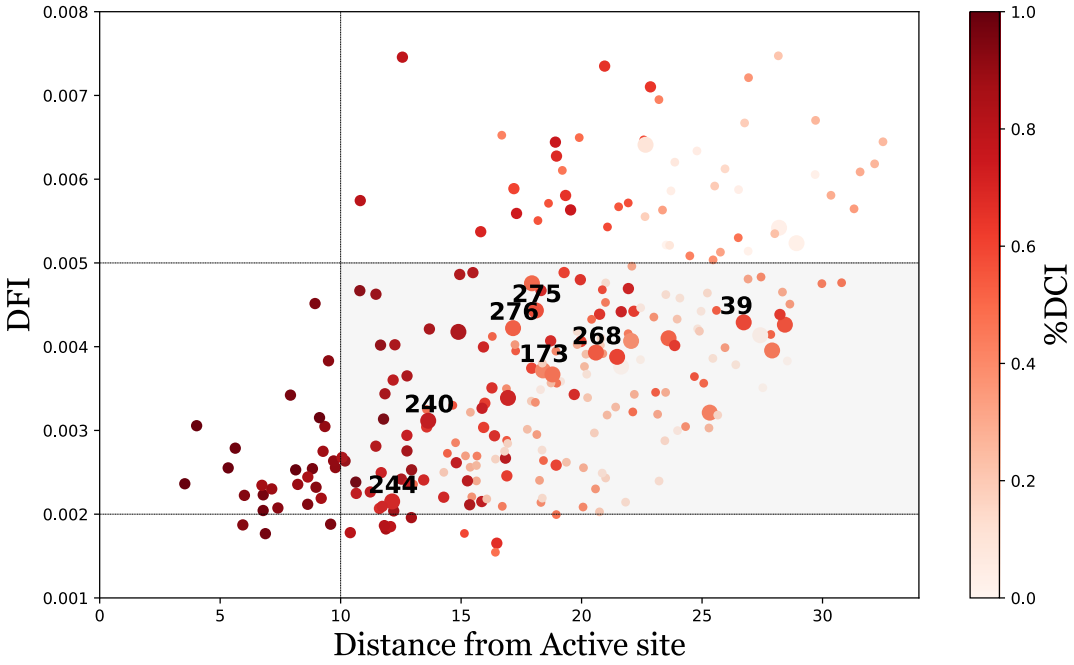

Figure 4.4: A distribution of flexibility of various residue locations in TEM-1 β-lactamase given by their DFI score (y-axis) with their distance from the catalytic region (x-axis). The residues are colored in accordance to the strength of their coupling with the active sites using their %DCI score with red being most coupled (%DCI = 1) to white without any coupling (%DCI = 0). We observed that a large number of the mutations analyzed in clinical isolates and laboratory evolution experiments of TEM-1 β-lactamase had a strong coupling (represented by bigger circles) regardless of being distal to the active sites. Moreover, these mutations, along with many other residue positions coupled to the active site, lay in a region of medium flexibility in the protein.

Moreover, upon observing the DCI scores of all of the residue locations in TEM-1 β-lactamase (Figure 4.4) we noticed that, in addition to the mutational sites found in clinical and laboratory isolates, there were a large number of other residue locations in TEM-1 β-lactamase with a high DCI coupling score with the active site (> 0.7). Many

of these highly coupled residue locations were distal from the active sites and also lie in a region of medium flexibility (gray region in Figure 4.4). Studies of dynamics of positions mutated in other protein systems using DFI and DCI analysis suggest that mutations at residue positions that exhibit higher coupling to the active site and are also situated distally with a medium flexibility generally serve as allosteric modulators for the function without compromising the fold or stability while making subtle changes in the dynamics of active sites (Nevin Gerek *et al.*, 2013; Butler *et al.*, 2015). Therefore, there is a need to perform an exhaustive mutagenesis scan of all possible mutations at each residue location in TEM-1 β-lactamase in order to analyze the impact of mutations at these sites on the activity of the protein. Doing such an analysis will enable us to verify if DCI analysis can be used as a predictor for functional impact upon mutation, i.e., substitutions at residue positions with high DCI typically lead to functional changes in the protein dynamics.

Similar in vivo experiments were performed by Stiffler *et al.* (2015), where they investigated the impact of all possible single point mutations (4997 mutations) in TEM-1 β-lactamase on the organism's fitness under the selection pressure for varying concentrations of antibiotic ampicillin resistance (i.e., the wild type function) and cefotaxime (CTX) resistance (i.e., evolving a new function) (see Figure 4.5). After the fitness data is mapped over the DCI profile of the protein, it was observed that a significant number of residues with mutations with a positive effect on the fitness for the CTX antibiotic were highly coupled to the active sites (%DCI > 0.7) (Figure 4.5B). This observation was in agreement with the analysis shown in the previous section (Figure 4.4) where the mutations that drive resistance to an antibiotic should exhibit high dynamic coupling with the active sites, irrespective of their distal location with the active sites.

This pattern can be further analysed by considering a hypothesis that residue

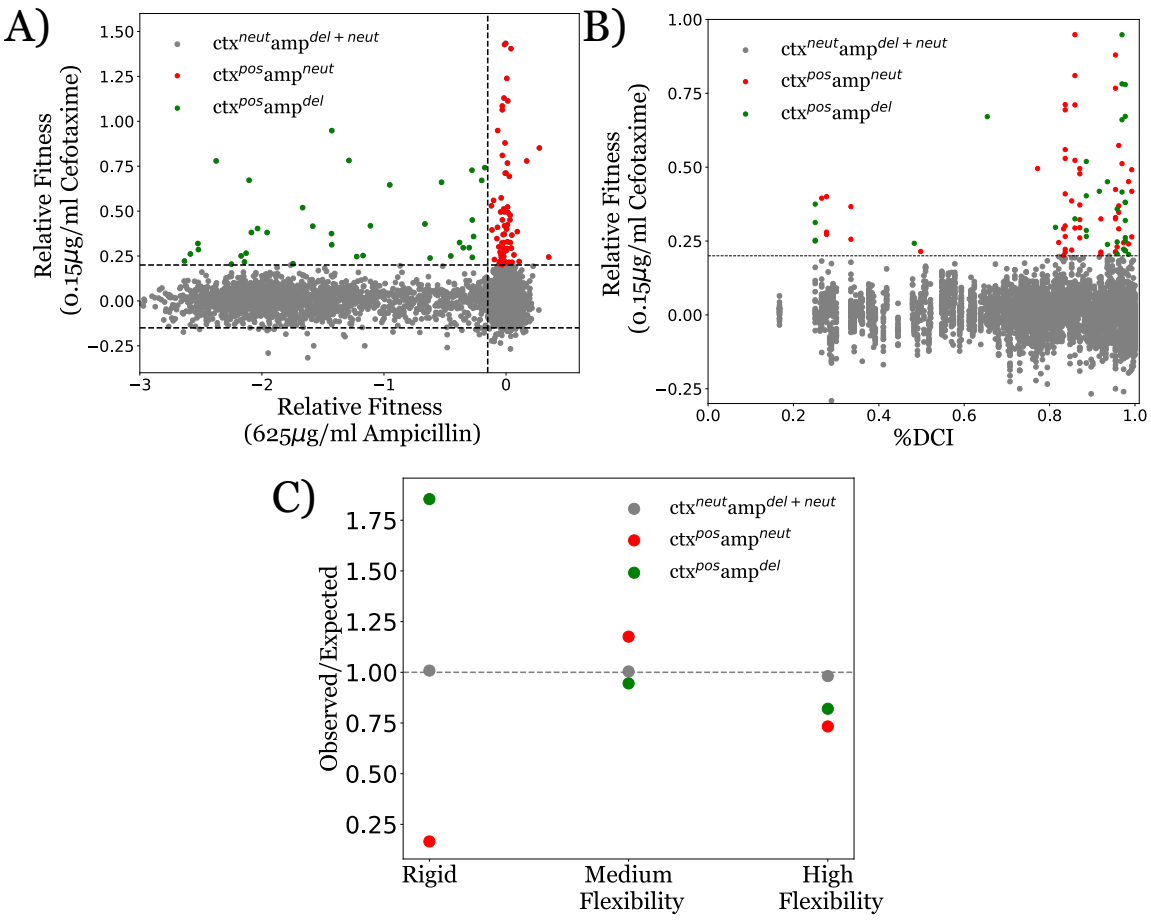

Figure 4.5: The Dynamic Coupling Index and flexibility of a residue position determines emergence of new function upon substitution. (A) Relative fitness values for all mutations under ampicillin (625 μg/mL (x-axis) versus CTX selection (0.15 μg/mL; y-axis) showed that most mutations conferring significant CTX resistance were mostly neutral (red), but some were deleterious (green) in ampicillin resistance, while the color grey shows the variants with no effect on CTX resistance. (B) Change in the relative fitness of the organism upon various mutations as a function of their %DCI couplings with the active site. We observed that a majority of the mutations with positive fitness in CTX were highly coupled to the active region (%DCI $>$ 0.7). (C) Observed to expected ratio of TEM-1 β-lactamase variants based on their flexibility category.

positions which a larger impact on the CTX fitness of the enzyme should predominantly belong to a group with medium flexibility. Hence, for this, a null hypothesis of no effect can be created according to which, the ratio of the observed to expected numbers of residue positions hosting variants with increasing CTX fitness should be close to 1.0. While the null hypothesis was observed to be true for the ampicillin

and CTX neutral mutations, it was completely rejected for the CTX positive variants (Figure 4.5C). It was observed that CTX-positive and ampicillin deleterious variants were over abundant at rigid sites (ratio = 1.86), which is in agreement with the earlier studies where mutations at rigid sites were more deleterious when compared to wild type for the function (in this case ampicillin resistance) (Butler *et al.*, 2015; Kumar *et al.*, 2015a). Moreover, in consensus with another previous human proteome wide study (Nevin Gerek *et al.*, 2013), it was witnessed that CTX-positive and ampicillin-neutral variants were over abundant at positions exhibiting medium flexibility (ratio = 1.21) (Figure 4.5C). Medium flexibility at these sites usually provides conformational freedom to accommodate the change in their interaction upon amino acid substitution. Furthermore, these also exhibit a higher dynamic coupling with the active sites as compared to the dynamic coupling between the active site and the sites with higher flexibility. Therefore, the mutations at these residue will be usually neutral for ampicillin degradation, but they help the enzyme to evolve a new function (or alter activity) which is CTX degradation in this case.

Similar conclusions were drawn in a more detailed analysis on the evolution of Thioredoxins (Modi *et al.*, 2018) which also indicated that evolution utilizes the residues belonging to mid-flexibility regions in proteins ($0.002 \leq \text{DFI} < 0.005$) to introduce mutations and adapt to the new environments to function at higher pH and cooler temperatures. In that study it was observed that the population of low flexibility rigid region ($\text{DFI} < 0.002$) and high flexibility regions ($0.005 \leq \text{DFI}$) increased with time, and that of the medium flexibility region decreased. The same mechanism was found to be operational for the mutations driving CTX resistance in this large dataset, as shown by the observed to expected ratios (Figure 4.5C) of the population of low, mid, and high flexibility residues in TEM-1 β-lactamase enzymes.

4.5 Conclusion

In this study, the impact of resistance driving single point mutations in TEM-1 β-lactamase enzyme was investigated. These mutations were observed in multiple clinical isolates of TEM-1 β-lactamase and also in *in vitro* laboratory evolution experiments compiled in (Salverda *et al.*, 2010). These mutations are responsible for providing antibiotic resistance to TEM-1 β-lactamase by either altering their activity with β-lactams or with inhibitors like β-lactam–inhibitor combination ampicillin–clavulanate (Vakulenko *et al.*, 1998), etc. Mapping the locations of mutational sites onto a 3D structure of TEM-1 β-lactamase revealed that a large number of these mutations are distal to the active site region (Figure 4.1) and hence these might be exploiting allosteric regulations in order to alter the activity of the TEM-1 β-lactamase remotely.

Moreover, analyzing the DFI profiles of the mutants and comparing them with that of the wild type TEM-1 β-lactamase revealed that the mutations altered the flexibility of the rigid parts of the protein. Particularly around the active site S70, which plays a key role in the activity of the protein (Figure 4.2 and Figure 4.3). While the flexibility of the positions near the mutational sites did not change significantly, it was observed that the mutations remotely modulated the network of interactions near the active sites. This fine-tuning of the active site dynamics in TEM-1 β-lactamase can be a possible mechanism responsible for the degradation of a novel substrate. This remote modulation between the allosteric mutational site and the active sites is actualized via long distance dynamic coupling as indicated by the observation that a majority of the resistance driving mutation sites exhibited high dynamic coupling (DCI) with the active sites (Figure 4.4).

These findings were further consolidated through an analysis of the dynamic cou-

pling and the flexibility of all the residue positions in TEM-1 β-lactamase. Each residue position in TEM-1 β-lactamase were characterized in another study by Stiffler *et al.* (2015) where all possible single point mutations were performed in the sequence of TEM-1 β-lactamase (4997 in total). The fitness score of each mutant thus created was further quantified based on their survivability under a selection pressure of CTX and ampicillin antibiotic. Upon mapping the DCI values of the mutational sites onto their fitness scores, it was observed that, in agreement with analysis above, the mutations that altered the activity to a new function (activity with CTX in this case) were found to be dynamically coupled to the active sites (Figure 4.5B). In addition, upon analyzing the flexibility of the mutating amino acid positions, it was also observed that the mutations belonging to the medium flexibility region led to the evolution of a new function (as seen from the observed to expected ratios, Figure 4.5C). On the other hand, mutations in low flexibility regions of the protein were generally deleterious for ampicillin resistance.

These results are in complete agreement with the investigation performed on the underlying mechanism of evolution of new function in Thioredoxins (Modi *et al.*, 2018). This indicates that there is a principal mechanism in play– evolution may utilize residues with medium flexibility to perform substitutions to fine-tune the flexibility of lowly and highly flexible residue positions in order to evolve a new function (or simply alter the activity). In the process of doing so, it can exploit allosteric interactions in the network to change flexibilities remotely and compensate for the damaging effects of deleterious mutations by performing other mutations (Modi *et al.*, 2018; McLeish *et al.*, 2015).

Chapter 5

HINGE-SHIFT MECHANISM: A PROTEIN DESIGN PRINCIPLE FOR THE EVOLUTION OF β-LACTAMASE FROM SUBSTRATE PROMISCUITY TO SPECIFICITY

This chapter is adapted from "Modi, T., Risso, V. A., Martinez-Rodriguez, S., Gavira, J. A., Mebrat, M. D., Horn, W. D. V., Ruiz, J. M. S.,& Ozkan, S. B. Hinge-Shift Mechanism: A Protein design principle for the evolution of β-lactamase from substrate promiscuity to specificity. Under review at Nature Communication."

Chapters 3 and 4 provide evidence of how nature alters the native state ensemble of a protein to evolve new functions. It describes how proteins adapt to a new environment or enhance enzymatic activity through a hinge shift mechanism where, the flexibility profile associated with function alters through substitutions. Particularly, loss in flexibility of some sites are compensated by enhancement in flexibility of other distal rigid sites. Here, we develop a novel conformational dynamics-based computational design approach which rationally molds the protein flexibility profile on the basis of a hinge-shift mechanism, then deliberately weigh and alter the conformational dynamics of engineered enzyme towards desired function.

In this study, we tested our method on ancestral Precambrian β-lactamase, a moderately efficient promiscuous enzyme capable of degrading a diversity of β-lactam antibiotics in contrast with the modern TEM-1 β-lactamase, that degrades only penicillin with higher efficiency. This generalist to specialist enzyme evolution has involved more than 100 mutational changes. Using our dynamics-based approach, we computationally designed an "engineered specialist Precambrian β-lactamase". With only

21 mutations identified by the flexibility profile, we not only convert a generalist Precambrian β-lactamase to a specific enzyme with better efficiency of degrading penicillin (i.e., mimicking the extant TEM-1 β-lactamase). This is further biophysically characterized by our collaborators at Ruiz lab . In addition, they also obtained the X-ray crystal structure of the engineered Precambrian β-lactamase and comparing it with the wild type β-lactamases also suggests that we have managed to keep the 3D fold conserved. More importantly, the NMR analysis performed by Van Horn Lab also validates the design approach such that the engineered Precambrian β-lactamase exhibits different dynamics despite sharing an identical 3D fold. These findings highlight the key role played by, computationally designed, distal allosteric mutations in fine-tuning the dynamics of the enzyme to evolve function.

5.1 Introduction

Proteins are cellular biomolecular machines with the capacity to participate in a wide variety of functions with remarkable efficiencies and specificities. Apart from being the efficient worker bees of the cell, proteins evolve and develop new functions over time, this process is critical for the evolution and survival of the organism. Proteins owe this remarkable capability to their 3D network of atomic interactions, which orchestrates the communication between different parts of the protein chain in order to accomplish their designated functions. A complete understanding of the "blueprint" of their functional behavior (i.e., the relationship between their sequence, structure, dynamics and function) and how it evolves with time could dramatically expand our ability to develop novel protein-based catalysts with potentially far-reaching applications to fields including chemistry, biotechnology and medicine.

As described in the previous chapter, TEM-1 β-lactamase is a popular target for evolutionary studies. These aid bacteria in their fight against antibiotics by hydrolyz-

ing β-lactam antibiotics like penicillin, cefotaxime, etc., rendering the antibiotic useless. In an attempt to understand the molecular mechanism of antibiotic resistance in bacteria, particularly how variations at sequence level impact function, these enzymes have been a target of a variety of evolutionary studies (Allen *et al.*, 2009; Atanasov *et al.*, 2000; Bradford, 2001; Cortina and Kasson, 2016; Figliuzzi *et al.*, 2016; Fisher *et al.*, 2005; Livermore, 1995; Medeiros, 1997; Modi and Ozkan, 2018; Zou *et al.*, 2015). After its discovery in 1963 (Ruiz, 2018), around 170 other variants of TEM-1 β-lactamase have been isolated, making it one of the best understood and investigated enzymes from an evolutionary perspective (Allen *et al.*, 2009; Livermore, 1995; Medeiros, 1997; Modi and Ozkan, 2018; Zou *et al.*, 2015; Ruiz, 2018; Risso *et al.*, 2018; Salverda *et al.*, 2010; Stiffler *et al.*, 2015; Risso *et al.*, 2013).

Through a Bayesian approach in a phylogenetic framework, the Precambrian nodes in the evolution of class-A β-lactamases have been resurrected (Risso *et al.*, 2013). This study provided us with the sequences and structures of enterobacteria (ENCA), the last common ancestor of various Gram-negative bacteria (GNCA), and the last common ancestor of gram-positive and gram-negative bacteria (PNCA) β-lactamases. Based on the estimates of divergence times, these enzymes existed about 1 Ga (ENCA), 1.5 Ga (GPBCA), 2 Ga (GNCA), and 3 Ga (PNCA). Comparison of ancestral β-lactamase enzymes with the extant TEM-1 β-lactamase revealed that these share several physical features— X-ray structures show that they share a common 3D fold (root mean square deviation (RMSD) 0.585Å); pairwise sequence alignment between the sequences of GNCA and TEM-1 β-lactamases indicates about 50% conservation in their amino acid sequence (see Figure 5.2). In addition, they also shared the same composition and the shape of their catalytic pocket (Modi and Ozkan, 2018; Zou *et al.*, 2015). Despite these striking similarities, these have very divergent catalytic activity. The resurrected Precambrian β-lactamases were found

to degrade both penicillin and third-generation antibiotics (such as cefotaxime) with moderate catalytic efficiency. Of course, third generation antibiotics are a human invention and did not exist in the Precambrian. The results, however, supports that Precambrian β-lactamases were moderately efficient promiscuous enzymes capable of degrading a diversity of β-lactam antibiotics. However, with time, these have evolved into highly specific enzymes that selectively degrade penicillin with about two magnitudes of higher activity (Zou *et al.*, 2015; Risso *et al.*, 2013). These results are in stark contrast to the common "structure-function" paradigm where the structure has a one-to-one relationship with function. However, this anomalous property of violation of "structure-function" paradigm is not unique to β-lactamases only as suggested by Osadchy and Kolodny (2011) where they describe the relative diversity in the functional space and the structural space of proteins. They focus their analysis on the structures and functions for the proteins available in the protein data bank and show how proteins closer to each other in their 3D structure can be very divergent in sequences as well as function.

In all the resurrected ancestral protein studies, it has been observed that the evolution of a new function and/or adaptation to a new environment is always accomplished while preserving the 3D structure (Risso *et al.*, 2018; Khersonsky and Tawfik, 2010; Khersonsky *et al.*, 2006; Tokuriki and Tawfik, 2009b; Harms and Thornton, 2013b; Ingles-Prieto *et al.*, 2013; Babtie *et al.*, 2010; Gerlt and Babbitt, 2009; Nguyen *et al.*, 2017). Our previous studies identify similar trends in other enzymes and proteins, including Thioredoxins (Modi *et al.*, 2018), GFP (Kim *et al.*, 2015), and others (Modi and Ozkan, 2018; Zou *et al.*, 2015; Campitelli *et al.*, 2020) which highlight the important role played by the conformational dynamics in the evolution of an enzyme where the structure-function model of the protein activity is replaced by the "ensemble model" (Modi and Ozkan, 2018; Campitelli *et al.*, 2020; Modi *et al.*, 2018; Kim *et al.*,

2015; Zou *et al.*, 2015). In this model, the native state of a protein is represented by a collection of different conformations visited by the protein. The protein then samples these conformations through a broad range of motions from atomic fluctuations and side-chain rotations to collective domain movements. Therefore, in this model, the function of a protein is governed by the dynamics of sampling through this ensemble as opposed to merely a dominant structure.

The ensemble model of protein dynamics also fits very well with protein evolution as it explains the emergence of a new function or modulation of a pre-existing function for adaptation to a new environment while conserving the dominant structure. Nature modulates function by performing a series of subtle modifications in the ensemble of the protein conformations such that the structure remains conserved, but the dynamics of the protein are now different by restricting the sampling of a group of conformers while allowing others. Indeed, studies on protein design through directed evolution have also highlighted the importance of conformational dynamics (Romero and Arnold, 2009; Wu *et al.*, 2019; Lane and Seelig, 2014; Otten *et al.*, 2018; Taylor *et al.*, 2015; Bloom *et al.*, 2006). However, the underlying molecular mechanism for evolution, particularly, which position to substitute in order to modulate the conformational dynamics of the protein still presents a major challenge. This challenge also addresses the issue that the activities of rationally designed enzymes are almost always universally lower efficiency than their naturally occurring counterparts by a couple of orders of magnitude (Wilding *et al.*, 2019; Mak and Siegel, 2014; Bar-Even *et al.*, 2011). Albeit, recent efforts in the designed enzyme methods have enabled the successful engineering of proteins with novel catalytic functions (Gerlt and Babbitt, 2009; Otten *et al.*, 2018; Mak and Siegel, 2014; Bar-Even *et al.*, 2011; Khersonsky *et al.*, 2010). A rigorous comparison shows that natural and engineered enzymes have essentially equivalent substrate-binding affinity, yet drastic differences

in catalytic rates. These low catalytic rates could only slightly be increased after rounds of directed evolution experiments that allows distal mutations (Khersonsky *et al.*, 2010). Therefore, the lack of fundamental knowledge of which mutations will modulate the conformational dynamics towards targeted functions currently prevents rational engineering of novel enzymes with catalytic efficiency that mirrors that of the native ones.

In our previous computational studies (Modi and Ozkan, 2018; Modi *et al.*, 2018; Zou *et al.*, 2015; Kim *et al.*, 2015) we have identified a common underlying "hinge-shift" mechanism that accounts for many functional features during protein evolution. Here, "hinges" are the regions in a protein with relatively lower flexibility, that help coordinate the motions in higher flexibility regions. These computational studies suggest that an enzyme evolves with subtle changes in dynamics, and concurrently its function, through a series of hinge-shift mutations by interchanging and altering distinct, flexible positions with rigid positions. In doing so, evolution exploits an allosteric network of interactions that can modulate the active site by making distal substitutions. Several other studies have also validated the critical role played by allosteric interactions for evolutionary trajectories of enzyme function (Saavedra *et al.*, 2018; Hadzipasic *et al.*, 2020; Kamp *et al.*, 2018; Wodak *et al.*, 2019b,b; Sinha and Nussinov, 2001).

In an effort to further validate the hinge-shift mechanism, we have engineered the minimum number of necessary substitutions in the ancestral GNCA β-lactamase, such that, the designed mutant is closer, in its activity, to that of the specialist extant TEM-1 β-lactamase. In order to do so, we use Dynamic Flexibility Index (DFI) (Modi *et al.*, 2018; Modi and Ozkan, 2018; Campitelli *et al.*, 2020; Kim *et al.*, 2015; Zou *et al.*, 2015; Nevin Gerek *et al.*, 2013; Glembo *et al.*, 2012) and DCI (Modi and Ozkan, 2018; Campitelli *et al.*, 2020; Kim *et al.*, 2015; Butler *et al.*, 2015; Larrimore *et al.*, 2017;

Kumar *et al.*, 2015b) as the metrics for quantifying the flexibility profile and the coupling profile of various residue positions in GNCA and TEM-1 β-lactamase. We propose a detailed mechanism, according to which, selected substitution mutations are applied to GNCA β-lactamase. These gradually drives its dynamics towards that of TEM-1 β-lactamase. This change in dynamics should cascaded further to a change in the activity of the designed mutant. This is then tested and validated by functional assays, where we observed that, as predicted by our analysis, the activity of the designed mutant to degrade benzylpenicillin increased by 3 fold; whereas, the activity for degrading cefotaxime showed a remarkable decrease of 10,000-fold. These results indicate that through mutations predicted by the hinge-shift mechanism, we have rationally engineered the promiscuous GNCA β-lactamase into a specialist enzyme that mimics TEM-1 β-lactamase. These results further highlight the significant role that allostery and conformational dynamics play in the functional evolution of enzymes.

5.2 Results and Discussion

5.2.1 Evolution Conserves the Three-Dimensional Structure of β-lactamase While Changing the Dynamics.

As the first step, to explain the functional differences between the ancestral and extant β-lactamases, we employ a comparatively recent and less explored dynamics–function paradigm which helps relate protein function with dynamics. We use all-atom molecular dynamics (MD) simulations (See section 2.1.1) to obtain 400ns of equilibrium dynamics of ancestral (GNCA, PDB id: 4B88 (Risso *et al.*, 2013)) and extant (TEM-1, PDB id: 1BTL (Jelsch *et al.*, 1993)) β-lactamases. Thereafter, using a covariance matrix of size 50ns sampled from a time slot of 100ns to 400ns from the trajectory, we compute the residue-specific flexibility profiles of the two proteins

using DFI analysis (see sections 2.2.2 and 2.2.3).

Several computational studies have been performed which predicts the relationship between the function and the flexibility of protein chains (Modi *et al.*, 2018; Nevin Gerek *et al.*, 2013; Gerek *et al.*, 2009; Gerek and Ozkan, 2011; Glembo *et al.*, 2012). Flexibility analysis using DFI reveals regions of low flexibility and high flexibility in a protein. Furthermore, the residues belonging to low flexibility region in a protein, called "hinges", have been shown to play a critical role in coordinating the collective dynamics of the protein. These typically act as hubs for communication between different parts of the protein (Kim *et al.*, 2015; Campitelli *et al.*, 2018; Li *et al.*, 2015; Khade *et al.*, 2020; Wriggers and Schulten, 1997). Various evolutionary studies on the flexibility profiles of proteins have revealed that residues with low flexibility have a higher propensity to be conserved in evolution and mutations in such low flexibility regions usually prove to be deleterious for the function, particularly when it comes to pathogenic mutations (Nevin Gerek *et al.*, 2013; Butler *et al.*, 2015; Glembo *et al.*, 2012; Kumar *et al.*, 2015b). On the other hand, residues found in the higher flexibility regions in a protein have a higher conformational/vibrational entropy and are able to sample the conformational landscape with a relative ease. Therefore, these participate in functions demanding a higher mobility such as ligand recognition, etc. Such regions are observed to be more prone to neutral or compensatory mutations throughout evolution and are more forgiving to the effects of amino acid substitutions.

Upon comparing the DFI profiles of the ancestral β-lactamase, GNCA with the extant TEM-1 β-lactamase, we identify various differences in their predicted flexibilities (see Figure 5.1). This result is in agreement with our previous DFI analysis, where we provided insights about the puzzling question of how ancestral β-lactamases can degrade a variety of antibiotics, exhibiting promiscuity, unlike the specific modern

homologs that can only inhibit penicillin, while maintaining the same structure (Modi and Ozkan, 2018; Zou *et al.*, 2015). The special structural dynamics associated with substrate promiscuity of ancestral β-lactamases was revealed by patterns of high DFI values in regions close to the active site, illuminating the flexibility required for the binding and catalysis of different ligands (Modi *et al.*, 2018; Zou *et al.*, 2015). Furhter on, we carefully examine the differences in the DFI profiles of GNCA and TEM-1 β-lactamases by comparing their percentile rankings (%DFI). The percentile ranking gives an idea for the relative ranking of each residue, i.e., a residue with a %DFI score of 0.1 would imply that the residue is among the 10% least flexible residues. Comparing these, we observe a number of low flexibility residues in GNCA and TEM-1 β-lactamases, (i.e., hinges exhibiting <0.2 %DFI) which retained their flexibility through evolution. We label these residues as common hinges. On the other hand, we also identified many hinge positions between GNCA and TEM-1 β-lactamases, which underwent a significant change in their flexibility by increased dynamics or enhanced rigidity through evolution. We label such residues as non-common hinges (see Figure 5.2).

5.2.2 Mimicking the Dynamics of TEM-1 β-lactamase by Introducing Hinge Shifts in GNCA β-lactamase.

In our previous analysis on various protein system like Thioredoxins (Modi *et al.*, 2018), β-lactamase (Modi and Ozkan, 2018), GFP proteins (Kim *et al.*, 2015), etc. we have shown that changes in the flexibility profile of a protein, through DFI, is able to accurately capture the changes in function of the protein. Through computational studies, we have observed that during evolution, nature manipulates the dynamics of the protein by shifting its hinge residues—the hinge-shift mechanism, where some flexible residues become more rigid evolving into hinges, on the other hand, other

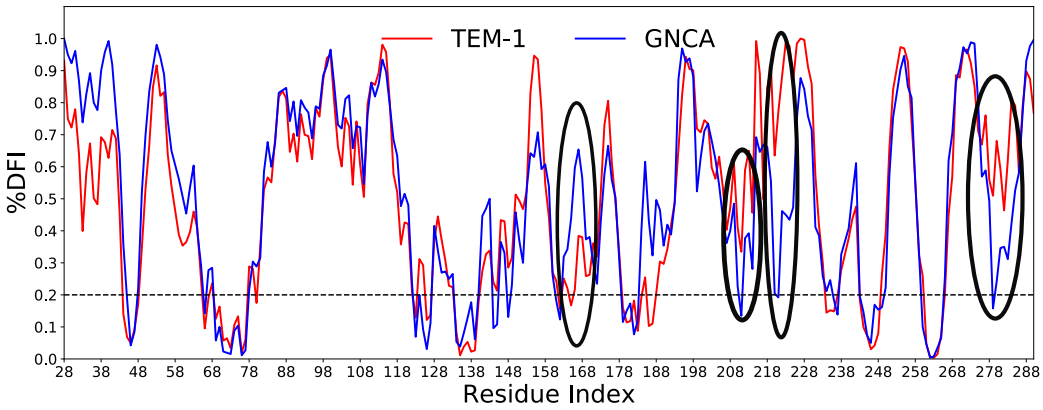

Figure 5.1: Comparison of the flexibility profile of ancestral β-lactamase (GNCA, blue) with the modern β-lactamase (TEM-1, red). We identify several regions focused around residues 166, 205, 223 and 280 where key differences (highlighted, black circles) in the dynamics of GNCA and TEM-1 β-lactamases, are observed. Typically, residues with the percentile rank of their DFI score (%DFI) less than 0.2 are deemed as rigid hinges. These residues have been observed to play a critical role in the functional dynamics of the protein.

rigid hinge positions give up their rigidity to become more flexible, leading to change in dynamics to adapt to a new environment or to evolve a new function (Modi and Ozkan, 2018; Modi *et al.*, 2018; Campitelli *et al.*, 2020; Kim *et al.*, 2015). In order to perform these hinge-shifts without sacrificing the 3D fold, the residues with moderate flexibility are usually substituted making them rigid. This process is accompanied by a loss of rigid regions in the ancestral proteins in the form of compensation in order to preserve the fold and stability of the protein. Here we attempt to manipulate the dynamics of the ancestral β-lactamase (GNCA) such that it emulates the dynamics and function of a modern β-lactamase (TEM-1). To achieve this, we target a minimum number of substitutions for the positions involved in hinge-shifts. This is done first in-silico and then computational predictions are characterized experimentally in order to validate and gain a deeper understanding of the underlying mechanism of evolution for modulation of the function from exhibiting promiscuous activity towards antibiotics to becoming a specialist.

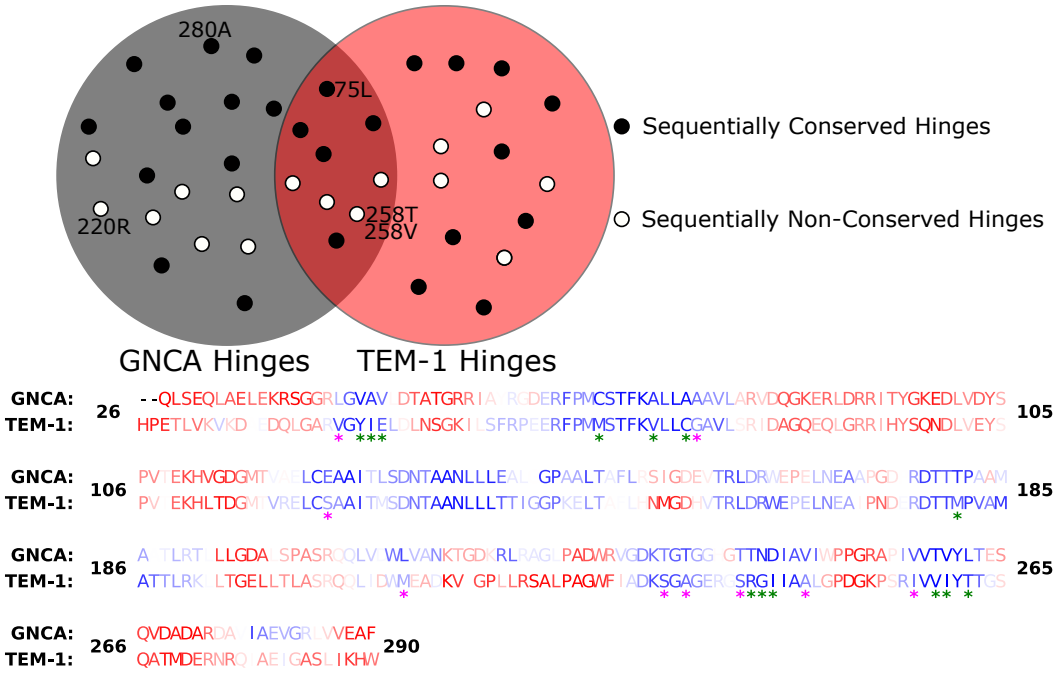

Figure 5.2: A Venn diagram showing the conservation of amino acids in the common and non-common hinges in GNCA and TEM-1 β-lactamase. The circles schematically (not to scale) represent residues in GNCA and TEM-1 β-lactamase with or without amino acid substitutions. In the amino acid sequence of TEM-1 and GNCA β-lactamases shown above, the letter codes are colored based on their DFI scores with blue being rigid and red as flexible. We observe several residue positions which have preserved their rigidity through evolution (common hinges, 41 in total) and also some with shifts in hinges (non-common hinges, 10 in GNCA and 11 in TEM-1 β-lactamase). Many of these residue positions have conserved the identity of their amino acids (i.e., sequentially conserved), while many have evolved into a different amino acid (hence, sequentially non-conserved). The residue positions which have maintained their rigidity without conserving their amino acid identity are highlighted by a green asterisk. The residues where flexibility increased or decreased beyond 0.2 (non-common hinges) along with a substitution are highlighted with a pink asterisk.

We first focus on the critical role played by hinges (i.e., low flexibility residue positions) in GNCA and TEM-1 β-lactamases. With the help of flexibility profile obtained by DFI, we have identified several hinge positions in ancestral, GNCA and the extant, TEM-1 β-lactamase which have preserved their low dynamic flexibility (common hinges) and also identified the positions that exhibit low-flexibility either in GNCA or TEM-1 β-lactamase, (non-common hinges) as shown in Figure 5.2.

In order to shift the dynamics of GNCA towards TEM-1 β-lactamase, we aim to use a rational design principle of the hinge-shift mechanism where we attempt to reproduce the shift in hinge locations made through evolution. We therefore substitute the positions of the residues in GNCA with the corresponding amino acids identified in TEM-1 β-lactamase in the following two sets:

- Non-common hinges and sequentially non-conserved residue positions (Set X): Here, we identify residue positions in GNCA β-lactamase which have been substituted in TEM-1 β-lactamase while becoming either flexible (hinge loss) or rigid (hinge gain) (residues highlighted with a pink asterisk in Figure 5.2). As discussed earlier, such residue positions are expected to play an important role in describing the functional landscape of the protein, and typically substitutions there lead to the modulation of dynamics and hence, the function (the hinge-shift mechanism for evolution (Modi and Ozkan, 2018; Kim *et al.*, 2015; Modi *et al.*, 2018; Zou *et al.*, 2015)). We follow this mechanism by considering sequentially non-conserved hinge residues in GNCA and TEM-1 β-lactamases for substitutions. In order to identify the minimum set of such hinges to replicate the desired change where it would shift the DFI profile of GNCA towards that of TEM-1 β-lactamase, we select only those non-common hinges in GNCA β-lactamase which are dynamically coupled with other non-common hinges in TEM-1 β-lactamase. The strength of coupling of a residue with another residue is quantified using DCI analysis, where we select only those residues which have a %DCI score of higher than a coupling threshold of 0.8 (Figure 5.3). Here, %DCI represents the percentile ranking of the DCI score of residues. Therefore, a residue with a %DCI score of greater than 0.8 would imply that its score is among the 20% of residues with the highest score. Further on, in the DCI analysis, similar to the calculation of DFI, the coupling between different residues of

GNCA and those of TEM-1 β-lactamase are calculated using covariance of fluctuations between pairs of residues obtained with the help of MD simulations (see section 2.2.4). For the analysis, we perturbed the sequentially non-conserved and non-common hinge positions and probed the coupling of the rest of the protein with these residues (see section 2.2.4). These selected residues in GNCA β-lactamase are then substituted with the amino acids at corresponding residue identities of TEM-1 β-lactamase. These residues are shown in Figure 5.4A.

- Common hinges and sequentially non-conserved residue positions (Set Y): As shown in Figure 5.1 and 5.2, there are a large number of residues which maintained their low flexibility, (i.e., remained as hinges) during the evolution from GNCA to TEM-1 β-lactamase (residues highlighted with a green asterisk in Figure 5.2). However, many of these positions are substituted in TEM-1 β-lactamase as we call them sequentially non-conserved common hinges. This suggests that these common hinges are crucial for the dynamics underlying the function, therefore, these positions must exhibit long distance communications with other non-common hinges within the 3D network of interactions. Because of the shifts in hinges in other part of the protein, some of these dynamically conserved positions (aka common hinges) need to be substituted in order to compensate the change in flexibility of the hinge-shift substitutions in other parts. Hence, we identify the sequentially non-conserved and common hinges in GNCA and TEM-1 β-lactamase which exhibit high dynamic coupling to the other sequentially non-conserved and non-common hinge positions obtained using the DCI analysis. In this case, we perturb the sequentially non-conserved and common hinges in GNCA and TEM-1 β-lactamase in order to calculate the DCI score of the rest of the chain with respect to these residues. From

the analysis, highly coupled residues are selected which exhibit a high %DCI score with a threshold value of 0.8 (Figure 5.3B). Afterwards, we substitute the residue positions selected in mutation set Y in GNCA β-lactamase with the amino acids in the corresponding residue ids in TEM-1 β-lactamase. These residues are shown in Figure 5.4B. Further details of residues selected in the Set Y are provided in the Figure A.1 in Appendix A.

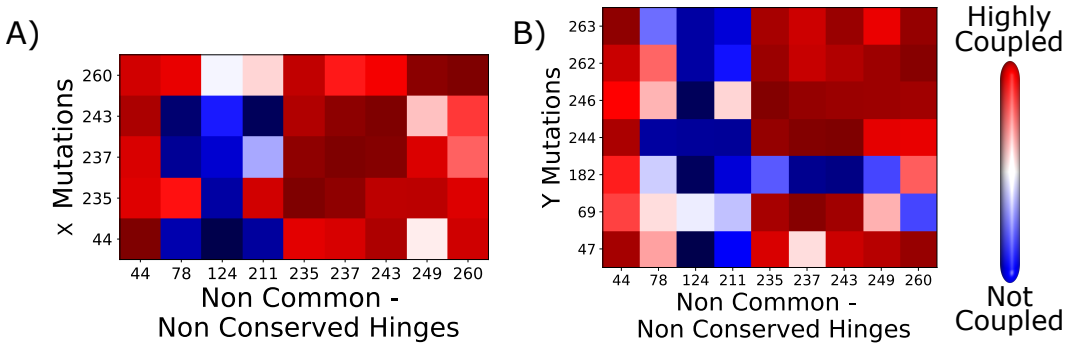

Figure 5.3: Criteria for selecting substituting residues in sets X and Y. The coupling of the selected common and sequentially non-conserved hinges comprising (A) X mutation set in GNCA and TEM-1 with other Non-common and sequentially non-conserved hinges in GNCA and TEM-1 β-lactamase, and (B) Y mutation set in GNCA and TEM-1 with other Non-common and sequentially non-conserved hinges in GNCA and TEM-1 β-lactamase. The hinge residues selected for the two sets are strongly coupled (%DCI > 0.8) to other such non-common and non-conserved hinge residues in both GNCA and TEM-1.

We analyzed the impact of the substitutions from set X and set Y on the flexibility profiles of GNCA β-lactamase. In order to do so, we synthesized the mutants *in silico* using PyMol (Schrödinger, LLC, 2015). The mutant containing substitutions from set X is called GNCA-X and the one with substitutions from set Y is called GNCA-Y. Afterwards, we performed MD simulations of the mutants (see section 2.1.1) to obtain 400ns of equilibrium dynamics. Thereafter, using a covariance matrix of size 50ns sampled from a time slot of 100ns to 400ns from the trajectory, we compute the residue-specific flexibility profiles of the two proteins using DFI (see sections 2.2.2 and 2.2.3).

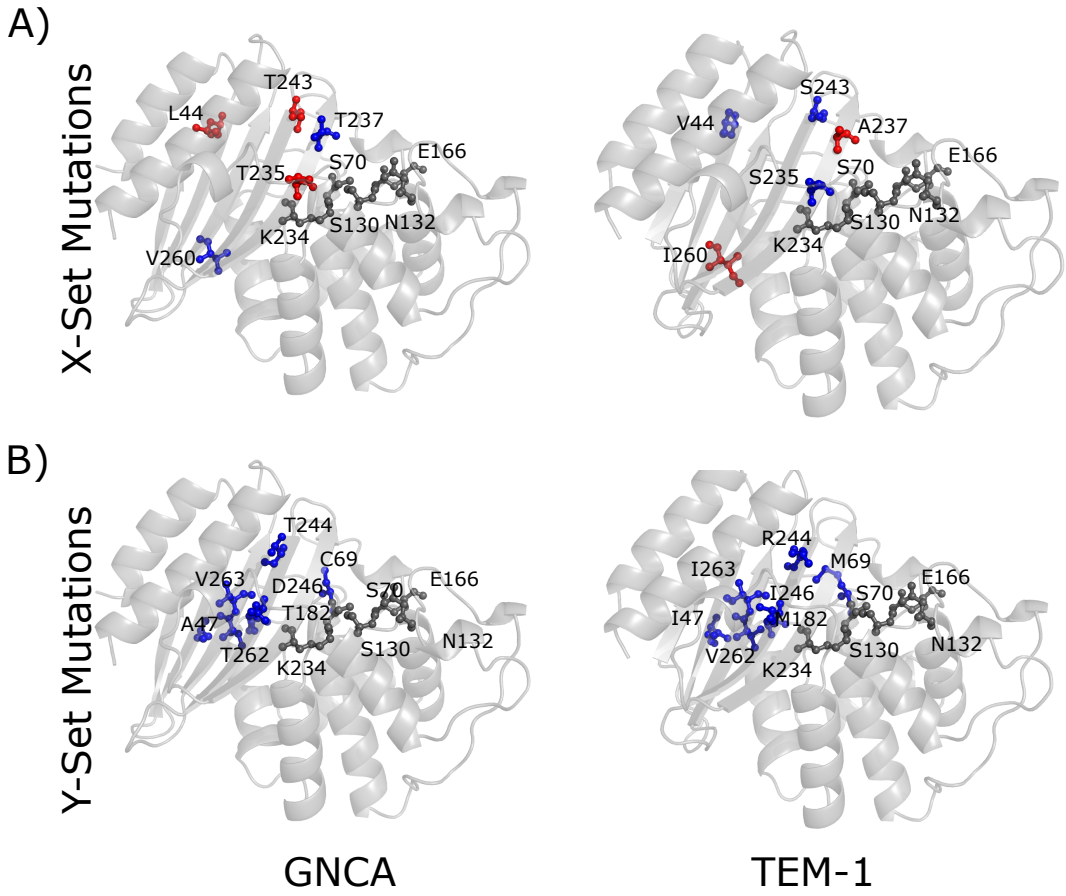

Figure 5.4: DFI scores of residues selected for substitution in sets X and Y. (A) Non-common and sequentially non-conserved residues substituted in set X shown as sticks on the cartoon representation of GNCA and TEM-1 β-lactamase. (B) Common and sequentially non-conserved residues substituted in set Y shown as sticks on the cartoon representation of GNCA and TEM-1 β-lactamase. The substituting residues are colored based on their DFI profile where blue sticks represent residues with low DFI (hinge) and the red sticks represents residues with high DFI (flexible/non-hinge). The catalytic positions are shown in dark gray.

We observe that, substitutions from set X has impacted the dynamics of GNCA such that a few of its residues share their dynamic flexibility with TEM-1 β-lactamase, creating hinge shifts particularly around residues 185 (see Figure 5.5). In addition, the critical role played by the dynamical coupling between the hinges selected in set X and the rest of the non-common and non-conserved can also be observed by performing substitutions at all the non-common and non-conserved hinge positions (see Figure A.3 in Appendix A). Mutating all the non-common and non-conserved

hinges in GNCA (forming the mutant GNCA-AllNN) have a major impact on the dynamics of the catalytic pocket of GNCA β-lactamase, driving the flexibility profile away from either of the wild types. On the other hand, performing the substitutions at residues from the set Y has changed the flexibility profile of GNCA quite significantly and as a result, GNCA-Y no longer shares the dynamic similarities with either of the two enzymes (see Figure A.2 in Appendix A).

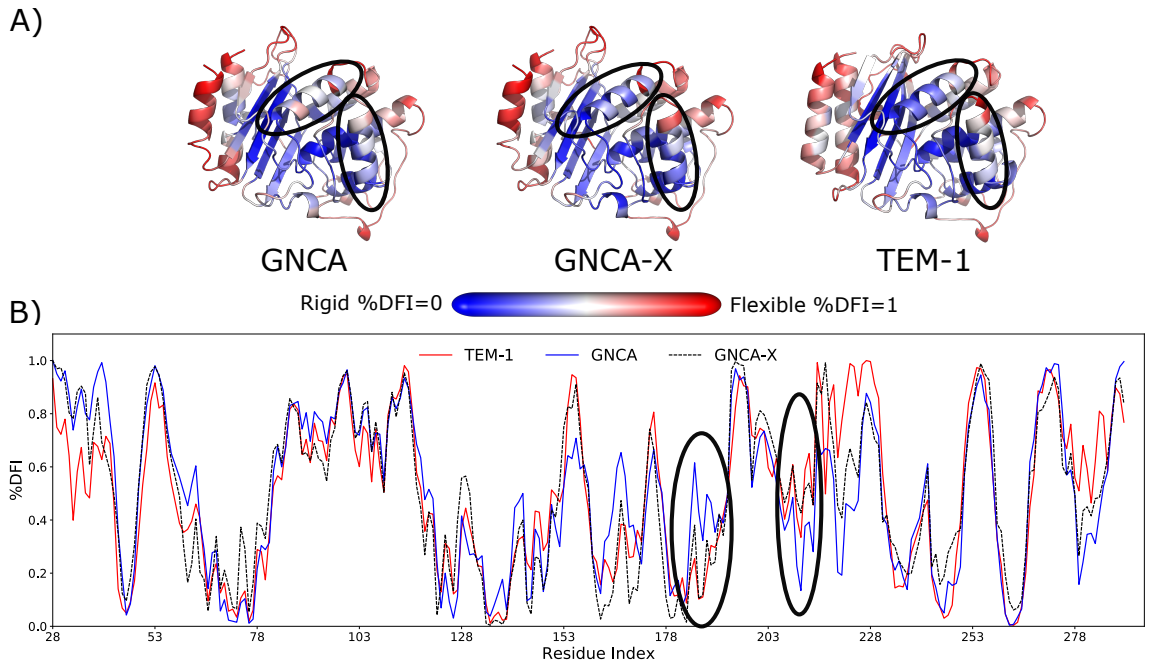

Figure 5.5: Impact of substitutions from set X on the flexibility profile of GNCA β-lactamase. (A) This is shown by comparison of their cartoon representations color coded with the flexibility profile. These are also shown on (B) where the %DFI profile of the mutant GNCA-X is compared with the wild types GNCA (blue) and TEM-1 (red) β-lactamase. The regions where GNCA-X mimics the flexibility profile of TEM-1 are highlighted.

Further on, the substitutions from set X and Y can be combined into one set (set XY) to create another mutant (GNCA-XY). We synthesized the mutant *in silico* and performed all atom equilibrium MD simulation (see section 2.1.1 to obtain 600ns of simulation data. Thereafter, using a covariance matrix of size 50ns sampled from a time slot of 200ns to 600ns from the trajectory, we compute the residue-specific

flexibility profiles of the two proteins using DFI analysis (see sections 2.2.2 and 2.2.3). Combining the substitutions from set X and Y brought the dynamics of GNCA closer to TEM-1 β-lactamase, Figure 5.6A,C. Particularly, set XY mutations was able to induce hinge-shifts in the GNCA β-lactamase around residues 220 and 280, that bring the flexibility profile XY mutant closer to TEM-1 β-lactamase.

A better approach to evaluate the degree of impact of these mutations on the dynamical flexibility is with the help of principal component analysis (PCA). For this, we align and arrange the %DFI profiles of the mutants along with those of GNCA and TEM-1 β-lactamase such that each protein can be represented by a vector of dimension N, where N represents the total number of residues in each protein after alignment. This data can be stored in a matrix which can be decomposed to give its principal components (the details of this method are described in section 2.2.5). Thereafter, we compare the projection of the vectors representing proteins along lowest principal components in the vector space in order to observe the salient features differentiating them. Comparing the lowest two principal components of the flexibility profiles of the proteins showed that, as expected, GNCA-XY is indeed closer to that of TEM-1 β-lactamase (Figure 5.6B) suggesting that these substitutions may make the designed ancestral enzyme more specific. In addition, this analysis also indicates that the mutations from set X and set Y have a non-additive impact on the dynamical landscape of GNCA β-lactamase. This points towards a possible role of epistasis (Romero and Arnold, 2009; Wu *et al.*, 2019; Miton and Tokuriki, 2016; Zhang *et al.*, 2012; Weinreich *et al.*, 2018) between the substitutions from sets X and Y on the background of GNCA β-lactamase.

Even though the PCA analysis showed similarities between the flexibility profiles of the designed mutant and wild type TEM-1 β-lactamase, we still see that there are some significant differences between the hinges of TEM-1 β-lactamase and mutant

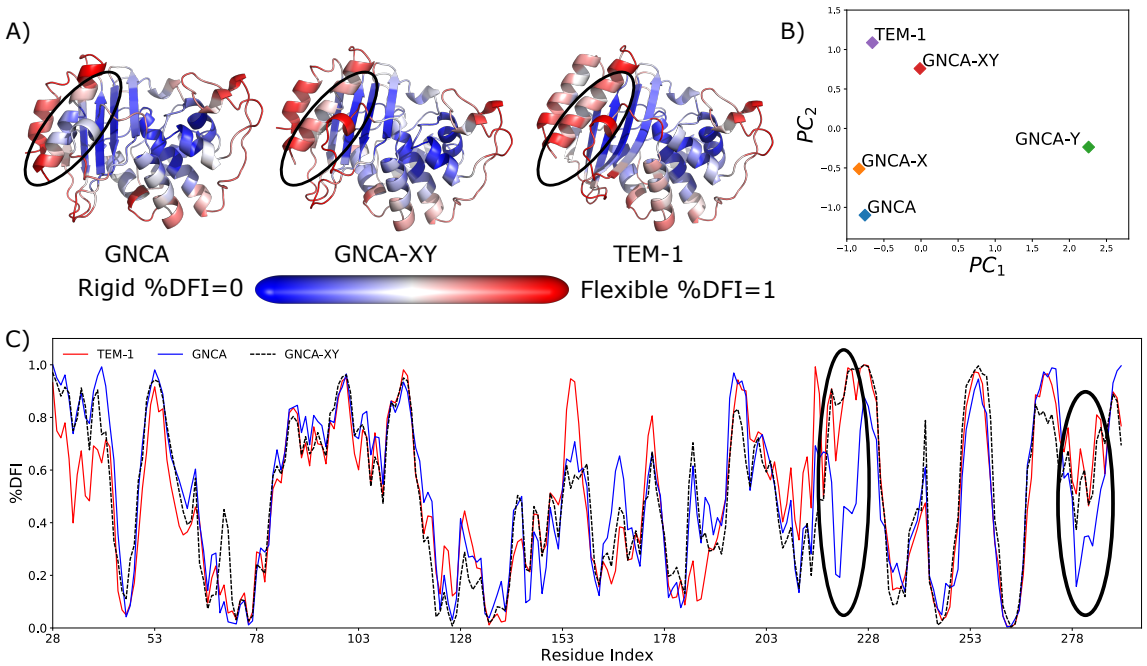

Figure 5.6: Comparing the flexibility profile of wild type GNCA and TEM-1 β-lactamase with the GNCA mutant created by performing mutations from sets X and Y together (GNCA-XY) using principal component analysis (B). It is observed that GNCA-XY mimics the flexibility profile of TEM-1 β-lactamase more closely as compared to the mutants created by mutation sets X and Y alone, particularly around residues 220 and 280 (highlighted) (C), also seen in their cartoon representations (A) color coded with the flexibility profile of their residues, red being flexible and blue rigid.

GNCA-XY. Especially in some regions around residues 203-215, 180-190 and the C terminal of the protein, the mutant was not able to replicate the flexibility profile of TEM-1 β-lactamase. Particularly, some of these positions where mutations failed to recapitulate flexibility profiles of TEM-1 β-lactamase correspond to the non-common hinges which are also conserved between GNCA and TEM-1 β-lactamases, hence we could not simply introduce substitutions by inferring the sequence variations between the ancestral and the extant β-lactamases as we did for the other non-conserved and non-common hinges. Therefore, in an attempt to bring the GNCA-XY mutant even closer to TEM-1 β-lactamase, we now focus on the flexible sites that exhibit allosteric dynamic coupling interactions with the catalytic site. These distal sites

which are highly coupled to the catalytic sites through an allosteric network of dynamic interactions are called Dynamic Allosteric Coupling (DARC) spots. These play a critical role in the evolution of protein dynamics towards new function (Modi and Ozkan, 2018; Campitelli *et al.*, 2020; Butler *et al.*, 2018; Kumar *et al.*, 2015b).

5.2.3 Role of Allostery Through Non-Invasive DARC Spots (Set Z) Brings Mutant GNCA-XY Closer to TEM-1 β-lactamase.

Through our previous ancestral studies, we have observed that nature introduces substitutions at the DARC spots that are distal from the active site, where these mutations act as small perturbative changes, and modulate the dynamics of the protein/catalytic site in order to evolve a new function or adapt to new environment (Modi and Ozkan, 2018; Modi *et al.*, 2018; Campitelli *et al.*, 2020). Indeed, this is also true for β-lactamases. First, we observed that a large fraction of the mutations (the clinically isolated mutations or those that emerged from directed evolution) are far from the active site, yet they modulate the equilibrium dynamics of the protein to confer resistance to antibiotics (Modi and Ozkan, 2018; Campitelli *et al.*, 2020). Furthermore, the DFI and DCI analysis of the exhaustive set of 5000 mutations in TEM-1 β-lactamase (Stiffler *et al.*, 2015) have shown that the sites exhibiting mid-range flexibility and high dynamic coupling with the active site contribute most to the emergence of degrading different antibiotics (Modi and Ozkan, 2018; Modi *et al.*, 2018; Campitelli *et al.*, 2020). Therefore, in order to emulate nature, we also identify mid-flexible residues that are allosterically coupled to the active site (i.e., are DARC spots (Modi and Ozkan, 2018; Campitelli *et al.*, 2020; Butler *et al.*, 2018; Kumar *et al.*, 2015b)). Since the active site positions are the rigid sites exhibiting low DFI throughout the evolution, they are the part of the sequentially conserved and common hinges. However, by introducing substitution at the DARC spots, we also aim

to induce hinge shifts at non-common hinges, which are also sequentially conserved. Hence, we only select the DARC spot residues that are not only distal (i.e., they are more than 8Åaway from the catalytic pocket) but also exhibit high dynamic coupling (i.e., %DCI > 0.8) with such non-common and sequentially conserved hinge sites (see Figure 5.7). These steps were taken in order to minimize the deleterious impact of the substitutions on the thermal stability and the kinetic activity of the protein. The residues closer to the active sites typically exhibits higher dynamic coupling with the active sites as they are directly interacting; however, the substitutions at those sites are typically more invasive, most likely impact the dynamics and hence, the function. Therefore, selecting DARC spots reduces the perturbative impact of mutations as they are relatively more flexible, which agrees with our earlier proteome wide analysis showing that evolving sites are usually flexible sites. Due to their high flexibility, they can compensate for changes upon substitution (Campitelli *et al.*, 2020; Butler *et al.*, 2015; Kumar *et al.*, 2015b). However, these DARC spots are not only flexible, but they also exhibit high dynamic coupling with the active or functionally important hinges sites. Therefore, the dynamics of the protein can tolerate the effect of substitutions at such residues allowing us to make fine changes by modulating the dynamics of the other functional sites (Modi and Ozkan, 2018; Modi *et al.*, 2018; Campitelli *et al.*, 2020; Nevin Gerek *et al.*, 2013). Furthermore, the fact that these DARC spots exhibit high dynamic coupling with the non-common and sequentially conserved hinges allow us to exploit their compensatory network interactions to modulate the flexibility of these hinges.

Thereafter, we performed MD simulations of the mutant generated by performing mutations from set Z on GNCA-XY mutant (GNCA-XYZ) following the protocol described in section 2.1.1. Using this, 1μs of equilibrium simulation trajectory was obtained. Thereafter, using a covariance matrix of size 50ns sampled from a time

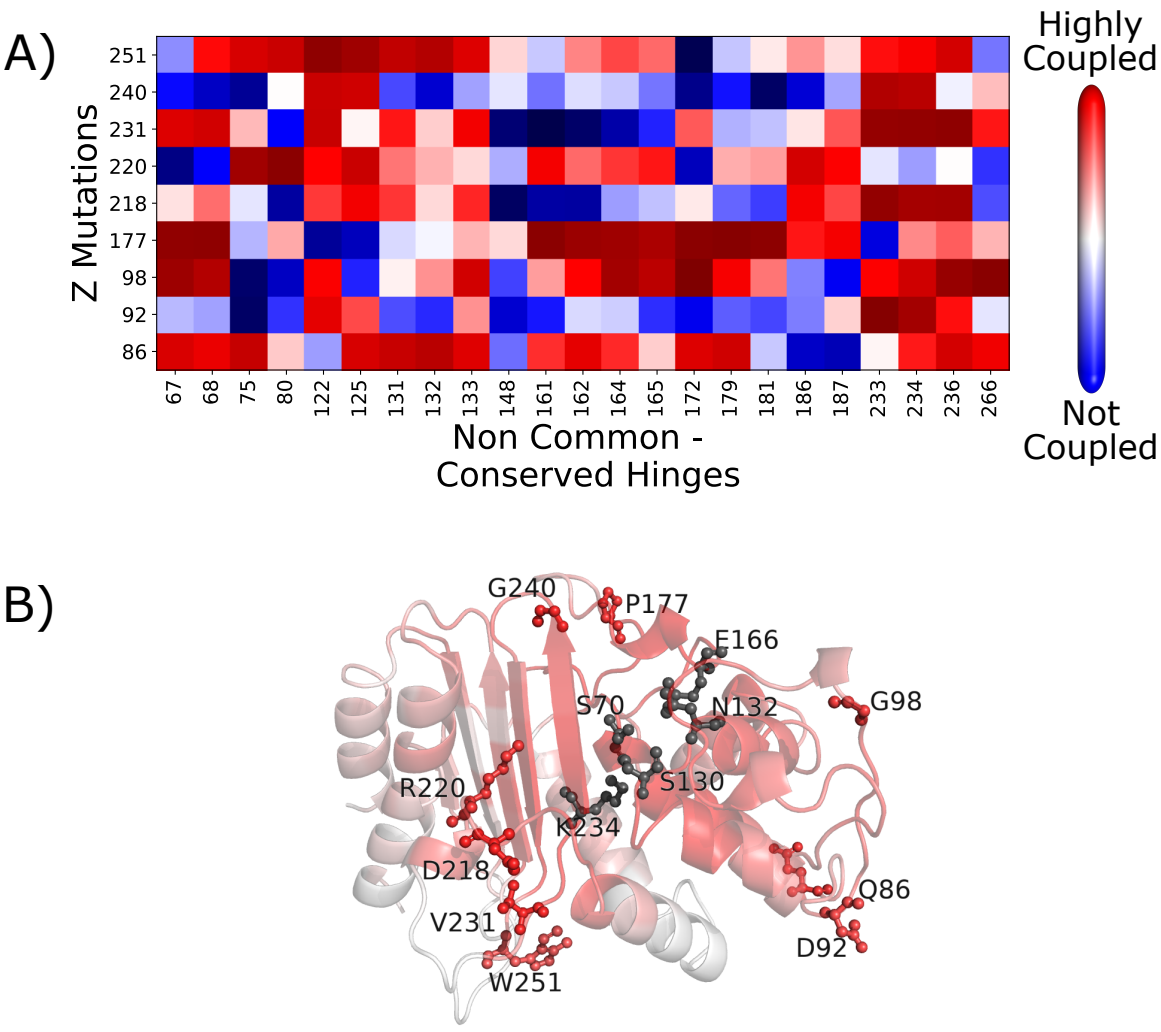

Figure 5.7: Criteria for selecting residues for substitution in Set Z. (A) The coupling of residues selected for substitutions in set Z with the non-common and sequentially non-conserved hinges. We have selected these as they exhibit higher coupling (%DCI > 0.8) with the non-common and sequentially non-conserved hinges. Apart from this, these residues also have a medium flexibility and are distally located from the active site (> 8Å). (B) In addition, the residues in set Z are also coupled to the active site as shown by the cartoon representation of GNCA β-lactamase where each residue is color coded with the %DCI score with respect to the active site. Red colored residues are the most coupled (%DCI=1) and white are the least coupled (%DCI = 0). The residues in set Z are shown as sticks. We can observe that these also show a high coupling with the active site (%DCI > 0.8).

slot of 300ns to 1000ns from the trajectory, we compute the residue-specific flexibility profiles of the two proteins using DFI (see sections 2.2.2 and 2.2.3).

Afterwards, we analyzed the impact of the substitutions from set X, Y and Z together by clustering this DFI profile with the DFI profiles of the wild type GNCA and TEM-1 β-lactamase and those of other mutants, as shown in Figure 5.8. The comparison of the flexibility profiles of the GNCA mutant with mutations from set XY and set Z together and the wild type GNCA and TEM-1 β-lactamases (Figure 5.8A,C) shows that the mutant successfully recapitulates the dynamics of TEM-1 β-lactamase. It is able to strengthen the shortcomings of GNCA-XY mutant in regions close to 180-190, 203-215 and 153-157. This is also indicated by the PCA analysis of the mutants from sets X, XY and XYZ with the wild type proteins by comparing their first two principal components, Figure 5.8C. We observe that, as expected, the mutant GNCA-XYZ lies very close to the wild type TEM-1 β-lactamase, which is an improvement over mutants with mutations from set X and XY suggesting that the ancestral variant GNCA-XYZ with merely 21 substitutions should degrade only penicillin with better efficiency than GNCA β-lactamase, mimicking the catalytic activity of TEM-1 β-lactamase. Therefore, using these 21 substitutions, we are able to dynamically replicate the effect of a total of 119 substitutions observed between GNCA and TEM-1 β-lactamase.

5.2.4 Experimental Characterization Confirms Antibiotic Specificity of GNCA-XYZ

Based on the mutation sets discussed in the last sections, the mutants GNCA-X, GNCA-XY and GNCA-XYZ were synthesized. Further on, the activity of the synthesized mutants against the antibiotics cefotaxime (CTX) and benzyl-penicillin (BZ) was also characterized (see Appendix A). The experimental characterization of wild type ancestral and extant β-lactamases shows that, as shown by previous studies (Risso *et al.*, 2013), GNCA β-lactamase is promiscuous in its activity towards

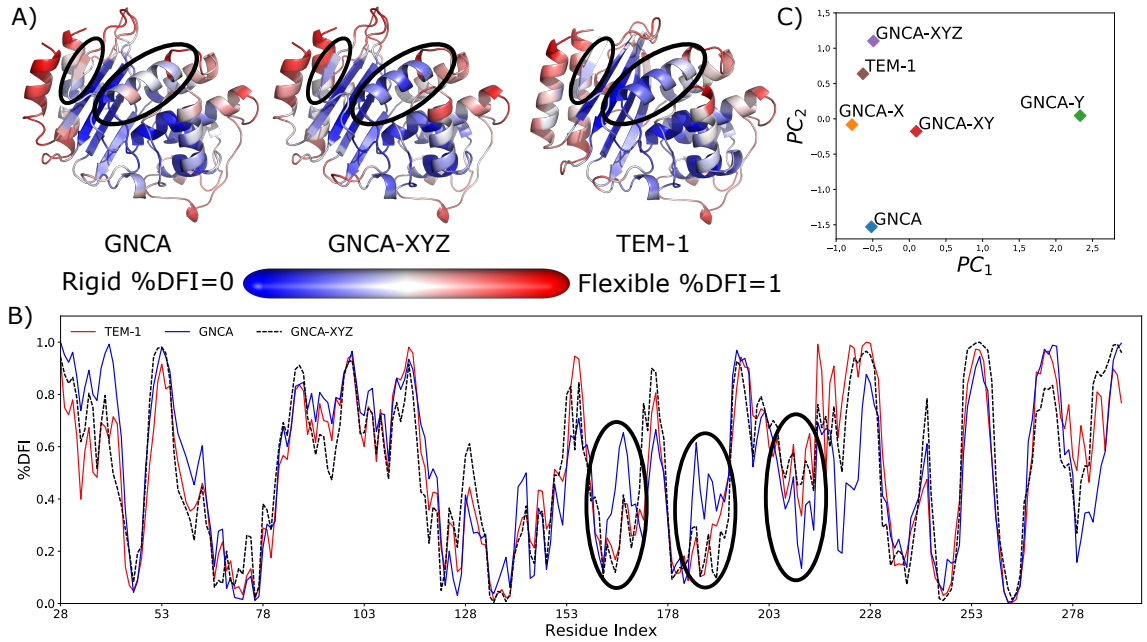

Figure 5.8: (A and B) Comparing the flexibility profile of wild type GNCA and TEM-1 β-lactamase with the GNCA mutants created by performing mutations from set Z over GNCA-XY (GNCA-XY). It is observed that GNCA-XYZ very closely mimics the flexibility profile of TEM-1 β-lactamase, particularly around 185, 155, 210, and the C-Terminus (highlighted regions in B). The cartoon representations are color-coded with the flexibility profile of their residues, red being flexible and blue being rigid. This can also be observed from the comparison of their lowest two principal components (C)

antibiotics BZ and CTX by not showing any selective preference towards any of them (with turnover rates of $0.3s^{-1}\mu M^{-1}$ and $1.2s^{-1}\mu M^{-1}$ respectively). On the other hand, evolution has turned TEM-1 β-lactamase into a specialist which preferentially catalyzes only BZ with a higher turnover rate ($26s^{-1}\mu M^{-1}$) in contrast with CTX ($2.6 \times 10^{-3}s^{-1}\mu M^{-1}$).

The mutants proposed here through our computational analysis by performing substitutions from set X (GNCA-X) reduced the turnover rates for CTX ($2.5 \times 10^{-4}s^{-1}\mu M^{-1}$). However, it has also rendered it ineffective towards BZ by reducing its turnover rate by a factor of ten ($0.03s^{-1}\mu M^{-1}$). We did not perform an experimental characterization of GNCA-Y as it was rejected as a standalone in our initial round of

computational analysis. The turnover rates of the mutant with the combined mutations from X and Y (GNCA-XY) showed an improvement over the GNCA-X (as predicted to be slightly closer to TEM-1 β-lactamase), where its turnover rate for BZ did not show appreciable change ($0.22 s^{-1}\mu M^{-1}$). The mutant however showed a remarkable decrease in its turnover rate for CTX ($2.6 \times 10^{-5} s^{-1}\mu M^{-1}$), thus agreeing with the predicted analysis that introduced substitutions slowly drives GNCA β-lactamase from promiscuity to specificity, by reducing the catalysis of CTX while preserving the turnover rate for catalysis of BZ. Interestingly, with the addition of substitutions at DARC spots from set Z, the mutant GNCA-XYZ becomes more specific in its activity towards BZ by a three-fold increase in its turnover rate ($0.9 s^{-1}\mu M^{-1}$). Moreover, its turnover rate towards CTX also showed a significant reduction ($< 1 \times 10^{-4} s^{-1}\mu M^{-1}$) making it more preferential towards BZ which is a functional characteristic property of TEM-1 β-lactamase. This data is shown in more detail in Table 5.1 and the method is described in more detail in Appendix A.

As a control, we also wanted to check if the substitutions that drove GNCA towards being a specific enzyme with enhanced degradation rate for BZ are due to the ones that are closer to the active site. Therefore, we also studied the specific mutants with substitutions in X, Y and Z which lie with relatively closer to the catalytic site in β-lactamase enzymes (T235S, T237A, T243S in GNCA$^{\text{T235S_T237A_T243S}}$ and T235S, T237A, T243S, C69M in GNCA$^{\text{T235S_T237A_T243S_C69M}}$). Upon experimental characterization of the activity of these new mutants, we observe that these mutations alone have rendered the mutant ineffective to catalyze both BZ as well as CTX as shown by their turnover rates (Table 5.1). This emphasizes the importance of allosteric interactions in modulating the function of an enzyme and the key part they play in evolution. It is also interesting to point out the turnover rates of GNCA-AllNN where again as predicted, we observe a striking reduction in the turnover rates for catalysis

Table 5.1: Experimental characterization of wild type β-lactamase GNCA and TEM-1, and the mutants created by mutation sets X, Y and Z by calculating their turnover rates for catalysis of antibiotics benzyl-penicillin (BZ) and cefotaxime (CTX) (all values are in $s^{-1}\mu M^{-1}$).

Protein	**BZ** (k_{cat}/k_M)	**CTX** (k_{cat}/k_M)
TEM-1	26 ± 4.7	$2.6 \times 10^{-3} \pm 1 \times 10^{-3}$
GNCA	0.3 ± 0.1	1.2 ± 0.3
GNCA-X	0.03 ± 0.01	$2.5 \times 10^{-4} \pm 5 \times 10^{-5}$
GNCA-XY	0.22 ± 0.1	$< 1 \times 10^{-4}$
GNCA-XYZ	0.9 ± 0.3	1×10^{-4}
$\text{GNCA}^{\text{T235S\_T237A\_T243S}}$	0.03 ± 0.01	$3.2 \times 10^{-4} \pm 1 \times 10^{-4}$
$\text{GNCA}^{\text{T235S\_T237A\_T243S\_C69M}}$	0.05 ± 0.01	$3.0 \times 10^{-4} \pm 1 \times 10^{-4}$
GNCA-AllNN	0.03 ± 0.006	$1.7 \times 10^{-4} \pm 1.7 \times 10^{-5}$

of BZ as well as CTX.

Subsequently, we also obtained the crystal structure of the final engineered enzyme, GNCA-XYZ (see Appendix A for crystallographic details). We observed that the substitutions from mutation set X, Y and Z has preserved the three-dimensional fold of the enzyme (RMSD < 1 Å) (Figure 5.9A). We computationally characterized the flexibility profile of the mutant by obtaining its dynamics through an MD simulation using the new crystal structure as the starting point and then calculating the DFI flexibility profile (GNCA-XYZ(X-ray)). Afterward, we compared this flexibility profile with the DFI profile predicted earlier of the mutant and that of wild type TEM-1 β-lactamase, (Figure 5.9B). We observe that, as predicted, the flexibility profile of the engineered mutant is very similar to the DFI profile of the wild type TEM-1 β-lactamase.

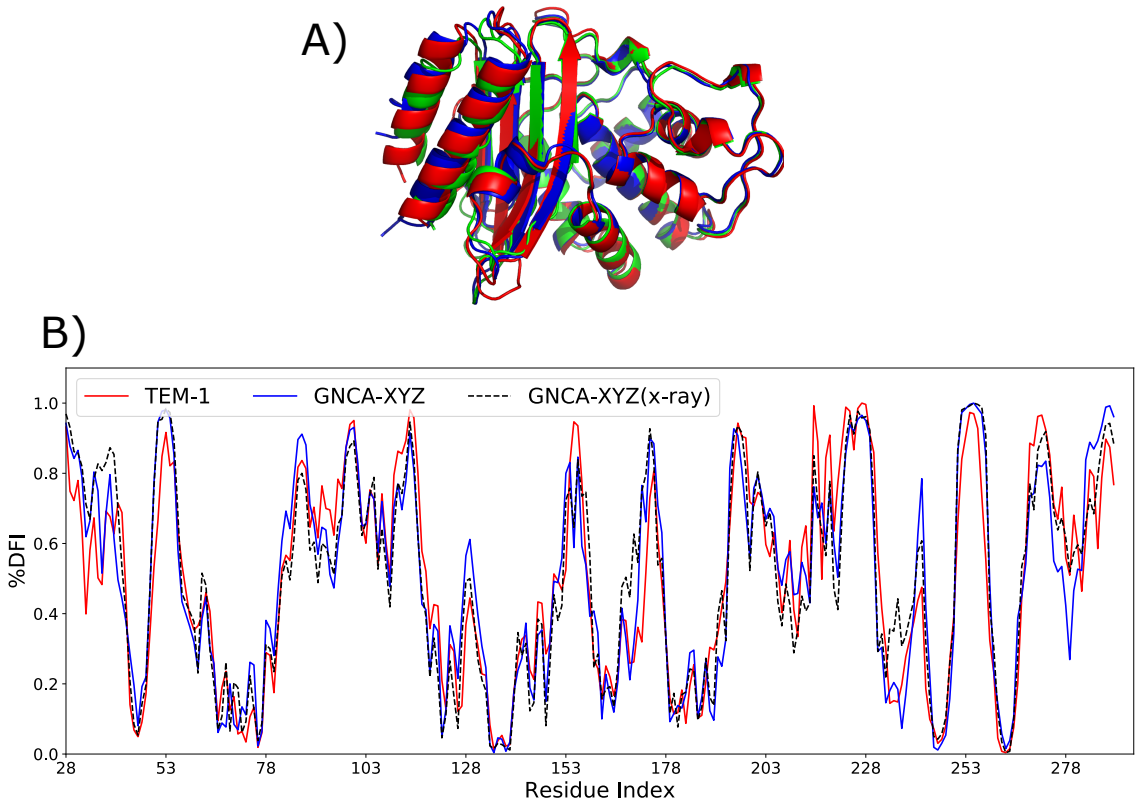

Figure 5.9: Comparing the structure and flexibility profile obtained from x-ray crystallography of the synthesized mutant GNCA-XYZ with the predicted flexibility profile. (A) Superimposing the cartoon representation of the crystal structure of GNCA-XYZ (blue) obtained from X-ray crystallography over the wild types GNCA (green) and TEM-1 β-lactamase (red) shows that it shares the same 3D fold as the wild type protein (RMSD $< 1\mathring{A}$). (B) Comparing the DFI profile of the mutant GNCA-XYZ calculated using MD simulation starting from the crystal structure obtained by X-ray crystallography (black broken line) with the DFI profile of the wild type TEM-1 β-lactamase (red). We observe that the mutant, as predicted (blue), is able to successfully mimic the rigid regions of TEM-1 β-lactamase, and also its flexible regions with remarkable accuracy. Moreover, the DFI profile of the mutant calculated using the structure obtained through X-ray crystallography also matches with our prediction obtained through *in silico* mutations.

5.2.5 NMR Analysis Shows Dynamical Differences between Wild Type GNCA And Mutant GNCA-XYZ

In the previous sections, we have computationally designed GNCA-XYZ function through attempts to modulate its dynamics by substituting hinge positions and re-

gions coupled to hinge positions with the goal of shifting dynamic flexibility profile of GNCA towards that of TEM-1 β-lactamase, (Figure 5.8). As DFI profiles have been shown to correlate with functional outcomes (Modi *et al.*, 2018; Campitelli *et al.*, 2020), our hypothesis, that the function of the engineered mutant (GNCA-XYZ) should behave more like TEM-1 β-lactamase, with regards to the specificity and catalytic activity. We have shown that the experimental characterization of GNCA-XYZ supports the hypothesis as we observe a significant difference in the turnover rates of GNCA-XYZ with antibiotics CTZ and BZ as compared to ancestral and extant enzymes (Table 5.1). Moreover, the GNCA-XYZ structure from X-ray crystallography confirms that the 3D structure is preserved, indicating that the engineered 21 substitutions have introduced changes only in dynamics. In order to further validate the computational predictions that the changes in dynamics govern the function of GNCA-XYZ, we performed solution NMR experiments. A standard protein NMR experiment couples amide proton and nitrogen atoms giving "probes" throughout the protein backbone, which results in a fingerprint type identifying 2D spectrum. These HSQC-based (heteronuclear single quantum coherence) experiments allow for a qualitative assay of the folded state and structural ensemble of a given protein in solution. One attribute of HSQC data is the proton dimension dispersion, consistent with the X-ray crystallography data, both GNCA and GNCA-XYZ have broad proton dispersion indicating that both proteins are well-folded in solution. A qualitative assessment of dynamics can also be gleaned by these experiments. For a well-structured rigid protein, there is nearly a one-to-one correlation between the number of NMR spectral resonances and the number of residues, which arises naturally as the HN bond is a de facto probe and gives rise to a discreet resonance. In the context of a protein with increased dynamics (which is GNCA-XYZ in this case), the correlation between the resonance number and residue number can diverge away from parity depending on

the timescales associated with conformational fluctuations between states. Consistent with a well-folded and relatively rigid protein, the HSQC data from GNCA identifies over 90% (253 resonances) of the expected number of resonances, which is consistent with GNCA being rigid. GNCA-XYZ, on the other hand, shows about 60% (158 resonances) of the expected resonances, which is consistent with protein dynamics on the intermediate NMR timescale (Jaudzems *et al.*, 2010) (see Figure 5.10).

Beyond resonance number and proton dispersion, the HSQC peak resolution and intensity heterogeneity can also serve as qualitative measures of protein dynamics. Comparatively, GNCA-XYZ has lower peak resolution and increased peak intensity heterogeneity, which is consistent with it having increased backbone dynamics over the relatively rigid GNCA protein. Taken together, the NMR data complement the crystallography by showing that GNCA-XYZ is well structured in solution. Similarly, the data support the computational predictions that GNCA-XYZ has increased dynamics compared to GNCA due to the loss of hinges after substitutions from the XYZ set as observed from the comparison of the DFI profile of the mutant with the wild type GNCA β-lactamase (see Fig. 3). This offer validation to the computationally predicted hypothesis of protein dynamics contributing to the regulation of protein function. The method is described in more detail in Appendix A.

5.3 Conclusion

In this study, we developed a design principle to evolve the antibiotic activity of ancestral β-lactamase (GNCA) to mimic the activity o its extant counterpart (TEM-1 β-lactamase) by focusing only on the differences in their dynamics obtained through DFI profiles. It has been observed that the two proteins share relatively high sequence identity (about 50%) as well as a common 3D structure (Zou *et al.*, 2015; Risso *et al.*, 2013). However, despite having these similarities, they are functionally divergent.

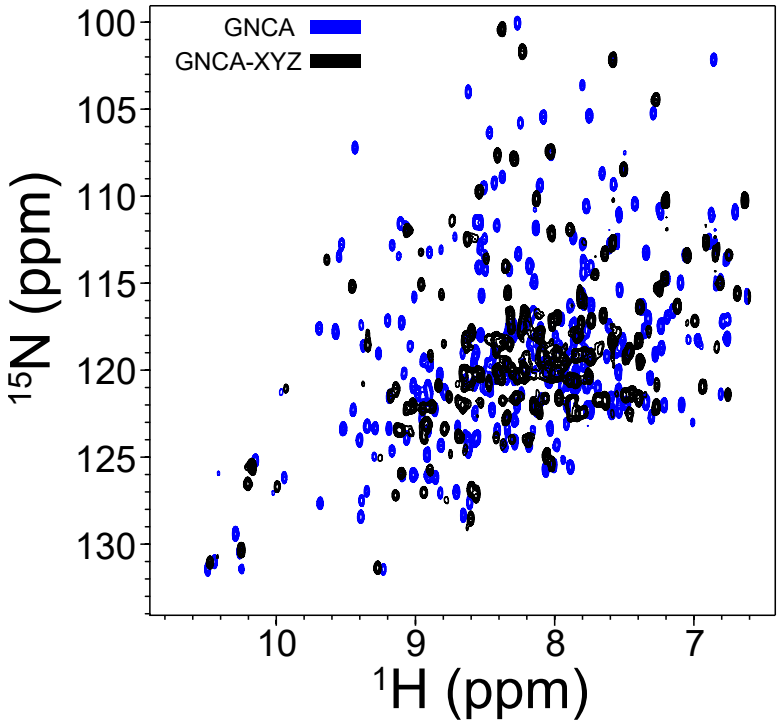

Figure 5.10: HSQC NMR data indicate that GNCA-XYZ is more dynamic than GNCA β-lactamase. An overlay of GNCA-XYZ (black) and GNCA (blue) ^{15}N-HSQC spectra. Both proteins are well-structured as noted in the >3 ppm proton dispersion and consistent with the X-ray structures of a mixed α-helix and β-sheet structure. However, GNCA-XYZ shows spectral features consistent with increased protein dynamics. Specifically, GNCA-XYZ shows fewer resonances, less peak resolution, and increased peak intensity heterogeneity; all are markers of increased protein dynamics relative to GNCA. Data were collected at 850 MHz ^{1}H frequency and 30 °C.

Specifically, ancestral β-lactamase exhibits promiscuity towards the catalysis of the two types of antibiotics (CTX and BZ) by having their turnover rates in the proximity of each other. On the other hand, extant β-lactamase shows more specificity towards its choice of antibiotic by having a 10000-fold difference in its turnover rate for CTX and BZ (Risso *et al.*, 2013).

As shown by our previous studies (Modi and Ozkan, 2018; Modi *et al.*, 2018; Campitelli *et al.*, 2020; Glembo *et al.*, 2012; Nevin Gerek *et al.*, 2013; Butler *et al.*, 2018; Kumar *et al.*, 2015a; Kim *et al.*, 2015), such functional diversity between ancestral and extant enzymes can be explained by focusing on the convergent as well

as divergent features in the flexibility profile of their residues obtained through our computational analysis (DFI). The flexibility profiles were obtained through DFI by performing molecular dynamic simulations of the enzymes. Through DFI, the enzyme residues are separated into rigid and flexible zones, and it was observed that the evolution has manipulated the catalytic activity of the enzyme by performing subtle shifts in the hinge regions of their DFI profiles. This mechanism of hinge-shift was also observed in the evolutionary history of other protein systems like Thioredoxin (Modi *et al.*, 2018) and GFP (Kim *et al.*, 2015). Following this, the hinges in GNCA and TEM-1 β-lactamase, were classified based on whether they are substituted between the two enzymes and if their rigidity is maintained through evolution.

In order to mimic the shift in hinges, we first focused on a subset of hinges that have altered their dynamics by gaining flexibility. On the other hand, some other flexible sites were rigidified and turned to hinges through substitutions (set X). Through this, we were able to successfully emulate hinge-shifts in several regions of the protein (see Figure 5.5). Second, we turned our attention to hinges, which have retained their rigidity despite having substitutions. This indicates the critical role played by these residues in the enzyme's 3D network of interactions to mediate dynamics. However, these common hinges need to be substituted to compensate for the new hinge formations observed by substitution in X set. Thus, we identified Y set substitutions by measuring the long-range dynamic interaction with the X set positions through our DCI metric. Substitutions from set Y, on their own, had a deteriorating impact on the dynamics of the protein (see Figure A.2). However, together with set X (called set XY), these substitutions compensate the impact of the previous substitutions and bring the flexibility profile of the mutant very close to the target profile of TEM-1 β-lactamase (see Figure 5.6).

Lastly, we emphasized on the critical role of dynamic allosteric interactions in

fine-tuning the function of the enzyme. We focused on the residues with medium flexibility (due to their less-invasive nature upon substitution (Modi and Ozkan, 2018; Butler *et al.*, 2015; Kumar *et al.*, 2015b)), which also exhibit high dynamic coupling with the catalytic sites (DARC spots) obtained through DCI. Of these, a subset was selected that show stronger coupling with other non-common hinge residues that are not substituted during evolution (set Z). This was the key step in our design strategy as it allowed us to distally modulate the flexibility of these hinge positions without directly mutating them. Adding substitutions in set Z with the previous substitutions (set XYZ), we managed to shift DFI profile of the engineered mutant to that of TEM-1 β-lactamase. Particularly bringing changes in certain hinge positions that were missing in the mutant with Set XY substitutions (see Figure 5.8). Experimental characterization of the mutant also corroborates with our prediction as we are able to introduce a 10000-fold disparity in the turnover rates for antibiotics by enhancing its turnover rate catalyzing BZ (by an order of 3, a much harder goal to achieve) and, at the same time suppressing its turnover rate for CTX (by an order of 10^4) (see Table 5.1). Thus, showing the success of the hinge-shift mechanism for engineering the desired functional activity.

Overall, we used an approach motivated by conformational dynamics to rationally designing a promiscuous enzyme, GNCA, toward an enzyme with better efficiency to a specific substrate similar to TEM-1 β-lactamase. First, our dynamics approach uses the hinge-shift mechanism, presented in our earlier studies (Modi and Ozkan, 2018; Modi *et al.*, 2018; Kim *et al.*, 2015; Campitelli *et al.*, 2018) highlighting how compensation of enhanced flexibility of rigid (hinge) sites with rigidification of flexible sites modulate the conformational dynamics toward the desired function. Second, our approach uses the protein flexibility profiles as an optimization criterion towards the desired function such that it checks whether the substitutions shift the flexibility pro-

file of the mutant from that of wild type to the flexibility profile of the enzyme with the desired function. The success of our dynamics-based design approach brings to light the importance of protein dynamics and allosteric interactions in engineering new dynamics for an enzyme which is a current Achilles' heel in enzyme design (Khersonsky *et al.*, 2006; Khersonsky and Tawfik, 2010; Tokuriki and Tawfik, 2009b; Bar-Even *et al.*, 2011; Khersonsky *et al.*, 2010). It also opens up novel computational design principles to enhance or fine-tune the activity of enzymes which is in stark difference from any other computational enzyme design that optimizes the interaction within a 3D structure near catalytic site (Romero and Arnold, 2009; Otten *et al.*, 2018).

Chapter 6

ION COMPETITION REDUCES THE FITNESS OF AN ANTIBIOTIC-RESISTANT RIBOSOME VARIANT

This chapter is adapted from "Moon, E. C., Modi, T., Dong-yeon, D. L., Gracia-Ojalvo, J., Ozkan, S. B & Süel, G. M. A magnesium tug-of-war impedes bacterial antibiotic resistance. Under writing process."

In previous chapters, we have explored protein evolution through the lens of allosteric regulations. Allostery, generally, is considered to be a phenomenon at the level of a protein. However, proteins do not function in isolation. They are part of a more elaborate machinery at a cellular level. Therefore, a question which begs to be answered is—do allosteric interactions in a protein also impact the function/activity of other molecules in the cell and if so how? Here, we address this question by studying the dynamics of ribosome molecule in an E. coli.

Ribosomes require a large amount of magnesium ions for the stability and functionality to neutralize the electrostatic repulsion and to spatially coordinate rRNA functional groups. In this study, we focus on the L22* variant of the ribosome which is responsible for providing antibiotic resistance to the bacteria. However, despite this advantage, the antibiotic-resistant ribosome variant has not established itself as the "wild type" form. We explore this by modeling the interactions within a wild type ribosome and its variant (L22*) using elastic network models. The dynamics of the two ribosomes through the Dynamic Flexibility Index analysis presents a stark difference in their flexibility profiles, particularly suggesting that L22* enhances the association of magnesium ions within the ribosome. This in turn disrupts the delicate

balance of binding between magnesium ions with ribosomes and with ATP molecules (as active ATP molecules also need to be complex with magnesium ions). It is further experimentally verified by our collaborators at Süel Lab that increased magnesium association with the L22* ribosome reduce the concentration of free magnesium ions in the cell, which may in turn constitute a physiological cost for the antibiotic-resistant ribosome variant. We thus introduce here the concept of Ionic Allostery, where ions transmit the effect of ribosome association to distant ATP molecules, allowing for long-range regulation of activity. This is also described with the help of a toy model of kinetics of the competition between ribosome and ATP molecules in a cell for magnesium ions by our collaborator Jordi Garcia-Ojalvo.

6.1 Introduction

Ribosomes are essential components of all living cells and are required for protein synthesis, cellular growth, and replication. These are also a major target for antibiotic drugs to control bacterial growth. Interestingly, studies have identified the spontaneous emergence of natural ribosome variants in bacteria that confer resistance to antibiotics, such as aminoglycosides (Sharrock *et al.*, 1981; Thorbjarnardóttir *et al.*, 1978; Buckel *et al.*, 1977; Nessar *et al.*, 2011; Criswell *et al.*, 2006). For example, in *Bacillus subtilis*, a spontaneously arising ribosome variant (L22*) has a sequence extension of the short loop comprising its L22 subunit. This L22* variant is associated with conferring resistance to erythromycin and aminoglycoside antibiotics (Criswell *et al.*, 2006). Given that such antibiotics are natural products, and antibiotic resistance confers a well-documented fitness advantage (Davies, 2010), it remains unclear why such ribosome variants are not the dominant form in nature, the so-called "wild type" (WT). One possibility is that the spontaneously arising ribosome variants carry some fitness cost that counters the benefit of antibiotic resistance (Andersson and

Hughes, 2010; Andersson, 2003). Therefore, understanding the balance between the cost and benefit of a given ribosome variant may reveal fundamental insights not only into ribosomes but also bacterial physiology.

However, the study of ribosomes is complex due to their necessary association with cations. These cations, such as magnesium ions are important for the stability of ribosomes and they also play a critical role in their function. Ribosomes require cations such as magnesium ions to neutralize the electrostatic repulsion and to spatially coordinate the interactions among rRNA (Ribosomal Ribonucleic acid) and other proteins. High-resolution crystal structures of bacterial ribosomes have revealed that each complex contains more than 170 structural magnesium ions (Schuwirth *et al.*, 2005), which are an integral part of the ribosome complex (Knight *et al.*, 2013). Therefore, a deep and comprehensive understanding of ribosomes can only emerge if we take into account the dynamic of its interactions with magnesium ions. However, these interactions have proven very difficult to measure, and thus, the interplay between magnesium (Mg) ions and ribosomes has remained elusive. To investigate and compare the functional dynamics of the WT and L22* ribosomes, we utilized a coarse-grained Elastic Network Model (ENM). ENM has been utilized in the past to successfully study the dynamics of inter-domain motions in ribosome (Chang *et al.*, 2015; Wang *et al.*, 2004; Kurkcuoglu *et al.*, 2016; Zimmerman *et al.*, 2016; Tama *et al.*, 2003). This approach is based on a simplified mechanical model of particles connected by springs (Ikeguchi *et al.*, 2005; Tirion, 1996; Bahar *et al.*, 2010). In the ENM, the dynamics of any molecular complex is dictated by a network of intra- and inter-molecular interactions among domains and subunits, and the intrinsic thermal fluctuations of atomic positions. ENM has proven effective in elucidating and predicting the biologically relevant motions of intra- and inter-molecular interactions that describe the structure-encoded dynamics of bio-molecules. By applying this modeling

approach to the ribosome complex that includes structural magnesium ions, we were able to study the dynamics between the ribosome and its Mg ions.

6.2 Extracting the Dynamic Flexibility Index (DFI) Profile of Ribosome Using ENM

DFI (Nevin Gerek *et al.*, 2013; Gerek *et al.*, 2009; Campitelli *et al.*, 2020; Modi *et al.*, 2018; Modi and Ozkan, 2018; Zou *et al.*, 2015) is a novel metric which quantifies the relative vibrational entropy of residues by calculating the resilience each residue experiences to perturbations in the molecule. Computation of DFI utilizes Perturbation Response Scanning (PRS) (Atilgan *et al.*, 2010) technique using random perturbative forces as probes to sample the local vibrational ensemble of each residue. These perturbative forces emulate the effect of stochastic nature of forces which exists in proteins due to its interaction with the thermal bath, solvent molecules and other small molecules in a cell. In order to investigate the association of the ribosome and its structural Mg ions, we utilized the coarse-grained ENM, which takes into account the interactions between different components of ribosome by incorporating all the C_α atoms from the protein chains, phosphorus atoms from the ribosomal nucleotides, and also the structural Mg ions. Specifically, atoms are modeled as nodes and interactions as springs connecting them. The interactions represented by harmonic springs have force constants which inversely scales with the 6^{th} power of the pairwise distance between them (i.e., $k \propto 1/r^{-6}$) (Nevin Gerek *et al.*, 2013; Gerek *et al.*, 2009). Therefore, the positions that are very far away from each other have a very small interaction energy between them and vice versa. Therefore, using this model, we calculate the Hessian matrix. This is then used to calculate the responses of the network of nodes to unit perturbations using the Linear response theory as described in section 2.2.2. However, this poses a problem. The dimensions of the Hessian matrix for a ribosome

will be $3N \times 3N$, where N are the total number of nodes in the ENM. For the E. coli ribosome we used (PDB id: 4v56 (Borovinskaya *et al.*, 2007)), N=10380. This makes the Hessian matrix very large (31140 x 31140) and expensive to store in the memory (nearly 40 GB if using float32). In addition, performing eigenvalue decomposition for inversion of such large matrices is not possible given the architecture of most high memory machines. Therefore, we used ARPACK (Lehoucq *et al.*, 1998) which is designed to compute a set of eigenvalues and corresponding eigenvectors numerically for a sparse matrix, which a Hessian is (see Appendix C.1 for the sub-routine used). It uses a method called Implicitly Restarted Lanczos Method (Calvetti *et al.*, 1994) to recursively estimate the largest 20 eigenvalues and corresponding eigenvectors within a lowest possible tolerance (based on machine precision). These are then used to estimate the inverse of the Hessian matrix. Following this, the protocol described in section 2.2.2 can be used to obtain the DFI score for each node in the ribosome (Figure 6.1).

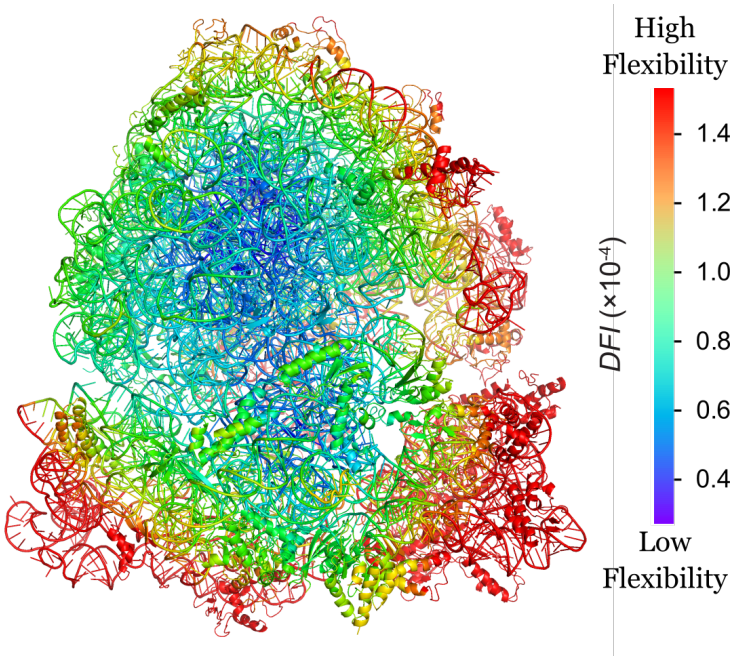

Figure 6.1: The cartoon representation of wild type ribosome color coded with Dynamic Flexibility Index score.

6.2.1 The Coarse-Grained Computational ENM Model Captures the Essence of All-Atom Molecular Dynamics (MD) Simulation.

In order to test the validity of our method, we compared the percentile scores of the root mean square fluctuations (RMSF) in the C_α and phosphorus atoms of the ribosome predicted by our coarse-grained ENM model with that observed in an all-atom MD simulation. For this, we used the data from MD trajectories calculated in a study by Warias *et al.* (2020). In the study, the authors have used the cryo-EM structure of an E. coli ribosome in complex with tRNAs, EF-Tu, and GTP as the starting structure for the MD simulation (PDB id: 5uym (Loveland *et al.*, 2017)). The simulations borrowed the force field parameters from amber99sb force field (Hornak *et al.*, 2006) and the SPC/E water model (Poole *et al.*, 1992). Two $2\mu s$ long independent simulations were ran initialized from the same initial structure (see (Warias *et al.*, 2020) for further details regarding the MD simulation).

Using the last $1\mu s$ of the two trajectories, we calculated the RMSF of each C_α and phosphorus atoms in the ribosome. These are then compared with the RMSF predicted by modelling the interactions in the structure from the same ribosome using the coarse-grained ENM network as described above (RMSF from MD and ENM are calculated following the method described in section 2.2.1). Through the coarse-grained ENM model, we are able to capture the salient features of the RMSF profile from MD. Particularly, we focused our attention on the following regions in the ribosomal assembly: EF-Tu Domain D3, and proteins L22 and L4. These were selected as the study (Warias *et al.*, 2020) provided the RMSF profile of domain D3, therefore, it can be used as a test of our approach. In addition, the proteins L22 and L4 constitute a part of the exit tunnel, hence, are critical for ribosomal function. Comparing the RMSF profile obtained from all-atom MD and our course-grained

ENM approach showed that we were able to accurately predict the rigid regions in these proteins (i.e., with the lowest RMSF). Out of the two independent MD trajectories, our approach was able to accurately predict the rigid regions observed in at least one of the them (see Figure 6.2). It should be noted that, despite these similarities, the RMSF profile obtained from ENM cannot capture all the features in RMSF obtained using the MD trajectories (highlighted in Figure 6.2). This indicates that ENM misses some rigid features possibly emanating from motions with longer time scales or bio-chemical specific interactions only sampled in MD.

6.2.2 Modelling the Dynamics of L22 Mutant Ribosome.*

The mutant L22* contains an insertion of 7 residues in the L22 protein. L22 protein is a globular protein which is embedded on the surface of the large ribosomal subunit (see Figure 6.3A). The L22 protein comprises of a loop which penetrates a cavity in the core of the ribosome. The insertion doubles the number of residues in this loop. Prior to insertion, the loop lies in a cavity inside the ribosome such that the radial distribution of the number of contacts (i.e., the number of contacts made within a cutoff distance) of this loop in the WT ribosome is significantly lower as compared to any other proteins in the core of the ribosome. After insertion in this extended loop, the radial distribution of the number of its interactions is hypothesized to increase as doubling the loop size should enhance its interactions. Therefore, in order to emulate the impact of this insertion that reflects the direct interaction of the loop with neighboring positions in the 3D structure, we scale up the interactions of the residues in L22 protein with insertion (between 80-100 residues), such that the new radial distribution of contacts follow the trend shown by any other protein in the core of the ribosome. Here, we used the protein L34 as a reference for these calculations (see Figure 6.3B). It should be noted that, here we have used this crude

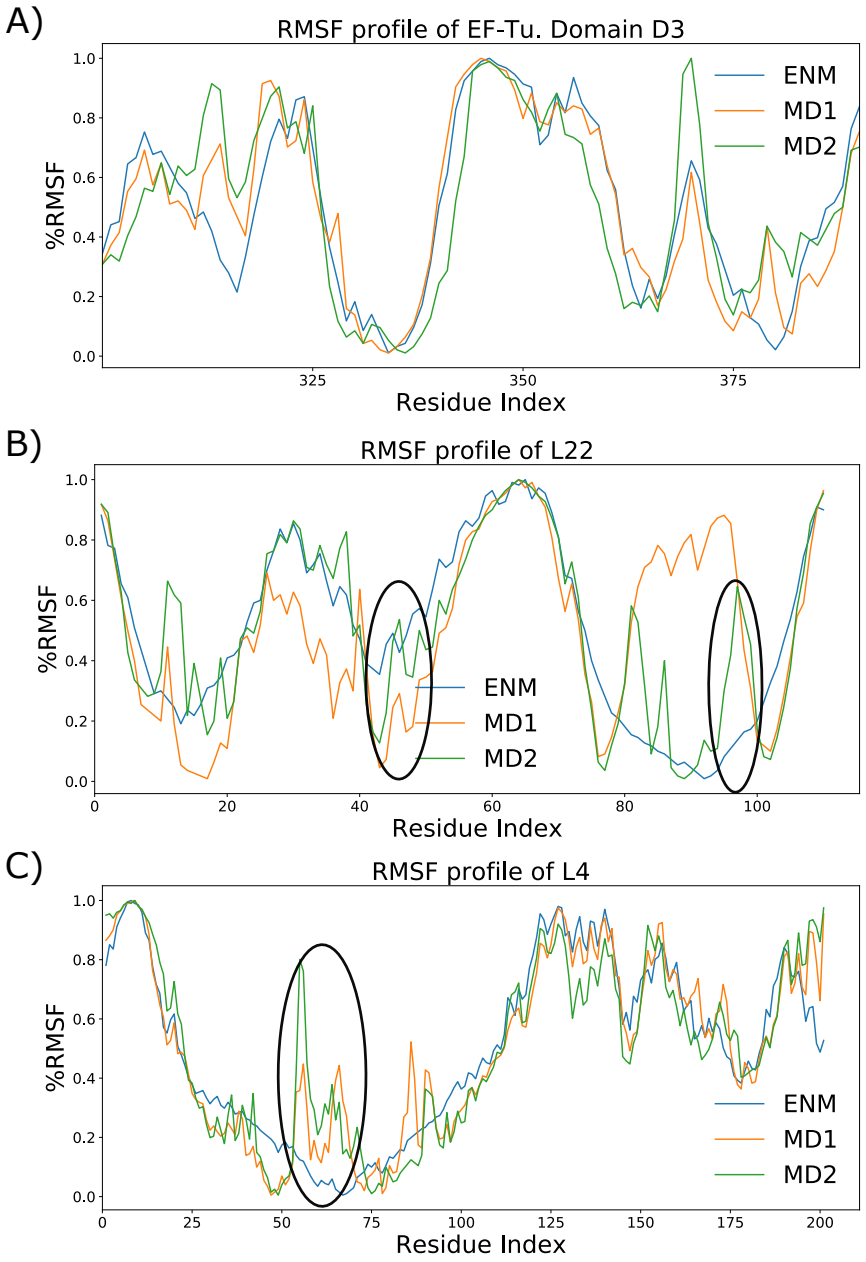

Figure 6.2: Comparing the percentile ranking of root mean square fluctuations (RMSF) obtained from two independent all-atom MD simulations of a ribosome (orange, MD1; green, MD2) with that predicted by a coarse-grained ENM model (blue). Here, as an example, we focus our comparison on three different regions of the ribosome, namely—(A) EF-Tu Domain D3, (B) protein L22 and (C) the protein L4. We observe that through ENM, we are able to accurately predict most of the rigid regions observed in at least one of the MD simulation. We also observe (highlighted) some disparities between RMSF from MD simulations and ENM.

approximation to emulate the effect of insertion at L22 protein as the structure of the mutant is not available and modelling the mutant by incorporating insertions in the sequence is computationally expensive and models typically used for this purpose (e.g. (Fiser *et al.*, 2000)) are not optimized for ribosomal proteins.

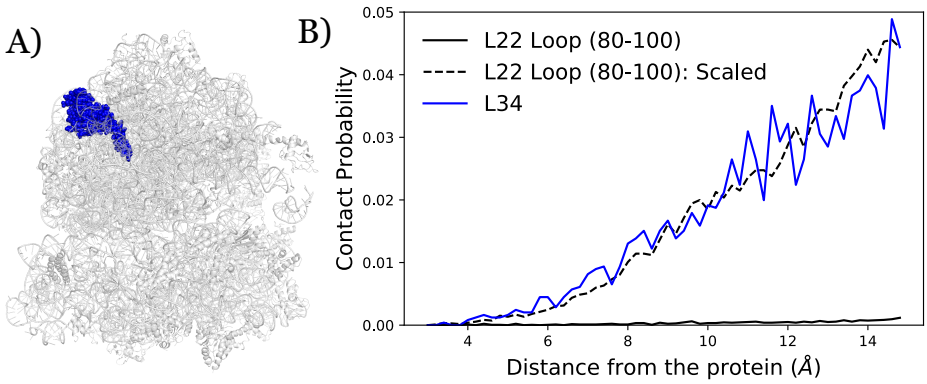

Figure 6.3: Modeling of radial distribution of contacts to emulate amino acid insertion in L22* variant. (A) L22 protein (shown in blue sphere) is embedded on the surface of ribosome (shown as cartoon representation in grey). (B) The radial distribution of contacts from the loop residues in L22 with insertion (80-100) in the ribosome (black solid line) and that of another protein buried in the core of ribosome, L34 (blue, solid). The interactions for these residues in L22 are scaled up such that the radial distribution emulates that of the proteins in bulk (black, broken). With this scaling factor, the extended loop in L22 with insertion now emulates the interaction energy of any other protein buried in the core of the ribosome.

6.3 Results and Discussion

6.3.1 L22 Mutations Changes the Association of Mg Ions With the Ribosome*

DFI as described earlier, is the net displacement of a given position relative to the net response of the whole complex when every node is individually perturbed. Therefore, the DFI score of a node can be interpreted as relative mobility upon physical perturbations (such as inter-molecular forces or stochastic thermal noise). Therefore, nodes with a lower DFI score have lower mobility, thereby suggesting a higher association with the network. Whereas, nodes with a higher DFI score would

be more susceptible to fluctuations, suggesting a lower association and higher mobility. Through these results, we observe that the Mg ions near the center of a ribosome have lower mobility, compared with those toward the edge (Figure 6.4). Consequently, DFI reveals the relative association strength (i.e., binding free energy) of Mg ions with the rest of the ribosome and makes it possible to map this data onto the ribosome's structure.

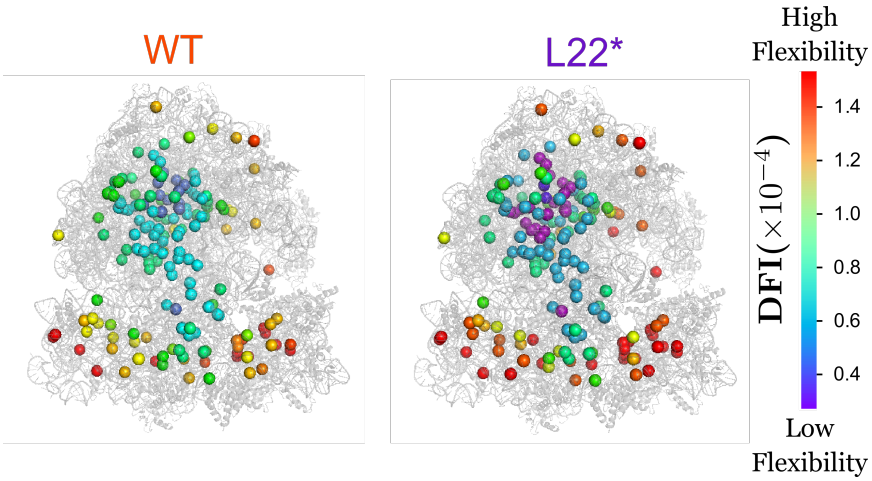

Figure 6.4: Cartoon representation of ribosome color coded with the DFI scores of magnesium ions in wild type (left) and L22* variant (right). We observe that the mutation causes a change in the mobility of magnesium ions.

Utilizing the DFI maps, we examined whether the ribosomal association with Mg ions differs between WT and the mutant L22* ribosomes. DFI analysis of the L22* ribosome reveals changes in the DFI values for all the structural Mg ions (Figure 6.4). To determine how the L22* ribosome variant affects its interaction with Mg ions, we subtracted the DFI of Mg ions in the WT ribosome from those in the L22* ribosome. We find a net loss in the mobility of the Mg ions in the mutant strain Figure 6.5A. As a control, when the same DFI analysis is done on carbon and phosphorus atoms, the change in the fraction of those nodes is much smaller in comparison (see Figure 6.5B). This demonstrates that the L22* mutation has a higher effect on the dynamics of Mg ions, compared to the rest of the ribosomal components. Specifically, rigid ions

become more rigid, while ions with medium mobility become more mobile upon L22* mutation Figure 6.5A. It is also notable that the increase in rigidity is larger than the gain in mobility. In other words, there is an overall increase in the association of Mg ions in the L22* ribosome. The increase in rigidity of Mg ions is particularly striking in the region surrounding the L22 protein. We find that 45.2% of Mg ions residing within a 50Åradius around the mutated L22 subunit have high rigidity, compared to only 32.1% of Mg ions that exhibit rigidity in the WT ribosome, Figure 6.5C. These results were consistent when we analyzed other ribosome structures Figure 6.6. Together, these data suggest that the L22* mutation increases the association of Mg ions with the ribosome. It should be noted that, this increased association of Mg ions is linked to the cascading effect of increased number of interactions due to the residues inserted in L22 protein.

Inspired by the results above, we hypothesized that the increased association of the Mg ions with the L22* ribosome would reduce the concentration of free Mg ions in the cell, which may in turn constitute a physiological cost (Figure 6.7A). To experimentally test this hypothesis, we set out to measure intra-cellular free Mg in the live bacterium. Specifically, we generated a genetic fluorescent reporter utilizing a previously characterized *B. subtilis* native M-box riboswitch that is sensitive to the free Mg levels in the cell (Figure 6.7B). The M-box riboswitch is part of the promoter region of the *mgtE* gene, which codes for a Mg importer, and its function is to reduce *mgtE* expression upon binding of free Mg ions. In this way, the M-box riboswitch enables cells to express the Mg transporter *MgtE* only when needed. By fusing the *mgtE* promoter region with a *yfp* gene, we created a reporter that indicates a shortage of free Mg ions in cells. As a control, we used the M-box with the previously characterized M3 mutation (Dann *et al.*, 2007) in the aptamer domain, which is known to make the riboswitch independent of the Mg levels by disabling its

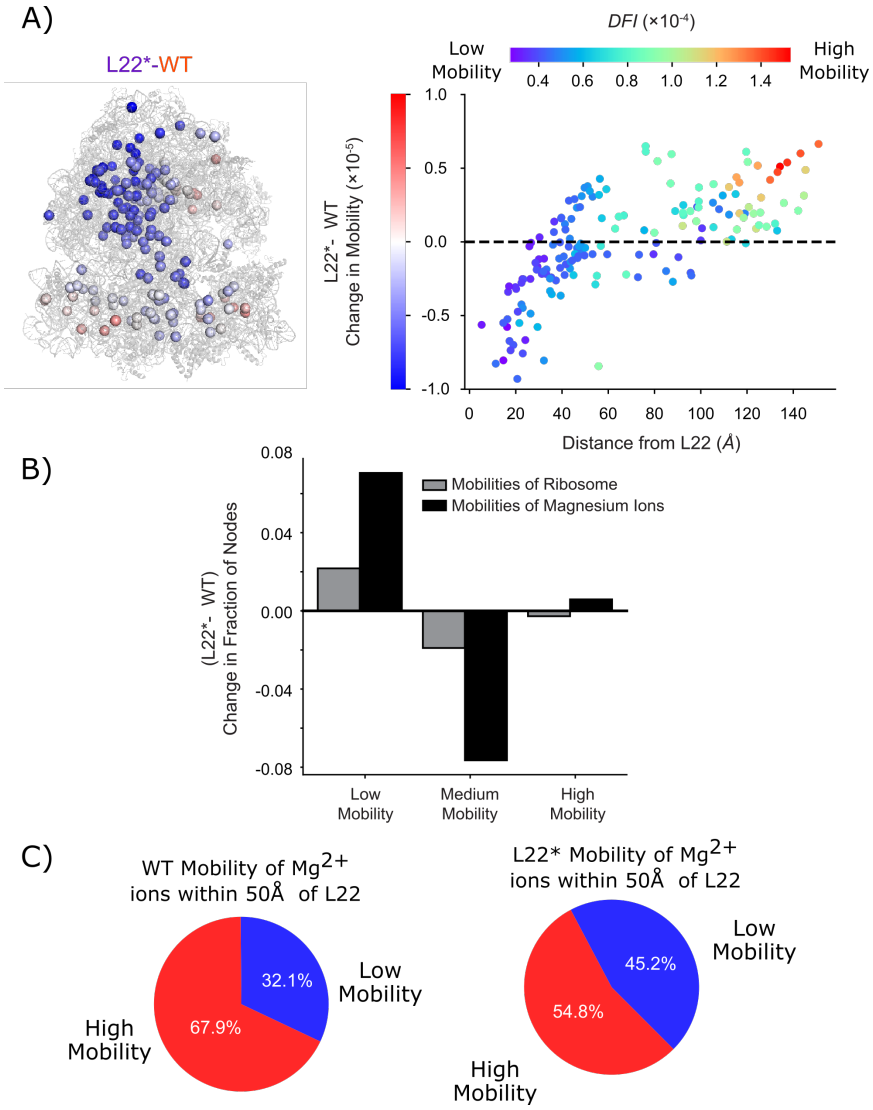

Figure 6.5: Change in the mobility of magnesium ions in E. coli ribosome upon L22* mutation. (A) Cartoon representation of ribosome where each magnesium ion is color coded with the differences of DFI upon L22* mutation. A color bar shared with panel on the right where the change in the mobility of magnesium ions is plotted by the distance from L22 ribosomal protein. Each data point is colored based on the WT DFI value. (B) Fractional change in the number of nodes from the low mobility, medium mobility, and high mobility groups upon L22* mutation in whole ribosome (gray) and only magnesium ions (black). Higher fraction of magnesium ions have undergone either an increase or decrease in the mobility compared to the rest of the ribosome containing the protein and rRNA components. (C) Pie chart illustration of the percentage of low and high mobility magnesium ions within 50Åof L22 ribosomal subunit in WT ribosome (left) and L22* variant (right). A threshold value for low and high mobility is 4×10^{-5}. The insertion at L22 protein has reduced the mobility of magnesium ions in the L22* variant ribosome.

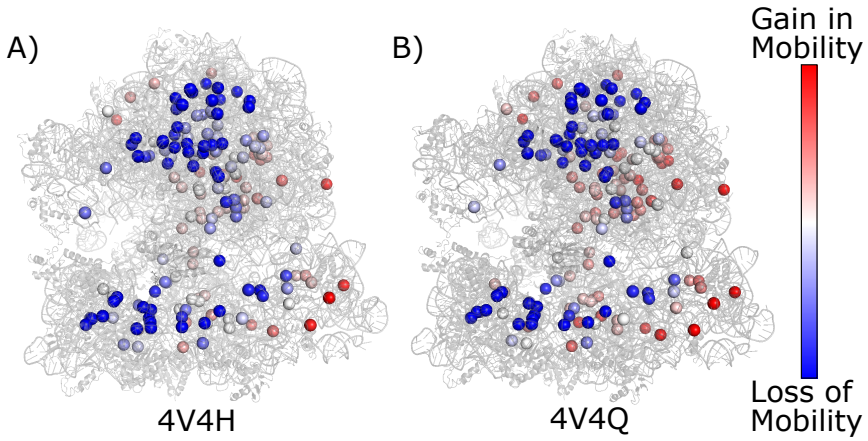

Figure 6.6: Change in the mobility of magnesium ions calculated from other bacterial ribosome structures mapped onto cartoon representations (A) PDB id: 4V4H and (B) PDB id: 4V4Q from E. coli.

binding to the riboswitch. As expected, YFP expression in the M3 control remained high, regardless of extracellular Mg concentrations. Consequently, the riboswitch-based reporter provides a means to measure relative changes in intra-cellular free Mg concentrations.

Using this riboswitch-based reporter described above, we then tested our hypothesis that free intra-cellular Mg levels would be lower in the L22* strain, relative to WT. In particular, we measured the activity of the reporter in the WT, L22*, and M3 strains at different extracellular Mg concentrations. Bacteria were grown in minimal defined media (MSgg) and then imaged on MSgg-agar pads with single-cell resolution (Figure 6.7C). As expected, we found that the L22* strain is deficient in maintaining free intra-cellular Mg at low extracellular Mg levels, when compared to WT. Specifically, when we reduced extracellular Mg concentrations from 0.2 to 0.02 mM, we observed a clear increase in the YFP signal from L22* cells (Figure 6.7C), which indicates a reduction in intra-cellular free Mg concentrations (Figure 6.7D). No such obvious trend was observed in WT cells, which displayed a similar YFP signal with increasing extracellular Mg concentrations. Furthermore, as expected, no clear

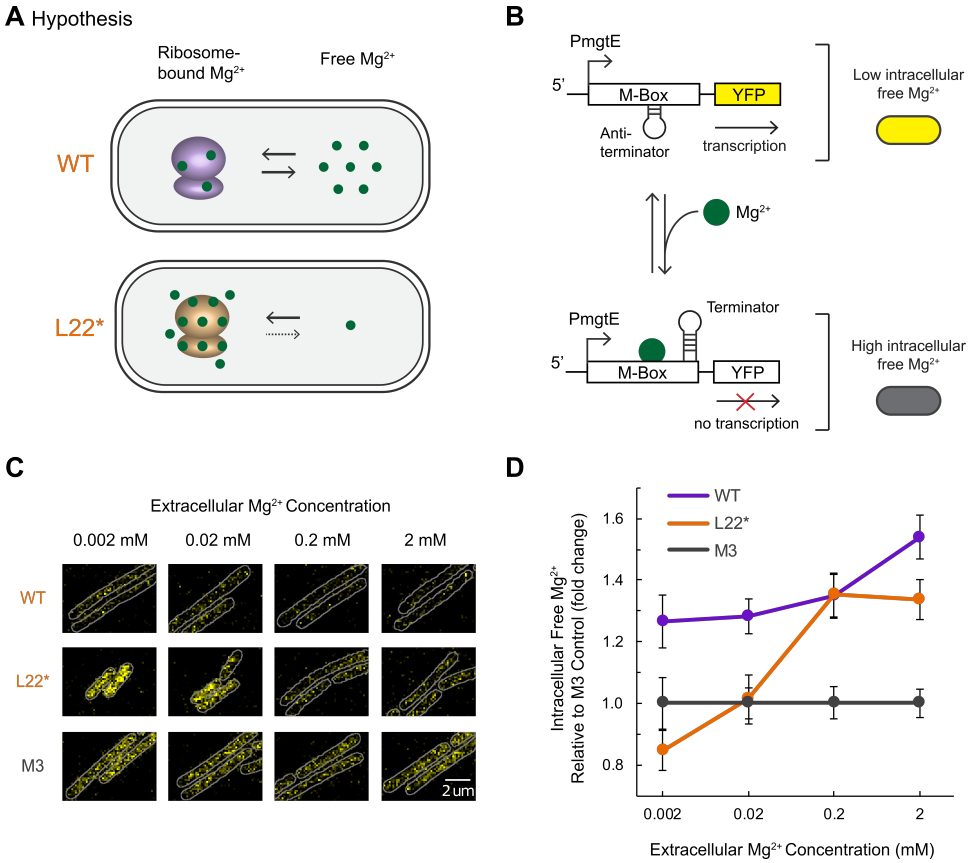

Figure 6.7: The L22* ribosome reduces free magnesium pool in a cell. (A) Cartoon representation of the hypothesis that L22* strain has less free magnesium ions compared to WT strain due to stronger magnesium ion association in L22* ribosome. (B) Schematic of M-box riboswitch-based reporter used in this study. A cell will turn on *yfp* gene when it experiences low intracellular free magnesium concentration. (C) Representative snapshots of YFP expression at different extracellular magnesium concentrations. Each cell is outlined in gray based on the corresponding phase image. WT means PmgtE-yfp (intact M-box riboswitch) expressed in WT *Bacillus subtilis*, L22* means the same promoter-reporter in L22* *B. subtilis* strain, and M3 means *PmgtEM3-yfp* (M-box mutant expresses YFP independent of the intracellular magnesium level) expressed in WT *B. subtilis*. Scale bar, 2 μm. (D) Intracellular free magnesium concentrations of each strain are plotted for different extracellular magnesium concentrations (mean $\pm$ 95% confidence interval; n $\geq$ 4 images from 3 experiments, $\geq$ 10 cells analyzed from each image). The values are inversed and normalized by mean M3 YFP signals.

trends in the YFP signal were observed in the M3 control. We quantified the fluorescence signal and normalized the values according to the M3 control strain to account for potential experimental variability (Figure B.1). In other words, while the WT

strain is robust and able to maintain free Mg levels despite changes in extracellular Mg concentrations, the L22* strain exhibits quantifiable sensitivity to low extracellular Mg levels. These data imply that the increased association of Mg ions with L22* ribosomes comes at the expense of the ability to maintain free intra-cellular Mg concentrations under diminishing extracellular Mg levels. This leads us to the question, what consequences would arise if cells could not maintain a sufficient free Mg concentrations in the cell.

6.3.2 Ribosome Competes With ATP to Bind With Intra-Cellular Mg Ions.

Mg ions serve as counter-ions for ATP, the energy currency of cells. In fact, Mg-bound ATP is the only biologically active form of ATP. Together with ribosomes, ATP constitutes the largest intra-cellular store of chelated Mg ions (Pontes *et al.*, 2015). We hypothesized that reduced free Mg levels would reduce the availability of Mg-ATP, the biologically active form of ATP (Figure 6.8A). To formulate our hypothesis more precisely and generate concrete testable predictions, we constructed a simple mathematical steady-state model described below.

We consider three potential states in which Mg ions can reside within *B. subtilis*: free, bound to ribosomes, and bound to ATPs. The corresponding chemical reactions are:

$$\mathrm{Mg} + \mathrm{R_i} \quad \underset{k_{rui}}{\overset{k_{rbi}}{\rightleftharpoons}} \quad \mathrm{R_{i+1}} \tag{6.1}$$

$$\mathrm{Mg} + \mathrm{ATP} \quad \underset{k_{au}}{\overset{k_{ab}}{\rightleftharpoons}} \quad \mathrm{Mg \cdot ATP \cdot} \tag{6.2}$$

Here reaction 6.1 runs for $i = 0, 1, \ldots, N-1$, with R_i representing the ribosome bound to i Mg ions. Given the indistinguishable nature of the Mg binding sites, the combinatorics of this reaction dictates that

$$k_{rbi} = (N-i)k_{rb}, \qquad k_{rui} = (i+1)k_{ru}\,, \tag{6.3}$$

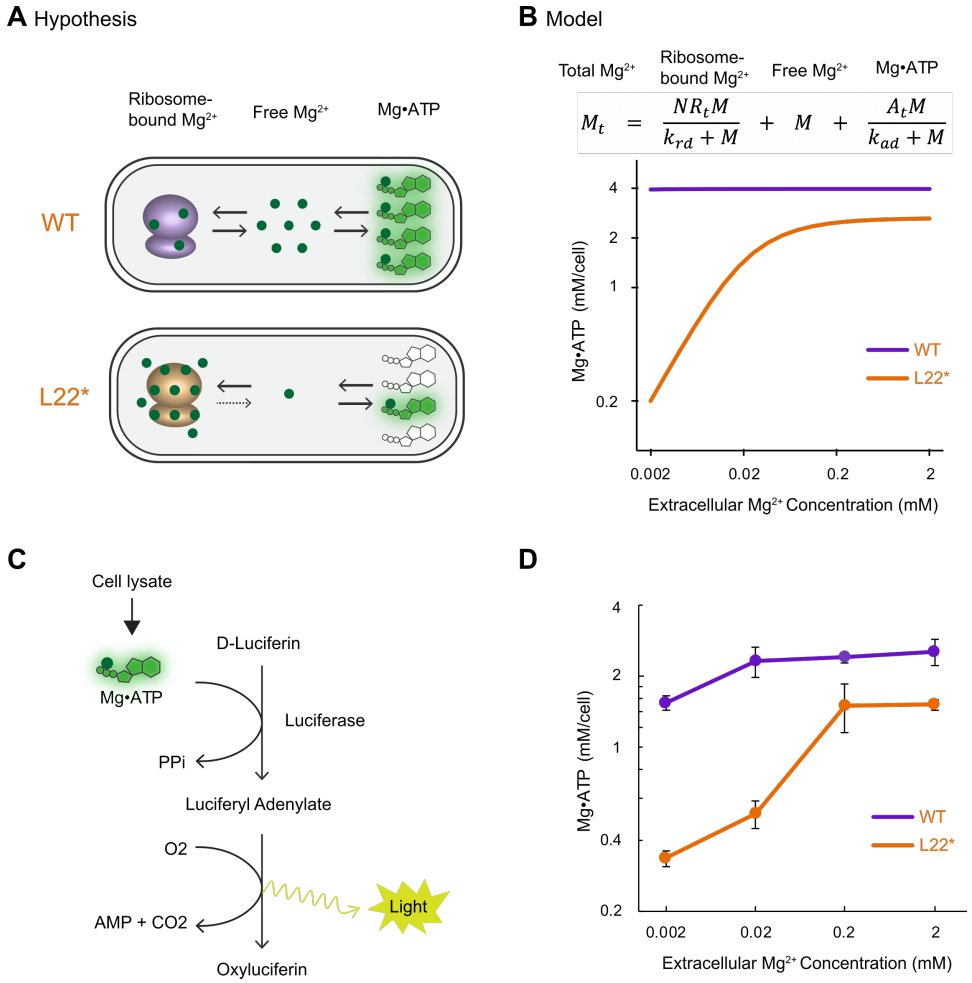

Figure 6.8: The L22* ribosome reduces Mg-ATP availability. (A) Cartoon description of the hypothesis that L22* ribosome reduces the availability of Mg-ATP due to reduced availability of free magnesium ions. (B) Simple mathematical model equation and prediction results based on the free magnesium level in WT and L22* strain. (C) Schematic of bioluminescence-based assay used in this study to determine Mg-ATP level. (D) Intracellular Mg-ATP concentrations of each strain at different extracellular magnesium concentrations (mean ± 95% confidence interval; n = 5 experiments, each experiment consisting of at least 2 samples). The values are calculated using the standard curve shown in Figure B.2

where N is the maximum number of Mg ions that the ribosome can hold, and we have defined k_{rb} and k_{ru} as basal binding and unbinding rates between ribosomes and Mg ions.

In what follows, we will use M to represent the concentration of free Mg inside the cell, with a denoting the concentration of free ATP, and A the concentration of

ATP bound to Mg. The following conservation laws must hold:

$$a + A = A_t \tag{6.4}$$

$$\sum_{i=0}^{N} R_i = R_t \tag{6.5}$$

$$M + A + \sum_{i=1}^{N} iR_i = M_t \tag{6.6}$$

Our goal is to determine how the total Mg in the cell is distributed among the three different pools listed above, which correspond to the three terms in the left-hand side of Eq. 6.6. In particular, we aim to determine how these three terms depend on the total Mg concentration in the cell, and eventually on the extracellular Mg concentration (which we can control experimentally).

We first note that the dynamics of the system can be described by means of the following set of coupled ordinary differential equations:

$$\frac{dA}{dt} = k_{ab}M(A_t - A) - k_{au}A \tag{6.7}$$

$$\frac{dR_0}{dt} = -k_{rb0}MR_0 + k_{ru0}R_1 \tag{6.8}$$

$$\frac{dR_i}{dt} = -k_{rbi}MR_i + k_{rui}R_{i+1} + k_{rb(i-1)}MR_{(i-1)} \tag{6.9}$$

$$-k_{ru(i-1)}R_i, \qquad i = 0, 1, \ldots, N-1 \tag{6.10}$$

$$\frac{dR_N}{dt} = k_{rb(N-1)}MR_{N-1} - k_{ru(N-1)}R_N \tag{6.11}$$

Solving these equations in the steady state leads to:

$$\bar{A} = \frac{A_t\bar{M}}{k_{ad} + \bar{M}} \tag{6.12}$$

$$\bar{R}_{i+1} = \frac{\bar{M}\bar{R}_i}{k_{rdi}} \tag{6.13}$$

where $k_{ad} \equiv k_{au}/k_{ab}$, while the expressions of k_{rbi} and k_{rui} given in Eq. 6.3 lead to $k_{rdi} \equiv k_{rd}(i+1)/(N-i)$ with $k_{rd} \equiv k_{ru}/k_{rb}$. With this, Eq. 6.13 transforms into:

$$R_{i+1} = \frac{M}{k_{rd}}R_i\frac{N-i}{i+1} \implies R_i = \left(\frac{M}{k_{rd}}\right)^i R_0\binom{N}{i} \tag{6.14}$$

where we have dropped the bars that denote steady state for simplicity. Introducing this expression for R_i into the conservation law Eq. 6.5 and using the binomial theorem leads to the following expression for the concentration of naked ribosomes as a function of the total ribosome and free Mg concentrations:

$$R_0 = \frac{R_t}{(1 + M/k_{rd})^N} \tag{6.15}$$

Eqs. 6.14 and 6.15 allow us to determine the concentration of Mg that is bound to ribosomes in the cell:

$$\sum_{i=1}^{N} iR_i = \frac{R_t}{(1 + M/k_{rd})^N} \sum_{i=1}^{N} i \left(\frac{M}{k_{rd}}\right)^i \binom{N}{i} \tag{6.16}$$

The sum in this expression can be computed through the moment-generating function of the binomial distribution, resulting in

$$\sum_{i=1}^{N} i \left(\frac{M}{k_{rd}}\right)^i \binom{N}{i} = N\frac{M}{k_{rd}} \left(1 + \frac{M}{k_{rd}}\right)^{N-1} \tag{6.17}$$

Introducing Eq. 6.17 into Eq. 6.16 leads to the final expression of the concentration of Mg bound to ribosomes:

$$\sum_{i=1}^{N} iR_i = \frac{NR_tM}{k_{rd} + M} \tag{6.18}$$

Using Eqs. 6.12 and 6.18, the Mg conservation law, Eq. 6.6 transforms into the final balance equation for the three Mg pools in the cell:

$$M_t = M + \frac{A_tM}{k_{ad} + M} + \frac{NR_tM}{k_{rd} + M}, \tag{6.19}$$

where, the first pool represents the free Mg ions in the cell, the second represents the concentration of Mg bound to ATP (Mg-ATP) and finally the third pool represents the ribosome bound Mg.

Our aim is to solve Eq. 6.19 to determine how Mg is distributed among these three different pools as a function of extracellular concentration. First we need to model

how the total intra-cellular Mg concentration M_t depends on the extracellular Mg concentration M_E. We assume a hyperbolic dependence between the two quantities due to the transport of extracellular Mg ions inside the cell, corresponding to a Michaelis-Menten-like kinetics that eventually saturates:

$$M_t = \frac{\alpha M_E}{K_m + M_E} \tag{6.20}$$

Using Eq. 6.20 we can solve numerically Eq. 6.19 to obtain the Mg-ATP as a function of extracellular Mg concentration, which we can then compare with the experimental observation, as shown in Figure 6.8B,D. Results are plotted for both the wild type and the L22* mutation. The effect of the mutation is to increase the affinity between Mg and ribosomes by two orders of magnitude, and the concentration of ribosomes 4-fold (see parameters in Table 6.1). It should be noted that the parameters are not fitted to the experiment, but are chosen to describe the relative behaviour between the concentration of Mg.ATP complex and extracellular Mg concentration. As long as k_{rd} for L22* variant is lower than that of WT, a lower concentration of Mg.ATP complex is observed.

According to this model, we predict a decrease in Mg-ATP upon reducing the extracellular Mg in the L22* strain, but not for WT (Figure 6.8B). Further on, we can also compute the energetic differences between the binding of Mg ions with ATP and ribosome at steady state conditions described in the model above. It can be expressed as:

$$\Delta\Delta G_{R_i-ATP} = -kT \ln \frac{k_{ad}}{k_{rdi}}, \tag{6.21}$$

and using Eq. 6.3:

$$\Delta\Delta G_{R_i-ATP} == -kT \ln \Big[\frac{k_{ad}}{k_{rd}} \frac{(N-i)}{(i+1)}\Big]. \tag{6.22}$$

Here, we have used the Arrhenius rate law to describe the relationship between the free energies and the respective rate constants. This can be calculated for the WT

Table 6.1: Parameters used for the magnesium ion pooling model describing the competition between the binding of free magnesium ions with ribosome and ATP.

Parameter	**Description**	**Value**
A_t	Total ATP concentration	4 mM
R_t	Total ribosome concentration	50 μM (WT), 200 μM (L22*)
k_{ad}	Mg-ATP Dissociation constant	0.1 mM
k_{rd}	Mg-Ribosome Dissociation constant	10 μM (WT), 0.1 μM (L22*)
N	Maximum number of Mg per Ribosome	200
α	Intra-cellular Mg concentration at saturation	42 mM
K_m	Michaelis-Menton constant of Mg import	0.8 μM

ribosome as well as the mutant L22* (Figure 6.9). Using the values for the parameters used earlier (see Table. 6.1), we observe that the binding of Mg ions is always more favorable to the mutant L22* as compared to the WT ribosome regardless of the number of Mg ions already bound to the ribosome. Moreover, we also observe that in the case of WT ribosome, there is a saturation limit of the number of Mg ions, which can be favourably bound to the ribosome. Beyond that, it is more favorable for those Mg ions to bind to ATP. However, this limit does not exist for the mutant L22*. It binds Mg ions more favorably than ATP even when the ribosome is already bound to a large number of Mg ions.

In other words, a free Mg ion that is introduced to the cell is more likely to associate with the L22* ribosome rather than an ATP molecule, even if the L22*

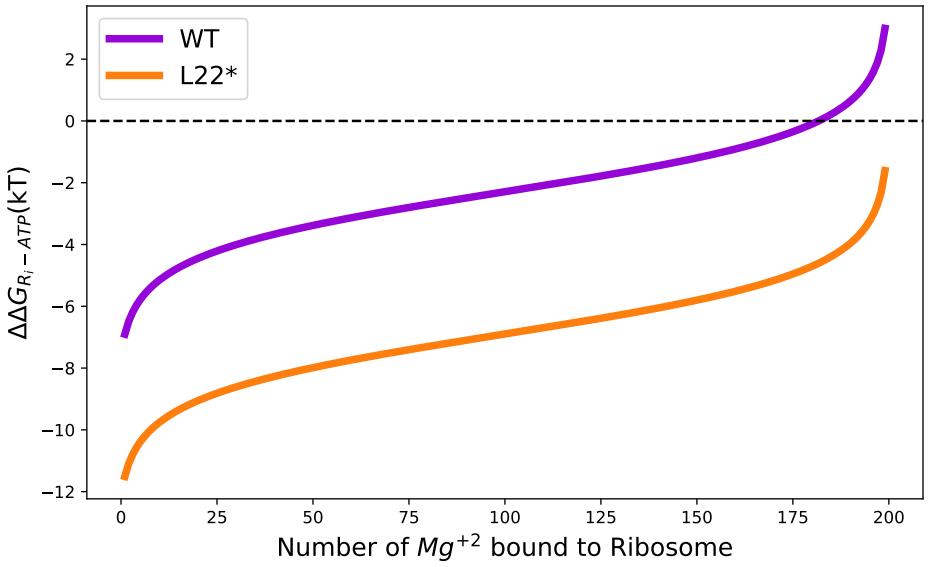

Figure 6.9: Variation of the free energy difference between the binding of magnesium ions with the WT ribosome and the ATP with the number of magnesium ions already bound to ribosome. (A) The WT ribosome (orange) and the L22* mutant (blue). We observe that in the case of the mutant, it is always more favorable to bind more magnesium ions to it as compared to ATP regardless of how many magnesium ions are already bound to it.

ribosome is fully occupied with Mg ions. This is not the case for WT ribosomes, which can only outcompete ATP molecules if less than 90% of their Mg sites are occupied (Figure 6.9). The model demonstrates that, at least for a simplified system considering three potential pools of Mg, changes in the affinity of ribosomes impact the levels of the other states of Mg in the cell, including Mg-ATP.

To experimentally test our modeling prediction that the L22* strain would have reduced Mg-ATP levels, we used a bioluminescence-based assay to quantify Mg-ATP in L22* and WT strains. The assay utilizes the Mg-ATP dependency of luciferase that catalyzes the oxidation of luciferin along with the emission of green light. Cell lysates containing Mg-ATP were exposed to the reaction, and the resulting light emission was quantitatively measured (Figure 6.8C and B.2). In agreement with our expectations, we find that the L22* strain has a lower Mg-ATP concentration when extracellular

Mg availability is low (Figure 6.8D). In particular, consistent with our model prediction, Mg-ATP levels drop significantly in the L22* strain when extracellular Mg concentrations are reduced from 0.2 mM to 0.02 mM. This increased sensitivity of ATP levels in the L22* strain to a reduction in extracellular Mg levels is consistent with reduced free levels of Mg ions (Figure 6.7D and 6.8D). In other words, we observe that both intra-cellular free Mg and Mg-ATP levels are low when cells with L22* ribosomes experience extracellular Mg limitation. From these multiple lines of evidence, we conclude that the higher association of the L22* ribosome variant with Mg ions constitutes a physiological cost to the L22* strain by increasing its dependence on extracellular Mg concentrations. Concurrently, our results also show that the WT ribosome allows cells to maintain a relatively stable level of both free Mg and active ATP across a broad range of extracellular Mg concentrations.

To more directly determine the physiological impact of the L22* ribosome variant, we investigated the growth of WT and L22* strains across a range of extracellular Mg concentrations. Cellular growth requires both ribosomes and Mg-ATP. This sets up the interesting question of whether the computationally predicted tighter association of the L22* ribosome with stabilizing Mg ions provides a growth advantage, or whether the ability of cells with WT ribosomes to maintain intra-cellular Mg-ATP levels would be more beneficial. To address this question, we tracked single cells over time and measured cell elongation to determine the growth rate (Figure B.3A). We find that both L22* and WT cells grow at similar rates with high extracellular Mg concentrations. However, for low extracellular Mg concentrations, the L22* cells exhibit reduced growth while the WT cells grow largely unaffected (Figure 6.10A). Notably, the concentration range of extracellular Mg over which the elongation rate of the L22* cells drops corresponds to the region where we also observe a drop in free Mg and Mg-ATP levels (compare Figure 6.8D and 6.10A). Therefore, when com-

pared to L22*, the growth of cells with WT ribosomes is more robust to changes in extracellular Mg concentrations. This result reveals a physiological cost for cells with L22* ribosomes in terms of their sensitivity to environmental conditions.

Given the environment-dependent growth cost of cells with the L22* ribosome, we next examined how the antibiotic resilience of WT and L22* strains compare over a range of extracellular Mg concentrations. We grew both strains on agar plates containing varying extracellular Mg concentrations and in the presence of commercially available antibiotic strips that contain a range of precisely defined concentrations of the aminoglycoside spectinomycin (Figure B.3B). The Maximum Inhibitory Concentration (MIC) was then determined by identifying the antibiotic concentration at which the growth of bacteria was cleared. We find that the specific value of MIC, and thus antibiotic tolerance, for both WT and L22* cells depends on extracellular Mg concentrations. More importantly, the MIC of cells with the L22* ribosome is never lower than WT for any tested Mg concentrations (Figure 6.10B). This data confirms the previously reported higher antibiotic tolerance of L22* ribosomes compared to WT ribosomes.

Quantifying the growth and antibiotic tolerance of cells with WT, or L22* ribosomes over the same range of extracellular Mg concentrations allows for a direct and comprehensive comparison. We generated a fitness phase diagram by plotting the growth robustness of each strain against its antibiotic tolerance (Figure 6.10C). The resulting plot visualizes the benefit provided to cells containing either the WT or L22* ribosomes. Cells with the L22* ribosome have a higher antibiotic tolerance, while cells with WT ribosomes are better at maintaining their growth rate. This results in each strain localizing in opposite corners of the fitness phase diagram. Consequently, antibiotic tolerance and robust growth appear to be competing benefits due to their shared dependence on Mg ions. This insight provides a possible answer to

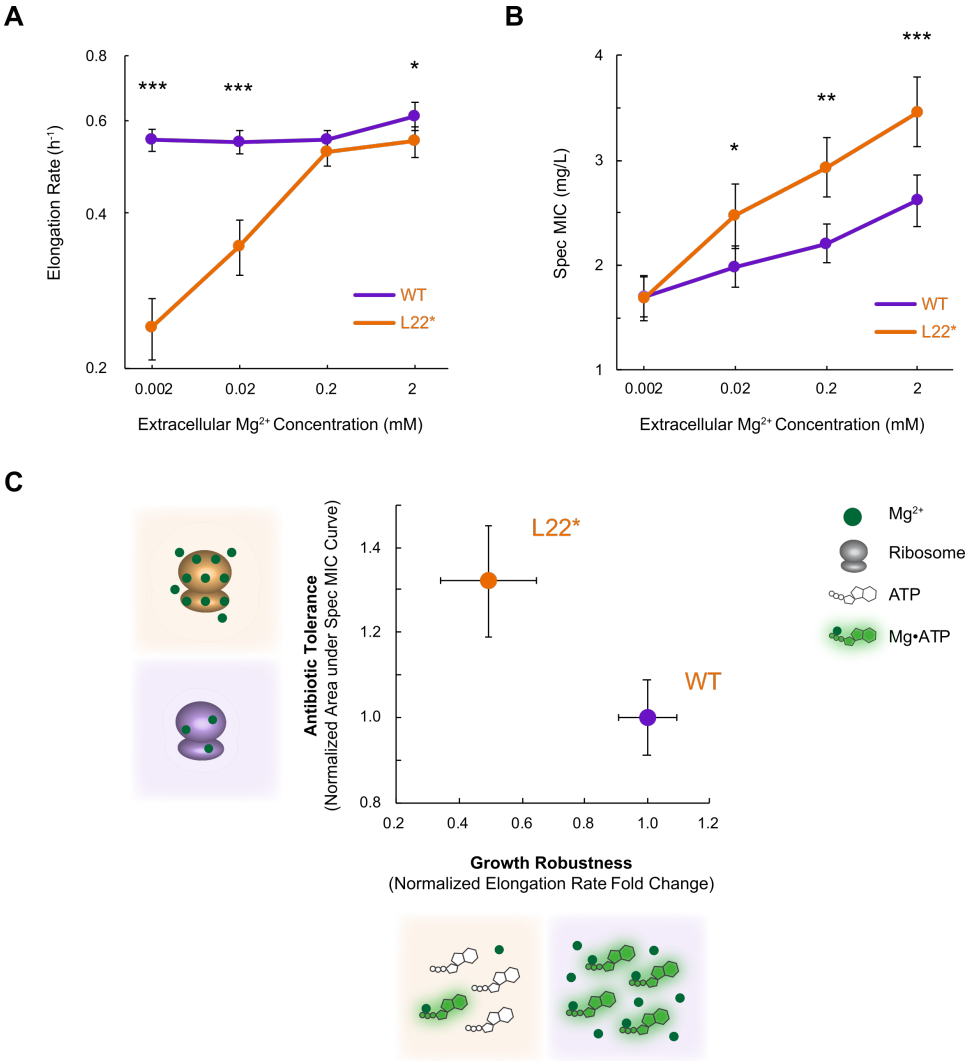

Figure 6.10: Environmental robustness and antibiotic tolerance reveal a cost-benefit metric for L22* strain. (A) Growth (elongation rate) of individual cells are monitored and quantified at various extracellular magnesium concentrations (mean ± 5% confidence interval; n ≥ 5 images from 3 experiments, at least 8 cells analyzed from each image). Asterisks indicate statistical significance (Student's t-test; *P < 0.05; **P < 0.01; ***P < 0.001). (B) Spec MIC test results of WT and L22* strains at various extracellular magnesium concentrations. A spectinomycin antibiotic strip is used for each plate (mean ± 5% confidence interval; n ≥ 8 plates from at least 2 experiments). Asterisks indicate statistical significance (Student's t-test; *P<0.05; **P<0.01; ***P<0.001). (C) A fitness phase diagram of the growth robustness against antibiotic tolerance (error bars, ± 95% confidence interval). Each metric is normalized by WT value. L22* strain (orange shading) has increased ribosome-magnesium ion association, likely attributing to higher antibiotic tolerance compared to WT strain (purple shading). L22* strain cannot maintain free and ATP-bound magnesium levels, likely attributing to deficiency in growth robustness.

the question of why L22* ribosomes have not been selected to become the dominant, or WT variant. Appendix B contains further details on the materials and methods used in the study.

6.4 Conclusion

Taken together, the results obtained in this study suggest that the tighter association of L22* ribosomes with Mg ions reduces the intra-cellular free Mg concentrations and thus Mg-ATP levels in cells. We have identified a physiological cost associated with the L22* ribosome variant. With this discovery, we can begin addressing the fundamental biological question as to why a spontaneously emerging natural ribosome variant that provides antibiotic resistance has not established itself as the "wild type". The relatively weak association of WT ribosomes with Mg ions allows better buffering of intra-cellular Mg levels, making cells more robust against changes in extracellular Mg concentrations. In this way, the benefit of antibiotic resistance provided by the L22* ribosome is pitted against the environmental robustness provided by the WT ribosome. Actual selection pressures that have shaped evolutionary history remain obscure, but our data suggest that bacteria such as *B. subtilis* may be more likely to face environmental changes in Mg levels, compared to naturally occurring antibiotics that target ribosomes. Therefore, the benefit of coping with variable Mg concentrations in the environment may outweigh the benefit of antibiotic resistance.

Ribosomes have been studied in great detail for many decades. However, studies have typically focused on the protein subunits and RNA components, and far less is known about the interactions with Mg ions. Furthermore, crystal structures of the ribosome are typically only available for wild type variants. Our coarse-grained modeling approach quantifies the association strengths for all structural Mg ions and

also allows us to model ribosome mutations and variants. For example, by modulating spring constants between nodes that represent residues and atoms, we can use Newton's laws to provide concrete testable predictions for how any given mutation can alter the biophysical properties of the ribosome. Additionally, our method is independent of the valency of ions, since we are not explicitly modeling electrochemical interactions. Consequently, our modeling approach circumvents the difficulty in capturing ion valency in simulations. The modeling framework presented here thus makes it possible to model and study interactions of structural ions within wild type and mutant ribosomes. This advance provides an important step forward in the quest to more broadly study and understand ionic interactions in cells, which have been referred to as "the dark matter of biology" (Ross, 2016).

It is well known that inorganic alkaline metal ions such as Mg are essential for life on our planet. However, the difficulty of measuring ions in cells has hindered our ability to understand the diverse roles these charged particles play in cell physiology. Here we integrated atomic-scale mathematical models with experimental measurements of ion concentrations and single-cell resolution physiological metrics to gain insight into the functional role of Mg ions. Our work indicates that Mg ions are not merely counter-ions for phosphate groups and charged amino acids but can also function as signaling molecules that directly regulate and coordinate between two of the most fundamental components of any living cell, namely ribosomes and ATP. There is an emerging understanding that ions can serve as signaling molecules, such as calcium ions acting as second messengers (Dominguez, 2004), and potassium ions enabling cell-to-cell signaling among bacteria in biofilms (Prindle *et al.*, 2015; Liu *et al.*, 2015). Here we show that Mg ions can relay intra-cellular information about the state of ribosomes. This points to the possibility that other ions that are also shared between different molecules and processes, could also function for coordinat-

ing signals. We therefore, propose here the concept of "Ionic Allostery", where ions transmit the effect of their association with one molecule (here, ribosome) to other distant molecules in the cell (here, ATP), thereby, allowing for a long-range regulation of activity. Our work is therefore likely to inspire future research to explore this concept in detail and the associated signaling roles for the many inorganic ions that are present in all cells.

Chapter 7

INFORMATION PROPAGATION IN TIME THROUGH ALLOSTERIC SIGNALING

This chapter is adapted from "Modi, T., Ozkan, S. B. & Pressé, S. Information propagation in time through allosteric signaling. Phys. Rev. Research 2, (2020)."

In the previous chapter, we observed the impact of allostery in ribosome, mediated throughout the cell via concentration of magnesium ions. Motivated by our findings, we now would like to explore the role of allosteric interactions in cell signalling pathways. For this purpose, we propose a toy model where the product of an enzyme activates another protein downstream in the signaling pathway. Using this model, we study how allosteric regulation alters the product formation rates and, thus, modulates the communication between the enzyme and the receiver protein via product molecules. We demonstrate that, the amount of information encoded downstream by an allosterically regulated protein depends on not only the kinetics of product generation and the binding kinetics of receiver, but also on the transition kinetics between allosterically regulated states of the sender.

7.1 Abstract

Naively, one expects the information communicated by an enzyme downstream within a signaling network, in which the enzyme is embedded, to grow monotonically with the enzyme's rate of product formation. However, here we observe that this does not necessarily hold true for allosterically regulated enzymes, often observed in signaling networks. In particular, we find that the mutual information between the catalytic

sites of an allosterically regulated enzyme and a receiver protein downstream in the signaling pathway depends on the transition kinetics between the different allosterically regulated states of the enzyme and their respective rates of product formation. Surprisingly, our work implies that allosteric down-regulation of an enzyme's rate of product formation may not only be used as a way to silence itself, as one would normally expect. Rather, down-regulation may also be used to increase the information communicated by this enzyme to a receiver protein downstream in a signaling pathway.

7.2 Introduction

Despite the fact that only a small fraction of a cell is composed of proteins (e.g., proteins constitute 17% of E. Coli (Milo, 2013)), proteins not only mediate the key processes in cells, but also give rise to spatio-temporal signaling to control a cell's response to its local environment underlying all critical decision-making (Wodak *et al.*, 2019a) such as a cell's development and metabolism (Link *et al.*, 2014), motility (Tafoya and Bustamante, 2018), immunity and cell-death (apoptosis) (Herr, 2018) and many more.

These spatio-temporally coordinated events are often achieved by proteins exhibiting allostery– a phenomenon by which the binding of a molecule at one site of a protein changes the binding affinity or catalytic activity at another distant site. Several models of allostery have previously been explored to study how allosteric interactions within one enzyme (i.e., proteins which act as catalysts) modulate its activity (Monod *et al.*, 1965; Cooper and Dryden, 1984; Koshland *et al.*, 1966). However, allostery goes beyond just remote modulation. It also provides the "circuit components" from which nature builds up complex signaling networks in enzymes. Here we employ an information theoretic framework to quantify how the impact of

allosteric interactions in an enzyme propagates downstream through protein signaling pathways in order to modulate cellular response.

Information theory has already been proven useful in studying various biological phenomena such as modeling protein interaction networks (Lenaerts *et al.*, 2008; LeVine and Weinstein, 2014; Voliotis *et al.*, 2014), evolution (Adami, 2012; de Vladar and Barton, 2011; Weiss *et al.*, 2000) and single molecule experiments (Turkcan and Masson, 2013; Tavakoli *et al.*, 2017). Moreover, it has also been instrumental in studying the effect of allostery on the transfer of information between the allosteric regulator and the catalytic site of allosteric proteins, particularly enzymes (Komorowski and Tawfik, 2019; Marzen *et al.*, 2013). However, these studies primarily explored the thermodynamics of substrate binding through allostery, focusing solely on the concentration of products generated. On the other hand, the activity of the receiver protein varies in time according to the availability and on/off binding of the product produced by the enzyme. Therefore, information conveyed is encoded not only in the total concentration of the enzymatic product generated by the "sender enzyme", but also in its time varying activity. Here, we apply information theory to quantify the role of allostery in a sender enzyme while it communicates with a receiver protein (a protein downstream in the signaling pathway which binds to the product molecules generated by the sender enzyme).

In order to apply the information theory, we first consider a model of an enzyme which allows for homotropic allosteric interactions between the two binding sites. This enzyme, labelled as 'sender enzyme', generates product molecules which bind with another receiver protein downstream in the signaling pathway. Afterwards, we construct simple master equations for this model which captures the essence of allostery in the arrival times of products at the receiver protein. Next, this model is simplified to a two state model which is studied in a discrete time domain with the

help of Hidden Markov Models (HMM) (Rabiner and Juang, 1986; Seymore *et al.*, 1999). Further on, we compute the probability of observing product arrival events accounting for all possible latent allosteric states ("hidden trajectories") of the sender enzyme. Using the probability distributions thus obtained, we then calculate the mutual information (MI) (Cover and Thomas, 2006) over the joint probability distribution of the state of the sender enzyme and the state of receiver protein. The joint probability distribution encodes both the allosterically induced state-switching transitions and the product formations by the sender enzyme's catalytic site.

Our work illustrates how allostery directly impacts the transfer of information within signaling pathways. It shows that the communication by a sender enzyme in a signaling pathway is not merely modulated by the number of products generated, but also significantly depends on the time signature of product arrivals at other receiver proteins present downstream. The analysis also suggests a broader role for allostery: a way to increase the information communicated within signaling pathways even, counter-intuitively, via down-regulation. This also provides a possible strategy which can be used to engineer allosteric interactions in a protein participating in a cellular signaling pathway in order to regulate the pathway and the associated function.

7.3 Model

We start with a minimal model of a sender enzyme with an allosteric regulator site (Figure 7.1) inspired by the KNF (Koshland-Nmethy-Filmer) model of protein allosteric regulation as described by Koshland *et al.* (1966). The model features a sender enzyme, E_{AX} which functions as a sender of information to another protein downstream in the signaling pathway (a receiver protein) via product molecules generated by the sender enzyme. Here, by information, we imply the influence the sender enzyme has on the probability of the receiver protein to be bound to the product

molecules in time. The sender enzyme has a catalytic site X which can bind or unbind to a substrate S to make complex E_{AX_S}, with rate constants k_{X+} and k_{X-}, respectively. In addition, the sender enzyme also contains an allosteric regulator site A. The sender enzyme with the bound complex at the catalytic site subsequently degrades to generate product, P, with a rate constant of d_p releasing its catalytic site X back to its original unbound form.

When the allosteric regulator site A reversibly binds to substrate, it creates a complex E_{A_SX} with the rate constants k_{A+} and k_{A-} respectively (see Figure 7.1a). The bound complex at the allosteric regulator can exploit the sender enzyme's network of interactions to influence the activity of the catalytic site (see Figure 7.1b). This allosteric interaction occurs between a bound allosteric regulator site and unbound catalytic site with a rate constant of h_p. The modified catalytic site (X^*) now performs its functions with different activity such that its rate constants for binding and unbinding with the substrates changes to k_{X^*+} and k_{X^*-} respectively. Moreover, the rate constant of the degradation of the complex at X^* with substrate to create product also changes to r_p (see Figure 7.1c). It should be noted that, these changes are only localized to the catalytic site, whereas, the allosteric regulator does not show any changes in its dynamics of interaction, i.e., the rate constants of binding and unbinding of substrate at the allosteric regulator site does not change as the catalytic site alters its state.

Finally, the catalytic site (X) after being allosterically modified to (X^*) can relax back to its original state in a process with a rate constant of s_p. This process can occur regardless of the state of the allosteric regulator (i.e., whether it is bound or unbound) (see Figure 7.1d).

The products, P, produced in processes a) and c) in the Figure 7.1 can interact with the receiver protein Y downstream in the signaling pathway with a rate constant

of k_Y. This protein acts as a receiver for the signal generated by the sender enzyme in the signaling pathway in the form of products. During the interaction of product P with the receiver Y, a binding event signifies a successful transfer of signal from the allosterically regulated sender enzyme, E to the receiver protein, Y.

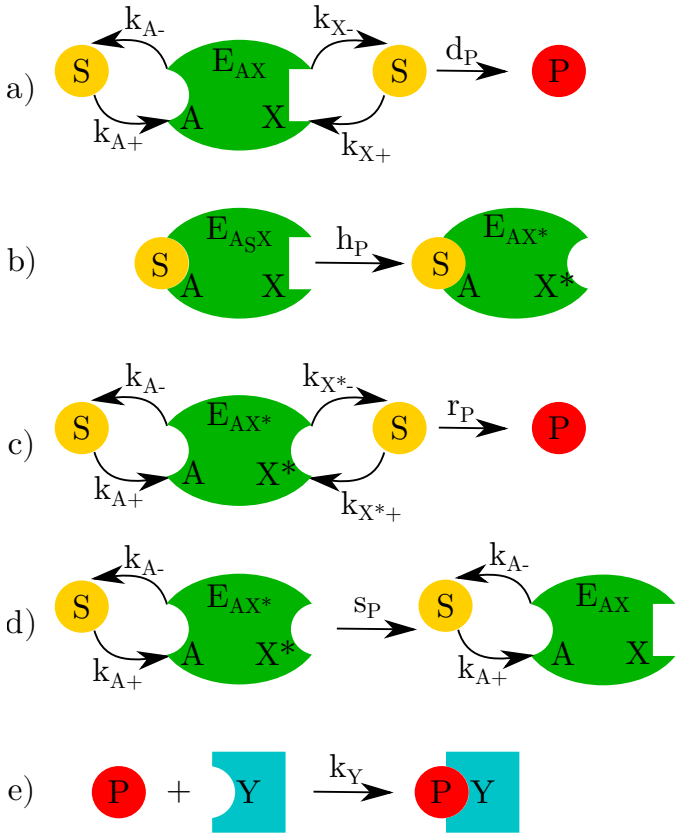

Figure 7.1: Schematic diagram of the model of an allosterically regulated sender enzyme. It shows E_{AX}, a sender enzyme, with catalytic site X interacting with substrates S to make product P (a) and allosteric regulator site A which can interact with the substrates to change the state of the sender enzyme to E_{AX} with catalytic site X^* (b). The catalytic site X^* can then also interact with the substrate to make products P however with different rate constants (c). The enzyme's catalytic site in its alternate state X^* can also relax back to its original state X (d). Finally, the products generated then bind to a receiver protein Y downstream in the signaling pathway (e).

For illustrative purposes, the reactions described in the processes a) to d) in Figure 7.1 can be simulated in a straightforward fashion using the Gillespie algorithm (Gillespie, 1977). Here, we simulate the product arrival events described in

Figure 7.1 with the help of a stochastic simulation for a sender enzyme in a system with a fixed large number of substrates. It should be noted that here, we are only simulating the product formation events. Therefore, the diffusion of product molecules is not included in the calculations. The sub-routine used for the stochastic simulation is given in Appendix C.2.

In the case of allosterically up-regulated sender enzyme (i.e., when the rate constant/s $k_{X^{*+}}$ or/and d_p is/are greater than rate constant/s $k_{X^{*+}}$ or/and r_p), we observed "bursts" of higher rates of product formation events in the midst of a lower rate of product formation, Figure 7.2. On the other hand, without any allosteric regulation, the waiting times between product arrivals are approximately exponentially distributed (as expected). This can be shown by calculating the ratio of the mean squared to the variance in the waiting times between product arrival events which is ≈ 1.00, for the choice of parameters specified in the caption of Figure 7.2. This is expected for exponentially distributed waiting times. This ratio strongly deviates from unity (≈ 17.40) in the presence of up-allosteric regulation in the sender enzyme, Figure 7.2.

Moreover, the stochastic simulations are also able to distinguish between the variation of product formation rates of sender enzymes with the amount of substrate present in the presence of allosteric regulation, and its absence (see Figure 7.3). Analogous results are observed in the presence of down-regulation in sender enzymes. In addition, our stochastic simulations also suggests that the model is able to replicate a "bursty" product formation in the presence of both K-type (i.e., when allosteric regulation manifests itself by modulating the catalytic site's binding affinity with the substrate), Figure 7.4 and V- type allosteric regulations (i.e., when allosteric regulation affects the rate by which the bound complex of catalytic site reduces to give products), Figure 7.5.

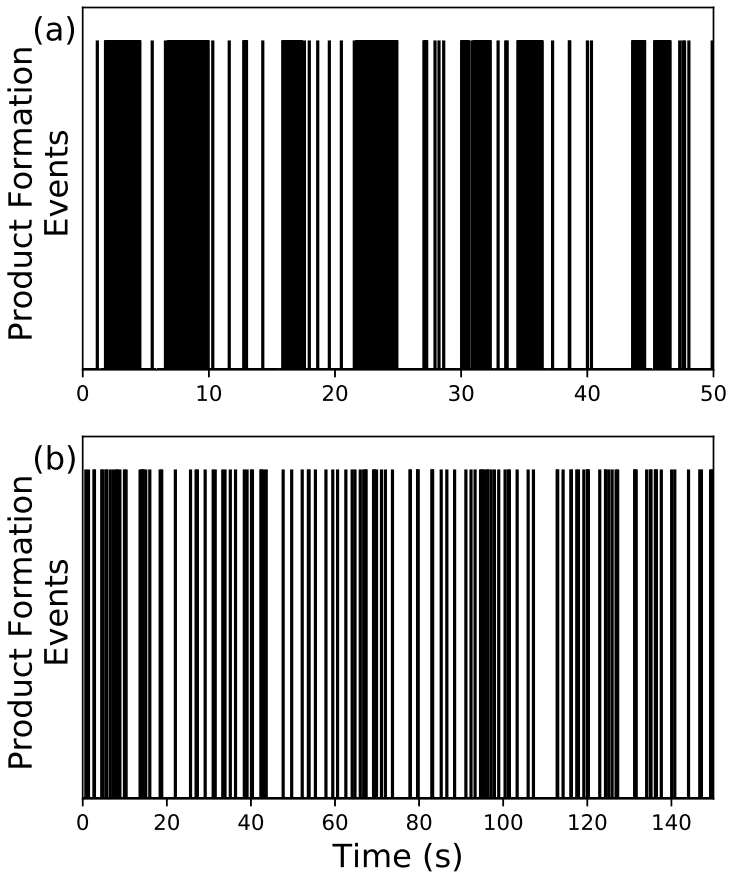

Figure 7.2: Time series of product arrivals stochastically simulated using Gillespie's algorithm from coupled chemical reactions shown in Figure 7.1 with an allosterically regulated (a) and unregulated (b) protein. The product arrivals are represented by vertical lines. The model with allostery exhibits a "bursty behavior". This behavior is not observed in the absence of allostery. Parameter values used are: $k_{X+} = 50s^{-1}, k_{X-} = 25s^{-1}, d_p = 0.9s^{-1}, k_{A+} = 50s^{-1}, k_{A-} = 25s^{-1}, k_{X^*+} = 75s^{-1}, k_{X^*-} = 25s^{-1}, r_p = 50s^{-1}, s_p = 70s^{-1}, h_p = 50s^{-1}$ for (a) and all the same except $h_p = 0s^{-1}$ for (b). For both the simulations we have an enzyme present in an excess of substrate (50 molecules).

The results of stochastic simulation suggest that in the presence of allosteric up/down regulation, the time of arrival of products exhibit a more complicated behavior as the catalytic site can exist in more than one type of state. Next, the products generated in bursts may bind to the receiver protein downstream transmitting the signal. Due to the bursty product formation, the amount of information encoded in the up-signal, propagated by the binding events between the products and the receiver

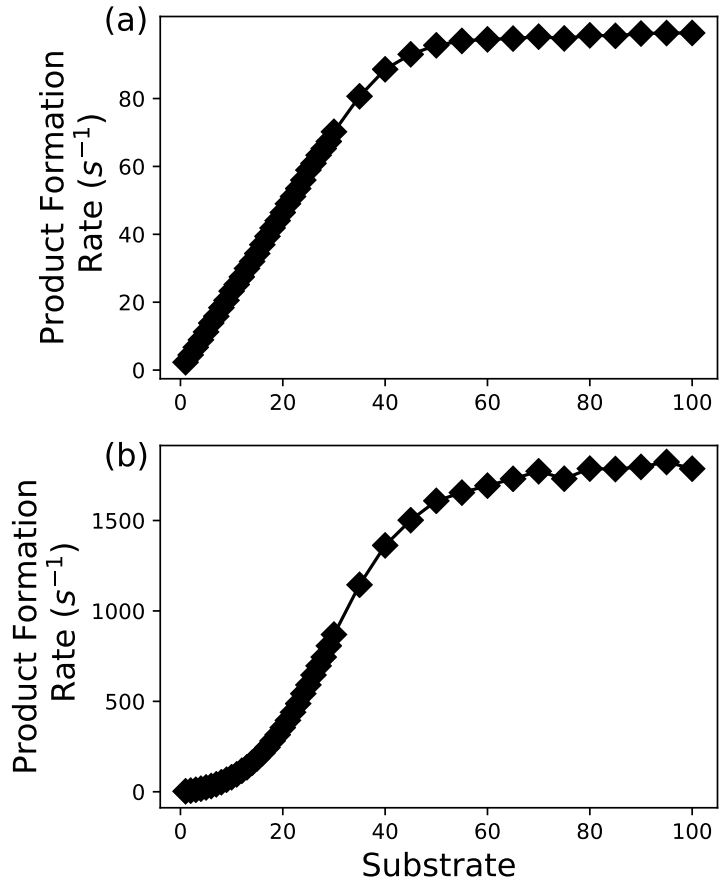

Figure 7.3: Variation in rates of product formation with substrate concentration. We observe that the model without allosteric regulation, (a) recapitulates the dynamics commonly observed in proteins following Michaelis-Menton Kinetics. On the other hand, when allosteric regulation is present, (b) the rate of product formation is non-linear at lower concentrations of substrate, a signature of cooperativity due to allosteric regulation. The following parameter values were used: $k_{X+} = 50s^{-1}, k_{X-} = 25s^{-1}, d_p = 5s^{-1}, k_{A+} = 50s^{-1}, k_{A-} = 25s^{-1}, k_{X^*+} = 75s^{-1}, k_{X^*-} = 50s^{-1}, r_p = 10000s^{-1}, s_p = 50s^{-1}, h_p = 0s^{-1}$ for (a) with a change of h_p to $h_p = 100s^{-1}$ for (b). The system was simulated for 30s and rates for each substrate concentration were averaged over 1000 independent runs. For both the simulations we have an enzyme present in an excess of substrate (50 molecules).

protein, is a complicated function that depends on the time-dependent interaction between the allosteric regulator site and the catalytic site of the sender enzyme.

In order to study the information encoded in the bursts of products in allosterically regulated sender enzymes, the model in Figure 7.1 can be further simplified under

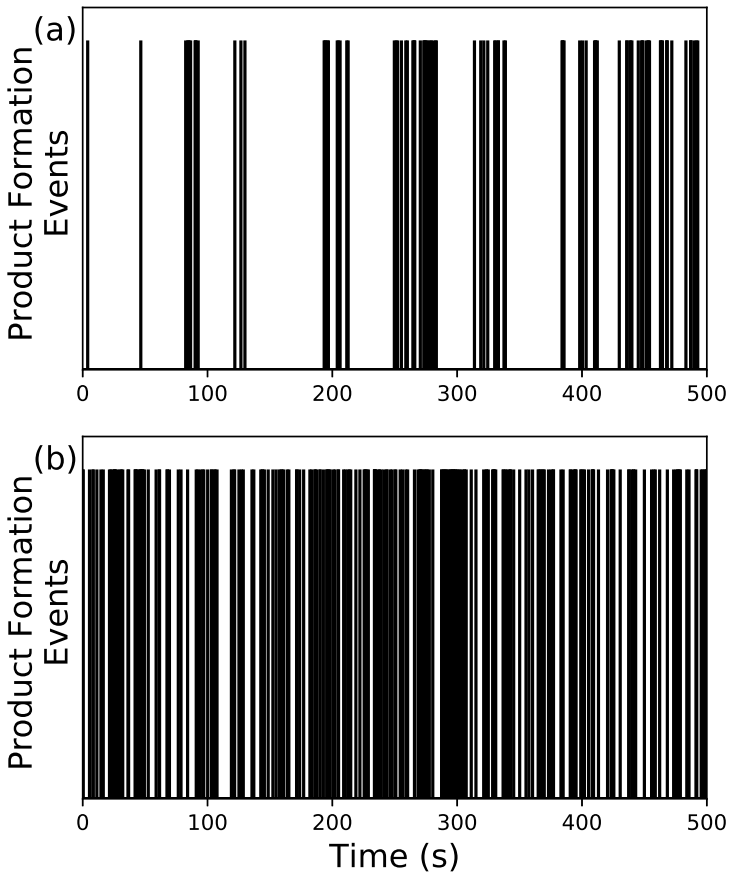

Figure 7.4: Bursts of products observed in K-type allostery. We observe that the model with K-type allosteric regulation, (a) shows a "bursty" formation of products which is not observed in the model without allosteric regulation, (b). Vertical lines represent product formation events. The following parameter values were used: $k_{X+} = 10000s^{-1}, k_{X-} = 10s^{-1}, d_p = 0.5s^{-1}, k_{A+} = 1000s^{-1}, k_{A-} = 25s^{-1}, k_{X^*+} = 0.01s^{-1}, k_{X^*-} = 1000s^{-1}, r_p = 0.5s^{-1}, s_p = 0.04s^{-1}, h_p = 5000s^{-1}$ for (a) and $h_p = 0s^{-1}$ for (b). For both the simulations we have an enzyme present in an excess of substrate (50 molecules).

the condition of detailed balance in steps (a) and (c) in Figure 7.1 as shown in the section below.

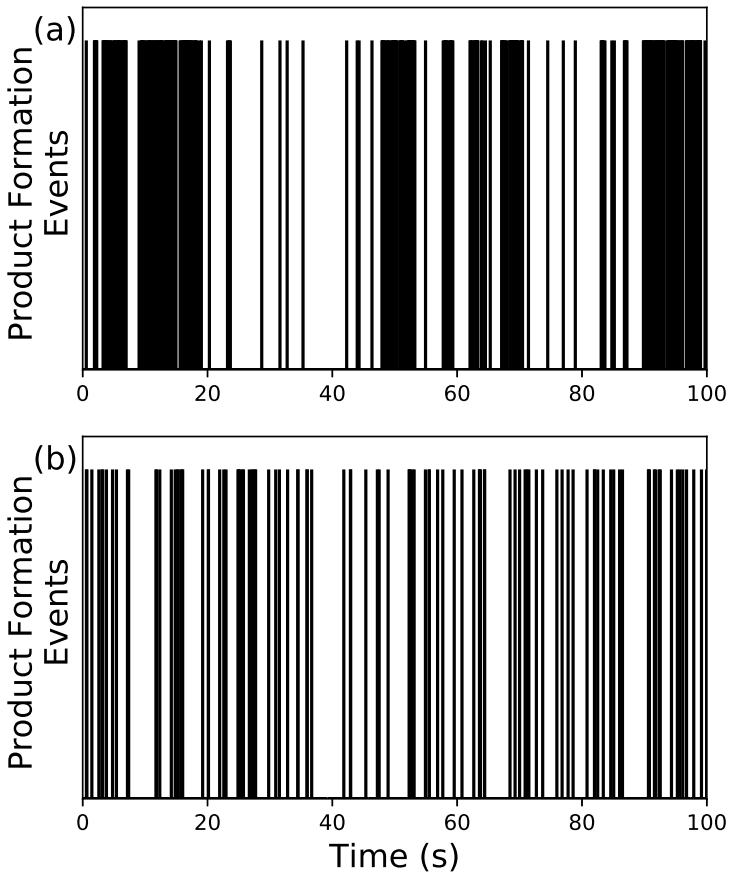

Figure 7.5: Bursts of products observed in V-type allostery. We observe that the model with V-type allosteric regulation, (a) shows a "bursty" formation of products which is not observed in the model without allosteric regulation, (b). Vertical lines represent product formation events. The following values of the parameters were used: $k_{X+} = 50s^{-1}, k_{X-} = 25s^{-1}, d_p = 0.3s^{-1}, k_{A+}50s^{-1}, k_{A-} = 25s^{-1}, k_{X^*+} = 50s^{-1}, k_{X^*-} = 25s^{-1}, r_p = 10s^{-1}, s_p = 70s^{-1}$ and $h_p = 50s^{-1}$ for (a) and $h_p = 0s^{-1}$ for (b). For both the simulations we have an enzyme present in an excess of substrate (50 molecules).

7.3.1 Coupled Chemical Reactions for Allosteric Regulation Reduce to a Two State System.

As the system reaches detailed balance, i.e., when the forward and backward reaction of complex formation between substrate and the sender enzyme is at equilibrium, the concentrations of bound complexes do not change with time. Therefore, for reactions in step (a) in Figure 7.1, when the rate of formation of $[E_{AX_S}]$ and $[E_{A_SX_S}]$,

and their rate of degradation are equal, i.e.,:

$$\begin{aligned}\frac{d[E_{AX_S}]}{dt} &= k_{X+}[E_{AX}]_{eq}[S]_{eq} - k_{X-}[E_{AX_S}]_{eq} \\ &= 0, \qquad (7.1)\\ \frac{d[E_{A_SX_S}]}{dt} &= k_{X+}[E_{A_SX}]_{eq}[S]_{eq} - k_{X-}[E_{A_SX_S}]_{eq} \\ &= 0 \qquad (7.2)\end{aligned}$$

respectively. Here, E_{AX_S} represents the sender enzyme E with the unbound allosteric site A and the catalytic site X bound to a substrate and the subscript 'eq' stands for the concentration of the corresponding biochemical species in detailed balance. Hence, we can now write:

$$\begin{aligned}\frac{[E_{AX_S}]_{eq}}{[E_{AX}]_{eq}[S]_{eq}} &= \frac{k_{X+}}{k_{X-}} \\ \frac{[E_{A_SX_S}]_{eq}}{[E_{A_SX}]_{eq}[S]_{eq}} &= \frac{k_{X+}}{k_{X-}} \qquad (7.3)\end{aligned}$$

Similarly, when the rate of formation of bound allosteric site, $[E_{A_SX}]$ and $[E_{A_SX_S}]$, and their rate of degradation are equal, we can write:

$$\begin{aligned}\frac{[E_{A_SX}]_{eq}}{[E_{AX}]_{eq}[S]_{eq}} &= \frac{k_{A+}}{k_{A-}} \\ \frac{[E_{A_SX_S}]_{eq}}{[E_{AX_S}]_{eq}[S]_{eq}} &= \frac{k_{A+}}{k_{A-}}. \qquad (7.4)\end{aligned}$$

Likewise, for reaction (c) in Figure 7.1 when then rate of formation of complex between the catalytic site in its alternate state (X^*) and the substrate, i.e., $[E_{AX_S^*}]$ and $[E_{A_SX_S^*}]$, and their rate of degradation are equal, we can write:

$$\begin{aligned}\frac{[E_{AX_S^*}]_{eq}}{[E_{AX^*}]_{eq}[S]_{eq}} &= \frac{k_{X^*+}}{k_{X^*-}} \\ \frac{[E_{A_SX_S^*}]_{eq}}{[E_{A_SX^*}]_{eq}[S]_{eq}} &= \frac{k_{X^*+}}{k_{X^*-}}. \qquad (7.5)\end{aligned}$$

Onward, for simplicity and clarity, we would refer to the sender enzymes with its unbound catalytic site in its original state (X) as R and the sender enzyme with its

unbound catalytic in its alternate state (X^*) as T (going by the historical representations of the two available states used by the likes of Monod *et al.* (1965) and Koshland *et al.* (1966)). Therefore, RS and TS would represent their bound forms of the sender enzyme with the substrate bound at the catalytic site respectively. In addition, given the initial concentration of the sender enzyme, $[E]_{initial}$, mass balance implies an additional condition on the concentrations of the chemical species as shown below:

$$\begin{aligned}
[E]_{initial} &= [E_{AX}]_{eq} + [E_{AX_S}]_{eq} + [E_{A_SX}]_{eq} + [E_{A_SX_S}]_{eq} + \\
&\quad [E_{AX^*}]_{eq} + [E_{AX^*_S}]_{eq} + [E_{A_SX^*}]_{eq} + [E_{A_SX^*_S}]_{eq} \\
[E]_{initial} &= [R]_{eq} + [RS]_{eq} + [T]_{eq} + [TS]_{eq}. \qquad (7.6)
\end{aligned}$$

Combining Eq. 7.6 with Eqs. 7.3, 7.4 and 7.5 with the condition that substrate is present in excess, i.e., $[S]_{eq} \approx [S]_{initial}$, we can reduce the model, to a two state model as shown below:

$$\mathrm{R} + \mathrm{S} \underset{k_{X-}}{\overset{k_{X+}}{\rightleftharpoons}} \mathrm{RS} \xrightarrow{d_p} \mathrm{P} + \mathrm{R} \qquad (7.7)$$

$$\mathrm{T} + \mathrm{S} \underset{k_{X^*-}}{\overset{k_{X^*+}}{\rightleftharpoons}} \mathrm{TS} \xrightarrow{r_p} \mathrm{P} + \mathrm{T}. \qquad (7.8)$$

Imposing the condition of detailed balance in reaction 7.7 the rate of formation of RS and its degradation will be equal, i.e.,:

$$k_{X+}[R]_{eq}[S]_{eq} = k_{X-}[RS]_{eq}. \qquad (7.9)$$

Therefore,

$$[RS]_{eq} = \frac{k_{X+}}{k_{X-}}[R]_{eq}[S]_{eq}. \qquad (7.10)$$

Also, the rate of product formation from step (a) in Figure 7.1 can be expressed using

reaction 7.7 and Eq. 7.10 as:

$$\frac{d[P]}{dt} = d_P[RS]_{eq} \tag{7.11}$$

$$\frac{d[P]}{dt} = \frac{k_{X+}d_P}{k_{X-}}[R]_{eq}[S]_{eq} \tag{7.12}$$

$$\frac{d[P]}{dt} = k_R[R]_{eq}[S]_{eq} \tag{7.13}$$

where, k_R can be treated as the effective forward rate for the formation of product molecules in reaction 7.7. Following a similar logic, reactions 7.7 and 7.8 can be expressed in the two state system as:

$$\mathrm{R} + \mathrm{S} \xrightarrow{k_R} \mathrm{P} + \mathrm{R} \tag{7.14}$$

$$\mathrm{T} + \mathrm{S} \xrightarrow{k_T} \mathrm{P} + \mathrm{T} \tag{7.15}$$

where, $k_R = \frac{k_{X+}}{k_{X-}}d_p$ and $k_T = \frac{k_{X*+}}{k_{X*-}}r_p$ are the effective product formation rate constants for the generation of products from the sender enzyme in the two states (R and T respectively). In addition, as shown in Figure 7.1, the switching between the states of the sender enzyme is facilitated by processes b) and d).

7.3.2 Hidden Markov Model Representation of the Coupled Chemical Reactions.

The model described here can be further simplified with the help of HMMs. In the language of HMMs, the state of the sender enzyme can be represented by a variable (s) which can take any value between 1 (for state R) and 2 (for state T). Further on, in order to represent the state of the sender enzyme as a Markov chain, we discretize time into intervals of δt. The time interval δt is selected to be small enough such that only one or no product molecules can be produced in any time interval regardless of the state of the sender enzyme. This choice of small δt also ensures that the sender enzyme retains the state of its catalytic site during δt. This approximation is also backed by several studies which suggest that allostery manifests itself on timescales

varying from the order of 10ps to several nanoseconds (McDonald *et al.*, 2013). As a result of the approximation, the time scales pertaining to the rates of transitions between the states of sender enzyme (i.e., R to T and T to R) are several orders slower than the time scales involved in the formation of products in a given state (see Figure 7.6), i.e., the waiting times of the enzyme in state R or T is much greater than the time involved in product formation.

Therefore, the probability distribution of the states of the sender enzyme during time interval i ($P(s_i)$) can be expressed as a function of the probability distribution of states during the time interval $i-1$, and the dynamics of the catalytic site and the allosteric site in the sender enzyme as (Kampen, 1992):

$$P(s_i|s_{i-1}) = \Gamma_{i-1\to i} \tag{7.16}$$

where, $\Gamma_{i-1\to i}$ represents the transition matrix whose elements describe the probability of transition to a state s_i at time interval i given the state of the sender enzyme at time interval $i-1$. For a sender enzyme in state 1 (s_i=1), the probability of remaining in state 1, $P_{1\to 1}$, is proportional to $[R]_{eq}[S]_{eq}k_R$ from Eq. 7.14. Whereas, the probability for switching its state from state 1 to 2, $P_{1\to 2}$ will proportional to h_p time the probability of having a bound allosteric site, Figure 7.1b, i.e., $h_p\frac{[E_{AX}]_{eq}[S]_{eq}k_{A+}}{k_{A-}}$. On the other hand, if the sender enzyme is in state 2, the probability to remain in the state $P_{2\to 2}$ is proportional to $[T]_{eq}[S]_{eq}k_T$ and the probability to switch, $P_{2\to 1}$ is proportional to s_p from Figure 7.1d. Using these, the transition probabilities can be

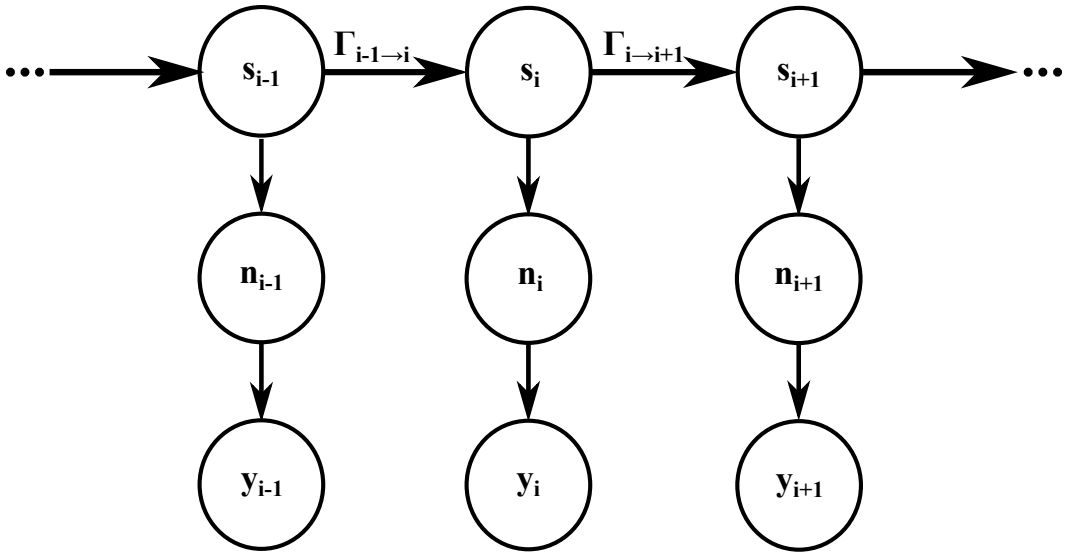

Figure 7.6: Graphical Model describing a two state system of a sender enzyme with Poissonian emission and a receiver protein. Here the observable, n_i, is the number of products produced by a catalytic site at time level i in the hidden state s_i. $\Gamma_{i-1\to i}$ represents the transition matrix of the states between time steps $i-1$ to i and y_i shows the status of the receiver protein at time interval i.

explicitly calculated using appropriate normalization conditions as:

$$\begin{aligned} P_{1\to 1} &= \frac{[R]_{eq}[S]_{eq}k_R}{[R]_{eq}[S]_{eq}k_R + h_p\frac{[E_{AX}]_{eq}[S]_{eq}k_{A+}}{k_{A-}}} \\ P_{1\to 2} &= \frac{h_p\frac{[E_{AX}]_{eq}[S]_{eq}k_{A+}}{k_{A-}}}{[R]_{eq}[S]_{eq}k_R + h_p\frac{[E_{AX}]_{eq}[S]_{eq}k_{A+}}{k_{A-}}} \\ P_{2\to 1} &= \frac{s_p}{[T]_{eq}[S]_{eq}k_T + s_p} \\ P_{2\to 2} &= \frac{[T]_{eq}[S]_{eq}k_T}{[T]_{eq}[S]_{eq}k_T + s_p}. \end{aligned} \tag{7.17}$$

Moreover, once the detailed balance is achieved in steps (a) and (c) in Figure 7.1, the dynamics of the above described process, Eq. 7.16 becomes independent in time, i.e., $\Gamma_{i+1\to i} = \Gamma$: $\forall i \in 1, 2, ..., M$ for M time steps where $M \to \infty$.

The HMM model in Figure 7.6 describes the dynamics of a catalytic site existing in state s_i at time step i and produces products n_i as described by the Poisson process

given below:

$$P(n_i|s_i) = \frac{e^{-\lambda_{s_i}} {\lambda_{s_i}}^{n_i}}{n_i!} \tag{7.18}$$

where, λ_{s_i} is the average number of products generated by the sender enzyme in state s_i at time interval i. This parameter is closely related to the rate constants for chemical processes described in Figure 7.1 as:

$$\lambda_{s_i} = \begin{cases} d_p[RS]_{eq}\delta t, & \text{if } s_i = 1 \\ r_p[TS]_{eq}\delta t, & \text{if } s_i = 2. \end{cases} \tag{7.19}$$

Here, $[RS]_{eq}$ and $[TS]_{eq}$ are the number of bound complexes between the substrate and the catalytic site in state 1 and 2 respectively when detailed balance exists in steps (a) and (c) in Figure 7.1. It should be noted that here we have used number of bound complexes as opposed to the traditional use of concentration. This change is also reflected in the units of the rate constants. Using this HMM framework, the probability of generating n_i products during time interval i can be written as:

$$P(n_i) = \sum_{s_i=1}^{2} P(n_i|s_i)P(s_i). \tag{7.20}$$

Finally, we must also describe how n_i changes the state of the receiver protein. As we select a time interval for the event to occur, at most one product can be produced in any time interval by the catalytic site. Consistent with the assumptions inherent to the Gillespie simulation, we assume that diffusion occurs on timescales vastly exceeding the rate of any chemical reaction (including the product formation rate). The product can then either bind to the receiver protein or disappear, by diffusing to a sink (such as an off-pathway, receiver), but it does not accumulate. Hence, the probability of the state of the receiver protein downstream in the signaling pathway (i.e., whether it is in complex with the product at time interval i or not) can be given

as a binomial process as:

$$P(y_i = 1|n_i) = \begin{cases} 0, & \text{if } n_i = 0 \\ p_Y, & \text{if } n_i = 1 \end{cases} \tag{7.21}$$

or,

$$P(y_i = 0|n_i) = \begin{cases} 1, & \text{if } n_i = 0 \\ 1 - p_Y, & \text{if } n_i = 1 \end{cases} \tag{7.22}$$

where, the state variable y_i reflects a binding event at the receiver protein during time interval i such that if $y_i = 1$ for a successful binding event and $y_i = 0$ otherwise. Here, p_Y is a real number less than 1 which depicts the probability of activation of receiver in the presence of the product. It should be noted that our choice of binomial process for activation of the receiver protein (i.e., binding with the product molecule) stems from our need for simplicity. Much more complicated models can be used for this process with no effect to the physics described here. In addition, for further simplicity, the dynamics of receiver (e.g., allostery in receiver protein) are not included in the model and is an interesting topic open for further research. Therefore, at the beginning of each time interval, the receiver is assumed to be in a refreshed state waiting to receive a new product. The probabilities shown above describes the dynamics of the model which will be used for calculating the amount of information encoded by the sender enzyme while communicating with the receiver protein downstream in the signaling pathway.

In summery, at a time interval i, the products generated by the sender enzyme which acts as a sender of a message in a state s_i are received by another receiver protein downstream in the signaling pathway which encodes the information onto the state y_i.

7.4 Results and Discussion

In the section above, we reduced the dynamical model of allosteric regulation of a sender enzyme to a simplified two-state model which can be represented by an HMM in discrete time. The model consists of a sender enzyme with a catalytic site and an allosteric regulator site. The sender enzyme binds with substrate which leads to the generation of products. The sender enzyme can exist in two different states as R (when the catalytic site exists as X) and T (when the catalytic site exists as X^*). The two states differ in their dynamics of substrate binding and the rate of product formation by the catalytic site.

The sender enzyme communicates with a receiver protein downstream in the signaling pathway with the help of products generated in time. The binding of the product with the receiver protein represents a successful reception of the signal. Finally, we can express the dynamics of switching of states of the sender protein as a function of allosteric interaction between the catalytic site and the allosteric regulator.

Now, exploiting tools from information theory, we quantify the amount of information conveyed by the sender enzyme to the receiver protein downstream at time interval i, as the mutual information (MI) (Cover and Thomas, 2006) between them.

$$MI = \sum_{s_i=1}^{2} \sum_{y_i=0}^{1} P(s_i, y_i) \log \frac{P(s_i, y_i)}{P(s_i)P(y_i)} \tag{7.23}$$

where, $P(s_i, y_i)$ represents the joint probability distribution between s_i and y_i at time interval i. According to the model, MI quantifies the amount of information encoded by the catalytic site dynamics in the form of arrival times of products which is transferred to a receiver protein in turn binding with it. For the calculation of MI from Eq. 7.23, the summations are performed over all the possible states of the sender enzyme and the receiver protein, i.e., $s_i \in [1, 2]$ and $y_i \in [0, 1]$.

In addition, to compute the expression above in Eq. 7.23, the joint probability

can be further expressed analytically as:

$$\begin{aligned} P(s_i, y_i) &= P(y_i|s_i)P(s_i) \\ &= \sum_{n_i=0}^{1} P(y_i, n_i|s_i)P(s_i) \end{aligned} \tag{7.24}$$

where, due to the strategic choice of δt, $n_i \in [0, 1]$. Further on,

$$\begin{aligned} P(s_i, y_i) &= \sum_{n_i=0}^{1} P(y_i, n_i|s_i)P(s_i) \\ &= \Big[\sum_{n_i=0}^{1} P(y_i|n_i)P(n_i|s_i)\Big]P(s_i). \end{aligned} \tag{7.25}$$

Here, $P(y_i|n_i)$ is the receiver process described earlier and $P(n_i|s_i)$ describes the process for the formation of products. $P(s_i)$ can be calculated as:

$$P(s_i) = \sum_{s_{i-1}=1}^{2} P(s_i|s_{i-1})P(s_{i-1}). \tag{7.26}$$

In this expression $P(s_i|s_{i-1})$ can be obtained from the switching dynamics represented by the Markov process representing the catalytic site dynamics of the sender enzyme, Eq. 7.16. This equation can be computed in an iterative fashion to calculate $P(s_i)$ for some given initial conditions of the active sites, i.e., $P(s_o)$. Moreover,

$$\begin{aligned} P(y_i) &= \sum_{n_i=0}^{1} P(y_i, n_i) \\ &= \sum_{n_i=0}^{1} P(y_i|n_i)P(n_i) \\ &= \sum_{n_i=0}^{1} P(y_i|n_i)\Big[\sum_{s_i=1}^{2} P(n_i, s_i)\Big] \\ &= \sum_{n_i=0}^{1} P(y_i|n_i)\Big[\sum_{s_i=1}^{2} P(n_i|s_i)P(s_i)\Big]. \end{aligned} \tag{7.27}$$

Therefore, with the help of Eqs. 7.25, 7.26 and 7.27 one can compute the MI described by Eq. 7.23 exactly for any time interval 'i' and given initial conditions $P(s_o)$. Here

we choose the initial condition that the sender enzyme at time 0 is in state 1, i.e., $s_o = 1$ and a large value of i, $i = 500$, is selected to ensure that the detailed balance is achieved in steps (a) and (c) in Figure 7.1 and that the MI no longer varies with i (this is equivalent to observing the long time behaviour of the model to minimize transient effects).

In the model, the modulation of the information conveyed through product arrivals can be achieved by two different mechanisms: (i) by changing the average number of products generated by the sender enzyme in the state R or T, and (ii) by altering the switching probabilities between these two states.

We first investigated the effect of varying the product formation rate of the first state R, λ_1, while keeping the product formation rate of the second state T and the switching probabilities between states constant, Figure 7.7. First, as a sanity check, we observe that without any allosteric regulation, i.e., when the state T is never visited by the sender enzyme (setting the transition probability of the sender enzyme from state R to T ($P_{1\to2}$) as zero), the MI between the state of the sender enzyme and the binding events at the receiver protein is zero. This is an expected result as, in this case, the sender protein is restricted to a single state, and alternate states can no longer influence the binding of product at the receiver protein. Therefore, no information is conveyed by the products from the sender enzyme to the receiver protein, regardless of the rate of product formation by the sender.

Next, when $\lambda_1 < \lambda_2$, i.e., in the case of up-regulation, the MI between the sender enzyme and the receiver protein is higher in contrast to the case when allosteric regulation is absent (i.e., when $P_{1\to2} = 0$). However, as λ_1 increases and comes closer to the value of λ_2, the difference between the product formation rates in the two states of the sender enzyme decreases and so does the MI until it drops down to zero when $\lambda_1 = \lambda_2$. This observation is counter-intuitive as it suggests that the MI

between the sender enzyme and the receiver protein decreases with an increase in the average rate of product formation. Equally counter-intuitively, the information communicated by the sender to the receiver rises once again when $\lambda_1 > \lambda_2$, i.e., for the case of down-regulation.

A physical interpretation of these observations lies in the implication of the meaning of MI. According to Eq. 7.23, MI is proportional to the joint probability of the state of the sender enzyme and the state of the receiver protein. For a time interval i, this depends on the conditional probability of y_i, dependent on s_i. As a result, MI quantifies the influence of the state of sender enzyme over the state of the receiver protein. Therefore, if y_i does not depend on s_i (one of such case would be in the lack of allosteric interactions where state s_i is now fixed and y_i is independent of s_i, and purely stochastic), the MI would be zero as shown below:

$$\begin{aligned} MI &= \sum_{s_i=1}^{2}\sum_{y_i=0}^{1} P(s_i, y_i) \log \frac{P(s_i, y_i)}{P(s_i)P(y_i)} \\ &= \sum_{s_i=1}^{2}\sum_{y_i=0}^{1} P(s_i, y_i) \log \frac{P(y_i|s_i)P(s_i)}{P(s_i)P(y_i)} \\ &= \sum_{s_i=1}^{2}\sum_{y_i=0}^{1} P(s_i, y_i) \log \frac{P(y_i)P(s_i)}{P(s_i)P(y_i)} \\ &= 0 \end{aligned} \tag{7.28}$$

On the other hand, as the influence of s_i over y_i increases, so does the MI. Therefore, during up-regulation as well as down-regulation, the non-zero MI implies that the sender enzyme is able to regulate the state of the receiver protein. This is counter-intuitive, particularly for the case of down-regulation as it predicts that the sender enzyme can regulate the state of the receiver protein despite generating a lower number of product molecules.

Second, we analyzed the variation of the information transmitted with allosteric regulation by varying the probability of switching of the state of the sender enzyme

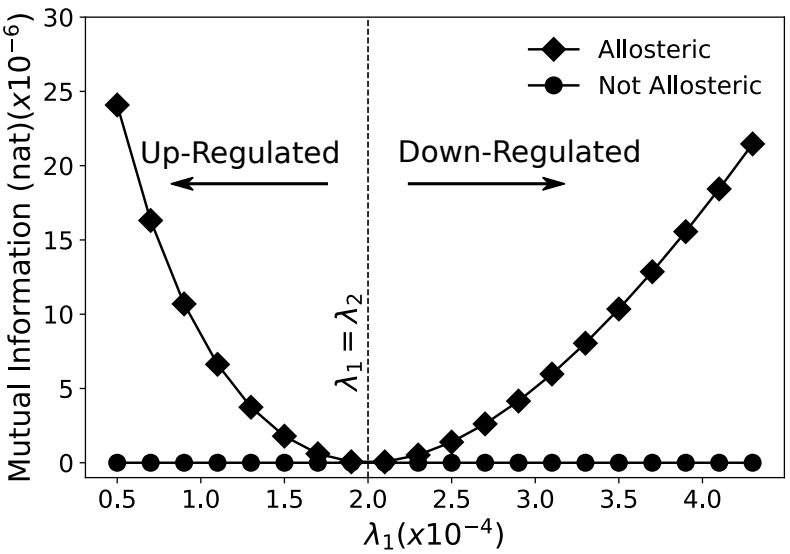

Figure 7.7: Effect of allostery, as captured by λ_1, on the information communicated by the sender enzyme. With allosteric up-regulation, i.e., when $\lambda_1 < \lambda_2$, the time series of product arrivals encodes a larger amount of information. The dotted line represents the value of λ_1 for which $\lambda_1 = \lambda_2$ and where the benefit of allostery on the MI vanishes. With allosteric down-regulation, i.e., when $\lambda_1 > \lambda_2$ the sender enzyme still communicates more information than in the absence of allostery. For the case of the allosteric sender enzyme, we used $\lambda_2 = 2 \times 10^{-4}$, and the transition probability of the sender enzyme from state R to T, $P_{1\to2}$ and vice versa, $P_{2\to1}$ as 0.2 and for the case of non-allosteric sender enzyme, $P_{1\to2} = 0$.

from R to T, i.e., $P_{1\to2}$, while holding other parameters (i.e., λ_1, λ_2, and $P_{2\to1}$) fixed, Figure 7.8. Once again, as a sanity check, we first observed the amount of information conveyed in the absence of allosteric regulation, i.e., when $\lambda_1 = \lambda_2$. As expected (similar to the case observed above), the signal no longer carries any MI when the effect of allostery is absent, and is also insensitive to the value of $P_{1\to2}$. Which implies that the binding events at the receiver protein would be independent of the state the sender enzyme and, as a result, there is no MI between them. On the other hand, for up- and down-regulated proteins, the sender enzyme conveys minimum information while being restricted to only one state (i.e., when $P_{1\to2} = 0$). However, as $P_{1\to2}$ grows, with up-regulation, the sender enzyme is able to access a state with higher average production and conveys a larger amount of information to the receiver

protein. Although, unexpectedly, with down-regulation, the sender enzyme is still able to transmit a larger amount of information for a range of switching probabilities. This counter-intuitive observation illustrates once more that information transmitted is not a mere function of the amount of product available to the receiver protein but also depends on the switching kinetics between the allosteric states of the sender enzyme.

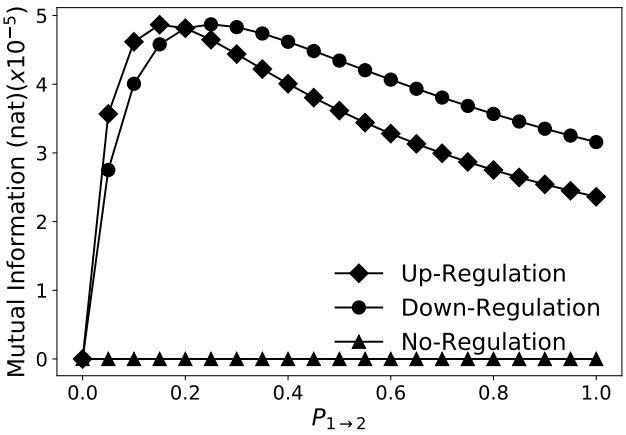

Figure 7.8: Effect of changing the switching probability from state R to T, $P_{1\to 2}$ on the information communicated by the sender enzyme for different types of allosteric regulations. We observe that without any allosteric regulation, the information encoded in the product time series do not vary with switching probabilities. However, with up-regulation, i.e., when $\lambda_1(1 \times 10^{-4}) < \lambda_2(4 \times 10^{-4})$, the sender enzyme communicates more information as the probability of it switching to an up-regulated state increases. Counter-intuitively, a down-regulated sender enzyme, i.e., when $\lambda_1(4 \times 10^{-4}) > \lambda_2(1 \times 10^{-4})$, is still able to communicate a larger amount of information for a wide range of switching probabilities. For all plots, we fix the transition probability from state T to R, $P_{2\to 1}$ to 0.08.

7.5 Conclusion

Here, we investigated a minimal model of allostery and quantified the information communicated by the catalytic site of an allosterically regulated sender enzyme to another receiver protein further down in the signaling pathway through time-varying product formation rates. Our choice of a two-state model is a matter of convenience;

our formalism readily generalizes to more complex kinetic models. More importantly, we found that an allosterically regulated enzymes may convey a larger amount of information as compared to an enzyme with no allosteric regulation by decreasing its net rate of product formation.

This suggests that allostery may provide the means to control the information encoded in the time of arrival of products in a way that goes beyond the energetically demanding "more product – better signal" exploitative paradigm. That is, allostery may provide an alternate "lower signal but higher information" regime. The possibility of parameter fine-tuning to communicate more information is especially relevant given allostery's key role in protein evolution (Modi *et al.*, 2018; Nussinov *et al.*, 2014; Townsend *et al.*, 2015). It also opens the possibility to study that nature may fine-tune allosteric parameters (including switching rates between states as well as production rates) to adapt/evolve its signaling pathways in response to external stimuli warranting exploratory (high information/low signal) or exploitative (low information/high signal) strategies (Hills *et al.*, 2015).

Chapter 8

FINAL REMARKS

8.1 Conclusions

The aim of this thesis was to study "the second secret of life", allostery by expanding beyond the traditional perspective where it describes how the binding of a ligand at a non-catalytic site can influence the dynamics of the distal catalytic site on the protein. Indeed, allostery, i.e., a distal perturbation by a ligand binding regulating the active site's response could take place via distal perturbation in terms of mutations. In this thesis, we studied, how allosteric interactions among residues within a protein influence the evolution of its function through mutations. This enabled us to describe the mechanistic principles involved in protein evolution of protein. Particularly, emergence of new functions or adaptation to a new environment. Furthermore, we also explored the role these interactions play beyond the scope of an isolated protein and at the level of a cell where proteins are allowed to interact with other proteins.

We approached the first problem in chapters 3 (Modi *et al.*, 2018) and 4 (Modi and Ozkan, 2018) where, we focused on the evolution of function in Thioredoxins and β-lactamases respectively. Using Dynamic Flexibility Index (DFI), we quantified the native state dynamics of ancestral and modern Thioredoxins, and showed how the flexibility profiles capture their functional evolution. Moreover, further distinctions in the allosteric network of interaction among residues were captured between the ancestral and extant Thioredoxins with the help of Dynamic Coupling Index (DCI) analysis. With the use of these metrics, a general mechanism can be observed where nature evolve the function of Thioredoxin proteins by fine-tuning their dynamics

through the "hinge shift" mechanism. According to this, a loss of rigidity in one part of the protein is compensated by a decrease in the flexibility of another distal part of the protein. Nature does so by performing mutations at residues with medium flexibilities which exhibit high dynamic allosteric coupling with the catalytic site.

Similar conclusions were drawn while observing the dynamical differences between the wild type TEM-1 β-lactamase and the resistance driving mutations observed from clinical and laboratory isolates. We observed that TEM-1 β-lactamase also evolved a new function by altering its native state dynamics, particularly through changes in the flexibility of the catalytic site residues. In addition, we also analyzed the data from exhaustive mutagenesis analysis performed on TEM-1 β-lactamase by Stiffler *et al.* (2015) and explored how the evolvability of a residue position towards emergence of antibiotic resistance can be described using its flexibility and allosteric dynamic coupling with the catalytic site.

Further on, using the mechanism of hinge shift observed in chapters 3 and 4, we developed a protein dynamics-based computational design approach to engineer an ancestral protein mutant. Our approach introduce the mutations on the ancestral protein sequence through the DFI and DCI profiles of ancestral and extant proteins to mold the dynamics of the engineered ancestral protein toward that of the target protein. We applied this approach to shift (*in silico*) the dynamics (DFI) of promiscuous ancestral GNCA β-lactamase, with moderate degrading activity towards cefotaxime and penicillin, to that of modern, penicillin specific TEM-1 β-lactamase. In this protocol, we selected mutations in two steps. First, we focused on introducing hinge shifts with the help of substitutions at residues which are not conserved between GNCA and TEM-1 β-lactamases based on their dynamic flexibilities and dynamic coupling with the rest of the protein. Second, we concentrated on the key role played by allosteric interactions in β-lactamase by selecting residues which exhibit medium

flexibilities and higher dynamic coupling with other sequentially conserved common hinges and the catalytic sites (these are called DARC spots). With substitutions at the residues selected in the two steps above, we were able to drive the flexibility profile of GNCA β-lactamase towards that of TEM-1 β-lactamase. In addition, the approach described here, was further verified experimentally by our collaborators, showing that engineered ancestral GNCA β-lactamase has become penicillin specific.

These analyses provided an understanding of the role played by allosteric interactions in regulating protein function in evolution. Afterwards, we addressed the other goal of the thesis– exploring the implications of allosteric interactions at a cellular level in chapters 6 and 7 (Modi *et al.*, 2020). In chapter 6, we directed our attention to E. coli ribosome where insertions in L22 protein of ribosome provides antibiotic resistance to the bacteria. However, despite the added benefit, this variant of ribosome (L22*) is not found to be dominant in nature. We approached this question by studying the differences in the dynamics introduced by mutations in L22 protein of an E. coli ribosome using coarse-grained models. Particularly, we focused on the magnesium ions which are critical for spatially coordinating the structure and function of ribosome. Through DFI analysis, it was observed that, L22* variant ribosome has a higher affinity to bind magnesium ions. Based on enhanced magnesium association of L22*, we hypothesize that there will be depletion of free magnesium ions in the cell, which could also diminish the amount of biologically active magnesium bound ATP in the cell. This has an impact on the growth rate of the organism with L22* variant. This was further verified by our experimental collaborators who showed that the L22* variant indeed has a lower concentration of activated ATP molecules and thus, exhibited a reduced growth rate of the cell. This provided the evidence of a more general phenomenon named as ionic allostery– the indirect interaction between two distinct and distal molecules in a cell which are propagated via the concentration

of free ions in the cell.

We further studied the implications of allostery in a signaling pathway in chapter 7 by proposing a simplistic toy model of an allosterically regulated enzyme. The enzyme communicates to a receiver protein downstream in the signaling pathway via product molecules. This enzyme also contains an allosteric regulator site which, upon binding to a substrate, can drive the transition of the catalytic site to another state with different kinetics. We described, using master equations and under the conditions of detail balance, how this enzyme can be reduced to a two state Hidden Markov Model in discrete time. This enabled us to calculate the mutual information between the state of enzyme's catalytic site and the state of the receiver protein (i.e., whether it is bound to a product molecule or not). This helps quantify the strength of control the enzyme has on the state of the receiver protein. Using this model, we showed that allostery efficiently regulates the time signature of the arrival of product molecules at the receiver protein. Further on, the model also predicts that the enzyme can effectively regulate the state of the receiver molecule not only via up-regulation, but through down-regulation as well. This was counter intuitive, as typically a loss in the product formation rate of enzymes is associated with a loss in the sensitivity of the signaling pathway due to a drop in the concentration of product molecules. Thereby, indicating a possible role played by allostery in regulating signaling pathways.

To summarize, the results we have discussed in the thesis describe the importance of understanding allosteric interactions in order to "crack the code" which governs the function of the protein, the role it plays in evolution and also how these interactions regulate other constituents in the rest of the cell. Thereby, emphasizing the significance of Monod's description of allostery– *"The second secret of Life"*.

8.2 Future Directions

The analyses described here explores a general mechanism of how nature exploits allostery in proteins to evolve their function. The evolutionary principles explored through the lens of allostery in chapters 3 and 4 provides the description of the principle of "hinge-shift" used by Thioredoxins and extant β-lactamase to evolve their functions. The applicability of this mechanism is further demonstrated by using it to identify mutations in ancestral β-lactamase enzyme to design a targeted function. This mechanism can be further applied to other protein system to study their evolution. It can also be used to motivate the design of new and novel drugs to counteract the infections developing resistance to existing drugs (e.g., antibiotic resistance bacteria, etc.).

Furthermore, analyzing the implications of allostery at cellular level, provides a platform to study the outcome of higher order interactions among different proteins in a cell, whose prediction is still an open problem in biology. In addition, the model used for studying the propagation of information in a pathway can be improved further by incorporating continuous time regime as opposed to discrete time. This can allow a better treatment of diffusion and accumulation of product molecules in the pathway.

Finally, the applicability of DFI analysis can be further expanded by incorporating time dependent dynamics in the picture. This is currently being explored by studying the vibrational spectra of C_α atoms in a protein. This can also provide a better mechanistic picture of how evolution alters the native state ensemble through allosteric mutations.

"Communication breakdown: Protein dynamics and drug design", Structure **17**, 3, 319 – 320 (2009).

"Molecular dynamics simulations of protein dynamics and their relevance to drug discovery", Current Opinion in Pharmacology **10**, 6, 738 – 744 (2010).

Adami, C., "The use of information theory in evolutionary biology", Annals of the New York Academy of Sciences **1256**, 49–65 (2012).

Adams, P. D., P. V. Afonine, G. Bunkóczi, V. B. Chen, I. W. Davis, N. Echols, J. J. Headd, L.-W. Hung, G. J. Kapral, R. W. Grosse-Kunstleve, A. J. McCoy, N. W. Moriarty, R. Oeffner, R. J. Read, D. C. Richardson, J. S. Richardson, T. C. Terwilliger and P. H. Zwart, "Phenix: a comprehensive python-based system for macromolecular structure solution", Acta Crystallographica Section D: Biological Crystallography **66**, 2, 213–221 (2010).

Afonine, P. V., M. Mustyakimov, R. W. Grosse-Kunstleve, N. W. Moriarty, P. Langan and P. D. Adams, "Joint x-ray and neutron refinement with phenix.refine", Acta Crystallographica Section D **66**, 11, 1153–1163 (2010).

Agozzino, L. and K. A. Dill, "Protein evolution speed depends on its stability and abundance and on chaperone concentrations", Proceedings of the National Academy of Sciences **115**, 37, 9092–9097 (2018).

Akanuma, S., Y. Nakajima, S.-i. Yokobori, M. Kimura, N. Nemoto, T. Mase, K.-i. Miyazono, M. Tanokura and A. Yamagishi, "Experimental evidence for the thermophilicity of ancestral life", Proceedings of the National Academy of Sciences of the United States of America **110**, 27, 11067–11072 (2013).

Alexander, S., A. Mathie and J. Peters, "Ion channels", British Journal of Pharmacology **164**, s1, S137–S174 (2011).

Allen, H. K., L. A. Moe, J. Rodbumrer, A. Gaarder and J. Handelsman, "Functional metagenomics reveals diverse beta-lactamases in a remote alaskan soil", The ISME journal **3**, 2, 243–251 (2009).

Anandakrishnan, R., B. Aguilar and A. V. Onufriev, "H++ 3.0: automating pk prediction and the preparation of biomolecular structures for atomistic molecular modeling and simulations", Nucleic Acids Research **40**, Web Server issue, W537–w541 (2012).

Andersen, H. C., "Molecular dynamics simulations at constant pressure and/or temperature", The Journal of chemical physics **72**, 4, 2384–2393 (1980).

Andersson, D., "Persistence of antibiotic resistant bacteria", Curr. Opin. Microbiol **6**, 5, 452–456 (2003).

Andersson, D. and D. Hughes, "Antibiotic resistance and its cost: is it possible to reverse resistance?", Nat. Rev. Microbiol **8**, 4, 260–71 (2010).

Appelbaum, P. C., "2012 and beyond: potential for the start of a second pre-antibiotic era?", The Journal of Antimicrobial Chemotherapy **67**, 9, 2062–2068 (2012).

Arganda-Carreras, I., V. Kaynig, C. Rueden, K. Eliceiri, J. Schindelin, A. Cardona and H. Seung, "Trainable weka segmentation: A machine learning tool for microscopy pixel classification", Bioinformatics **33**, 15, 2424–2426 (2017).

Arnér, E. S. J. and A. Holmgren, "Physiological functions of thioredoxin and thioredoxin reductase", European Journal of Biochemistry **267**, 20, 6102–6109 (2000).

Atanasov, B. P., D. Mustafi and M. W. Makinen, "Protonation of the beta-lactam nitrogen is the trigger event in the catalytic action of class a beta-lactamases", Proceedings of the National Academy of Sciences of the United States of America **97**, 7, 3160–3165 (2000).

Atilgan, A. R., S. R. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin and I. Bahar, "Anisotropy of fluctuation dynamics of proteins with an elastic network model", Biophysical Journal **80**, 1, 505–515 (2001).

Atilgan, C., Z. N. Gerek, S. B. Ozkan and A. R. Atilgan, "Manipulation of conformational change in proteins by single-residue perturbations", Biophysical Journal **99**, 3, 933–943 (2010).

Babtie, A., N. Tokuriki and F. Hollfelder, "What makes an enzyme promiscuous?", Current Opinion in Chemical Biology **14**, 2, 200–207 (2010).

Bahar, I., A. R. Atilgan, M. C. Demirel and B. Erman, "Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability", Physical Review Letters **80**, 12, 2733–2736 (1998).

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

Bahar, I., C. Chennubhotla and D. Tobi, "Intrinsic enzyme dynamics in the unbound state and relation to allosteric regulation", Current opinion in structural biology **17**, 6, 633–640 (2007).

Bahar, I., T. Lezon, L. Yang and E. Eyal, "Global dynamics of proteins : Bridging between structure and function", Annu. Rev. Biophys **39**, 23–42 (2010).

Balsera, M. A., W. Wriggers, Y. Oono and K. Schulten, "Principal component analysis and long time protein dynamics", The Journal of Physical Chemistry **100**, 7, 2567–2572 (1996).

Bar-Even, A., E. Noor, Y. Savir, W. Liebermeister, D. Davidi, D. S. Tawfik and R. Milo, "The moderately efficient enzyme: Evolutionary and physicochemical trends shaping enzyme parameters", Biochemistry **50**, 21, 4402–4410 (2011).

Bar-Rogovsky, H., A. Hugenmatter and D. S. Tawfik, "The evolutionary origins of detoxifying enzymes: the mammalian serum paraoxonases (pons) relate to bacterial homoserine lactonases", The Journal of Biological Chemistry **288**, 33, 23914–23927 (2013).

Barbosa, C., V. Trebosc, C. Kemmer, P. Rosenstiel, R. Beardmore, H. Schulenburg and G. Jansen, "Alternative evolutionary paths to bacterial antibiotic resistance cause distinct collateral effects", Molecular Biology and Evolution **34**, 9, 2229–2244 (2017).

Bayly, C. I., P. Cieplak, W. Cornell and P. A. Kollman, "A well-behaved electrostatic potential based method using charge restraints for deriving atomic charges: the resp model", The Journal of Physical Chemistry **97**, 40, 10269–10280 (1993).

Berendsen, H. J. C., J. P. M. Postma, W. F. van Gunsteren, A. DiNola and J. R. Haak, "Molecular dynamics with coupling to an external bath", **81**, 8, 3684–3690 (1984).

Berggård, T., S. Linse and P. James, "Methods for the detection and analysis of protein–protein interactions", PROTEOMICS **7**, 16, 2833–2842 (2007).

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, "The Protein Data Bank", Nucleic Acids Research **28**, 1, 235–242 (2000).

Bershtein, S., M. Segal, R. Bekerman, N. Tokuriki and D. S. Tawfik, "Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein", Nature **444**, 7121, 929 (2006).

Bloom, J. D., S. T. Labthavikul, C. R. Otey and F. H. Arnold, "Protein stability promotes evolvability", Proceedings of the National Academy of Sciences of the United States of America **103**, 15, 5869–5874 (2006).

Boehr, D. D., H. J. Dyson and P. E. Wright, "An nmr perspective on enzyme dynamics", Chemical Reviews **106**, 8, 3055–3079 (2006).

Born, M. and K. Huang, *Dynamical Theory of Crystal Lattices*, Oxford Classic Texts in the Physical Sciences (Oxford University Press, Oxford, New York, 1998).

Borovinskaya, M. A., S. Shoji, J. M. Holton, K. Fredrick and J. H. D. Cate, "A steric block in translation caused by the antibiotic spectinomycin", ACS Chemical Biology **2**, 8, 545–552 (2007).

Boucher, J. I., J. R. Jacobowitz, B. C. Beckett, S. Classen and D. L. Theobald, "An atomic-resolution view of neofunctionalization in the evolution of apicomplexan lactate dehydrogenases", eLife **3** (2014).

Bowman, G. R., E. R. Bolin, K. M. Hart, B. C. Maguire and S. Marqusee, "Discovery of multiple hidden allosteric sites by combining markov state models and experiments", Proceedings of the National Academy of Sciences **112**, 9, 2734–2739 (2015).

Bowman, G. R. and P. L. Geissler, "Equilibrium fluctuations of a single folded protein reveal a multitude of potential cryptic allosteric sites", Proceedings of the National Academy of Sciences **109**, 29, 11681–11686 (2012).

Bradford, P. A., "Extended-spectrum beta-lactamases in the 21st century: characterization, epidemiology, and detection of this important resistance threat", Clinical Microbiology Reviews **14**, 4, 933–951, table of contents (2001).

Breneman, C. M. and K. B. Wiberg, "Determining atom-centered monopoles from molecular electrostatic potentials. the need for high sampling density in formamide conformational analysis", Journal of Computational Chemistry **11**, 3, 361–373 (1990).

Bridgham, J. T., E. A. Ortlund and J. W. Thornton, "An epistatic ratchet constrains the direction of glucocorticoid receptor evolution", Nature **461**, 7263, 515–519 (2009).

Brooks, B. R., R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan and M. Karplus, "CHARMM: A program for macromolecular energy, minimization, and dynamics calculations", Journal of Computational Chemistry **4**, 2, 187–217 (1983).

Buchenberg, S., F. Sittel and G. Stock, "Time-resolved observation of protein allosteric communication", Proceedings of the National Academy of Sciences **114**, 33, E6804–e6811 (2017).

Buckel, P., A. Buchberger, A. Böck and H. Wittmann, "Alteration of ribosomal protein l6 in mutants of escherichia coli resistant to gentamicin", MGG Mol. Gen. Genet **158**, 1, 47–54 (1977).

Buckingham, R. A. and J. E. Lennard-Jones, "The classical equation of state of gaseous helium, neon and argon", Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences **168**, 933, 264–283 (1938).

Butler, B. M., Z. N. Gerek, S. Kumar and S. B. Ozkan, "Conformational dynamics of nonsynonymous variants at protein interfaces reveals disease association", Proteins: Structure, Function, and Bioinformatics **83**, 3, 428–435 (2015).

Butler, B. M., I. C. Kazan, A. Kumar and S. B. Ozkan, "Coevolving residues inform protein dynamics profiles and disease susceptibility of nsnvs", PLOS Computational Biology **14**, 11, e1006626 (2018).

Cabrera, A. C., P. A. Sánchez-Murcia and F. Gago, "Making sense of the past: hyperstability of ancestral thioredoxins explained by free energy simulations", Physical Chemistry Chemical Physics **19**, 34, 23239–23246 (2017).

Calvetti, D., L. Reichel and D. C. Sorensen, "An implicitly restarted lanczos method for large symmetric eigenvalue problems", Electron. Trans. Numer. Anal. **2**, 1–21 (1994).

Campbell, E. C., G. J. Correy, P. D. Mabbitt, A. M. Buckle, N. Tokuriki and C. J. Jackson, "Laboratory evolution of protein conformational dynamics", Current Opinion in Structural Biology **50**, 49–57 (2018).

Campitelli, P., J. Guo, H.-X. Zhou and S. B. Ozkan, "Hinge-shift mechanism modulates allosteric regulations in human pin1", The Journal of Physical Chemistry B **122**, 21, 5623–5629 (2018).

Campitelli, P., T. Modi, S. Kumar and S. Ozkan, "The role of conformational dynamics and allostery in modulating protein evolution", Annu. Rev. Biophys **49**, 1, 267–288 (2020).

Capitani, G., Z. Markovic-Housley, J. N. Jansonius, G. D. Val, M. Morris, P. Schürmann and L. Biochimie, "Crystal structures of thioredoxins f and m from spinach chloroplasts", in "Photosynthesis: Mechanisms and Effects", pp. 1939–1942 (Springer, Dordrecht, 1998).

Carr, H. Y. and E. M. Purcell, "Effects of diffusion on free precession in nuclear magnetic resonance experiments", Physical review **94**, 3, 630 (1954).

Carroll, S. M., J. T. Bridgham and J. W. Thornton, "Evolution of hormone signaling in elasmobranchs by exploitation of promiscuous receptors", Molecular Biology and Evolution **25**, 12, 2643–2652 (2008).

Case, D., V. Babin, J. Berryman, R. Betz, Q. Cai, D. Cerutti, T. Cheatham, T. Darden, R. Duke, H. Gohlke, A. Goetz, S. Gusarov, N. Homeyer, P. Janowski, J. Kaus, I. Kolossváry, A. Kovalenko, T. Lee, S. LeGrand, T. Luchko, R. Luo, B. Madej, K. Merz, F. Paesani, D. Roe, A. Roitberg, C. Sagui, R. Salomon-Ferrer, G. Seabra, C. Simmerling, W. Smith, J. Swails, Walker, J. Wang, R. Wolf, X. Wu and P. Kollman, {*Amber 14*} (2014).

Cavanagh, J., W. J. Fairbrother, A. G. Palmer III and N. J. Skelton, *Protein NMR spectroscopy: principles and practice* (Elsevier, 1995).

Chakrabarti, A., S. Srivastava, C. Swaminathan, A. Surolia and R. Varadarajan, "Thermodynamics of replacing an α-helical pro residue in the p40s mutant of escherichia-coli thioredoxin", Protein Science **8**, 11, 2455–2459 (1999).

Chang, K., J. Wen and L. Yang, "Functional importance of mobile ribosomal proteins", Biomed Res. Int **2015**, 1–11 (2015).

Changeux, J. P., "The feedback control mechanisms of biosynthetic l-threonine deaminase by l-isoleucine.", Cold Spring Harbor symposia on quantitative biology **26**, 313–8 (1961).

Chen, V. B., I. I. I. Arendall, J. J. Headd, D. A. Keedy, R. M. Immormino, G. J. Kapral, L. W. Murray, J. S. Richardson and D. C. Richardson, "Molprobity: all-atom structure validation for macromolecular crystallography", Acta Crystallographica. Section D: Biological Crystallography **66**, Pt 1 (2010).

Chockalingam, K., M. Blenner and S. Banta, "Design and application of stimulus-responsive peptide systems", Protein Engineering, Design and Selection **20**, 4, 155–161 (2007).

Choi, I.-G. and S.-H. Kim, "Evolution of protein structural classes and protein sequence families", Proceedings of the National Academy of Sciences **103**, 38, 14056–14061 (2006).

Cooper, A. and D. T. Dryden, "Allostery without conformational change. A plausible model", European biophysics journal: EBJ **11**, 2, 103–109 (1984).

Copp, J. N., D. W. Anderson, E. Akiva, P. C. Babbitt and N. Tokuriki, "Exploring the sequence, function, and evolutionary space of protein superfamilies using sequence similarity networks and phylogenetic reconstructions", Methods in Enzymology **620**, 315–347 (2019).

Cortina, G. A. and P. M. Kasson, "Excess positional mutual information predicts both local and allosteric mutations affecting beta lactamase drug resistance", Bioinformatics p. btw492 (2016).

Cortina, G. A. and P. M. Kasson, "Predicting allostery and microbial drug resistance with molecular simulations", Current Opinion in Structural Biology **52**, 80–86 (2018).

Cover, T. M. and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)* (Wiley-Interscience, New York, NY, USA, 2006).

Criswell, D., V. Tobiason, J. Lodmell and D. Samuels, "Mutations conferring aminoglycoside and spectinomycin resistance in borrelia burgdorferi", Antimicrob. Agents Chemother **50**, 2, 445–452 (2006).

Cusack, M. P., B. Thibert, D. E. Bredesen and d. G. Rio, "Efficient identification of critical residues based only on protein structure by network analysis", Plos One **2**, 5, e421 (2007).

Dann, C., C. Wakeman, C. Sieling, S. Baker, I. Irnov and W. Winkler, "Structure and mechanism of a metal-sensing regulatory rna", Cell **130**, 5, 878–892 (2007).

Darden, T., D. York and L. Pedersen, "Particle mesh ewald: An n.log(n) method for ewald sums in large systems", The Journal of Chemical Physics **98**, 12, 10089–10092 (1993).

Davies, J., "Origins and evolution of antibiotic resistance", Microbiologia **12**, 3, 9–16 (2010).

Day, W. H. E. and H. Edelsbrunner, "Efficient algorithms for agglomerative hierarchical clustering methods", Journal of Classification **1**, 1, 7–24 (1984).

de Vladar, H. P. and N. H. Barton, "The contribution of statistical physics to evolutionary biology", Trends in Ecology & Evolution **26**, 8, 424–432 (2011).

Dima, R. I. and D. Thirumalai, "Determination of network of residues that regulate allostery in protein families using sequence analysis", Protein Science: A Publication of the Protein Society **15**, 2, 258–268 (2006).

Dominguez, D., "Calcium signalling in bacteria", Mol. Microbiol **54**, 2, 291–297 (2004).

Doron, D., A. Kohen, K. Nam and D. T. Major, "How accurate are transition states from simulations of enzymatic reactions?", Journal of Chemical Theory and Computation **10**, 5, 1863–1871 (2014).

Echave, J. and F. M. Fernández, "A perturbative view of protein structural variation", Proteins **78**, 1, 173–180 (2010).

Eklund, H., F. K. Gleason and A. Holmgren, "Structural and functional relations among thioredoxins of different species", Proteins **11**, 1, 13–28 (1991).

Emsley, P., B. Lohkamp, W. G. Scott and K. Cowtan, "Features and development of coot", Acta Crystallographica Section D: Biological Crystallography **66**, 4, 486–501 (2010).

Essmann, U., L. Perera, M. L. Berkowitz, T. Darden, H. Lee and L. G. Pedersen, "A smooth particle mesh ewald method", The Journal of Chemical Physics **103**, 19, 8577–8593 (1995).

Evans, P., "Scaling and assessment of data quality", Acta Crystallographica. Section D, Biological Crystallography **62**, Pt 1, 72–82 (2006).

Ewig, C. S., R. Berry, U. Dinur, J.-R. Hill, M.-J. Hwang, H. Li, C. Liang, J. Maple, Z. Peng, T. P. Stockfisch, T. S. Thacher, L. Yan, X. Ni and A. T. Hagler, "Derivation of class II force fields. VIII. Derivation of a general quantum mechanical force field for organic compounds", Journal of Computational Chemistry **22**, 15, 1782–1800 (2001).

Eyal, E., L.-W. Yang and I. Bahar, "Anisotropic network model: systematic evaluation and a new web interface", Bioinformatics **22**, 21, 2619–2627 (2006).

Eyring, H., "The activated complex in chemical reactions", The Journal of Chemical Physics **3**, 2, 107–115 (1935).

Figliuzzi, M., H. Jacquier, A. Schug, O. Tenaillon and M. Weigt, "Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase tem-1", Molecular Biology and Evolution **33**, 1, 268–280 (2016).

Finnigan, G. C., V. Hanson-Smith, T. H. Stevens and J. W. Thornton, "Evolution of increased complexity in a molecular machine", Nature **481**, 7381, 360–364 (2012).

Fiser, A., R. K. G. Do and A. Šali, "Modeling of loops in protein structures", Protein Science **9**, 9, 1753–1773 (2000).

Fisher, J. F., S. O. Meroueh and S. Mobashery, "Bacterial resistance to beta-lactam antibiotics: compelling opportunism, compelling opportunity", Chemical Reviews **105**, 2, 395–424 (2005).

Fraser, J. S., M. W. Clarkson, S. C. Degnan, R. Erion, D. Kern and T. Alber, "Hidden alternative structures of proline isomerase essential for catalysis", Nature **462**, 7273, 669–673 (2009).

Fuchs, G., "Alternative Pathways of Carbon Dioxide Fixation: Insights into the Early Evolution of Life?", Annual Review of Microbiology **65**, 1, 631–658 (2011).

Fuglebakk, E., J. Echave and N. Reuter, "Measuring and comparing structural fluctuation patterns in large protein datasets", Bioinformatics (Oxford, England) **28**, 19, 2431–2440 (2012).

Gerek, Z., O. Keskin and S. Ozkan, "Identification of specificity and promiscuity of pdz domain interactions through their dynamic behavior", Proteins **77**, 4, 796–811 (2009).

Gerek, Z. N. and S. B. Ozkan, "A flexible docking scheme to explore the binding selectivity of pdz domains", Protein Science: A Publication of the Protein Society **19**, 5, 914–928 (2010).

Gerek, Z. N. and S. B. Ozkan, "Change in allosteric network affects binding affinities of pdz domains: Analysis through perturbation response scanning", PLOS Computational Biology **7**, 10, e1002154 (2011).

Gerlt, J. A. and P. C. Babbitt, "Enzyme (re)design: lessons from natural evolution and computation", Current Opinion in Chemical Biology **13**, 1, 10–18 (2009).

Gillespie, D. T., "Exact stochastic simulation of coupled chemical reactions", The Journal of Physical Chemistry **81**, 25, 2340–2361 (1977).

Glembo, T. J., D. W. Farrell, Z. N. Gerek, M. F. Thorpe and S. B. Ozkan, "Collective dynamics differentiates functional divergence in protein evolution", PLOS Computational Biology **8**, 3, e1002428 (2012).

Godoy-Ruiz, R., R. Perez-Jimenez, B. Ibarra-Molero and J. M. Sanchez-Ruiz, "Relation between protein stability, evolution and structure, as probed by carboxylic acid mutations", Journal of Molecular Biology **336**, 2, 313–318 (2004).

Goldstein, H., C. Poole and J. Safko, *Classical Mechanics, 3rd ed.*, vol. 70 (2002).

González, M.A., "Force fields and molecular dynamics simulations", JDN **12**, 169–200 (2011).

González-Ramírez, L. A., C. R. Ruiz-Martínez, R. A. Estremera-Andújar, C. A. Nieves-Marrero, A. García-Caballero, J. A. Gavira, J. López-Garriga and J. M. García-Ruiz, "Efficient screening methodology for protein crystallization based on the counter-diffusion technique", Crystal Growth & Design **17**, 12, 6780–6786 (2017).

Gordon, J. C., J. B. Myers, T. Folta, V. Shoja, L. S. Heath and A. Onufriev, "H++: a server for estimating pkas and adding missing hydrogens to macromolecules", Nucleic Acids Research **33**, Web Server issue, W368–371 (2005).

Gruber, R. and A. Horovitz, "Allosteric mechanisms in chaperonin machines", Chemical Reviews **116**, 11, 6588–6606 (2016).

Hadzipasic, A., C. Wilson, V. Nguyen, N. Kern, C. Kim, W. Pitsawong, J. Villali, Y. Zheng and D. Kern, "Ancient origins of allosteric activation in a ser-thr kinase", Science (New York, N.Y.) **367**, 6480, 912–917 (2020).

Halabi, N., O. Rivoire, S. Leibler and R. Ranganathan, "Protein Sectors: Evolutionary Units of Three-Dimensional Structure", Cell **138**, 4, 774–786 (2009).

Halgren, T. A., "Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94", Journal of Computational Chemistry **17**, 5-6, 490–519 (1996).

Hänggi, P., P. Talkner and M. Borkovec, "Reaction-rate theory: fifty years after kramers", Rev. Mod. Phys. **62**, 251–341 (1990).

Harms, M. J. and J. Thornton, "Evolutionary biochemistry: revealing the historical and physical causes of protein properties", Nature Reviews Genetics **14**, 559–571 (2013a).

Harms, M. J. and J. W. Thornton, "Evolutionary biochemistry: revealing the historical and physical causes of protein properties", Nature Reviews. Genetics **14**, 8, 559–571 (2013b).

Harris, C. R., K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del R'ıo, M. Wiebe, P. Peterson, P. G'erard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke and T. E. Oliphant, "Array programming with NumPy", Nature **585**, 7825, 357–362 (2020).

Hart, K. M., M. J. Harms, B. H. Schmidt, C. Elya, J. W. Thornton and S. Marqusee, "Thermodynamic system drift in protein evolution", PLoS biology **12**, 11, e1001994 (2014).

Hartley, H., "Origin of the Word 'Protein"', Nature **168**, 4267, 244–244 (1951).

Henzler-Wildman, K. A., M. Lei, V. Thai, S. J. Kerns, M. Karplus and D. Kern, "A hierarchy of timescales in protein dynamics is linked to enzyme catalysis", Nature **450**, 7171, 913–916 (2007).

Hernández-García, E. and T. Rosenbaum, "Chapter Six - Lipid Modulation of Thermal Transient Receptor Potential Channels", in "Current Topics in Membranes", edited by L. D. Islas and F. Qin, vol. 74 of *Thermal Sensors*, pp. 135–180 (Academic Press, 2014).

Herr, A. B., "Evolution of an allosteric "off switch" in apoptotic caspases", Journal of Biological Chemistry **293**, 15, 5462–5463 (2018).

Hills, T. T., P. M. Todd, D. Lazer, A. D. Redish, I. D. Couzin and Cognitive Search Research Group, "Exploration versus exploitation in space, mind, and society", Trends in Cognitive Sciences **19**, 1, 46–54 (2015).

Hockney, R. W. and J. W. Eastwood, *Computer Simulation Using Particles* (Taylor & Francis, Inc., Bristol, PA, USA, 1988).

Holmgren, A., "Thioredoxin", Annual Review of Biochemistry **54**, 1, 237–271 (1985).

Hoover, W. G., "Canonical dynamics: Equilibrium phase-space distributions", Physical review A **31**, 3, 1695 (1985).

Horn, J. R. and B. K. Shoichet, "Allosteric inhibition through core disruption", Journal of Molecular Biology **336**, 5, 1283–1291 (2004).

Hornak, V., R. Abel, A. Okur, B. Strockbine, A. Roitberg and C. Simmerling, "Comparison of multiple amber force fields and development of improved protein backbone parameters", Proteins: Structure, Function, and Bioinformatics **65**, 3, 712–725 (2006).

Ikeguchi, M., J. Ueno, M. Sato and A. Kidera, "Protein structural change upon ligand binding : Linear response theory", Phys. Rev. Lett **94**, 7, 1–4 (2005).

Ingles-Prieto, A., B. Ibarra-Molero, A. Delgado-Delgado, R. Perez-Jimenez, J. M. Fernandez, E. A. Gaucher, J. M. Sanchez-Ruiz and J. A. Gavira, "Conservation of protein structure over four billion years", Structure (London, England: 1993) **21**, 9, 1690–1697 (2013).

Jarymowycz, V. A. and M. J. Stone, "Fast time scale dynamics of protein backbones: Nmr relaxation methods, applications, and functional consequences", Chemical reviews **106**, 5, 1624–1671 (2006).

Jaudzems, K., M. Geralt, P. Serrano, B. Mohanty, R. Horst, B. Pedrini, M. A. Elsliger, I. A. Wilson and K. Wüthrich, "Nmr structure of the protein np_247299.1: comparison with the crystal structure", Acta Crystallographica. Section F, Structural Biology and Crystallization Communications **66**, Pt 10, 1367–1380 (2010).

Jeener, J., B. Meier, P. Bachmann and R. Ernst, "Investigation of exchange processes by two-dimensional nmr spectroscopy", The Journal of chemical physics **71**, 11, 4546–4553 (1979).

Jelsch, C., L. Mourey, J. M. Masson and J. P. Samama, "Crystal structure of escherichia coli tem1 beta-lactamase at 1.8 a resolution", Proteins **16**, 4, 364–383 (1993).

Jones, J. E. and S. Chapman, "On the determination of molecular fields.—i. from the variation of the viscosity of a gas with temperature", Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character **106**, 738, 441–462 (1924).

Jorgensen, W. L., J. Chandrasekhar, J. D. Madura, R. W. Impey and M. L. Klein, "Comparison of simple potential functions for simulating liquid water", The Journal of Chemical Physics **79**, 2, 926–935 (1983).

Joung, I. S. and T. E. Cheatham, "Determination of alkali and halide monovalent ion parameters for use in explicitly solvated biomolecular simulations", The Journal of Physical Chemistry B **112**, 30, 9020–9041 (2008).

Kabsch, W., "Xds", Acta Crystallographica Section D: Biological Crystallography **66**, 2, 125–132 (2010).

Kamp, v. d. M. W., E. J. Prentice, K. L. Kraakman, M. Connolly, A. J. Mulholland and V. L. Arcus, "Dynamical origins of heat capacity changes in enzyme-catalysed reactions", Nature Communications **9**, 1, 1–7 (2018).

Kampen, N. G. V., *Stochastic Processes in Physics and Chemistry* (Elsevier, 1992).

Kar, G., O. Keskin, A. Gursoy and R. Nussinov, "Allostery and population shift in drug discovery", Current Opinion in Pharmacology **10**, 6, 715–722 (2010).

Katti, S. K., D. M. LeMaster and H. Eklund, "Crystal structure of thioredoxin from escherichia coli at 1.68 a resolution", Journal of Molecular Biology **212**, 1, 167–184 (1990).

Keedy, D. A., Z. B. Hill, J. T. Biel, E. Kang, T. J. Rettenmaier, J. Brandão-Neto, N. M. Pearce, F. von Delft, J. A. Wells and J. S. Fraser, "An expanded allosteric network in ptp1b by multitemperature crystallography, fragment screening, and covalent tethering", eLife **7**, e36307 (2018).

Keskin, O., R. L. Jernigan and I. Bahar, "Proteins with similar architecture exhibit similar large-scale dynamic behavior", Biophysical Journal **78**, 4, 2093–2106 (2000).

Khade, P. M., A. Kumar and R. L. Jernigan, "Characterizing and predicting protein hinges for mechanistic insight", Journal of Molecular Biology **432**, 2, 508–522 (2020).

Khersonsky, O., C. Roodveldt and D. S. Tawfik, "Enzyme promiscuity: evolutionary and mechanistic aspects", Current Opinion in Chemical Biology **10**, 5, 498–508 (2006).

Khersonsky, O., D. Röthlisberger, O. Dym, S. Albeck, C. J. Jackson, D. Baker and D. S. Tawfik, "Evolutionary optimization of computationally designed enzymes: Kemp eliminases of the ke07 series", Journal of Molecular Biology **396**, 4, 1025–1042 (2010).

Khersonsky, O. and D. S. Tawfik, "Enzyme promiscuity: a mechanistic and evolutionary perspective", Annual Review of Biochemistry **79**, 471–505 (2010).

Kim, H., T. Zou, C. Modi, K. Dörner, T. J. Grunkemeyer, L. Chen, R. Fromme, M. V. Matz, S. B. Ozkan and R. M. Wachter, "A hinge migration mechanism unlocks the evolution of green-to-red photoconversion in gfp-like proteins", Structure (London, England : 1993) **23**, 1, 34–43 (2015).

Kim, H.-S., G. Fernandes and C.-W. Lee, "Protein Phosphatases Involved in Regulating Mitosis: Facts and Hypotheses", Molecules and Cells **39**, 9, 654–662 (2016).

Kister, J., C. Poyart and S. Edelstein, "Oxygen-organophosphate linkage in hemoglobin a. the double hump effect", Biophysical Journal **52**, 4, 527–535 (1987).

Knies, J. L., F. Cai and D. M. Weinreich, "Enzyme efficiency but not thermostability drives cefotaxime resistance evolution in tem-1 β-lactamase", Molecular Biology and Evolution **34**, 5, 1040–1054 (2017).

Knight, A. M., P. H. Culviner, N. Kurt-Yilmaz, T. Zou, S. B. Ozkan and S. Cavagnero, "Electrostatic Effect of the Ribosomal Surface on Nascent Polypeptide Dynamics", ACS Chemical Biology **8**, 6, 1195–1204 (2013).

Knudsen, M. and C. Wiuf, "The cath database", Human Genomics **4**, 3, 207 (2010).

Komorowski, M. and D. S. Tawfik, "The Limited Information Capacity of Cross-Reactive Sensors Drives the Evolutionary Expansion of Signaling", Cell Systems **8**, 1, 76–85.e6 (2019).

Koshland, D. E., G. Némethy and D. Filmer, "Comparison of experimental binding data and theoretical models in proteins containing subunits", Biochemistry **5**, 1, 365–385 (1966).

Kumar, A., B. M. Butler, S. Kumar and S. B. Ozkan, "Integration of structural dynamics and molecular evolution via protein interaction networks: a new era in genomic medicine", Current Opinion in Structural Biology **35**, Supplement C, 135–142 (2015a).

Kumar, A., T. J. Glembo and S. B. Ozkan, "The role of conformational dynamics and allostery in the disease development of human ferritin", Biophysical Journal **109**, 6, 1273–1281 (2015b).

Kurkcuoglu, Z., I. Bahar and P. Doruker, "Clustenm: Enm-based sampling of essential conformational space at full atomic resolution", J. Chem. Theory Comput **12**, 9, 4549–4562 (2016).

Ladbury, J. E., N. Kishore, H. W. Hellinga, R. Wynn and J. M. Sturtevant, "Thermodynamic effects of reduction of the active-site disulfide of escherichia coli thioredoxin explored by differential scanning calorimetry", Biochemistry **33**, 12, 3688–3692 (1994).

Lane, M. D. and B. Seelig, "Advances in the directed evolution of proteins", Current opinion in chemical biology **0**, 129–136 (2014).

Larrimore, K. E., I. C. Kazan, L. Kannan, R. P. Kendle, T. Jamal, M. Barcus, A. Bolia, S. Brimijoin, C.-G. Zhan, S. B. Ozkan and T. S. Mor, "Plant-expressed cocaine hydrolase variants of butyrylcholinesterase exhibit altered allosteric effects of cholinesterase activity and increased inhibitor sensitivity", Scientific Reports **7**, 1, 1–14 (2017).

Lee, D.-y. D., L. Galera-Laporta, M. Bialecka-Fornal, E. C. Moon, Z. Shen, S. P. Briggs, J. Garcia-Ojalvo and G. M. Süel, "Magnesium flux modulates ribosomes to increase bacterial survival", Cell **177**, 2, 352–360.e13 (2019).

Lehoucq, R. B., D. C. Sorensen and C. Yang, *ARPACK Users' Guide* (Society for Industrial and Applied Mathematics, 1998).

Lenaerts, T., J. Ferkinghoff-Borg, F. Stricher, L. Serrano, J. W. Schymkowitz and F. Rousseau, "Quantifying information transfer by protein domains: Analysis of the Fyn SH2 domain structure", BMC Structural Biology **8**, 1, 43 (2008).

Levin-Reisman, I., I. Ronin, O. Gefen, I. Braniss, N. Shoresh and N. Q. Balaban, "Antibiotic tolerance facilitates the evolution of resistance", Science (New York, N.Y.) **355**, 6327, 826–830 (2017).

LeVine, M. V. and H. Weinstein, "NbIT - A New Information Theory-Based Analysis of Allosteric Mechanisms Reveals Residues that Underlie Function in the Leucine Transporter LeuT", PLOS Computational Biology **10**, 5, e1003603 (2014).

Levitt, M., "Molecular dynamics of native protein. I. Computer simulation of trajectories", Journal of Molecular Biology **168**, 3, 595–617 (1983).

Levitt, M. H., *Spin dynamics: basics of nuclear magnetic resonance* (John Wiley & Sons, 2013).

Leyde., *Bulletin des Sciences Physiques et Naturelles en Nèerlande.*, vol. 1838 (Leyde., 1838).

Li, Z., A. Bolia, J. D. Maxwell, A. A. Bobkov, G. Ghirlanda, S. B. Ozkan and C. J. Margulis, "A rigid hinge region is necessary for high-affinity binding of dimannose to cyanovirin and associated constructs", Biochemistry **54**, 46, 6951–6960 (2015).

Link, H., D. Christodoulou and U. Sauer, "Advancing metabolic models with kinetic information", Current Opinion in Biotechnology **29**, 8–14 (2014).

Liu, J., A. Prindle, J. Humphries, M. Gabalda-Sagarra, M. Asally, D.-y. D. Lee, S. Ly, J. Garcia-Ojalvo and G. M. Süel, "Metabolic codependence gives rise to collective oscillations within biofilms", Nature **523**, 7562, 550–554 (2015).

Liu, T., S. T. Whitten and V. J. Hilser, "Functional residues serve a dominant role in mediating the cooperativity of the protein ensemble", Proceedings of the National Academy of Sciences of the United States of America **104**, 11, 4347–4352 (2007).

Livermore, D. M., "beta-lactamases in laboratory and clinical resistance", Clinical Microbiology Reviews **8**, 4, 557–584 (1995).

Loveland, A. B., G. Demo, N. Grigorieff and A. A. Korostelev, "Ensemble cryo-EM elucidates the mechanism of translation fidelity", Nature **546**, 7656, 113–117 (2017).

Lynch, M., "Evolution of the mutation rate", TRENDS in Genetics **26**, 8, 345–352 (2010).

Maguid, S., S. Fernández-Alberti, G. Parisi and J. Echave, "Evolutionary conservation of protein backbone flexibility", Journal of Molecular Evolution **63**, 4, 448–457 (2006).

Maier, J. A., C. Martinez, K. Kasavajhala, L. Wickstrom, K. E. Hauser and C. Simmerling, "ff14sb: Improving the accuracy of protein side chain and backbone parameters from ff99sb", Journal of Chemical Theory and Computation **11**, 8, 3696–3713 (2015).

Mak, W. S. and J. B. Siegel, "Computational enzyme design: Transitioning from catalytic proteins to enzymes", Current Opinion in Structural Biology **27**, 87–94 (2014).

Martínez, J. L., F. Baquero and D. I. Andersson, "Predicting antibiotic resistance", Nature Reviews Microbiology **5**, 12, 958–965 (2007).

Marzen, S., H. G. Garcia and R. Phillips, "Statistical mechanics of Monod-Wyman-Changeux (MWC) models", Journal of Molecular Biology **425**, 9, 1433–1460 (2013).

Mazal, H., H. Aviram, I. Riven and G. Haran, "Effect of ligand binding on a protein with a complex folding landscape", Physical Chemistry Chemical Physics (2017).

McConnell, H. M., "Reaction rates by nuclear magnetic resonance", The Journal of chemical physics **28**, 3, 430–431 (1958).

McCoy, A. J., R. W. Grosse-Kunstleve, P. D. Adams, M. D. Winn, L. C. Storoni and R. J. Read, "Phaser crystallographic software", Journal of Applied Crystallography **40**, 4, 658–674 (2007).

McDonald, L. R., M. J. Whitley, J. A. Boyer and A. L. Lee, "Colocalization of fast and slow timescale dynamics in the allosteric signaling protein chey", Journal of Molecular Biology **425**, 13, 2372–2381 (2013).

McLeish, T. C., M. J. Cann and T. L. Rodgers, "Dynamic transmission of protein allostery without structural change: Spatial pathways or global modes?", Biophysical Journal **109**, 6, 1240–1250 (2015).

McLeish, T. C. B., T. L. Rodgers and M. R. Wilson, "Allostery without conformation change: modelling protein dynamics at multiple scales", Physical Biology **10**, 5, 056004 (2013).

Medeiros, A. A., "Evolution and dissemination of beta-lactamases accelerated by generations of beta-lactam antibiotics", Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America **24 Suppl 1**, S19–45 (1997).

Meiboom, S. and D. Gill, "Modified spin-echo method for measuring nuclear relaxation times", Review of scientific instruments **29**, 8, 688–691 (1958).

Milo, R., "What is the total number of protein molecules per cell volume? A call to rethink some published values", BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology **35**, 12, 1050–1055 (2013).

Miton, C. M., S. Jonas, G. Fischer, F. Duarte, M. F. Mohamed, v. B. Loo, B. Kintses, S. C. L. Kamerlin, N. Tokuriki, M. Hyvönen and F. Hollfelder, "Evolutionary repurposing of a sulfatase: A new michaelis complex leads to efficient transition state charge offset", Proceedings of the National Academy of Sciences p. 201607817 (2018).

Miton, C. M. and N. Tokuriki, "How mutational epistasis impairs predictability in protein evolution and design", Protein Science: A Publication of the Protein Society **25**, 7, 1260–1272 (2016).

Modi, T., J. Huihui, K. Ghosh and S. Ozkan, "Ancient thioredoxins evolved to modern- day stability – function requirement by altering native state ensemble", B Biol. Sci **373**, 1749, 1–10 (2018).

Modi, T. and S. B. Ozkan, "Mutations utilize dynamic allostery to confer resistance in tem-1 β-lactamase", International Journal of Molecular Sciences **19**, 12 (2018).

Modi, T., S. B. Ozkan and S. Pressé, "Information propagation in time through allosteric signaling", Phys. Rev. Research **2**, 023367 (2020).

Monod, J. and F. Jacob, "Teleonomic mechanisms in cellular metabolism, growth, and differentiation", Cold Spring Harbor Symposia on Quantitative Biology **26**, 389–401 (1961).

Monod, J., J. Wyman and J.-P. Changeux, "On the nature of allosteric transitions: A plausible model", Journal of Molecular Biology **12**, 1, 88–118 (1965).

Moult, J., "Comparative Modeling in Structural Genomics", Structure **16**, 1, 14–16 (2008).

Murzin, A. G., S. E. Brenner, T. Hubbard and C. Chothia, "Scop: a structural classification of proteins database for the investigation of sequences and structures.", Journal of molecular biology **247**, 4, 536–540 (1995).

Mustacich, D. and G. Powis, "Thioredoxin reductase", Biochemical Journal **346**, 1, 1–8 (2000).

Myers, J., G. Grothaus, S. Narayanan and A. Onufriev, "A simple clustering algorithm can be accurate enough for use in calculations of pks in macromolecules", Proteins: Structure, Function, and Bioinformatics **63**, 4, 928–938 (2006).

Neria, E., S. Fischer and M. Karplus, "Simulation of activation free energies in molecular systems", The Journal of Chemical Physics **105**, 5, 1902–1921 (1996).

Nessar, R., J. Reyrat, A. Murray and B. Gicquel, "Genetic analysis of new 16s rrna mutations conferring aminoglycoside resistance in mycobacterium abscessus", J. Antimicrob. Chemother **66**, 8, 1719–1724 (2011).

Neuwald, A. F., "The CHAIN program: forging evolutionary links to underlying mechanisms", Trends in Biochemical Sciences **32**, 11, 487–493 (2007).

Nevin Gerek, Z., S. Kumar and S. Banu Ozkan, "Structural dynamics flexibility informs function and evolution at a proteome scale", Evolutionary Applications **6**, 3, 423–433 (2013).

Nguyen, V., C. Wilson, M. Hoemberger, J. B. Stiller, R. V. Agafonov, S. Kutter, J. English, D. L. Theobald and D. Kern, "Evolutionary drivers of thermoadaptation in enzyme catalysis", Science (New York, N.Y.) **355**, 6322, 289–294 (2017).

Nijhout, H. F., "The control of growth", Development **130**, 24, 5863–5867 (2003).

Nosé, S., "A molecular dynamics method for simulations in the canonical ensemble", Molecular physics **52**, 2, 255–268 (1984).

Number 4, C. C. P., "The ccp4 suite: programs for protein crystallography", Acta Crystallographica Section D **50**, 5, 760–763 (1994).

Nussinov, R., C.-J. Tsai and J. Liu, "Principles of Allosteric Interactions in Cell Signaling", Journal of the American Chemical Society **136**, 51, 17692–17701 (2014).

Ohki, S.-y. and M. Kainosho, "Stable isotope labeling methods for protein nmr spectroscopy", Progress in Nuclear Magnetic Resonance Spectroscopy **53**, 4, 208–226 (2008).

Oldfield, C. J. and A. K. Dunker, "Intrinsically disordered proteins and intrinsically disordered protein regions", Annual Review of Biochemistry **83**, 1, 553–584 (2014).

Ortlund, E. A., J. T. Bridgham, M. R. Redinbo and J. W. Thornton, "Crystal structure of an ancient protein", Science (New York, N.Y.) **317**, 5844, 1544–1548 (2007).

Osadchy, M. and R. Kolodny, "Maps of protein structure space reveal a fundamental relationship between protein structure and function", Proceedings of the National Academy of Sciences **108**, 30, 12301–12306 (2011).

Otten, R., L. Liu, L. R. Kenner, M. W. Clarkson, D. Mavor, D. S. Tawfik, D. Kern and J. S. Fraser, "Rescue of conformational dynamics in enzyme catalysis by directed evolution", Nature Communications **9**, 1, 1314 (2018).

Overhauser, A. W., "Polarization of nuclei in metals", Phys. Rev. **92**, 411–415 (1953).

Palmer 3rd, A., C. D. Kroenke and J. P. Loria, "Nuclear magnetic resonance methods for quantifying microsecond-to-millisecond motions in biological macromolecules.", Methods in enzymology **339**, 204 (2001).

Papaleo, E., L. Riccardi, C. Villa, P. Fantucci and L. De Gioia, "Flexibility and enzymatic cold-adaptation: a comparative molecular dynamics investigation of the elastase family", Biochimica Et Biophysica Acta **1764**, 8, 1397–1406 (2006).

Pauling, L., "The oxygen equilibrium of hemoglobin and its structural interpretation", Proceedings of the National Academy of Sciences **21**, 4, 186–191 (1935).

Peña, E. D. L., A. Mälkiä, H. Cabedo, C. Belmonte and F. Viana, "The contribution of TRPM8 channels to cold sensing in mammalian neurones", The Journal of Physiology **567**, 2, 415–426 (2005).

Pearlman, D. A., D. A. Case, J. W. Caldwell, W. S. Ross, T. E. Cheatham, S. DeBolt, D. Ferguson, G. Seibel and P. Kollman, "Amber, a package of computer programs for applying molecular mechanics, normal mode analysis, molecular dynamics and free energy calculations to simulate the structural and energetic properties of molecules", Computer Physics Communications **91**, 1-3, 1–41 (1995).

Perez-Jimenez, R., A. Inglés-Prieto, Z.-M. Zhao, I. Sanchez-Romero, J. Alegre-Cebollada, P. Kosuri, S. Garcia-Manyes, T. J. Kappock, M. Tanokura, A. Holmgren, J. M. Sanchez-Ruiz, E. A. Gaucher and J. M. Fernandez, "Single-molecule paleoenzymology probes the chemistry of resurrected enzymes", Nature Structural and Molecular Biology **18**, 5, nsmb.2020 (2011).

Perez-Jimenez, R., J. Li, P. Kosuri, I. Sanchez-Romero, A. P. Wiita, D. Rodriguez-Larrea, A. Chueca, A. Holmgren, A. Miranda-Vizuete, K. Becker, S.-H. Cho, J. Beckwith, E. Gelhaye, J. P. Jacquot, E. A. Gaucher, J. M. Sanchez-Ruiz, B. J. Berne and J. M. Fernandez, "Diversity of chemical mechanisms in thioredoxin catalysis revealed by single-molecule force spectroscopy", Nature Structural & Molecular Biology **16**, 8, 890 (2009).

Phillips, R., J. Kondev and J. Theriot, *Physical Biology of the Cell* (Garland Science, Taylor & Francis Group, New York, 2008).

Pontes, M. H., A. Sevostyanova and E. A. Groisman, "When Too Much ATP Is Bad for Protein Synthesis", Journal of Molecular Biology **427**, 16, 2586–2594 (2015).

Poole, P. H., F. Sciortino, U. Essmann and H. E. Stanley, "Phase behaviour of metastable water", Nature **360**, 6402, 324–328 (1992).

Porter, J. R., A. Meller, M. I. Zimmerman, M. J. Greenberg and G. R. Bowman, "Conformational distributions of isolated myosin motor domains encode their mechanochemical properties", eLife **9**, e55132 (2020).

Price, D. J. and C. L. Brooks, "A modified tip3p water potential for simulation with ewald summation", The Journal of Chemical Physics **121**, 20, 10096–10103 (2004).

Prindle, A., J. Liu, M. Asally, S. Ly, J. Garcia-Ojalvo and G. M. Süel, "Ion channels enable electrical communication in bacterial communities", Nature **527**, 7576, 59–63 (2015).

Rabi, I. I., J. R. Zacharias, S. Millman and P. Kusch, "A New Method of Measuring Nuclear Magnetic Moment", Physical Review **53**, 4, 318–318 (1938).

Rabiner, L. and B. Juang, "An introduction to hidden Markov models", IEEE ASSP Magazine **3**, 1, 4–16 (1986).

Raynes, Y., C. S. Wylie, P. D. Sniegowski and D. M. Weinreich, "Sign of selection on mutation rate modifiers depends on population size", Proceedings of the National Academy of Sciences **115**, 13, 3422–3427 (2018).

Reed, J. C., "Mechanisms of Apoptosis", The American Journal of Pathology **157**, 5, 1415–1430 (2000).

Risso, V. A., J. A. Gavira, E. A. Gaucher and J. M. Sanchez-Ruiz, "Phenotypic comparisons of consensus variants versus laboratory resurrections of precambrian proteins", Proteins **82**, 6, 887–896 (2014).

Risso, V. A., J. A. Gavira, D. F. Mejia-Carmona, E. A. Gaucher and J. M. Sanchez-Ruiz, "Hyperstability and substrate promiscuity in laboratory resurrections of precambrian β-lactamases", Journal of the American Chemical Society **135**, 8, 2899–2902 (2013).

Risso, V. A., F. Manssour-Triedo, A. Delgado-Delgado, R. Arco, A. Barroso-delJesus, A. Ingles-Prieto, R. Godoy-Ruiz, J. A. Gavira, E. A. Gaucher, B. Ibarra-Molero and J. M. Sanchez-Ruiz, "Mutational studies on resurrected ancestral proteins reveal conservation of site-specific amino acid preferences throughout evolutionary history", Molecular Biology and Evolution **32**, 2, 440–455 (2015).

Risso, V. A., S. Martinez-Rodriguez, A. M. Candel, D. M. Krüger, D. Pantoja-Uceda, M. Ortega-Muñoz, F. Santoyo-Gonzalez, E. A. Gaucher, S. C. L. Kamerlin, M. Bruix, J. A. Gavira and J. M. Sanchez-Ruiz, "De novo active sites for resurrected precambrian enzymes", Nature Communications **8**, 1, 1–13 (2017).

Risso, V. A. and J. M. Sanchez-Ruiz, "Resurrected Ancestral Proteins as Scaffolds for Protein Engineering", in "Directed Enzyme Evolution: Advances and Applications", edited by M. Alcalde, pp. 229–255 (Springer International Publishing, Cham, 2017).

Risso, V. A., J. M. Sanchez-Ruiz and S. B. Ozkan, "Biotechnological and protein-engineering implications of ancestral protein resurrection", Current Opinion in Structural Biology **51**, 106–115 (2018).

Romero, P. A. and F. H. Arnold, "Exploring protein fitness landscapes by directed evolution", Nature reviews. Molecular cell biology **10**, 12, 866–876 (2009).

Romero-Romero, M. L., V. A. Risso, S. Martinez-Rodriguez, B. Ibarra-Molero and J. M. Sanchez-Ruiz, "Engineering ancestral protein hyperstability", The Biochemical Journal **473**, 20, 3611–3620 (2016).

Ross, J., "The dark matter of biology", Biophys. J **111**, 5, 909–16 (2016).

Ruiz, J., "Etymologia: Tem", Emerging Infectious Diseases **24**, 4, 709 (2018).

Rule, G. S. and T. K. Hitchens, *Fundamentals of protein NMR spectroscopy*, vol. 5 (Springer Science & Business Media, 2006).

Saavedra, H. G., J. O. Wrabl, J. A. Anderson, J. Li and V. J. Hilser, "Dynamic allostery can drive cold adaptation in enzymes", Nature **558**, 7709, 324–328 (2018).

Sablowski, R. and M. Carnier Dornelas, "Interplay between cell growth and cell cycle in plants", Journal of Experimental Botany **65**, 10, 2703–2714 (2014).

Salomon-Ferrer, R., A. W. Götz, D. Poole, S. Le Grand and R. C. Walker, "Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald", Journal of Chemical Theory and Computation **9**, 9, 3878–3888 (2013).

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

Sawle, L., J. Huihui and K. Ghosh, "All-atom simulations reveal protein charge decoration in the folded and unfolded ensemble is key in thermophilic adaptation", Journal of Chemical Theory and Computation **13**, 10, 5065–5075 (2017).

Schrödinger, E., *What is Life? The Physical Aspect of the Living Cell* (Cambridge University Press, 1944).

Schrödinger, LLC, "The PyMOL molecular graphics system, version 1.8", (2015).

Schuwirth, B. S., M. A. Borovinskaya, C. W. Hau, W. Zhang, A. Vila-Sanjurjo, J. M. Holton and J. H. D. Cate, "Structures of the Bacterial Ribosome at 3.5 Å Resolution", Science **310**, 5749, 827–834 (2005).

Sethi, A., P. O'Donoghue and Z. Luthey-Schulten, "Evolutionary profiles from the QR factorization of multiple sequence alignments", Proceedings of the National Academy of Sciences **102**, 11, 4045–4050 (2005).

Seymore, K., A. Mccallum and R. Rosenfeld, "Learning Hidden Markov Model Structure for Information Extraction", in "In AAAI 99 Workshop on Machine Learning for Information Extraction", pp. 37–42 (1999).

Sharrock, R., T. Leighton and H. Wittmann, "Macrolide and aminoglycoside antibiotic resistance mutations in the bacillus subtilis ribosome resulting in temperature-sensitive sporulation", MGG Mol. Gen. Genet **183**, 3, 538–543 (1981).

Sinha, N. and R. Nussinov, "Point mutations and sequence variability in proteins: Redistributions of preexisting populations", Proceedings of the National Academy of Sciences **98**, 6, 3139–3144 (2001).

Skoff, D. R. and M. T. Zanni, *Protein Dynamics Studied with 1D and 2D Infrared Spectroscopy*, pp. 2000–2007 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2013), URL `https://doi.org/10.1007/978-3-642-16712-6_163`.

Smith, S. D., S. Wang and M. D. Rausher, "Functional evolution of an anthocyanin pathway enzyme during a flower color transition", Molecular Biology and Evolution **30**, 3, 602–612 (2013).

Socolich, M., S. W. Lockless, W. P. Russ, H. Lee, K. H. Gardner and R. Ranganathan, "Evolutionary information for specifying a protein fold", Nature **437**, 7058, 512–518 (2005).

Soskine, M. and D. S. Tawfik, "Mutational effects and the evolution of new protein functions", Nature Reviews Genetics **11**, 8, 572–582 (2010).

Stewart, B. W., "Mechanisms of Apoptosis: Integration of Genetic, Biochemical, and Cellular Indicators", JNCI: Journal of the National Cancer Institute **86**, 17, 1286–1296 (1994).

Stiffler, M. A., D. R. Hekstra and R. Ranganathan, "Evolvability as a function of purifying selection in tem-1 β-lactamase", Cell **160**, 5, 882–892 (2015).

Still, W. C., A. Tempczyk, R. C. Hawley and T. Hendrickson, "Semianalytical treatment of solvation for molecular mechanics and dynamics", Journal of the American Chemical Society **112**, 16, 6127–6129 (1990).

Tafoya, S. and C. Bustamante, "Molecular switch-like regulation in motor proteins", Philosophical Transactions of the Royal Society B: Biological Sciences **373**, 1749, 20170181 (2018).

Tama, F. and Y.-H. Sanejouand, "Conformational change of proteins arising from normal mode calculations", Protein Engineering, Design and Selection **14**, 1, 1–6 (2001).

Tama, F., M. Valle, J. Frank and C. Iii, "Dynamic reorganization of the functionally active ribosome explored by normal mode analysis and cryo-electron microscopy", Pnas **100**, 16, 9319–9323 (2003).

Tan, L., S. Serene, H. X. Chao and J. Gore, "Hidden randomness between fitness landscapes limits reverse evolution", Physical Review Letters **106**, 19, 198102 (2011).

Tavakoli, M., J. N. Taylor, C.-B. Li, T. Komatsuzaki and S. Pressé, *Single Molecule Data Analysis: An Introduction*, chap. 4, pp. 205–305 (John Wiley & Sons, Ltd, 2017).

Taylor, J. L., J. E. Price and M. D. Toney, "Directed evolution of the substrate specificity of dialkylglycine decarboxylase", Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics **1854**, 2, 146–155 (2015).

Thorbjarnardóttir, S., R. Á. Magnúsdóttir, G. Eggertsson, S. A. Kagan and Ó. S. Andrésson, "Mutations determining generalized resistance to aminoglycoside antibiotics in Escherichia coli", Molecular and General Genetics MGG **161**, 1, 89–98 (1978).

Thorpe, M. F., "Comment on elastic network models and proteins", Physical Biology **4**, 1, 60–63; discussion 64–65 (2007).

Tirion, M., "Large amplitude elastic motions in proteins from a single-parameter, atomic analysis", Phys. Rev. Lett **77**, 9, 1905–1908 (1996).

Tobi, D. and I. Bahar, "Structural changes involved in protein binding correlate with intrinsic motions of proteins in the unbound state", Proceedings of the National Academy of Sciences of the United States of America **102**, 52, 18908–18913 (2005).

Tokuriki, N., F. Stricher, L. Serrano and D. S. Tawfik, "How protein stability and new functions trade off", PLOS Computational Biology **4**, 2, e1000002 (2008).

Tokuriki, N. and D. S. Tawfik, "Protein dynamism and evolvability", Science (New York, N.Y.) **324**, 5924, 203–207 (2009a).

Tokuriki, N. and D. S. Tawfik, "Stability effects of mutations and protein evolvability", Current Opinion in Structural Biology **19**, 5, 596–604 (2009b).

Townsend, P. D., T. L. Rodgers, L. C. Glover, H. J. Korhonen, S. A. Richards, L. J. Colwell, E. Pohl, M. R. Wilson, D. R. W. Hodgson, T. C. B. McLeish and M. J. Cann, "The Role of Protein-Ligand Contacts in Allosteric Regulation of the Escherichia coli Catabolite Activator Protein", Journal of Biological Chemistry **290**, 36, 22225–22235 (2015).

Trudeau, D. L., M. Kaltenbach and D. S. Tawfik, "On the potential origins of the high stability of reconstructed ancestral proteins", Molecular Biology and Evolution **33**, 10, 2633–2641 (2016).

Truhlar, D. G., W. L. Hase and J. T. Hynes, "Current status of transition-state theory", The Journal of Physical Chemistry **87**, 15, 2664–2682 (1983).

Tsai, C.-J. and R. Nussinov, "A unified view of "how allostery works"", PLOS Computational Biology **10**, 2, 1–12 (2014).

Turkcan, S. and J.-B. Masson, "Bayesian decision tree for the classification of the mode of motion in single-molecule trajectories", Plos One **8**, 12, 1–14 (2013).

Tzul, F. O., D. Vasilchuk and G. I. Makhatadze, "Evidence for the principle of minimal frustration in the evolution of protein folding landscapes", Proceedings of the National Academy of Sciences **114**, 9, E1627–e1632 (2017).

Ullmann, A., "In Memoriam: Jacques Monod (1910–1976)", Genome Biology and Evolution **3**, 1025–1033 (2011).

Urbanc, B., "Protein Actions: Principles and Modeling by I. Bahar, R.L. Jernigan, and K.A. Dill", Journal of Biological Physics **43**, 4, 585–589 (2017).

Vakulenko, S. B., B. Geryk, L. P. Kotra, S. Mobashery and S. A. Lerner, "Selection and characterization of beta-lactam-beta-lactamase inactivator-resistant mutants following pcr mutagenesis of the tem-1 beta-lactamase gene", Antimicrobial Agents and Chemotherapy **42**, 7, 1542–1548 (1998).

Vallurupalli, P. and L. E. Kay, "Complementarity of ensemble and single-molecule measures of protein motion: a relaxation dispersion nmr study of an enzyme complex", Proceedings of the National Academy of Sciences **103**, 32, 11910–11915 (2006).

Van Kampen, N., "Chapter ix - the langevin approach", in "Stochastic Processes in Physics and Chemistry (Third Edition)", edited by N. Van Kampen, North-Holland Personal Library, pp. 219–243 (Elsevier, Amsterdam, 2007), third edition edn.

Vazquez, D. S., I. E. Sánchez, A. Garrote, M. P. Sica and J. Santos, "The e. coli thioredoxin folding mechanism: the key role of the c-terminal helix.", Biochimica et biophysica acta **1854**, 2, 127–137 (2015).

Vella, F., "The cell. A molecular approach; Edited by G H Cooper. pp 673. ASM Press, Washington DC, Sinauer Associates, Sunderland, MA. 1997 ISBN 0-87893-119-8", Biochemical Education **26**, 1, 98–99 (1998).

Virtanen, P., R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S. J. van der Walt, M. Brett, J. Wilson, K. J. Millman, N. Mayorov, A. R. J. Nelson, E. Jones, R. Kern, E. Larson, C. J. Carey, İ. Polat, Y. Feng, E. W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E. A. Quintero, C. R. Harris, A. M. Archibald, A. H. Ribeiro, F. Pedregosa, P. van Mulbregt and SciPy 1.0 Contributors, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python", Nature Methods **17**, 261–272 (2020).

Voliotis, M., R. M. Perrett, C. McWilliams, C. A. McArdle and C. G. Bowsher, "Information transfer by leaky, heterogeneous, protein kinase signaling systems", Proceedings of the National Academy of Sciences **111**, 3, E326–e333 (2014).

Vranken, W. F., W. Boucher, T. J. Stevens, R. H. Fogh, A. Pajon, M. Llinas, E. L. Ulrich, J. L. Markley, J. Ionides and E. D. Laue, "The ccpn data model for nmr spectroscopy: Development of a software pipeline", Proteins: Structure, Function, and Bioinformatics **59**, 4, 687–696 (2005).

Wagner, I. and H. Musso, "New Naturally Occurring Amino Acids", Angewandte Chemie International Edition in English **22**, 11, 816–828 (1983).

Wang, S.-W., A.-F. Bitbol and N. S. Wingreen, "Revealing evolutionary constraints on proteins through sequence analysis", PLOS Computational Biology **15**, 4, e1007010 (2019).

Wang, Y., A. Rader, I. Bahar and R. Jernigan, "Global ribosome motions revealed with elastic network model", J. Struct. Biol **147**, 3, 302–314 (2004).

Wang, Z., D. Antoniou, S. D. Schwartz and V. L. Schramm, "Hydride transfer in dhfr by transition path sampling, kinetic isotope effects, and heavy enzyme studies", Biochemistry **55**, 1, 157–166 (2016).

Warias, M., H. Grubmüller and L. V. Bock, "trna dissociation from ef-tu after gtp hydrolysis: Primary steps and antibiotic inhibition", Biophysical Journal **118**, 1, 151–161 (2020).

Weichsel, A., J. R. Gasdaska, G. Powis and W. R. Montfort, "Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer", Structure **4**, 6, 735–751 (1996).

Weiner, S. J., P. A. Kollman, D. A. Case, U. C. Singh, C. Ghio, G. Alagona, S. Profeta and P. Weiner, "A new force field for molecular mechanical simulation of nucleic acids and proteins", Journal of the American Chemical Society **106**, 3, 765–784 (1984).

Weinreich, D. M., N. F. Delaney, M. A. Depristo and D. L. Hartl, "Darwinian evolution can follow only very few mutational paths to fitter proteins", Science (New York, N.Y.) **312**, 5770, 111–114 (2006).

Weinreich, D. M., Y. Lan, J. Jaffe and R. B. Heckendorn, "The influence of higher-order epistasis on biological fitness landscape topography", Journal of Statistical Physics **172**, 1, 208–225 (2018).

Weinreich, D. M., Y. Lan, C. S. Wylie and R. B. Heckendorn, "Should evolutionary geneticists worry about higher-order epistasis?", Current Opinion in Genetics & Development **23**, 6, 700–707 (2013).

Weiss, O., M. A. Jiménez-monta no and H. Herzel, "Information Content of Protein Sequences", Journal of Theoretical Biology **206**, 3, 379–386 (2000).

Wigner, E., "Calculation of the rate of elementary association reactions", The Journal of Chemical Physics **5**, 9, 720–725 (1937).

Wilding, M., N. Hong, M. Spence, A. M. Buckle and C. J. Jackson, "Protein engineering: the potential of remote mutations", Biochemical Society Transactions p. Bst20180614 (2019).

Wilson, C., R. V. Agafonov, M. Hoemberger, S. Kutter, A. Zorba, J. Halpin, V. Buosi, R. Otten, D. Waterman, D. L. Theobald and D. Kern, "Kinase dynamics. using ancient protein kinases to unravel a modern cancer drug's mechanism", Science (New York, N.Y.) **347**, 6224, 882–886 (2015).

Wodak, S. J., E. Paci, N. V. Dokholyan, I. N. Berezovsky, A. Horovitz, J. Li, V. J. Hilser, I. Bahar, J. Karanicolas, G. Stock, P. Hamm, R. H. Stote, J. Eberhardt, Y. Chebaro, A. Dejaegere, M. Cecchini, J.-P. Changeux, P. G. Bolhuis, J. Vreede, P. Faccioli, S. Orioli, R. Ravasio, L. Yan, C. Brito, M. Wyart, P. Gkeka, I. Rivalta, G. Palermo, J. A. McCammon, J. Panecka-Hofman, R. C. Wade, A. D. Pizio, M. Y. Niv, R. Nussinov, C.-J. Tsai, H. Jang, D. Padhorny, D. Kozakov and T. McLeish, "Allostery in its many disguises: From theory to applications", Structure **27**, 4, 566–578 (2019a).

Wodak, S. J., E. Paci, N. V. Dokholyan, I. N. Berezovsky, A. Horovitz, J. Li, V. J. Hilser, I. Bahar, J. Karanicolas, G. Stock, P. Hamm, R. H. Stote, J. Eberhardt, Y. Chebaro, A. Dejaegere, M. Cecchini, J.-P. Changeux, P. G. Bolhuis, J. Vreede, P. Faccioli, S. Orioli, R. Ravasio, L. Yan, C. Brito, M. Wyart, P. Gkeka, I. Rivalta, G. Palermo, J. A. McCammon, J. Panecka-Hofman, R. C. Wade, A. Di Pizio, M. Y. Niv, R. Nussinov, C.-J. Tsai, H. Jang, D. Padhorny, D. Kozakov and T. McLeish, "Allostery in its many disguises: From theory to applications", Structure **27**, 4, 566–578 (2019b).

Wriggers, W. and K. Schulten, "Protein domain movements: detection of rigid domains and visualization of hinges in comparisons of atomic coordinates", Proteins: Structure, Function, and Bioinformatics **29**, 1, 1–14 (1997).

Wright, P. and H. Dyson, "Intrinsically disordered proteins in cellular signalling and regulation", Nature Reviews Molecular Cell Biology **16**, 18–29 (2014).

Wu, Z., S. B. J. Kan, R. D. Lewis, B. J. Wittmann and F. H. Arnold, "Machine learning-assisted directed protein evolution with combinatorial libraries", Proceedings of the National Academy of Sciences **116**, 18, 8852–8858 (2019).

Zeeb, M. and J. Balbach, "Protein folding studied by real-time nmr spectroscopy", Methods **34**, 1, 65–74 (2004).

Zhang, S., H. Li, J. M. Krieger and I. Bahar, "Shared Signature Dynamics Tempered by Local Fluctuations Enables Fold Adaptability and Specificity", Molecular Biology and Evolution **36**, 9, 2053–2068 (2019).

Zhang, W., D. F. A. R. Dourado, P. A. Fernandes, M. J. a. Ramos and B. Mannervik, "Multidimensional epistasis and fitness landscapes in enzyme evolution", The Biochemical Journal **445**, 1, 39–46 (2012).

Zhao, S., W. Xu, W. Jiang, W. Yu, Y. Lin, T. Zhang, J. Yao, L. Zhou, Y. Zeng, H. Li, Y. Li, J. Shi, W. An, S. M. Hancock, F. He, L. Qin, J. Chin, P. Yang, X. Chen, Q. Lei, Y. Xiong and K.-L. Guan, "Regulation of Cellular Metabolism by Protein Lysine Acetylation", Science **327**, 5968, 1000–1004 (2010).

Zheng, W., B. R. Brooks and D. Thirumalai, "Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations", Proceedings of the National Academy of Sciences of the United States of America **103**, 20, 7664–7669 (2006).

Zimmerman, M. I., K. M. Hart, C. A. Sibbald, T. E. Frederick, J. R. Jimah, C. R. Knoverek, N. H. Tolia and G. R. Bowman, "Prediction of new stabilizing mutations based on mechanistic insights from markov state models", ACS Central Science **3**, 12, 1311–1321 (2017).

Zimmerman, M. T., K. Jia and R. L. Jernigan, "Ribosome Mechanics Informs about Mechanism", Journal of molecular biology **428**, 5 Pt A, 802–810 (2016).

Zou, T., V. A. Risso, J. A. Gavira, J. M. Sanchez-Ruiz and S. B. Ozkan, "Evolution of conformational dynamics determines the conversion of a promiscuous generalist into a specialist enzyme", Molecular Biology and Evolution **32**, 1, 132–143 (2015).

APPENDIX A

METHODS USED TO EXPERIMENTALLY CHARACTERIZE ENGINEERED GNCA-XYZ

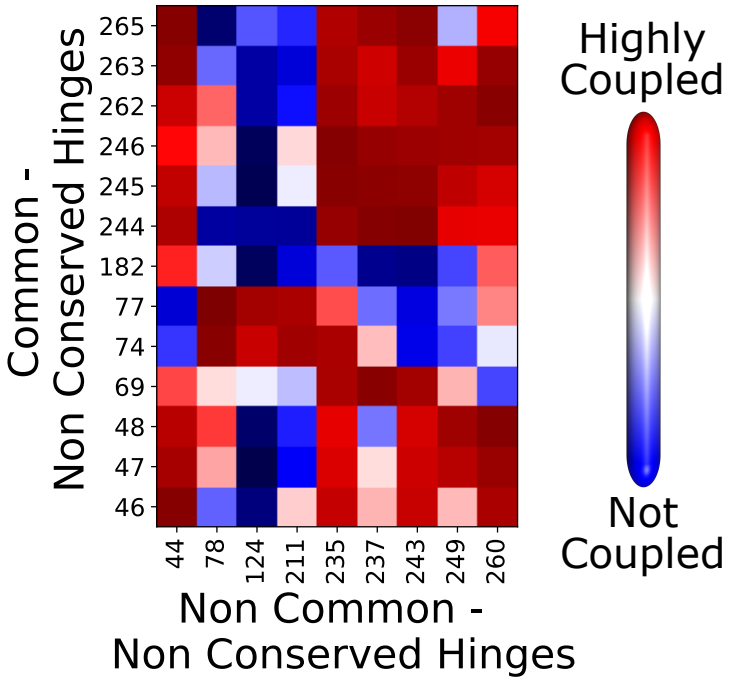

Figure A.1: The coupling of all common and sequentially non-conserved (NC) hinges with other non-common and sequentially non-conserved (NN) hinges in GNCA β-lactamase. Common and sequentially non-conserved hinges which exhibit a higher coupling (%DCI>0.8) with other non-common and sequentially non-conserved hinges are selected for substitutions in set Y. For this analysis, we selected such residues from each region in the protein. Therefore, out of residues 262, 263 and 265, only 262 and 263 are selected due to their high coupling with maximum number of NN hinges; out of 244, 245 and 246, only 244 and 246 are selected, and similarly for others. However, 182 is the only NC hinge in its vicinity which showed higher coupling to few with NN hinges (44 and 260), as a result we included in the set Y. It should be noted that the residue 182 was found to be of importance in other studies focusing on emergence of new function in β-lactamases (Salverda *et al.*, 2010; Weinreich *et al.*, 2006). We also compared the impact of mutations from set Y excluding the residue 182 as shown in Figure A.2

A.1 Experimental Characterization of the Mutants

Proteins studied in this work were prepared as we have previously described (Risso *et al.*, 2017). Briefly, genes cloned into a pET24 vector with resistance to kanamycin were transformed into E. coli BL21(DE3) cells. Proteins were purified by NTA affinity chromatography, taking advantage of the presence of a His-tag at the C-terminal. Protein solutions were prepared by exhaustive dialysis against 50 mM Hepes buffer 50 mM and protein concentrations were determined spectrophotometrically using a known value of the extinction coefficient at 280 nm.

Catalytic parameters for the hydrolysis of lactam antibiotics were determined at 25 ^{o}C, as we have previously described (Risso *et al.*, 2013). Briefly, initial rates were

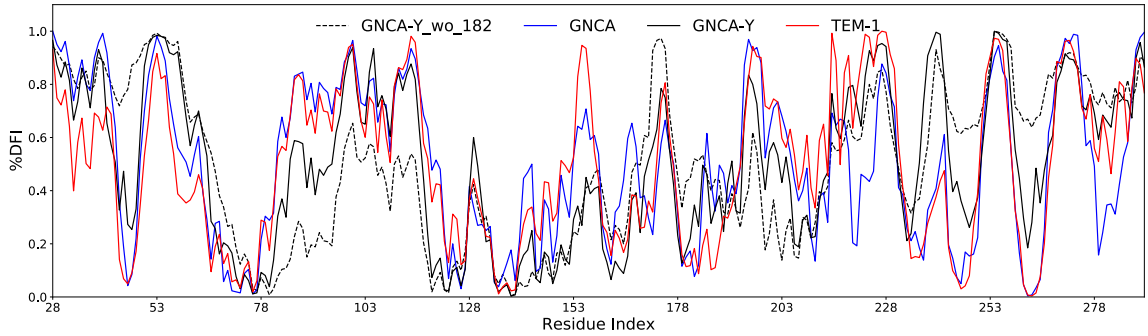

Figure A.2: The effect of mutations from set Y with (GNCA-Y) and without the mutation T182M (GNCAY_wo_182) is observed by comparing their DFI profiles with the wild types GNCA and TEM-1 β-lactamases. We observe that without the mutation T182M, the mutant (broken black line) exhibits very different dynamics particularly in several regions 78-110, around residue 200 and 240-260. On the other hand, the impact of the mutation set Y on dynamics in these regions is not so severe when we incorporate the mutation (black solid line).

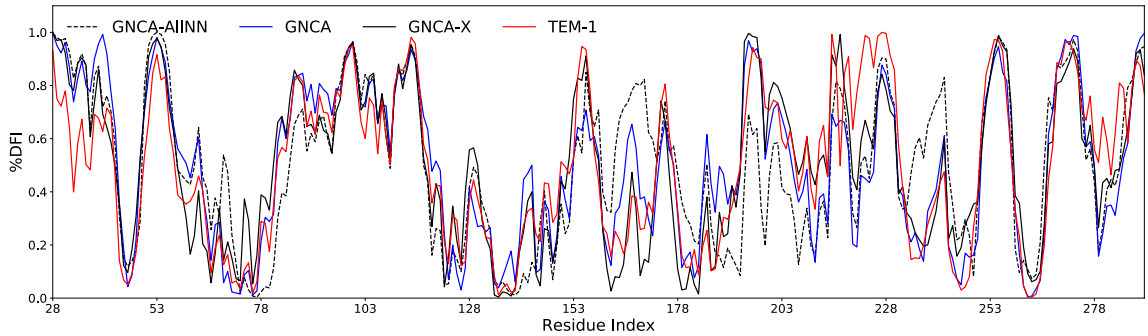

Figure A.3: The comparison of the effect of mutations from set X and the mutant created by performing mutations at all the non-common and non-conserved hinges in GNCA and TEM-1 β-lactamase (GNCA-AllNN) by comparing their DFI profiles with the wild types GNCA and TEM-1 β-lactamases. We observe that by performing mutations at all the non-common and non-conserved hinges (broken black line) has a drastic impact on the dynamics particularly in regions around the catalytic site 70, 166 and 234. On the other hand, the impact of the mutation from set X, which contains only a select number of non-common and non-common hinges (which exhibit high coupling with other non-common and non-conserved hinges), on dynamics in these regions is not so severe (black solid line). This indicates the critical role played by the coupling among these hinge residues which compensate the deleterious effect of mutations on the catalytic pocket.

determined from the change in UV absorbance that accompanies substrate hydrolysis and values of the Michaelis constant, turnover number and catalytic efficiency were calculated by fitting the Michaelis–Menten equation to the profiles of rate versus substrate concentration. In some cases, linear plots were observed, indicating a very large value of the Michaelis constant. In those cases, only the value of the catalytic

efficiency was calculated from the experimental profiles.

Denaturation temperatures were determined from differential scanning calorimetry experiments, as we have previously described (Risso *et al.*, 2013). Briefly, protein solutions were exhaustively dialyzed and the buffer from the last dialysis step was used as the reference solution in the calorimetric experiments. Equilibration of the instrument was ensured by recording several baselines prior to the experiment with the protein sample. Denaturation temperature values were determined as the temperature corresponding to the maximum of the heat capacity profiles.

A.2 Crystallization, Data Collection, and Structure Determination

Crystallization of GNCA-XYZ was done by the counter diffusion technique using the 24 conditions minimum crystallization screening kit together with the mixPEG at pH form 4.0 to 9.077 (González-Ramírez *et al.*, 2017). Protein solution, at 30 mg/ml in 25 mM Hepes pH 7.0, was loaded in capillaries of 0.2 mm inner diameter, sealed at the top of the capillary and confronted to the precipitant solutions. After several days first crystals appeared at the bottom of the capillary, but crystals were let set until data collection. Crystals were extracted from the capillary and cryo-protected by the equilibration with 15% (v/v) glycerol or 20% PEG 200 prepared in the mother liquid prior to flash-frozen in liquid nitrogen for transportation and data collection. Crystals grown in several conditions, but the best diffracting crystals were obtained in the mix of PEG at pH 9.0.

Crystals were diffracted at the beamline ID30B of the European Synchrotron Radiation Facility (ESRF, France). Data were indexed and integrated with XDS (Kabsch, 2010) and scaled with SCALA (Evans, 2006) of the CCP4 program suite (Number 4, 1994). Molecular replacement was done using as the search model the coordinates of GNCA, PDB ID. 4B88 in Phaser (McCoy *et al.*, 2007). Refinement was initiated with phenix.refine (Afonine *et al.*, 2010) of the PHENIX suite (Adams *et al.*, 2010) followed by manual building, water inspection and ligands identification in Coot (Emsley *et al.*, 2010). The final refinement was assessed, including Titration-Libration-Screw (TLS) parameterization. The model was verified with Molprobity (Chen *et al.*, 2010) prior deposit at the PDB (ID. 6YRS). Additional details provided in the Table A.1.

A.3 Nuclear Magnetic Resonance Spectroscopy

The GNCA and GNCA-XYZ genes were incorporated into a pET_24B (+) vector and transformed into BL21 (DE3) E. coli cells. Starter cultures were prepared with one colony in 5 mL LB with 38 μg/ml Kanamycin and incubated overnight at 37 ^{o}C with shaking. The starter culture was used to inoculated 1 L of minimal M9 media (12.8 g $Na_2HPO_4.7H_2O$, 3.0 g KH_2PO_4, 0.5 g NaCl, 1g $^{15}NH_4Cl$, 20 mL of 20% w/v D-Glucose, 10 mL 100$\times$ MEM vitamin solution, 1 mM $MgSO_4$, 100 μM $CaCl_2$). Protein expression was induced at 0.6 OD_{600} nm with 400 μM IPTG at 37 ^{o}C for 3 hrs. The resulting cells were harvested at 6000 $\times$g for 15 min 4 ^{o}C.

The cell pellet containing over expressed GNCA-XYZ were re-suspended in 20 mL of lysis buffer (20 mM Na_2HPO_4, 500 mM NaCl, pH 7.4) per 1 L of cell pellet, 1 mM phenylmethanesulfonyl fluride (PMSF), 5 mM magnesium acetate, 23 μg/ml lysozyme, 2.3 μg/ml DNase, and 2.3 μg/ml RNase. The sample was tumbled at room temperature for 20 min followed by sonication on ice with S-4000 Ultrasonic Processor

Table A.1: Data collection and refinement statistics. Statistics for the highest-resolution shell are shown in parentheses.

Source	**ESRF ID30B**
PDB ID	6YRS
Resolution range ($\AA$)	48.73 - 1.7 (1.76 - 1.7)
Space group	P 1 21 1
a, b, c ($\AA$) γ(°)	47.35, 81.40, 61.04 94.58
No. unique reflections	49578 (4871)
Multiplicity	3.0 (2.9)
Completeness (%)	97.68 (96.72)
Mean I/sigma(I)	12.03 (1.30)
Wilson B factor ($\AA^2$)	29.69
R-merge (%)	4.46 (77.17)
CC1/2 (%)	99.8 (58.8)
Reflections in working / test sets	49571 / 2531
R-work (%)	17.60 (30.96)
R-free (%)	20.85 (33.13)
Atoms (non-H)	4441
macromolecules	4049
ligands	136
solvent	256
Protein residues	504
RMS (bonds) ($\AA$)	0.012
RMS (angles) (°)	1.11
Ramachandran (%) ligands ligands	 98.19 1.81
Average B factor ($\AA^2$)	41.92
biomolecules	40.78
ligands	66.09
solvent	47.08

(Qsonica) at a 3 s on and 5 s off pulse cycle and 65% power. The resulting lysate was centrifuged at 38,500 ×g for 20 min at 4 °C. The supernatant was collected and was 0.45 μm filtered and loaded on to 5 ml QIAGEN Ni-NTA Superflow column at 1 mL/min rate. The column was equilibrated with 5 column volumes (CV) of binding buffer (20 mM Na_2HPO_4, 500 mM NaCl, 20 mM Imidazole, pH 7.4) and washed with 2.5 CV of 8% (58.4 mM imidazole) of elution buffer (20 mM Na_2HPO_4, 500 mM NaCl, 500 mM Imidazole, pH 7.4). GNCA-XYZ was eluted by a linear gradient of the binding and elution buffers over 9 CVs and ranging from 8-70% elution buffer concentration. Fractions were identified by SDS-PAGE and buffer exchanged into NMR buffer (25 mM Na_2HPO_4, 250 mM NaCl, pH 6.7) and concentrated to 0.5 mL using 10 kDa cutoff (Millipore Amicon Ultra-4 10K) for gel filtration chromatography (16XK column with Superdex 200 Prep Grade Resin, GE Healthcare Life Science).

SDS-PAGE was used to analyze fractions with high A280 readings, and selected fractions were combined and concentrated.

The NMR sample was prepared in a 3 mm NMR tube with 2.7% D_2O, 0.5 mM EDTA, and 550 μM protein concentration in final volume of 180 μL. All ^{1}H-^{15}N heteronuclear single quantum coherence (HSQC) experiments were recorded on a Bruker Avance III HD 850 MHz spectrometer at 303.15 K and equipped with a cryogenically cooled probe. Data were processed in NMRPipe86 and anazlyed with CcpNMR Analysis software (Vranken *et al.*, 2005).

APPENDIX B

MATERIALS AND METHODS USED TO CHARACTERIZE E. COLI WILD TYPE AND L22* RIBOSOME STRAINS

B.1 Bacterial Strains

The *Bacillus subtilis* strains used in this study are listed in Table B.1. SacA:: PmgtE-yfp strain is generated using the ECE174-based integration vector (ECE174-PmgtE-yfp) to insert PmgtE-yfp into the sacA locus. It was constructed by Gibson assembly of two PCR products. ECE174-yfp (CmR) vector was PCR amplified using the primers GS450 and GS351. The promoter region of mgtE was PCR amplified using the primers GS2462 and GS2463. The overlapping ends that facilitate Gibson assembly are indicated with lower case letters. The M3 mutation in the promoter (PmgtEM3) was created using the primers GS2465 and GS2472 (mutation site indicated with lower case letters). The DNA sequences of the primers are listed in Table B.2. The transformed strains were confirmed by sequencing.

Table B.1: Organisms and strains used in the study.

Name	**Source**	**BGSC ID**
Bacillus subtilis NCIB 3610 (WT)	Wade Winkler, University of Maryland	3A1
Bacillus subtilis NCIB 3610 *rplV94* (L22*)	Lee *et al.* (2019)	N/A
Bacillus subtilis NCIB 3610 *sacA::*P_{mgtE}-*yfp*	This Study	N/A
Bacillus subtilis NCIB 3610 *rplV94, sacA::*P_{mgtE}-*yfp*	This Study	N/A
Bacillus subtilis NCIB 3610 *sacA::*$P_{mgtE}{}^{M3}$-*yfp*	This Study	N/A

Table B.2: Primers used to generate vectors for reporter strains.

Name	**Sequence**	**Source**
GS450	ATGAGCAAAGGTGAAGAACTGTTCACC	This Study
GS351	GTCGCTACCATTACCAGTTGGTCTGG	This Study
GS2462	ccaactggtaatggtagcgacGTTTTGTTCCGTAATTGTGATGTAAGCGC	This Study
GS2463	aacagttcttcacctttgctCATCGGGACTCGTACCTCCTCTACG	This Study
GS2465	ATCGACATAAccaGATTTTTAATGCAGCTGG	This Study
GS2472	GCATTAAAAATCtggTTATGTCGATGATTTCTGTTGACCCATTGGCGTCT	This Study

B.2 Growth Conditions

For YFP fluorescence and growth rate measurements, desired *B. subtilis* strains were streaked on a fresh LB agar plate (with 5 mg/L chloramphenicol or 5 mg/L

erythromycin when appropriate) a day before the experiment and incubated at 37 $_{o}$C overnight. A few colonies of the desired strain were used for inoculation in 2 mL MSgg media [5 mM potassium phosphate (pH 7.0), 100 mM 3-(N-morpholino) propanesulfonic acid (pH 7.0), 700 μM $CaCl_2$, 50 μM $MnCl_2$, 100 μM $FeCl_3$, 1 μM $ZnCl_2$, 2 μM thiamine, 0.5% glycerol, 0.5% glutamate] with indicated final $MgCl_2$ concentration and grown at 37°C with shaking for 4.5 hours. Media were made from stock solutions immediately before experiments, and the stock solution of glutamate were made fresh every two days. The OD_{600} was adjusted to 0.02 with the same growth media, and the culture was applied onto a 1.5% carrageenan pad made with MSgg medium with desired final $MgCl_2$ concentration. The pads were covered, left to air dry for 1 hour at 30 °C, and placed into a coverslip-bottom Willco dish for imaging.

For luciferase driven Mg-ATP measurements and Spectinomycin MIC test, a few colonies of the desired strain were inoculated into 2 mL MSgg medium with indicated final $MgCl_2$ concentration and grown at 37 °C with shaking for 6 hours prior to test.

B.3 Time-Lapse Microscopy

Elongation rate and PmgtE-yfp signal of *B. subtilis* cells were monitored with time-lapse fluorescence microscopy. We recorded phase-contrast and YFP fluorescence images at 30 °C using Olympus confocal laser scanning microscope FV3000 with a motorized stage (ASI). Single layers of cells were imaged every 20 min under 60× objective lens with 1.5× zoom. Collected images were processed with ImageJ (National Institutes of Health, http://imagej.nih.gov/ij/).

B.4 Free Magnesium Measurements

Trainable Weka segmentation plugin from FIJI (ImageJ, (Arganda-Carreras *et al.*, 2017)) was used to segment cells from each phase-contrast image. Multilayer of cells, spores, and background noise were eliminated through size filtering and screening. A mask created for each image was obtained as ROIs and applied to the corresponding YFP images to extract fluorescence intensity for each segmented cell. The fluorescence intensity of cells was averaged per image. Then, the mean fluorescence intensity under a certain condition was calculated as the average of all images for the condition. Fold change relative to M3 control was calculated by dividing by the mean fluorescence intensity of M3 control at a given extracellular Mg^{2+} concentration.

B.5 Elongation Rate Calculation

Single cells were tracked for 3 hours 20 min by custom software developed in MATLAB (MathWorks) using time-lapse phase images. For each lineage's time trace of cell length, an exponential line was fitted between consecutive divisions (for each generation). A function $f(x) = ae^{bx}$ was used for the fitting, where b is the growth rate of a cell. The elongation rate of cells was averaged per image. Then, the mean elongation rate of a certain condition was calculated as the average of all images for the condition. Statistical analysis was performed using GraphPad Prism.

B.6 Mg-ATP Measurements

Steps described in ATP Bioluminescence Assay Kit HS II (Roche Scientific Cat #: 11-699-709-001) protocol were followed for the sample and standard preparation. Specifically, 5 × 105 cells were taken in a 50 μl volume to 96 well plate. 50 μl of lysis buffer was added to each sample for room temperature incubation for 5 minutes. 100μl of luciferase reagent was added to each sample, immediately followed by luminescence measurements using Spark Multimode Microplate Reader (Tecan). Standard curve was generated every experiment concurrently with samples. Luminescence intensity of samples were converted to concentration values using the standard curve. Mg-ATP concentration per cell was calculated with the assumption that OD_{600} of 1.0 is 5 × 109 cells/ml, and that volume of a cell is 0.9 fL.

B.7 MIC Tests

A total of 5 × 108 cells was spread on MSgg-agar plate supplemented with indicated final $MgCl_2$ concentration. A single test strip of 0.064-1024 mg/L Spectinomycin (Liofilchem s.r.l. Cat #:920141) was placed at the center of each plates and were incubated at 37 °C overnight. Statistical analysis was performed using GraphPad Prism.

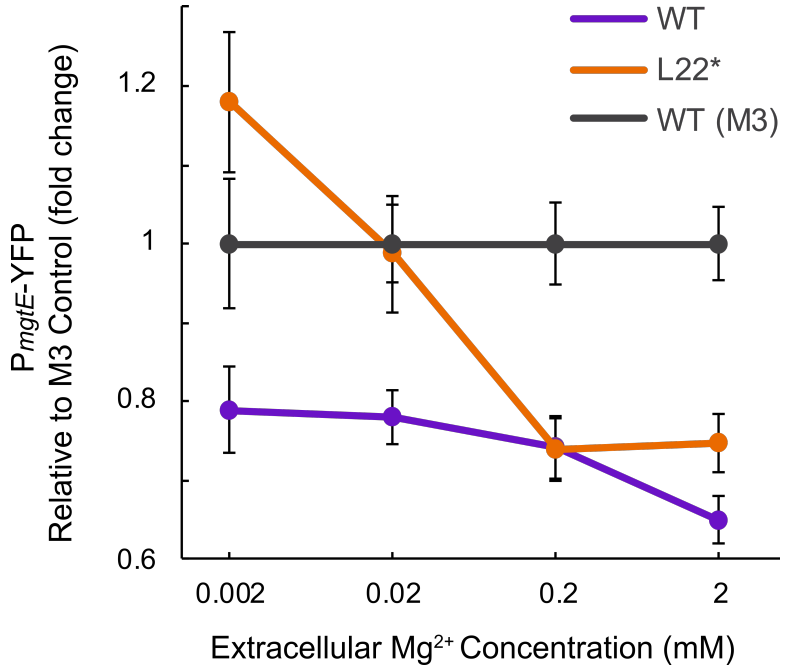

Figure B.1: Inversed free magnesium concentration in a cell, related to Figure. 6.9. *PmgtE-yfp* signals in WT and L22* strains relative to $PmgtE^{M3}$-*yfp* control at various extracellular magnesium concentrations.

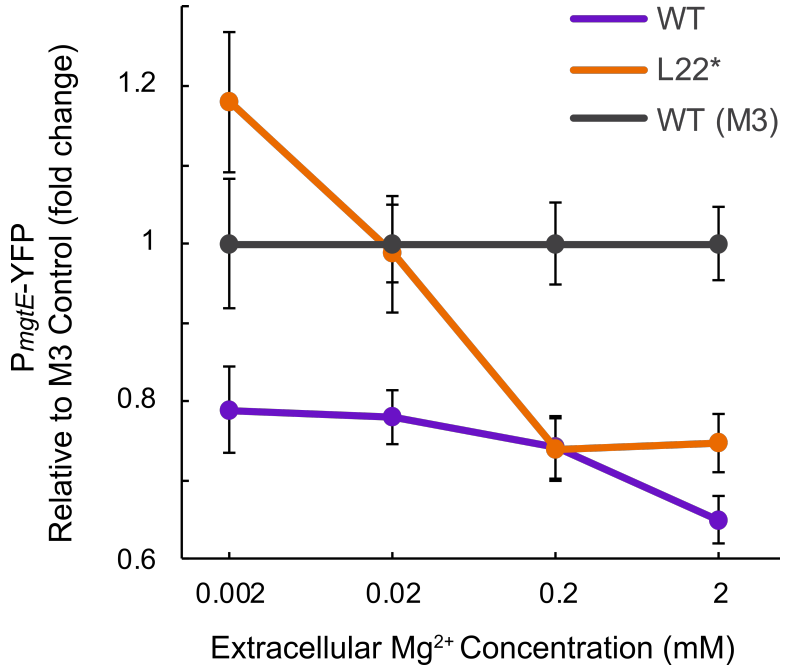

Figure B.2: A standard curve for Mg-ATP measurements, related to Figure. 6.8. Luminescence values of known Mg-ATP concentration generate a standard curve, which then fit with data.

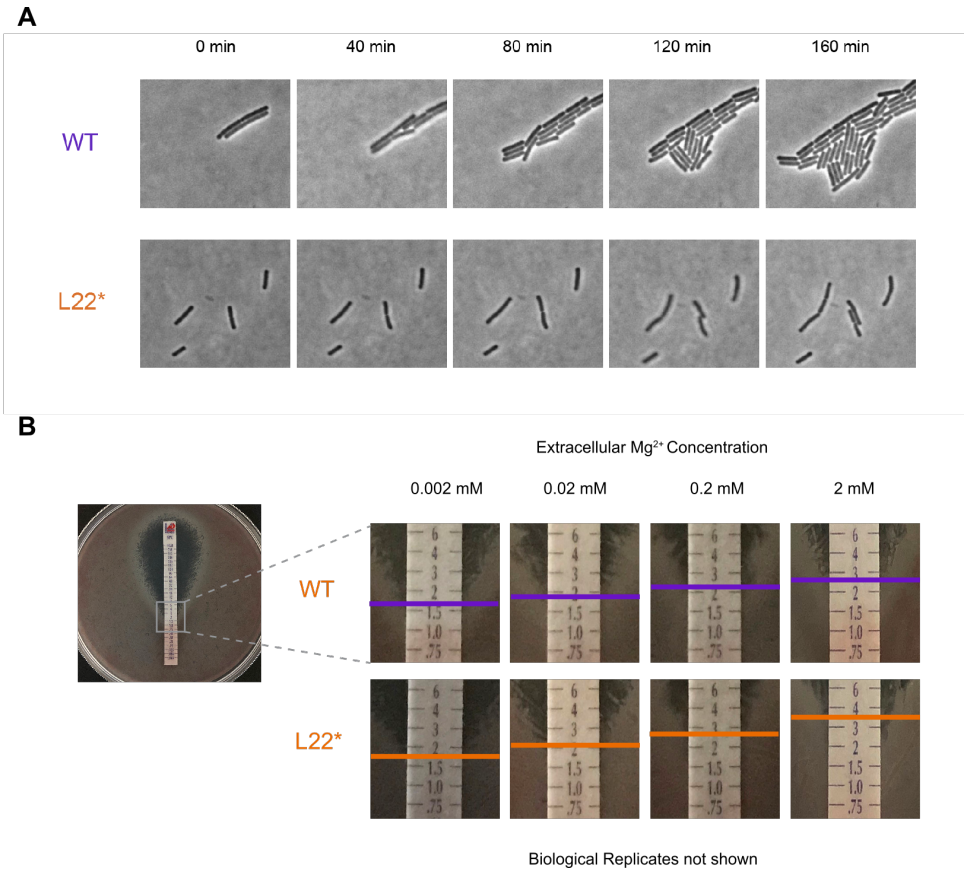

Figure B.3: Elongation rate and Spec MIC, related to Figure. .6.10 (A) Representative single-cell time trace images of WT and L22* cells for elongation rate measurements. Only 0.002 mM extracellular magnesium condition shown for simplicity. (B) Representative Spec MIC test pictures of WT and L22* strains at various extracellular magnesium concentrations.

APPENDIX C

SUB-ROUTINES USED IN THESIS

C.1 Sub-Routine to Diagonalize a Large Matrix.

Given below is a sub-routine in Python3.7 used to diagonalize the hessian generated from the elastic network model of the ribosome. For diagonalization, I used the sub-routines from ARPACK (Lehoucq *et al.*, 1998) implemented in the library SciPy (Virtanen *et al.*, 2020). Apart from that, I also used the library NumPy (Harris *et al.*, 2020) for optimizing array operations. It takes the hessian generated using coarse grained model of the ribosome as an input.

```
import numpy as np
import scipy.sparse.linalg as spl
def invert_hessian_shiftinvert(hess,tol_conv,no_modes=30):
    ei,evec=spl.eigsh(hess,k=no_modes,which='LM',tol=tol_conv,
                      sigma=0)
    print("Inverting the Hessian using ARPACK")
    tol = 1e-6
    singular = ei < tol
    invw = 1/ei
    invw[singular] = 0.
    invHrs = np.dot(evec,np.dot(np.diag(invw),evec.transpose()
                    ))
return invHrs
```

C.2 Using Gillespie Algorithm to Run a Stochastic Simulation of the Model of Allosteric Regulation in an Enzyme.

Given below is a sub-routine in Python3.7 used to implement the Gillespie algorithm to run a stochastic simulation of allosteric regulation in an enzyme. Here, I have used the library NumPy (Harris *et al.*, 2020) for optimizing array operations.

```
# Number of enzymes
N=1
# Number of substrate units present
M=100
# This function simulates stick breaking
def reac_prop(kp,km,dp,tp,tm,rp,hp,sp,Kp,Km,S,Enz):
    a=[]
    i=0 ### Over enzymes\
    while i<N:
        a1=float(S[-1]*Enz[i][0])*kp
        a2=float(Enz[i][2])*km
        a3=float(Enz[i][2])*dp
        a4=float(S[-1]*Enz[i][4])*Kp
        # In Kp and Km I am using the same variable as j
        # here as n==L
        a5=float(Enz[i][5])*Km
        a6=float(Enz[i][1]*S[-1])*tp
        a7=float(Enz[i][3])*tm
```

```
        a8=float(Enz[i][3])*rp
        a9=float(Enz[i][1])*sp
        a10=float(Enz[i][5]*Enz[i][0])*hp
        a=np.append(a,[a1,a2,a3,a4,a5,a6,a7,a8,a9,a10])
        i=i+1
    return a

# This function updates the concentrations
def reaction_update(Enz,P,P_id,S,mu):
    #Enzyme reacting
    i=int(mu/((10)))
    #Type of reaction
    u=int(mu%(10))
    if u==0:
        S.append(S[-1]-1)
        Enz[i][0]=0
        Enz[i][2]=1
        P.append(P[-1])
        P_id.append(0)

    elif u==1:
        S.append(S[-1]+1)
        Enz[i][0]=1
        Enz[i][2]=0
        P.append(P[-1])
        P_id.append(0)

    elif u==2:
        S.append(S[-1]+1)
        Enz[i][2]=0
        Enz[i][0]=1
        P.append(P[-1]+1)
        P_id.append(1)

    elif u==3:
        S.append(S[-1]-1)
        Enz[i][4]=0
        Enz[i][5]=1
        P.append(P[-1])
        P_id.append(0)

    elif u==4:
        S.append(S[-1]+1)
        Enz[i][4]=1
        Enz[i][5]=0
        P.append(P[-1])
        P_id.append(0)
```

```
    elif u==5:
        Enz[i][1]=0
        P.append(P[-1])
        S.append(S[-1]-1)
        #maintaining equal numbers of S
        Enz[i][3]=1
        P_id.append(0)

    elif u==6:
        S.append(S[-1]+1)
        Enz[i][3]=0
        Enz[i][1]=1
        P.append(P[-1])
        P_id.append(0)

    elif u==7:
        S.append(S[-1]+1)
        Enz[i][3]=0
        Enz[i][1]=1
        P.append(P[-1]+1)
        P_id.append(1)

    elif u==8:
        S.append(S[-1])
        Enz[i][1]=0
        Enz[i][0]=1
        P.append(P[-1])
        P_id.append(0)

    elif u==9:
        S.append(S[-1])
        Enz[i][0]=0
        Enz[i][1]=1
        P.append(P[-1])
        P_id.append(0)
return

# Function to run the simulation
# Returns the product formation event series in time
def run_sim(T,kp,km,dp,tp,tm,rp,sp,Kp,Km,hp):
    print("hp = %f"%hp)
    print("Kp = %f"%Kp)
    Enz=[]
    S=[M]
    i=0  # These are enzymes
    while i<N:
```

```
        #Initialization of active sites
        Enz.append([1,0,0,0,1,0])
        i=i+1
    P=[0]
    P_id=[0]
    time=[0]
    while time[-1]<=T:
        if ((hp==0) and (Enz[-1][1] ==1)):
            print("SHOUT")
        check=0
        a=reac_prop(kp,km,dp,tp,tm,rp,hp,sp,Kp,Km,S,Enz)
        a0=np.sum(a)
        if a0 == 0:
            print("Reaction died")
            break
        else:
            for f in a:
                if f <0:
                    print("Something Wrong")
                    check=1
                    break
            if check == 1:
                break
            random.seed()
            r1=random.uniform(0,1)
            r2=random.uniform(0,1)
            tau=-1*math.log(r1)/a0
            time.append(time[-1]+tau)
            u=1
            r2a0=r2*a0
            while u<=len(a):
                low_lim=np.sum(a[0:u])
                if low_lim>=r2a0:
                    mu=u-1
                    break
                else:
                    if u>=len(a):
                        check=1
                        print("Something Wrong")
                        break
                    u=u+1
            if check == 1:
                break
            reaction_update(Enz,P,P_id,S,mu)
    return P_id,time
# Some sample values for the parameters
kp=float(10000)
```

```
km=float(10)
dp=float(0.5)
tp=float(0.01)
tm=float(1000)
rp=float(0.5)
sp=float(0.04)
Kp=float(1000)
Km=float(25)
P_id_all,time_all=run_sim(T=500,kp=kp,km=km,\
            dp=dp,tp=tp,tm=tm,rp=rp,sp=sp,Kp=Kp,\
            Km=Km,hp=float(5000))
```

APPENDIX D

STATEMENT OF CO-AUTHOR PERMISSIONS

The Co-Authors have granted their consent to use the published articles in chapter 3 (Modi *et al.*, 2018), chapter 4 (Modi and Ozkan, 2018) and chapter 7 (Modi *et al.*, 2020). In addition, my collaborators in the studies mentioned in chapters 5 and 6 have also granted their consent for sharing of data.