UNIVERSITY OF CALIFORNIA, SAN DIEGO

Computational tools and insights to dynamical networks of an allosteric protein

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Chemistry

by

Adam Van Wart

Committee in charge:

Professor Rommie Amaro, Chair
Professor Michael Gilson
Professor Andrew McCammon
Professor Francesco Paesani
Professor Ross Walker

2013

The dissertation of Adam Van Wart is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____
Chair

University of California, San Diego

2013

DEDICATION

This work is dedicated to my grandfather William Boyce.

EPIGRAPH

Keep it cool.
Keep it light.
—Eric Van Wart

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express appreciation to the University of California, San Diego and Irvine for the opportunities and resources provided. Thank you Dr. Rommie Amaro for guidance and support that has been invaluable to the graduate program. I would also like to thank the Amaro lab for their wonderful attitude and exemplary work ethic. Within the Amaro Lab, special appreciation goes to Rob Swift, Lane Votopka, and Jacob Durrant for extensive collaboration.

Thank you Dr. Craig Martens at University of California, Irvine for your support and continued friendship. I would also like to acknowledge Patrick Hogan, a member of the Martens group, for useful discussions in quantum mechanics. Lastly, a special thanks to Dr. Donald Blake and Dr. Ioan Andricioaei for guidance and support that has proven crucial to the graduate outcome.

Chapter 2 is a minimally modified reprint of the material as it appears in **Van Wart A. T.**, Eargle J., Luthey-Schulten Z., and Amaro R. E., '"Exploring residue component contributions to dynamical network models of allostery"', *J. Chem. Theory Comput.*, (2012), 8, 2949-296. I was the primary investigator and author of this paper.

Chapter 3 is a minimally modified reprint of the material as it appears in **Van Wart A. T.**, Durrant J., Votapka L., and Amaro R. E., "Weighted Implementation of Suboptimal Paths (WISP): An optimized algorithm and tool for dynamical network analysis", *J. Chem. Theory Comput.*-(in revision 2013)

Chapter 4 is a minimally modified reprint of the material as it appears in Hogan P., **Van Wart A. T.**, Donoso A., Martens C. C., "Solving evolution equations using interacting trajectory ensembles" *Chemical Physics* (2010), 370, 2028. I was the primary investigator and author of the computational aspects of this paper.

VITA

2004	B. S. in Chemistry, University of California, Davis
2005-2007	Graduate Student and Teaching Assistant, California Polytechnic University, Pomona
2011	M. S. in Chemical and Material Physics, University of California, Irvine
2013	Ph. D. in Chemistry, University of California, San Diego

PUBLICATIONS

Hogan P., **Van Wart A. T.**, Donoso A., Martens C. C., "Solving evolution equations using interacting trajectory ensembles" *Chemical Physics* (2010), 370, 2028

Van Wart A. T., Eargle J., Luthey-Schulten Z., and Amaro R. E., '"Exploring residue component contributions to dynamical network models of allostery"', *J. Chem. Theory Comput.*, (2012), 8, 2949-2961

Van Wart A. T., Durrant J., Votapka L., and Amaro R. E., "Weighted Implementation of Suboptimal Paths (WISP): An optimized algorithm and tool for dynamical network analysis", *J. Chem. Theory Comput.*-(in revision 2013)

ABSTRACT OF THE DISSERTATION

Computational tools and insights to dynamical networks of an allosteric protein

by

Adam Van Wart

Doctor of Philosophy in Chemistry

University of California, San Diego, 2013

Professor Rommie Amaro, Chair

Molecular dynamics simulations can be used to probe a wide range of biologically relevant problems. One such problem is the mechanism of allostery. Allosteric regulation in biological systems is of considerable interest given the vast number of proteins that exhibit such behavior. Designing a consistent framework for allostery that can be rapidly quantified could pave the way for synthetic catalysts, allosteric inhibitors, and a more general understanding of protein function. To this end, this work explores some of the conventional uses of dynamical network analysis as applied to a model of allostery and shows that dynamical networks are highly dependent on what constituents of a protein are correlated. The work goes onto to show that the commonly used α-carbon of an amino acid, although a nexus

joining the backbone and side-chain, is an inferior handle to use for correlated motions and does not model side-chain interactions as well as the center of mass of an amino acid. The α-carbon is shown to be highly correlated to the amino acid's backbone center of mass. Other node choices are explored within the framework of dynamical network analysis and signaling pathways are calculated between the two stereospecific receptor sites known to exhibit allostery in HisH-HisF Thermotoga Maritima.

Further work expands around the optimal signaling pathway within HisH-HisF to include sub-optimal pathways. The Weighted Implementation of Suboptimal Paths (WISP) is developed as an efficient algorithm and tool for generating a statistical distribution of pathways between residues known to play in allostery. For HisH-HisF the degeneracy, or number of times an amino acid participates in a signaling pathway, shifts between the allosterically active and allosterically inactive forms (holo-state and apo-state respectively). Furthermore, all of the paths are tighter and exhibit more correlation for the allosterically active state. This shows evidence of a loss of entropy along the signaling pathways when the effector molecule ,PRFAR, is bound.

Lastly, a general approach to solving evolution equations for probability densities is described using interacting trajectory ensembles. This is done by deriving the general equations of motion for the trajectories in the kinematic space (e.g., configuration or phase space). The time rate of change of the trajectory members depends on both external forces and on the probability density itself. The dependence the trajectory has on the probability density lead to interactions between the ensemble members and a loss of each members independence. The result is illustrated by a number of numerical examples.

Chapter 1

Background material

1.1 Modern molecular dynamics

1.1.1 The potential energy surface

In molecular dynamics, the central problem to be solved is the trajectory. The atomic positions of a molecular dynamics simulation evolve as a function of time. In order to solve the trajectory of a system of atoms over reasonable time scales, certain approximations are used. For instance, the atom and force acting on it are both reduced representations of their true nature. In classical molecular dynamics the atom does not observe quantum behavior such as excited states or the formation and breaking of molecular bonds. Instead, the atom is propagated in time through a potential energy surface that provides the force acting on its average position of mass at any given time. The force acting on each atom i is equal and opposite to the gradient of the potential in three dimensional space:

$$F_i = -\nabla V(r) \tag{1.1}$$

The force given by equation 1.1 is a vector field often referred to as the molecular force field although ions, free of molecular bonds, are also included.

$$V(r)_{\text{bonded}} = \sum_{\text{bond i}} K_i^{\text{bond}}(r_i - r_{0i})^2 + \sum_{\text{angle i}} K_i^{\text{angle}}(\theta_i - \theta_{0i})^2 + \sum_{\text{dihedral i}} k_i^{\text{dihedral}} cos(n_i\phi_i - \gamma_i) \tag{1.2}$$

Forces that are due to non-bonded interactions are also included as van der Waals and electrostatic potential energy terms.

$$V(r)_{\text{non-bonded}} = \sum_{\text{van der Waals}} \epsilon_{ij} \left[(\frac{\sigma_{ij}}{r_{ij}})^{12} - (\frac{\sigma_{ij}}{r_{ij}})^{6} \right] + \sum_{\text{Coulomb}} \frac{q_i q_j}{4\pi\epsilon_0 r_{ij}} \tag{1.3}$$

The bonded and non-bonded terms together form $V(r)$.

$$V(r) = V(r)_{bonded} + V(r)_{non-bonded} \tag{1.4}$$

The force field is formed from separable potential energy terms where the bonded terms are argued to work well under relatively small deviations from equilibrium bond lengths, bond angles and dihedral angles. Since the non-bonded terms are pairwise among all atoms of a system, it is often not feasible to calculate all pairwise interactions and instead a cutoff distance is used for which non-bonded forces are turned off. Simulations that use periodic boundary conditions employ a particle-mesh Ewald summation[1, 2, 3, 4] to model long-range electrostatic interactions in conjunction with the van der Walls forces, Coulomb forces, and artificial cutoff distance.

Further approximations are numerical in nature. For instance, the use of a numerically stable symplectic integrator[5, 6] in the form of a velocity-Verlet algorithm[7] may be used to provide stability to the system being simulated at the expense of cumulative errors for long simulation times. Time steps depend on bond resonance times and are usually one to two femtoseconds.[8] A highly accurate trajectory may not be solved for long time scales, however important dynamic properties such as energy, momentum, time-reversibility, and symplecticness are conserved. This is useful for simulating the (NVE) ensemble. Using a symplectic integrator is believed to provide the best sampling of phase space in accordance to the Liouville theorem which requires a conservation of the phase space volume along the trajectory. In this way, the simulation of a molecule models the time evolution of a strict Hamiltonian system[5, 6].

Certain fluctuations may be added to the atoms to simulate the stochastic nature of bath dynamics. Langevin friction is used to dampen motion proportional to an atoms velocity and random energy terms, following a Boltzmann distribution, are redistributed among atoms to preserve energy. This process follows the

Langevin equation and is useful for modeling molecular statistics in the cononical (NVT) ensemble.

$$M\dot{v} = F(r) - \gamma v + \sqrt{\frac{2\gamma k_b T}{M}} R(t) \tag{1.5}$$

Here M is the mass of the atom and v is the velocity of the atom at position r. The term $F(r)$ is the force which is modified by a dissipative term γv and a perturbative term $\sqrt{(2\gamma k_b T)/(M)}R(t)$ where γ is a friction coefficient, k_B is the Bolzmann constant, T is the temperature, and R(t) is a univariate Gaussian random process.

The Langevin-piston method[9] and Hoover's method[10, 11, 12] for constant pressure simulations inspire tools for the (NPT) ensemble.[13] All of these methods play a part in the modern approach to solving molecular trajectories. Only a small sample of what has been done is touched upon here and the whole breadth of tools at our disposal is beyond the scope of this work. Software packages such as NAMD[14] are routinely used to provide solutions to the molecular trajectory. However, these methods still approximate the molecular trajectory and a natural question arises as to how well the trajectory performs as compared to available experiments.

1.1.2 The conformational ensemble

Equilibrium constraints confine a simulated molecule to a range of possible conformations often referred to as the conformational ensemble. If the molecular trajectory has sampled sufficient conformational space it is said to be Ergodic. Ergodicity requires that the fractional population assigned to each conformation has converged. Other more stringent requirements can be invoked on the fractional probabilities that a given conformation moves to another have also converged. Simply put, a system is Ergodic over a given length of time when all of the statistics of the system is completely defined such that repeating the trajectory again gives no new information.

Nature already computes the trajectories of molecules perfectly. It should not matter if the system is in the (NVE), (NVT), or (NPT) ensemble representa-

tion. Nature provides a range of possible conformations for a molecule consistent with the degrees of freedom and energy available to it.

Nuclear Magnetic Resonance is one of the most powerful experimental tools available to probe natures quantum trajectories in situ. Although limitations exist in what signals are available, experiments are routinely done in combination with molecular dynamics simulations.[15, 16, 17] The results of a computer simulation and natural experiment continue to converge in close agreement making molecular dynamic simulations a powerful tool for predicting the natural behavior of molecules.

The force field parameters have evolved to better capture statistical distributions of conformations for small molecules. The results have been accumulated into software packages such as AMBER[18] and CHARMM27[19, 20]. Larger molecules have also made great strides at the level of large scale motions such as protein folding. Capturing accurate large scale motions is a hallmark feature of modern molecular dynamics. As the field matures, the attempt to model large cell structures accurately will undoubtedly take place. However, verifying the exact accuracy of the simulated molecular dynamics trajectory is still a fundamental open ended problem. In many cases it is powerfully sufficient enough that the trajectory correlates well with experiment.

1.1.3 Computational benchmarks

Large strides have been made since the first molecular dynamics simulation of a simplified biological folding process in 1975.[21] Super computers have been purchased, leased, and even designed explicitly for molecular dynamics simulations. Within the last five years, simulations of protein structures have gone from nanoseconds of simulation time per week to microseconds per week as routinely accessible to the broader scientific community.[22]

In the area of protein folding, millisecond timescales have been reported for folding dynamics of lambda repressor in explicit solvent.[23] With the invention of the supercomputer, ANTON, single trajectories of millisecond length were acquired for the first time toward the end of 2010.[24] This makes it possible to predict

folding on time scales approaching 100 microseconds from a single trajectory.

1.2 Biology and the allosteric protein

1.2.1 General allostery

Within the study of proteins, allostery is broadly used to describe a phenomena where an effector molecule binds to a protein affecting the dynamics, or conformational landscape at another site of the protein that is often distant to the effector binding site. The word allostery has root words "allos" meaning other and "stereos" meaning solid or body. Allosteric phenomena is particularly noticeable when the effector molecule up-regulates or down-regulates protein activity.

The effector molecule takes on the abstract form of a signal and modulates protein activity. A few categories are given in the current literature outlining the mechanism of an allosteric signal via an effector molecule. The effector molecule may induce a concerted motion effecting the protein's conformation such as that outlined in the MWC model.[25] The effector molecule could propagate conformational changes upon binding such as the sequential KNF model.[26] Lastly, the conformation may not significantly change at all and instead the effector molecule induces subtle shifts in the dynamics.[27]

Metabolic processes are closely regulated through a network of interactions as a living organism would otherwise be susceptible to all manner of metabolic failure. Organisms have most likely evolved to build diverse process control mechanisms where the creation or destruction of proteins, metabolites, etc. are an everyday occurrence aimed at the organisms preservation. Having feedback loops in the form of molecular signals is believed to be evolutionarily pressured and therefore structural elements among homologous proteins exhibiting allostery have been evolutionarily conserved.[28]

1.2.2 Allostery in HisH-HisF

The protein HisH-HisF exhibits classic allosteric behavior. This protein is expressed in the hyperthermophilic, bacteria Thermotoga Maritima and is involved in the fifth step of a nine step biosynthetic pathway resulting in the essential amino acid, Histidine. HisH-HisF has two receptor sites. The receptor site (effector site) is located at the base of HisF near the C-terminal cyclase domain. The other site is located in HisH roughly 30 Å away.[29]

The effector molecule is the metabolite N1-[(5'-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR) which up-regulates the turnover of glutamine to glutamate by a factor of roughly 4900.[30]

It should be possible to study the signaling effect of PRFAR on the system HisH-HisF using a molecular dynamics simulaiton. Central to understanding how this might be done is understanding what signaling mechanism may take place. The HisH-HisF protein does not change very much in conformation between the PRFAR-bound and PRFAR-unbound (holo and apo-states respectively). IT may be a good approximation to consider small changes in dynamics as a way of comparing the active and inactive forms of the protein dimer.

1.3 Allostery and rational drug design

1.3.1 Central lethality in protein networks

Proteins may act as catalysts, signaling molecules, or building blocks in cells and microorganisms. This view often looks at the protein's individual attributes. However, a protein often affects another protein's function. Protein networks that form a measure of protein interdependence have become a useful tool when quantifying the centrality of a protein's role within an organism. Recent work in yeast (Saccharomyces cerevisiae) has shown a correlation between a protein's centrality and the lethal consequences to the organism if the protein is deleted (not expressed) in the protein network.[31, 32] In rational drug design, a small molecule would negate protein function.

1.3.2 Disrupting the allosteric state

If a protein is central to the organisms survival, then it is a candidate to be targeted. If the protein is also exhibits allostery, then targeting it is equivalent to impeding function by locking it in either its active or inactive form. It may be possible to compete with the natural signal induced by the effector molecule. An understanding of the protein's dynamics would most likely be necessary. Competing with the natural signal is a strategy that could be used in many beneficial circumstances such as anti-fungal, anti-microbial, and general anti-biotics.

1.3.3 Dynamical networks as a plausible strategy

Developing insight to changes induced by an effector molecule likely comes from the overall conformational change and the change in dynamics due to local shifts in dynamics. Finding dynamical correlations among protein components would allow for a comparison between an allosterically active and inactive protein. Any molecular dynamics simulation comparing these states would need to be performed at equilibrium as it may be highly problematic to simulate the non-equilibirum shift from the apo-state to the holo-state (allosterically active).

Chapter 2

Exploring residue component contributions to dynamical network models of allostery

2.1 Abstract

Allosteric regulation in biological systems is of considerable interest given the vast number of proteins that exhibit such behavior. Network models obtained from molecular dynamics simulations have been shown to be powerful tools for the analysis of allostery. In this work, different coarse-grain residue representations (nodes) are used together with a dynamical network model to investigate models of allosteric regulation. This model assumes that allosteric signals are dependent on positional correlations of protein substituents, as determined through molecular dynamics simulations, and uses correlated motion to generate a signaling weight between two given nodes. We examine four types of network models using different node representations in Cartesian coordinates: the (i) residue α-carbons, (ii) the side chain center of mass, (iii) the backbone center of mass, and the entire (iv) residue center of mass. All correlations are filtered by a dynamic contact map that defines the allowable interactions between nodes based on physical proximity. We apply the four models to imidazole glycerol phosphate synthase (IGPS), which pro-

vides a well-studied experimental framework in which allosteric communication is known to persist across disparate protein domains (e.g., a protein dimer interface). IGPS is modeled as a network of nodes and weighted edges. Optimal allosteric pathways are traced using the Floyd Warshall algorithm for weighted networks, and community analysis (a form of hierarchical clustering) is performed using the Girvan-Newman algorithm. Our results show that dynamical information encoded in the residue center of mass must be included in order to detect residues that are experimentally known to play a role in allosteric communication for IGPS. More broadly, this new method may be useful for predicting pathways of allosteric communication for any biomolecular system in atomic detail.

2.2 Introduction

Proteins and enzymes act as catalytic waypoints along a natural products assembly line. They are the molecular machinery of the cell designed to assist complex biological synthesis and mediate biosynthetic regulation. Allostery is broadly categorized as the mechanism, or set of mechanisms, for which an event at a distinct site affects the activity, function, conformational ensemble, or general attributes at another site within an enzyme or protein.[33] As an example, in the case of allosteric modulation, the catalytic efficiency of an allosterically active enzyme or protein is affected through the binding of a metabolite to an allosteric site. The metabolite is said to be an effector molecule that may propagate an activation or inhibition signal to a distant catalytic site that in turn regulates the catalytic activity of the enzyme or protein. Furthermore, a metabolite may take part in the catalytic reaction itself.[34, 25] In this way, allostery mechanizes regulation within metabolic pathways as an intrinsic characteristic of a variety of proteins and enzymes. Further understanding of allostery may lead to advances in practical applications, such as rational drug design, protein engineering, and development of synthetic catalysts.

Allosteric regulation usually involves at least two different stereospecific receptor sites.[34, 25, 35] As metabolic pathways are critical to living organisms,

it follows that communication between both receptor sites must be a protected feature of allosteric modulation. Consequently, homologous proteins often preserve structural elements that conserve communication between receptor sites, the central theme being that structural elements important to protein function, and thus allosteric signaling, have been evolutionarily conserved.[28, 36, 37, 30] A notable proponent of this theme can be found in Statistical Coupling Analysis first introduced by Lockless and Ranganathan.[38]

Much effort has been put forth to develop rigorous theories of allosteric modulation. Some of the early foundations of allostery are proposed in the concerted MWC model (Monod, Wyman, and Changeux [25]) and sequential KNF model (Koshland, Nemethy, and Filmer [26]). These models conceptualize the mechanism within a protein modulated by an allosteric effector as a conformational transition that is either concerted or induced sequentially.[39] Since then, mechanistic details have become rooted in an ensemble view of the protein where cooperativity between sites is governed by a density of states that is intimately connected to entropy and enthalpy.[27, 40] The accessibility of an allosteric state may be through an effector substrate inducing a large conformational change within the protein complex as well as small local conformational changes.[41] However, allosteric regulation can also occur in the absence of large conformational changes, as found in works by Cooper and Dryden.[42] In such a situation, an enzyme or protein would have the propensity to facilitate allosteric signals due to small vibrations inherent to the protein's state and not due to a change in shape or fold.

Experimental constructs to explore allostery have included Xray crystallography with site-directed mutagenesis, ^{1}H-^{15}N heteronuclear single quantum correlation (HSQC) NMR, kinetic assays with site directed mutagenesis, inhibition assays, and others.[43] These tests probe a range of questions associated with small molecule effectors, protein-protein interfaces (hot spots), allosteric pathways, and general protein functionality. NMR provides increasing potential for a direct view into the mechanics of allostery. The ^{1}H-^{15}N amide bond vector is an inherent characteristic of polypeptides and thus has been used to characterize fast backbone and side-chain dynamics. In NMR experiments, thermal motions are present

in all proteins above the glass transition temperature, and there is a direct link between molecular motions and local magnetic field fluctuations. These dynamics are commonly associated with the Lipari-Szabo order parameter (S^2).[44, 45] It is generally accepted that picosecond to nanosecond time scales capture allostery in the absence of large conformational shifts for a majority of proteins (see review [46]). The parameter S^2 is related to the Gibbs free energy through the canonical partition function and provides a means of deriving conformational entropy from bond vector fluctuations.[47]

Molecular dynamics seeks to give accurate virtual results that compete with experiments at a fraction of the cost. With the use of molecular dynamics simulations,[14, 20] details of conformational changes are sampled to clarify and categorize mechanisms of allostery that may otherwise be elusive. Quasiharmonic analysis is a popular analytical tool that has been used in molecular dynamics to quantify motions in the proteins trajectory ensemble.[48, 49, 50] This approach aligns trajectories to a reference frame, often the average of the protein, and can be used to deduce correlations among variables (e.g., dihedral angles, position) that account for the configurational space. However, harmonic motions may be an oversimplified assumption for protein dynamics, and the need for an anharmonic metric arises. Mutual information methods outlined by Lange and Grubmuller provide a more generalized framework that captures anharmonic motions in Cartesian coordinates.[51] Recent works by Killian et al. derive configurational entropies using internal coordinates for a trajectory ensemble.[52] The mutual information values are the secondorder terms derived from the entropy expansion and directly link correlated conformations within the protein to molecular configurational entropy values. This work inspired McClendon and co-workers to develop an entropy-based approach to harmonic and anharmonic correlations via mutual information (MutInf) among dihedral angles;[53] this method includes Bayesian filtering as well as sampling penalties to generate more statistically robust values.

Dynamical network analysis is another method used to explore putative allosteric communication pathways.[54, 55] Within the realm of biological computation, dynamical network analysis samples correlations among time dependent

variables for predefined components of a biological system (i.e., protein). These variables are associated with the trajectory ensemble and provide a distribution at equilibrium. As examples, a variable could be the position of an atom, center of mass (c.o.m.) of a residue, or a vector modeling the orientation of a residue. These variables serve to model the configuration of the protein at a given time. As developed by Luthey-Schulten and colleagues,[54, 55] the variable was the position of each α-carbon atom within an aminoacyl-tRNA synthetase system. MD simulations sample the system dynamics, which are used to generate correlation values among all variables (nodes) of the biological system. The absolute value of a given correlation is then incorporated into a function that models the signaling between nodes as a distance. The behavior of nodes that are highly correlated and within close physical proximity (aggregates) can be analyzed in terms of community behavior and structure. Allosteric signaling between community aggregates occurs through particular node pairs that exhibit a high degree of network degeneracy (i.e., community cross talk). These critical nodes transfer a relatively large degree of information across their edge. As critical node pairs carry out intercommunity cross talk, they may provide favorable sites for signal disruption.

In this work, we extend the dynamical network analysis method and propose allosteric signaling pathways in imidazole glycerol phosphate synthase (IGPS; HisH-HisF), which belongs to the glutamine amidotransferase (GAT) family of enzymes (Figure 2.1). This multidomain globular protein regulates the fifth step of the histidine biosynthetic pathway in plants, fungi, and microbes. Histidine is considered an essential amino acid, as humans do not have a histidine biosynthetic pathway. In bacteria, HisH-HisF is a protein dimer, whereas in eukaryotes it exists as one single polypeptide chain, with a $\sim$40 amino-acid-residue tether connecting the two domains. The HisH-HisF dimer has two receptor sites: one receptor site (effector site) is positioned at the C-terminal cyclase domain of HisF, and the other (allosteric site) is located in HisH (the glutaminase domain) roughly 30 Å away.[29] The triad GAT subfamily strictly conserves these catalytic residues across all members of the set, although the allosteric effector, remote binding site, upregulation, and allosteric site vary between different enzymes.[56, 57, 58, 59] Another impor-

tant conserved subfamily structure in GATs is a $(\beta/\alpha)_8$ barrel which is located in HisF. The $(\beta/\alpha)_8$ barrel is believed to assist in transport of the ammonia moiety to the effector or substrate molecule and may also play a role in allostery.[60, 61, 62]

Attachment of the primary metabolite N1-[(5'-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR, which is also the allosteric effector) to the cyclase domain induces several global structural changes in the protein dimer. A flexible loop on the C-terminal end of the $(\beta/\alpha)_8$ barrel appears to modulate access to the PRFAR substrate, and a hinge motion at the dimer interface changes the HisF and HisH domains by roughly 7.1^o to form a tighter interface.[30] These large scale conformational changes indicate that the PRFAR-bound (holo) and PRFAR-unbound (apo) states are distinct. Other induced motions consist of smaller conformational changes in the side chain orientations.[61] The catalytic turnover of glutamine to glutamate is increased $\sim$4900 fold when PRFAR is bound; however, the details of the allosteric mechanism are not completely understood.[30] All members of the GAT family of enzymes exhibit similar allosteric regulation to control the production and utilization of ammonia in the subsequent catalytic reaction.[63] After the catalytic turnover of glutamine, nascent ammonia is shuttled across the dimer interface and through the center of the $(\beta/\alpha)_8$ barrel along the barrels principal axis to the cyclase domain where PRFAR is broken into 5'-(5-aminoimidazole-4-carbox-amide) ribonucleotide (AICAR) and ImGP (Figure 2.1).[64] Previous studies have also found that AICAR and ImGP stimulate the glutaminase half reaction, although at a reduction of 45 to 188 times less, respectively, than the PRFAR substrate molecule.[61] Moreover, the two independent products also stimulate the glutaminase half reaction 15 times less than PRFAR. These results suggest that AICAR and ImGP both contribute to allostery, but ImGP may be more involved in the allosteric signaling network. Furthermore, the chemical bond between the AICAR and ImGP moieties is necessary for enhancement of the allosteric state.

We explore extensions of the original dynamical network analysis to ascertain new models for allosteric communication, which are benchmarked against experimental site-directed mutagenesis studies in HisH-HisF. The impact of uti-

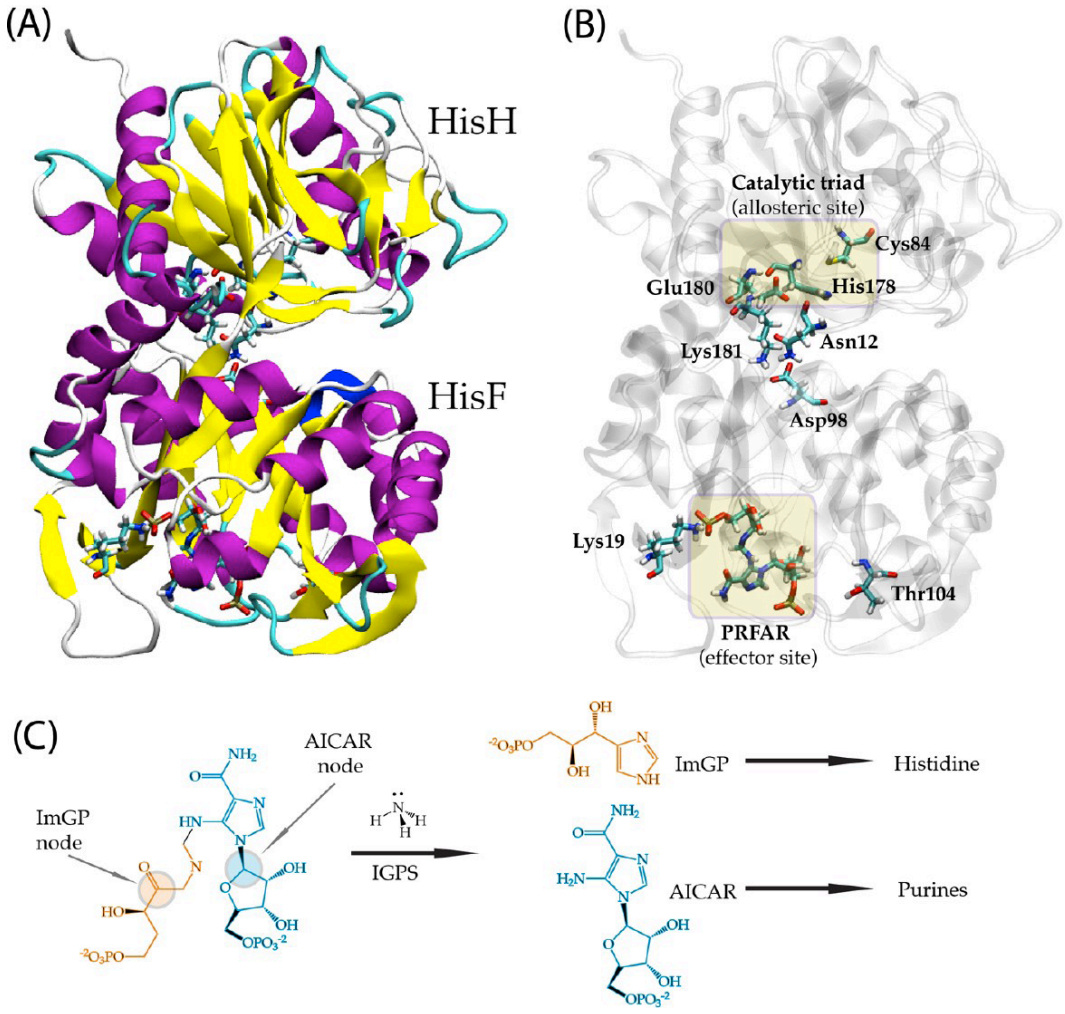


Figure 2.1: (A) The two domains,HisH (glutaminase domain) and HisF (cyclase domain), are shown colored by secondary structural elements. (B) The catalytic triad, residues tested experimentally for allostery, and the substrate PRFAR are presented again in ball and stick representation for clarity with the same orientation as A, showing the allosteric and effector sites, which are over 30 Åapart. (C) The overall reaction catalyzed by HisH and HisF produces AICAR and ImGP.

lizing different residue contributions (i.e., nodal methods) within the dynamical network analysis framework is investigated using HisH-HisF as a model system. The Floyd Warshall algorithm determines the shortest (optimal) path between two nodes in a network,[65, 66] and the Girvan-Newman algorithm,[67] a method without free parameters, generates aggregate nodal clusters, or communities, that optimize intracommunication pathways within the protein. For each nodal method, we determine the optimal signaling pathways. We then compare and investigate the corresponding community structures for the α-carbon and residue c.o.m. methods. Ultimately, we propose novel allosteric communication pathways for HisH-HisF and show that using the residue c.o.m. provides advantages over other nodal methods within the dynamical network analysis framework.

2.3 Methods

2.3.1 System Setup

The initial coordinates for the (holo) HisH-HisF were obtained as PDB 1GPW (Thermotoga maritima) from the Protein Data Bank and included resolved phosphate groups still bound to the active site. Furthermore, the PDB 1GPW included three states of HisH-HisF categorized by chains A-B, C-D, and E-F. In order to prepare an accurate residue environment, PROPKA was used to determine protonation states of titratable residues based on an environment at pH 7.[68] The δ-nitrogen was protonated for the HisF domain for residues His84, His209, and His244 and His73, His120, and His141 in HisH. An ϵ-nitrogen protonation state was assigned for HisF:His228, and for HisH:His53. Both the δ-nitrogen and ϵ-nitrogen of the residue HisF:His151 were protonated.

The crystal structure 1OX5 (from Saccharomyces cerevisiae) included all PRFAR heavy atoms and was used to build in the PRFAR substrate for 1GPW. Using the VMD plugin Multi-Seq,[69] the crystal structure chain A 1OX5 was aligned to 1GPW chains C and D via secondary structures in the HisF domain using the STAMP structural alignment algorithm.[70, 71] The result was that the PRFAR substrate aligned correctly to the 1GPW:HisF domain for the holo-state of HisH-

HisF. This alignment is consistent with other works.[62, 72] The 1GPW:crystal waters, 1OX5:PRFAR metabolite, 1GPW:HisF, and 1GPW:HisH segments were processed by PSFGEN to generate psf and pdb files. The parametrization of the PRFAR substrate followed standards established in the CHARMM protocol22 and used results from previous works.[73] The formal charge for the protein dimer was -8.0e, and the formal charge for the primary metabolite, PRFAR, was -5.0e. The formal charge of the system was -13.0e in total. The system was solvated using SOLVATE[74] to add explicit TIP3 water molecules to a box with a water pad of 10 Å around the protein. Thirteen sodium counterions were added with the VMD plugin Autoionize, bringing the overall formal charge of the entire system to zero.

The (apo) HisH-HisF was also derived from 1GPW using chains A and B. The system setup for the (apo) HisH-HisF followed the exact same procedure as (holo) HisH-HisF without the addition of the PRFAR ligand.

2.3.2 Molecular Dynamics Simulations

A four-step energy minimization was employed for both apo and holo structures to ensure maximal retention of experimental structural information. The four steps constrained heavy atoms for 1000 fs, heavy atoms excluding water for 1000 fs, backbone atoms excluding all other molecular groups for 2500 fs, and finally a constraintfree motion of all atoms for 5000 fs. Harmonic constraints were induced to allow minimal backbone motion to achieve water equilibration for five time frames of 25 000 fs each and different backbone constraint constants of 1.0, 0.75, 0.50, 0.25, and 0.0 kcal mol^{-1} Å.

Molecular dynamics for the holo structure were run for 30 ns (including an initial 10 ns to ensure the mutual equilibration of PRAFR and HisH-HisF) with an integration time step of 1 fs using NAMD2.7 and the CHARMM27 force field.[14, 19] Production molecular dynamics for the (apo) HisH-HisF were run for 20 ns with an initial 1 ns harmonic constraints equilibration phase. All other simulation parameters and benchmarks were the same as the (holo) HisH-HisF simulation. Periodic boundary conditions were used with the NPT ensemble, a pressure of 1 atm using a Langevin piston, and a temperature set to 310 K. A

particle mesh Ewald summation was used for the electrostatic approximations.[1] The long-range van der Waals interactions were given a cutoff distance of 12 Å and a switching distance at 10 Å. Simulations were performed on the TACC Ranger cluster using 256 processors. The benchmark for the ?45 000 atom systems were 0.0136156 s/step and 0.157588 days/ns.

Protein snapshots were extracted at 500 fs intervals over the 20 ns production phases of both the holo-state and apo-state, resulting in two sets of 40 000 structures for subsequent analysis. Each snapshot was aligned via RMSD to the average position of all atoms within the protein for both systems over the entire 20 ns simulation of production dynamics, in order to remove any rotational or translational motions.

2.3.3 Dynamical Network Model

We explore the use of different sources of correlation information, also known as nodal methods, including (i) α-carbon, and the c.o.m. of (ii) the backbone, (iii) the side chain, and (iv) whole-residue representations (residue c.o.m.). In each case, the covariance matrix is normalized, and as a consequence, the position, and not the mass of the node, is important. In order to model PRFAR as a set of nodes, the PRFAR substrate is broken into two moieties, which represent the reaction end products AICAR and ImGP (Figure 2.1). The centers of mass for both the AICAR and ImGP moieties are defined as separate nodes for the analysis.

Methods are carried out using a similar protocol to that of refs [54] and [55]. A contact map is used to determine Boolean values for node?node interactions. If a residue is within 4.5 Å of another residue for 75% of the 40 000 frame (20 ns) trajectory, then the two residues are considered to be in contact with one another (Supporting Information (SI) Figure 2.1). Nodes that do not fit these requirements are filtered out. The correlation values are obtained by trajectory data using the program CARMA.[75] The correlation is defined by eqs 2.1 and 2.2, where C_{ij} is the correlation value between two nodes over a given simulation time (Figure 2.2). The value of C_{ij} is unity when i = j and satisfies the condition that a node

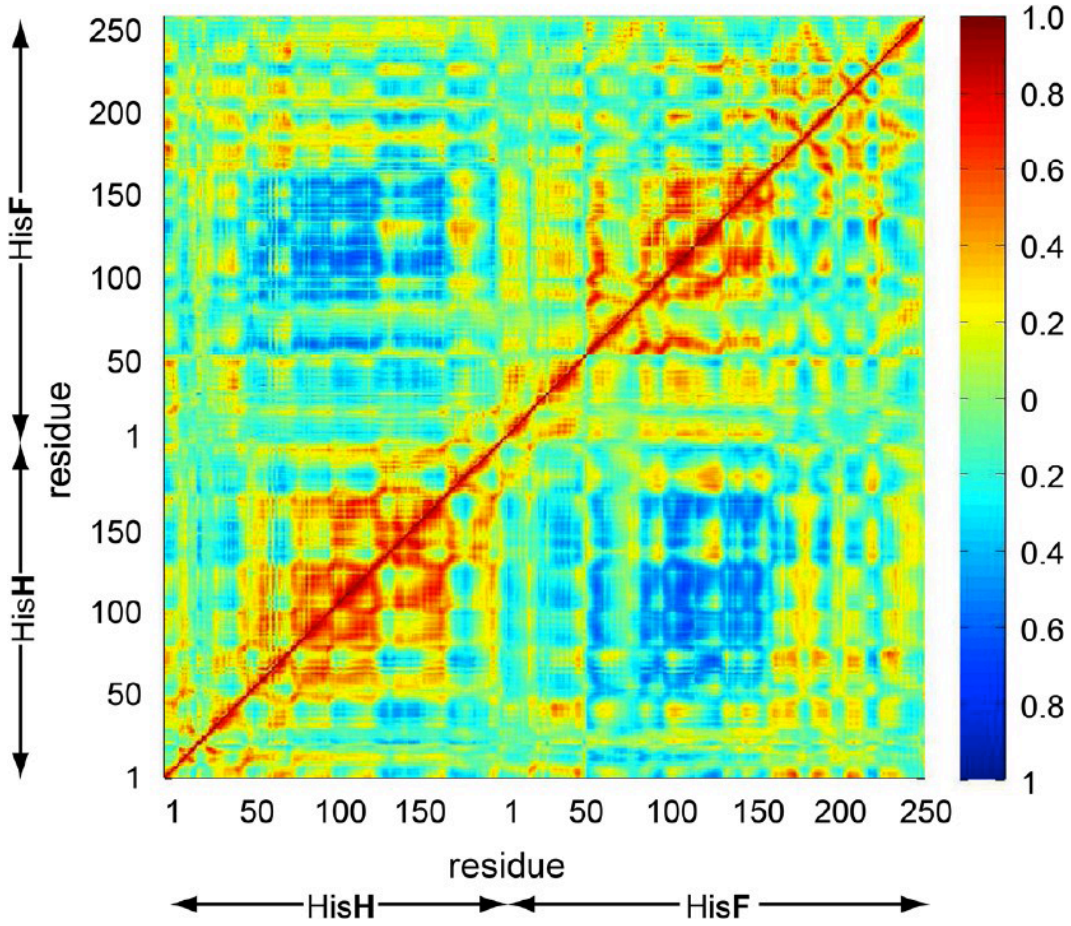


Figure 2.2: Entries in the lower right triangle show pairwise correlation values using the α-carbon atoms; the upper left triangle shows the correlation values using the c.o.m. of the entire residue.

correlates completely with itself. In eq 2.2, the position of a given node at time t is subtracted from the average position of the node over the entire production dynamics (20 ns) and requires that the average position of the node sampled is approximately stationary over time. When comparing the motion of one nodal method to another, eq 2.1 is also used (Figure 2.3); these are correlations among different nodal methods within the same residue.

$$C_{ij} = \frac{\langle \Delta\vec{r_i}(t) \cdot \Delta\vec{r_j}(t) \rangle}{(\langle \Delta\vec{r_i}(t)^2 \rangle \langle \Delta\vec{r_j}(t)^2 \rangle)^{1/2}} \tag{2.1}$$

and

$$\Delta\vec{r_i}(t) = \vec{r_i}(t) - \langle \vec{r_i}(t) \rangle \tag{2.2}$$

The use of eqs 2.1 and 2.2 as well as interaction criteria determined through the contact maps generates a filtered correlation matrix.

The edge between a pair of nodes is then given a length in network space by functionalizing the correlation values. A positive or negative correlation is viewed equally in that residue interactions are coherent whether they are correlated or anticorrelated, and so the edge lengths take the value w_{ij} in eq 2.3 between node i and node j:[54, 55]

$$w_{ij} = -log(|C_{ij}|) \tag{2.3}$$

Although the values of each length have meaning in network diagrams, they can be mapped to the physical protein space by weighting each edge with a thickness or cylindrical diameter. Therefore, thick edges between nodes are indicative of edges close to one another in network space, while thin edges and edges that do not exist are far or infinitely separated, respectively (Figures 2.4-7). Such edges represent the signaling strength between nodes.

One aim of our analysis is to identify putative allosteric communication pathways in proteins, using HisH-HisF as a model system. Evidence of cooperation between the substrate PRFAR (i.e., the source of allosteric signal) and the catalytic triad (i.e., where the allosteric signal ends) is captured within this framework. In this model, we assume that the most likely or biologically relevant pathways are those that minimize the network distance. Finding a singular optimal path is equivalent to minimizing the distance traveled between nodes in the network. The

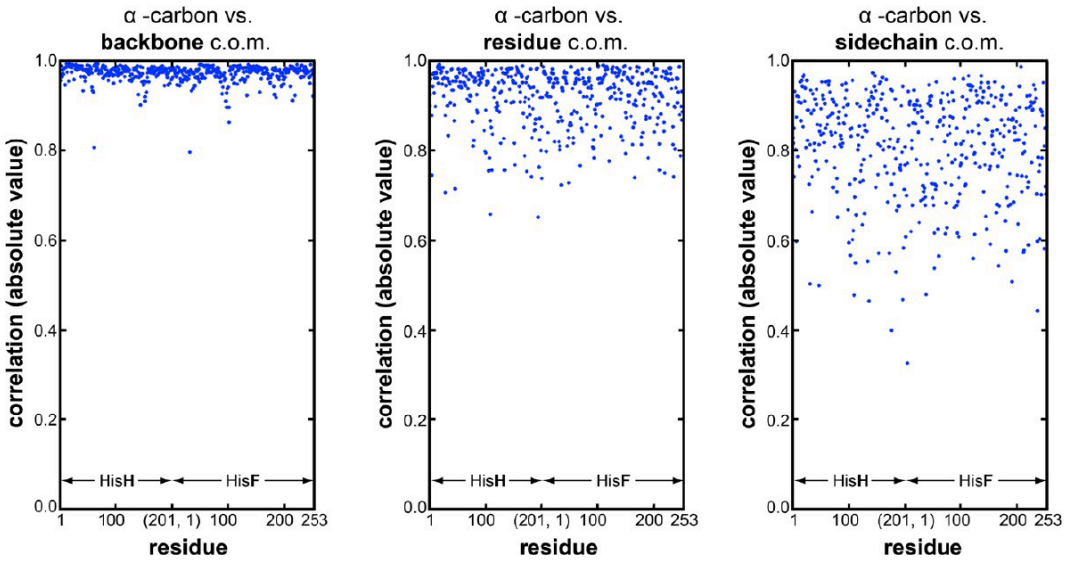


Figure 2.3: (A) Here, correlation between the α-carbon and the backbone c.o.m. is shown for each residue. The α-carbon is strongly correlated to the backbone c.o.m. and not well correlated to the side chain c.o.m. suggesting that information given by the α-carbon is highly biased toward backbone motion. (B) Next, correlation between the α-carbon and the c.o.m. for the entire residue is shown. (C) Lastly, comparisons between the α-carbon and side chain c.o.m. show the lowest correlation values and greatest spread, which indicates that information from side chain motion may be lost in an α-carbon representation.

length of such a path is the sum of the individual path lengths between each node set i, j (eq 2.4), starting with the source nodes in PRFAR and ending with the three sink nodes in the catalytic triad:

$$D_{ij} = \sum_{k,l} w_{k,l} \tag{2.4}$$

All other paths are deemed suboptimal if they fall within an acceptable deviation from the optimal path. Such pathways are likely to present highly degenerate signaling along the optimal path. Although it is possible to determine numerous suboptimal pathways within the dynamical network analysis framework, we do not investigate suboptimal pathways in this work. Yet, we acknowledge that their analysis may likely lead to further insight into the complexities of allosteric regulation.

2.3.4 Entropy Calculations

The entropy for both holo and apo states was calculated using CARMA,[75] which follows the derivation outlined by Andricioaei and Karplus.[76]

The mass weighted covariance matrix was used for the α- carbons, and each mass was 12.0107 g mol^{-1}. A mass weighted covariance matrix was also used for the residue c.o.m. method, and each mass varied depending on the selected residue. The PRFAR molecule was omitted when comparing entropies of the apo and holo-states in order to maintain identical degrees of freedom between the ligand bound and ligand free systems.

2.4 Results

2.4.1 Molecular dynamics simulations

The MD simulations for HisH-HisF must be sampled over equilibrium conditions in order to use correlations since the average position of a node is required to be approximately stationary over time. To assess each nodal method, 30 ns of MD were run for the holo-state. The RMSD for all 30 ns of molecular dynamics

was calculated at each frame for the holo-state and indicates system equilibration after the first 10 ns (SI Figure 2.2); the remaining 20 ns is considered production dynamics and was utilized in the dynamical network analyses. The RMSD for the 20 ns molecular dynamics of the apo-state also showed convergence. Trajectories were assessed at 5 ns, 10 ns, 15 ns, and 20 ns to determine the time dependence and convergence of the optimal paths.

2.4.2 Correlation and Contact Maps

The full correlation matrix, representing both the residue c.o.m. and α-carbon methods, is presented (Figure 2.2). A contact map indicating which residues were closely interacting was generated over the production MD phase for the holo-state. If residues were within 4.5 Å of another residue for 75% of the 40 000 frame (20 ns) trajectory, the correlation value in the correlation matrix was kept; otherwise the correlation value was set to zero. This contact-map filtering prevents edges from being drawn between residues that are physically far apart. The same contact map was used for the holo-state across each time assessment at 5 ns, 10 ns, 15 ns, and 20 ns (SI Figure 2.1). The contact map for the apo-state was generated from a separate 20 ns trajectory and followed the same criteria as the holo-state. This contact map was used to assess the apo-state over the entire 20 ns trajectory for comparison to the holostate.

2.4.3 Correlation of Nodal Methods in Holo HisH-HisF

The correlation of the different nodal methods was analyzed for the holo-state over 20 ns of production dynamics by determining the positional correlation between the α-carbon and the c.o.m. of the (a) backbone, (b) side chain, and (c) entire residue. This analysis indicates that the motion exhibited by the α-carbon is most highly correlated to the backbone motion as a general trend for same residue selections (Figure 2.3). The position of the α-carbon correlates strongest with the position of the backbone c.o.m. and weakest with the position of the c.o.m. of the side chain. Between these two extremes, the position of the residue c.o.m., which

represents both the backbone and side chain atoms, exhibits moderate correlation to the α-carbon position.

2.4.4 Optimal Pathways

Comparison of Different Nodal Methods within the Holo-State

The optimal communication path is the shortest path between the source of the allosteric signal, in this case PRFAR, and the end-point of the allosteric signal, or sink, which is the catalytic triad (Figure 2.4). The catalytic triad is comprised of three residues HisH:Glu180, HisH:His178, and HisH:Cys84 and therefore three potential sink candidates. Each residue within the triad was thus investigated in the search for the optimal paths. For simplicity, we focus on the sink residue HisH:Glu180 (Table 2.1). However, we note differences in the optimal pathways among all combinations of sources and sinks and provide this additional information (SI Table 2.1).

The optimal path between source node ImGP and sink node Glu180 captured the same residues at 15 and 20 ns for the different nodal methods, and the optimal path for the residue c.o.m. method was surprisingly the same for all time assessments (Table 2.1). The pathway between the ImGP moiety and residue Glu180 yielded similar results to the pathway between the AICAR moiety and residue Glu180 when using the α-carbon nodal method (SI Table 2.1).

The side chain c.o.m. method gives an optimal path of ImGP, HisF:Leu50, HisF:Val48, HisF:Thr78, HisF:Lys99, HisF:Asp98, HisH:Lys181, and HisH:Glu180. The residue c.o.m. method gives an optimal path of ImGP, HisF:Leu50, HisF:Gly80, HisF:Val79, HisF:Lys99, HisF:Asp98, HisH:Lys181, and HisH:Glu180. In contrast, the pathway predicted by the α-carbon method suggests an optimal pathway connecting the ImGP source through HisF:Leu50, HisF:Phe49, HisF:Phe77, HisF:Pro76, HisF:Lys181, and HisH:Glu180 (sink). The optimal pathway predicted by the backbone c.o.m. method traces through ImGP (source), HisF:Leu50, HisF:Phe49, HisF:Thr78, HisF:Phe77, HisF:Lys181, and HisH:Glu180 (sink).

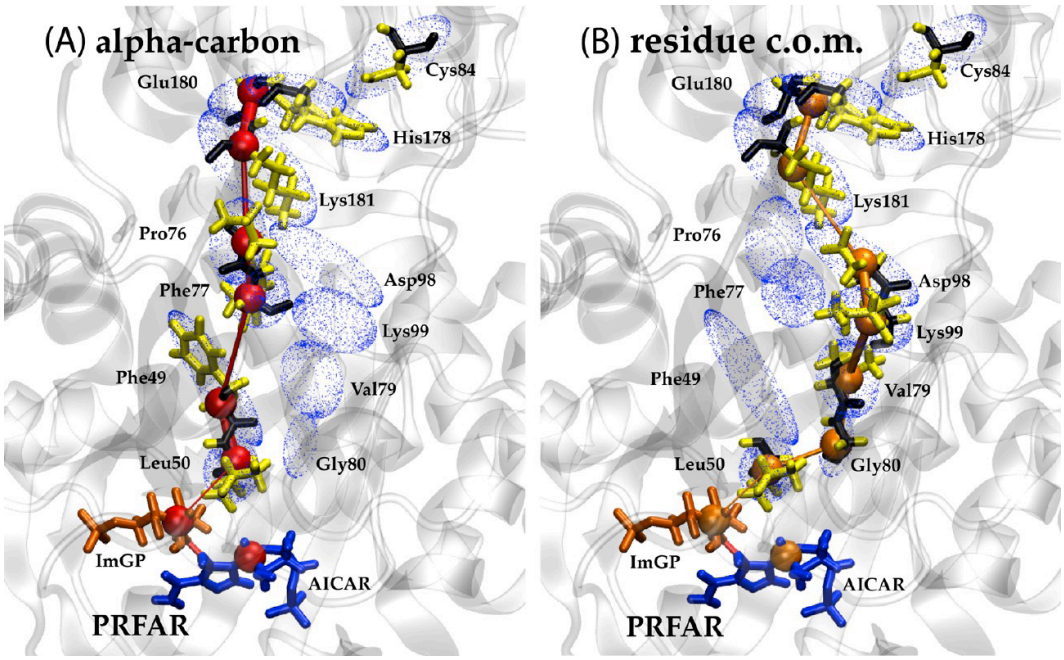


Figure 2.4: (A and B) Here, we show the optimal pathway generated from the Floyd-Warshall algorithm using (A) the α-carbon method (red spheres) and (B) the residue c.o.m. (orange spheres) method. In both methods, the source is the ImGP moiety, and the sink is catalytic residue HisH:Glu180. The ellipsoids (blue dots) are aligned along the principal axes of a given residue to illustrate a residue's orientation and spatial displacement. Edges are drawn between nodes, and the edge thickness is radially proportional to the weight given by eq 3.

Table 2.1: The optimal path from source ImGP to sink HisH:Glu180 is calculated on the basis of 20 ns of production dynamics after an initial 10 ns to allow PRFAR and HisH-HisF to equilibrate. Each entry in the table represents a residue or moiety of PRFAR. The residue name and dimer superset of each residue is indicated above the table. Number entries are associated with the crystal structure 1GPW from the protein data bank. The resulting correlation matrix for each nodal method is calculated at 5 ns, 10 ns, 15 ns, and 20 ns. However, the same contact map is used for each and is based on the 20 ns trajectory. For each nodal method shown, the optimal paths appear to converge to a consistent set of residues. Evolutionarily conserved (100% sequence identity) residues are both underlined and bold, whereas partially conserved (greater than 50% sequence identity) residues are just underlined in reference to previous homology studies.

ImGP-(HisH:Glu180) ◄—— HisF ——►◄► HisH ►

Node Method	time (ns)	ImGP	Leu	**Phe**	Val	Leu	Ile	Gly	Val	Thr	Phe	**Pro**	Ile	Asp	**Lys**	**Asp**	**Lys**	**Glu**
Side-chain	< 5 >	**ImGP**	**50**		**48**	**47**							**75**	**74**				**180**
c.o.m.	< 10 >	**ImGP**	**50**	**49**			**73**							**74**				**180**
	< 15 >	**ImGP**	**50**		**48**					**78**					**99**	**98**	**181**	**180**
	< 20 >	**ImGP**	**50**		**48**					**78**					**99**	**98**	**181**	**180**
Residue	< 5 >	**ImGP**	**50**					**80**	**79**						**99**	**98**	**181**	**180**
c.o.m.	< 10 >	**ImGP**	**50**					**80**	**79**						**99**	**98**	**181**	**180**
	< 15 >	**ImGP**	**50**					**80**	**79**						**99**	**98**	**181**	**180**
	< 20 >	**ImGP**	**50**					**80**	**79**						**99**	**98**	**181**	**180**
alpha-carbon	< 5 >	**ImGP**	**50**	**49**	**48**	**47**						**76**					**181**	**180**
	< 10 >	**ImGP**	**50**	**49**						**78**	**77**	**76**					**181**	**180**
	< 15 >	**ImGP**	**50**	**49**						**78**	**77**	**76**					**181**	**180**
	< 20 >	**ImGP**	**50**	**49**							**77**	**76**					**181**	**180**
Backbone	< 5 >	**ImGP**	**50**	**49**						**78**	**77**	**76**					**181**	**180**
c.o.m.	< 10 >	**ImGP**	**50**	**49**						**78**	**77**	**76**					**181**	**180**
	< 15 >	**ImGP**	**50**	**49**						**78**	**77**						**181**	**180**
	< 20 >	**ImGP**	**50**	**49**						**78**	**77**						**181**	**180**

Comparison of Holo- and Apo-States Using the Residue Center of Mass Method

Both the holo-state, with allosteric effector PRFAR bound, and the apo-state simulations sample equilibrium conformations of the protein and provide a comparison between the ligand activated holo-state and the basal (inactive) apo-state. The optimal path for the holo-state is compared to the optimal path of the apo-state from ImGP (source) to HisH:Glu180 (sink) using the residue c.o.m. as the preferred nodal method (Table 2.2). The residue c.o.m. method traces an optimal path for the holo-state as ImGP, HisF:Leu50, HisF:Gly80, HisF:Val79, HisF:Lys99, HisF:Asp98, HisH:- Lys181, and HisH:Glu180. In contrast, the apo-state gives an optimal path of HisF:Leu50, HisF:Phe49, HisF:Phe77, HisF:Pro76, HisH:Lys181, and HisH:Glu180. A noteworthy switch between the two states is the HisF:Leu50?HisF:Gly80 in the holo-state versus the HisF:Leu50-HisF:Phe49. The apostate HisF:Leu50-HisF:Phe49 is clearly a backbone dominated correlation

Table 2.2: Here, we show the optimal pathway between ImGP and HisH:Glu180 for both the apo-state and holo-state simulation with only the c.o.m. method being used (the preferred method for this paper). The apo-state and holo-state of HisH-HisF is compared here and shows a clear switch between both states. The holo-state traces out the salt bridge HisF:Asp98-HisH:Lys181 that is verified experimentally to be part of the allosteric mechanism. The apo-state does not provide this bridge. The distance values on the right are calculated according to eq 2.4.

ImGP-(HisH:Glu180)

			HisF								HisH		
HisH-HisF	time (ns)	ImGP	Leu	Phe	Gly	Val	Phe	Pro	Lys	Asp	Lys	Glu	distance
Holo-residue c.o.m.	< 20 >	ImGP	50		80	79			99	98	181	180	287
Apo-residue c.o.m.	< 20 >	none	50	49			77	76			181	180	298

where the holo-state HisF:Leu50?HisF:Gly80 suggests a correlation that may depend on side chain interaction. Another important difference is seen at the interface of the protein dimer, where the apo-state HisH:181Lys-HisF:Pro76 communication path switches to HisH:181Lys-HisF:Asp98 in the holo-state. The HisF:Asp98 residue is a component of the salt bridge verified by experimental results to be a contributing residue in the allosteric mechanism of HisH-HisF.[30, 77] Although the optimal apo-state path has fewer nodes than the holo-state path, in part due to

the absence of ImGP, it has a longer overall distance. This indicates that the coupled residue?residue motions involved in molecular signaling between the effector site and the active site are stronger when PRFAR is bound.

2.4.5 Community Structure Analysis

Comparison between Different Nodal Methods (Holo-State)

Community structures depict high regional correlations among nodes within the protein, and these regions are modeled as node aggregates or communities. The community subsets may give insights to allosteric signaling, and in particular, node pairs that perform intercommunity signaling, so-called critical nodes, provide possible sites for signal disruption. Although similarities between the α-carbon method and residue c.o.m. were exhibited in the community structures, the residue c.o.m. method indicated a significant difference regarding the community with which ImGP was associated. Furthermore, certain communities were drastically altered (blue in Figure 2.5) between the two methods.

The α-carbon method indicated an intercommunity connection between AICAR and ImGP, and the AICAR moiety was associated with a portion of the $(\beta/\alpha)_8$ barrel (green in Figure 2.6A). The α-carbon method suggested AICAR and ImGP to be constituents of two separate communities (orange and green in Figure 2.6A) and composed of proportional parts of the $(\beta/\alpha)_8$ barrel. In the α-carbon method, an intracommunity connection was exhibited between ImGP and HisF:Cys9, HisF:Gly202, HisF:Gly203, HisF:Leu222, HisF:Ala223, HisF:Ala224, and HisF:Ser225. Intracommunity connections were also detected between AICAR and HisF:Gly82, and HisF:Gly81, which was similar to the c.o.m. method.

In the residue c.o.m. method, the ImGP moiety is predicted to belong to the nodal community comprised of a loop near the C-terminus (black in Figure 2.6B) with intracommunity connections to HisF:Lys19 and HisF:Asp11. The AICAR moiety for the c.o.m. method belonged to the nodal community composed of a portion of the $(\beta/\alpha)_8$ barrel (green in Figure 2.6B) with intracommunity connections to HisF:Gly82, HisF:Gly81, and HisF:Leu50. Although the α-carbon and residue c.o.m. methods differ in important details of the community partitioning,

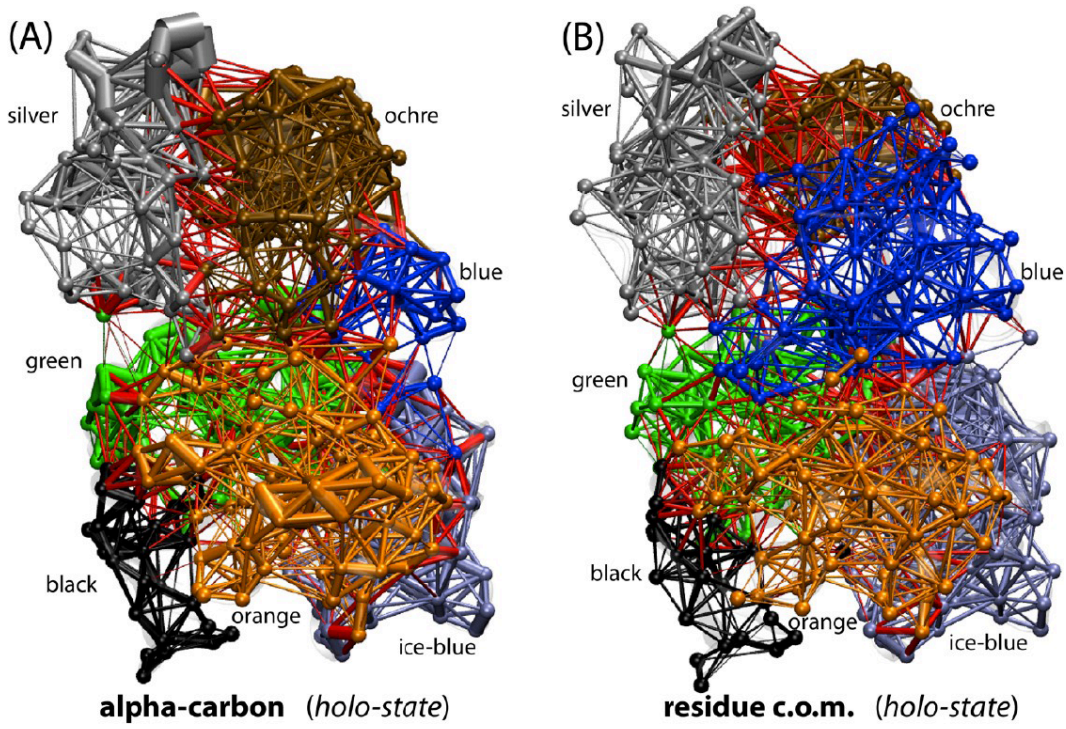


Figure 2.5: (A and B) The optimal communities for communication among residues are depicted, grouped via the Girvan-Newman algorithm. Note the significant change in community structure between the two methods: residue α-carbon (A) and residue c.o.m. (B). There are seven communities in the protein dimers: three for HisH (silver, ochre, and blue) and four for HisF (black, orange, green, and ice blue). The edges that attach critical nodes between communities are shown as red lines and depict high levels of degenerate signaling between community structures. Edge thickness is radially proportional to the weight given by eq 3.

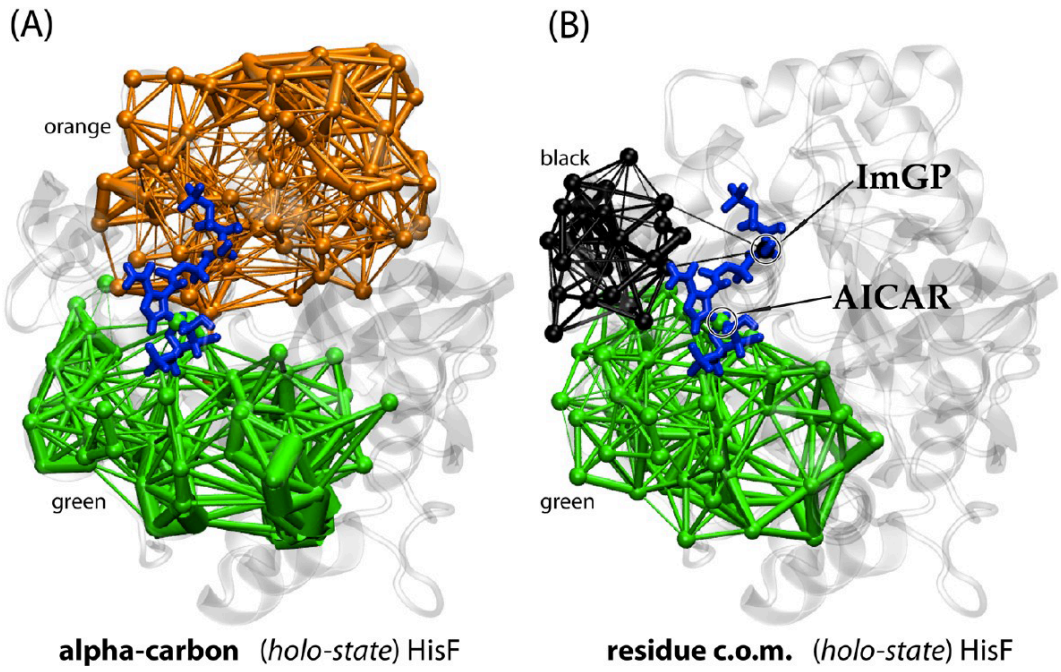


Figure 2.6: (A and B) The change in community structure for connections between PRFAR and other immediate residues is depicted. The α-carbon method (A) shows ImGP (blue) as a member of the orange community and AICAR (blue) as a member of the green community. Both the green and orange communities largely include two halves of the $(\beta/\alpha)_8$ barrel in HisF. The residue c.o.m. method (B) shifts the ImGP moiety (blue) into the black community comprising the loop at the C-terminal end of the $(\beta/\alpha)_8$ barrel.

both show a split between the ImGP and AICAR moieties of PRFAR in agreement with the catalytic mechanism of HisF.

Community Structure Comparison between the Apo- and Holo-States

The residue c.o.m. method was used to derive community structures for both the apo and holo conformational ensembles. Considerable shifts in community aggregates are seen in the HisH domain (Figure 2.7A and B, blue regions). The set of community aggregates also rearranged themselves in the HisF domain between apo and holo states (Figure 2.7A and B, black regions). Both the apo- and holo-states provided seven community aggregates: three community aggregates in the HisH domain and four in HisF. However, differences in community crossover at the dimer interface are present.

The black community in both the apo- and holo-states contains the flexible loop at the C-terminal end of the $(\beta/\alpha)_8$ barrel. The black community is clearly altered between the apo and holo-states (Figure 2.7). The residues that both states have in common are HisF:Asp11 thru HisF:Glu34 and residue HisF:Ile52. Root mean squared deviation (RMSD) analysis was used to quantify the structural differences for this set of common residues between both states over the course of the trajectories (SI Figures 2.3 and 2.4). The average RMSD for the holo-state is 2.166 Å with a standard deviation of 0.412 Å and for the apo-state is 1.416 Å with a standard deviation of 0.322 Å. The addition of PRFAR at the effector site (holo-state) increased the RMSD as well as the fluctuations about the average protein structure of the loop. This corresponds to a decrease in correlations within the black community and provides a distinct shift to the holo-state as seen by the extension of the green community into regions occupied by the black community in the apo-state (Figure 2.7A and B).

Critical Node Comparison between the Apo- and Holo-States

Critical nodes represent strong signaling conduits between communities. Critical nodes were analyzed for both the apo-state and holo-state simulations (Table 3). Four critical nodes were the same across both the apo-state and holo-state

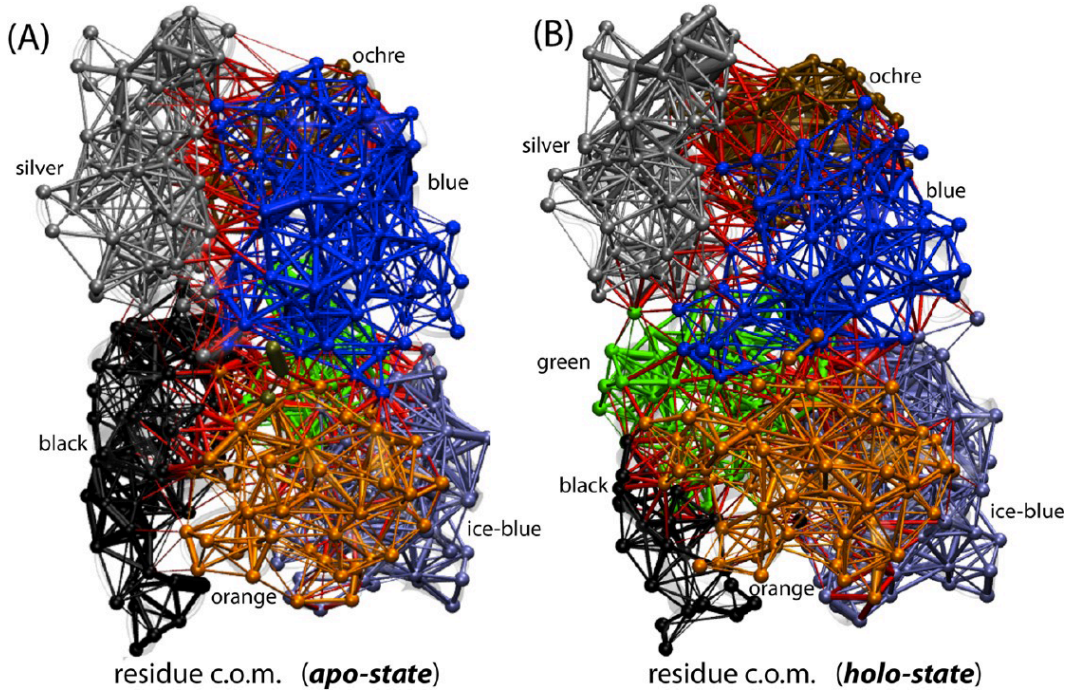


Figure 2.7: (A and B) The optimal communities for communication among residues are depicted for both the apo-state and holo-state using the residue c.o.m. method. There are seven communities in the protein dimers for both the apo-state and holo-state: three largely associated with HisH (silver, ochre, and blue) and four associated with HisF (black, orange, green, and ice blue). However, these communities do extend across the dimer interface. The edges that attach critical nodes between communities are shown as red lines. The black community in the apo-state (A) is larger than that in the holo-state (B). The ImGP moiety is a member of the black community for the holo-state, and this may be the direct cause of the community shift from the larger black community associated with the apo-state to the smaller black community in the holo-state.

in HisF. These nodes were HisF:Arg5, HisF:Glu46, HisF:Lys99, and HisF:Val127. All of these nodes are evolutionarily conserved, except for HisF:Val127, which is partially conserved.[30] For the HisH domain, six critical nodes were the same.

Table 2.3: The critical nodes in a community structure present high signal degeneracy between community aggregates. Here, we compare the critical nodes calculated for the apo-state and holo-state. The preservation of critical nodes across both states (grey columns) may be important to protein structure and function. Residues that are evolutionarily conserved are both bold and underlined, while partial conservation is indicated by just an underline. Residues that are not evolutionarily conserved are written in the column headers as plain text.

HisF Critical Nodes

	Arg	Ile	**Val**	**Gly**	Pro	Leu	Ser	**Asp**	**Glu**	Val	Ala	Glu	Ile	**Pro**	**Asp**	**Lys**	Ile	Ile	Ala	**Gly**
Holo	5		17	30	32	35			46	69		71				99	113	116	117	121
Apo	5	7					40	45	46		70		75	76	98	99				

	Gln	Val	Val	Trp	**Glu**	**Asp**	Thr	Ser	Leu	Ala	**Leu**	Val	Asn	**Arg**	Leu	Glu	Gly
Holo	123	126	127		167	176								249			
Apo			127	156			178	180	211	213	222	246	247		250	251	252

HisH Critical Nodes

	Arg	Val	Gly	Met	Leu	Ile	Ser	**Glu**	Arg	**Trp**	**Tyr**	Tyr	Tyr	**His**	**Lys**	Lys	Ile	Leu	Val	ImGP/ AICAR
Holo	18	51		58		67	94		117	123	138	143	158					189	193	ImGP/ AICAR
Apo	18	51	57		61			96			138	143		178	181	184	185	189	193	none

The evolutionarily conserved node is HisH:Tyr138. The partially conserved nodes are HisH:Val51, HisH:Tyr143, and HisH:Val193.[30] The two nodes that did not show conservation are HisH:Arg18 and HisH:Leu189. subsectionEntropy Calculations between Apo and Holo-States The entropies of the apo- and holo-states were calculated for both the α-carbon and c.o.m. methods using a quasiharmonic approximation.[76] The α-carbon method provided entropy values of 13,192.54 J mol^{-1} K^{-1} and 13,664.46 J mol^{-1} K^{-1} for the respective apo- and holo-states. Whereas the residue c.o.m. method provided entropy values of 27,540.41 J mol^{-1} K^{-1} and 27,905.33 J mol^{-1} K^{-1}. These results show a positive entropy gain, upon PRFAR binding, with values $\Delta S_{\alpha c}$ = 471.92 J mol^{-1} K^{-1} and ΔS_{com}= 364.92 J mol^{-1} K^{-1} for the α-carbon and c.o.m. methods, respectively.

2.5 Discussion

In this work, we have examined four types of networks by varying node representations in Cartesian coordinates: the (i) residue α-carbons, (ii) the side chain c.o.m., (iii) the backbone c.o.m., and the entire (iv) residue c.o.m. We applied these to imidazole glycerol phosphate synthase (IGPS) in order to test for allosteric signaling pathways within the framework of dynamical network analysis. The analysis was carried out at equilibrium conditions, and the resulting pathways were sensitive to which node representation was used.

In our results, the α-carbon method failed to indicate interplay between HisH:Lys181 and HisF:Asp98 for the holostate optimal path ImGP to HisH:Glu180, whereas the residue c.o.m. and side chain c.o.m. methods succeeded in showing this particular interaction at the interface (Table 2.1). Detecting HisH:Lys181 and HisF:Asp98 is a promising feature of the residue c.o.m. methodology because this salt bridge has been experimentally validated to participate in allostery for the cyclase half reaction.[30, 77] Furthermore, HisH:Lys181 and HisF:Asp98 are coupled to the active site of the catalytic triad through close physical proximity and strong correlations. The holo-state of HisH:Lys181 and HisF:Asp98 forms a longlasting contact across the dimer interface, confirmed by our simulations (data not shown), and both residues are directly adjacent to the glutaminase active site.

Mutations of HisF:Thr104 and HisF:Lys19, assessed in previous works,[30, 77] affected the glutaminase half reaction and may be attributed to destroying the functionality of the receptor site where PRFAR binds. In our simulation, the residue HisF:Lys19 is adjacent to the ImGP moiety of PRFAR and the Girvan?Newman algorithm clusters both into the same community when based on the residue c.o.m. methodology (black, Figure 2.7). Trajectories are distinct between the apostate, where HisF:Lys19 is not linked to PRFAR, and the holostate, where HisF:Lys19 is directly linked to the ImGP moiety of PRFAR and indirectly to the $(\beta/\alpha)_8$ barrel (green and orange, Figure 2.7). When analyzing the RMSD for both the apostate and holo-state (SI Figures 2.3 and 2.4), the variations are more stationary for the apo-state and less so for the holo-state over the 20 ns trajectory. The holo-state may comprise specific harmonic and anharmonic motions between the

loop structure and components of the $(\beta/\alpha)_8$ barrel that is mediated by the ImGP moiety of PRAFR. This would imply that HisF:Lys19 plays a role, for the holo-state, via signaling attributes of the loop structure by channeling motion through a direct connection into the optimal path from ImGP to HisH:Glu180. The apo- and holo-state RMSD time series are consistent with previous NMR studies for this loop structure and may indicate signaling attributes enveloped on the millisecond time scale.[78]

Comparison of the different nodal methods indicates that the α-carbon and backbone c.o.m. methods are strongly correlated (Figure 2.3). This analysis suggests a clear bias of the α-carbon to detect the same signals as the residue backbone.

Moreover, using the covariance matrix to calculate quasiharmonic entropy values gives notably different results when using the α-carbon versus the residue c.o.m. method. Collectively, these results suggest that biologically relevant information may be lost if side chain substituents are not taken into account and emphasize the importance of accounting for side chain dynamics in allosteric signaling captured through the use of the residue c.o.m. Including side chain information in dynamical network models may be essential to providing an accurate analysis of allosteric pathways as well as hierarchical clustering of dynamical networks.

2.5.1 Evolved Interdependence of the Two Domains

Interdependence between HisH and HisF exists as a result of protein function. Elegant biochemical kinetic experiments for HisH-HisF have decoupled the two enzyme reactions in vitro. Constructs in bacteria, where the two domains are encoded on separate polypeptide chains, have shown that the cyclase reaction can occur via HisF under nonbiological conditions (50 mM Tris-acetate buffer, pH 8.5, at 25 C), whereas the glutaminase reaction of isolated HisH has no detectable activity.[79] Consequently, residues that form a foundation for the dependence of the two domains may be highly conserved due to evolutionary pressures. Some of these residues are likely along the signaling pathway.

Every residue along the optimal path determined with our residue c.o.m.

analysis, from ImGP to HisH:Glu180, is known to be evolutionarily conserved.[30] This result is in contrast to the α-carbon method, which also included a nonconserved residue HisF:Thr78 in its optimal path (Table 2.1). Furthermore, many residues along the $(\beta/\alpha)_8$ barrel were highly selected for in both the apo-state and holo-state along the optimal paths for all methods. This result is consistent with the conserved nature of the barrel structure across the sub-family of GAT enzymes and is a quantitative argument against possible circuitous routes via other secondary structural elements of the protein. The pathway comparison between the apo- and holo-states (Table 2.2) may also provide insights to basal turnover as HisH-HisF produces low levels of glutamate in the half reaction when PRFAR is not bound.

2.5.2 Community Structure as an Allosteric Descriptor

Overall community structure is a function of intracommunication among a set of residues and consequently provides a coarse grained intercommunication between a set of communities. The community results are derived from different nodal methods and indicate that the choice used to determine the correlation matrices is important (Figures 2.5-7). When using the residue c.o.m. method, a clear shift among community structures is seen between the apo- and holo-states. The loop structure at the base of the $(\beta/\alpha)_8$ barrel directly influences the ImGP moiety of PRFAR as previously mentioned. However, other shifts in community structure may be more relevant. At the dimer interface, clear changes exist between the apo- and holo-state (Figure 2.7).

Critical nodes at the interface of the communities may be adequate descriptors of a proteins state. Even though critical nodes may not be part of the optimal pathway, they are assumed to be important to the configurational ensemble. For example, a critical node may represent a pharmaceutically relevant site for allosteric inhibition. The critical nodes between apo- and holo-states also shift (Table 2.3). However, some residues preserve this critical node quality between both apo and holo-states. For instance, HisF:Lys99 is evolutionarily conserved and quantified as critical for both states. Whereas, HisF:Asp98 and HisH:Lys181

are also conserved but are only critical nodes for the apo-state. The allosteric importace of HisF:Asp98 and HisH:Lys181 has already been established, and their split into separate communities is a feature of the apo state that quantifies the separation and extent of influence between the two residues.

From the community analysis, ImGP is also considered a critical node and carries information from the black community to both the green and orange communities, which are commensurate halves of the $(\beta/\alpha)_8$ barrel. Furthermore, AICAR is also a critical node, and this reflects the flexible characteristics of the PRFAR metabolite as both moieties correlate with different communities (Figure 2.6B).

2.5.3 Limitations of the Model

The optimal pathways for allosteric signaling were based on a contact-map-dependent methodology defined previously.[54, 55] Residues within a particular distance of another residue for some percentage of the simulation are assumed to influence the communication pathway directly, and residues that do not satisfy these constraints are removed from consideration. However, other methods for generating a contact map may improve definitive connections within the protein network. For example, a reasonable alternative contact map may use values that are not Boolean, instead relying on a continuous function that varies depending on the distance between nodes. Alternate methods for defining contacts among residues is beyond the scope of this paper, though we note such efforts may impact results and should be considered for future dynamical network models.

The available high-resolution crystal structures of the glutaminase domain have shown that the oxyanion hole required for glutaminase activity is not properly formed in the absence of the cyclase domain or the effector molecule.[29, 80, 81, 82] In fact, the rearrangement of the active site to a catalytically competent conformation is a hallmark feature of the regulation of nitrogen metabolism across the entire family of glutamine amidotransferases. The energy barrier required to overcome this oxyanion strand rearrangement in HisH-HisF has been estimated at over 20 kcal/mol.[78] Notably, these crystal structures, as well as the simulations presented

here, did not contain the glutamine substrate at the active site of HisH. Our work assesses the effects of PRFAR alone. Further studies are warranted in order to determine whether the inclusion of glutamine within the active site of HisH-HisF, either covalently bound to the active site, cysteine, or simply present locally, would affect the networks and community structures determined from the PRFAR-bound state.

Using a network approach to allostery is both intuitive and instructive for garnering information as to what residues may play crucial roles in an allosteric mechanism. However, there are limitations on how our correlation metrics may provide insight. Not all correlations are taken into account, as there may be correlations that exist but are computationally orthogonal to one another due to bond constraints. Furthermore, the covariance matrix assumes fluctuations about an average structure, but the structural average may not be adequate for anharmonic type motions.

2.6 Conclusions

In this work, two of five residues (HisH:Lys181 and HisF:Asp98), experimentally found to facilitate allostery in HisH-HisF, were determined to be along the optimal pathway of our network model when using correlation information determined from the residue c.o.m. method. The other three residues appear to be related to the disruption of the allosteric state, for which we find strong qualitative and quantitative arguments that HisF:Lys19 is a crucial link between ImGP and the loop structure at the base of the $(\beta/\alpha)_8$ barrel. This loop structure is distinctly different for the apo-state and holo-state and is arguably a feature of the allostery exhibited in HisH? HisF. The α-carbon method was clearly similar to the backbone c.o.m. method and traced out an optimal pathway that did not emphasize the dimer side chain interactions for ImGP (source) through HisH:Glu180 (sink). Using the covariance matrix to calculate quasiharmonic entropy values gives notably different results when using the α-carbon versus the residue c.o.m. method. In principal, quasiharmonic vibrational energies are also dependent on

this choice and may lead to inadequate values for properties of the system, such as heat capacity. As the residue c.o.m. is a composite set of the backbone and side chain groups, our works suggest that this method provides superior information in contrast to the popular α-carbon approach.

2.7 Acknowledgements

This work was funded in part by the National Institutes of Health through the NIH Directors New Innovator Award Program 1-DP2-OD007237 and through the NSF XSEDE Supercomputer resources grant LRAC CHE060073N to REA

Chapter 2 is a minimally modified reprint of the material as it appears in **Van Wart A. T.**, Eargle J., Luthey-Schulten Z., and Amaro R. E., '"Exploring residue component contributions to dynamical network models of allostery"', *J. Chem. Theory Comput.*, (2012), 8, 2949-296. I was the primary investigator and author of this paper.

Chapter 3

Weighted Implementation of Suboptimal Paths (WISP): An optimized algorithm and tool for dynamical network analysis

3.1 Abstract

Allostery can occur by way of subtle cooperation among protein residues (e.g., amino acids) even in the absence of large conformational shifts. Dynamical network analysis has been used to model this cooperation, helping to computationally explain how binding to an allosteric site can impact the behavior of a primary site many angstroms away. Traditionally, computational efforts have focused on the most optimal path of correlated motions leading from the allosteric to the primary active site. We present a program called Weighted Implementation of Suboptimal Paths (WISP) capable of rapidly identifying additional suboptimal pathways that may also play important roles in the transmission of allosteric signals. Aside from providing signal redundancy, suboptimal paths traverse residues that, if disrupted through pharmacological or mutational means, could modulate the allosteric regulation of important drug targets.

To demonstrate the utility of our program, we present a case study describing the allostery of HisH-HisF, an amidotransferase from T. maritima thermotiga. WISP and its VMD-based graphical user interface (GUI) can be downloaded from http://www.nbcr.net/wisp.

3.2 Introduction

Allosteric regulation is a key mechanism whereby proteins respond to environmental stimuli that modulate their activity.[83, 84, 85, 86, 87]. Classic models of allostery (e.g., the MWC[25] and KNF[26] models) suggest that a binding event at an allosteric site induces substantial conformational changes in the primary catalytic site. However, allostery has since been observed in the absence of large-scale conformational changes,[42, 27] suggesting that subtle alterations in protein dynamics can induce a population shift in the conformational ensemble without substantially changing the mean conformation of the protein. This subtle form of allosteric communication can be modeled by dynamical network analysis.

Recent advances in both correlated-residue clustering and dynamical network analysis have helped computationally quantify allosteric states. [88, 89, 54, 90, 91, 92, 93, 94, 95, 96] Dynamical network models of allostery often focus on the single most direct path of residues leading from the allosteric to the primary active site. However, few researchers have considered the state changes of slightly longer (suboptimal) allosteric pathways. The statistical distribution of these additional pathways may be useful for locating accessible residues that, if disrupted via pharmacological or mutational means, could modulate the allosteric regulation of important drug targets.

In this paper, we introduce Weighted Implementation of Suboptimal Paths (WISP), a tool that compliments current dynamical network models of allostery by rapidly calculating the primary communicating path between two residues as well as the slightly longer suboptimal paths. To facilitate use, we have also created a WISP plugin for the popular Visual Molecular Dynamics (VMD) package[97]. WISP has been specifically tested on several operating systems, using several ver-

sions of Python, NumPy, SciPy, and NetworkX (Table 1).[98, 99, 100, 101, 102, 103] The program is open source and can be downloaded from http://www.nbcr.net/wisp.

3.3 Materials and methods

3.3.1 Molecular-Dynamics trajectory input

As input, WISP accepts an aligned molecular dynamics trajectory in the common multi-frame PDB format[104]. Trajectory post-processing is necessary prior to WISP analysis, as most trajectories are not initially aligned or PDB formatted. We often use the freely available Visual Molecular Dynamics (VMD) software package[97] to perform the necessary alignment and conversion.

3.3.2 Generating the correlation matrix

WISP, similar to other dynamical network analysis tools[55], is based on the dynamic interdependence among protein constituents (e.g., amino acids). A protein system is first simplified by representing each constituent as a single node. For example, depending on user-specified WISP parameters, an amino acid can be represented by a node positioned at the residue center of mass, the side-chain center of mass, the backbone center of mass, or the alpha-carbon. As a default, the residue center of mass is used.

The interdependence among nodes is represented as a connecting edge with an associated numeric value that reflects its strength. There are numerous methods for describing the interdependence among nodes in a protein network. Typically, this interdependence is represented by a matrix $\boldsymbol{C}$ with values corresponding to the weights of each edge. By default, WISP generates a N^2 matrix $\boldsymbol{C}$ by calculating the correlated motion among node-node pairs as shown in equations 1 and 2:

$$c_{ij} = \frac{\langle \Delta\vec{r}_i(t) \cdot \Delta\vec{r}_j(t) \rangle}{(\langle \Delta\vec{r}_i(t)^2 \rangle \langle \Delta\vec{r}_j(t)^2 \rangle)^{1/2}} \tag{3.1}$$

$$\Delta\vec{r}_i(t) = \vec{r}_i(t) - \langle \vec{r}_i(t) \rangle \tag{3.2}$$

where N is the number of nodes, i and j are indices corresponding to individual nodes, $\vec{r_i}(t)$ is the location of node i at time t, and c_{ij} is the matrix element at position (i, j).

The absolute value of c_{ij} is larger when the motions of two nodes are highly correlated or anticorrelated. In order to compute signaling pathways, it is useful to construct a matrix where the opposite is true, i.e., where small values indicate highly correlated or anti-correlated motions. Consequently, the correlation matrix is functionalized according to eq. 3, as outlined in previous works.[54, 90]

$$w_{ij} = -log\left(|c_{ij}|\right) \tag{3.3}$$

As a point of clarification, each w_{ij} can be thought of as a distance in functionalized correlation space. Throughout the remainder of this paper, concepts like length and distance will refer to spans in this space, unless specifically described as Cartesian or physical. We further note that, while WISPs default functionalized correlation matrix is generally useful, any user-specified matrix that defines signaling strength as inversely proportional to edge length can be used.

3.3.3 Reducing the complexity of the functionalized correlation matrix

In order to improve the speed of subsequent path-finding steps, the complexity of the functionalized correlation matrix $\boldsymbol{W}$ must be reduced. To this end, two techniques are used. First, a contact-map matrix $\boldsymbol{M}_{\text{contact}}$ identifies entries in $\boldsymbol{W}$ that represent physically distant residues that are unlikely to be truly correlated. By default, $\boldsymbol{M}_{\text{contact}}$ is constructed using p_{cutoff}, a user-specified Cartesian cutoff distance. The average location of each atom over the course of the aligned molecular dynamics trajectory is first calculated, followed by a pairwise Cartesian distance comparison. Two nodes are considered to be in physical contact if the average locations of any of their associated residue atoms come within p_{cutoff} of one another. $\boldsymbol{M}_{\text{contact}}$ entries are set to zero for all node-node pairs that are not

in physical contact. A simplified, functionalized correlation matrix $\boldsymbol{W}_{\text{simp}}$ is then constructed by multiplying $\boldsymbol{W}$ and $\boldsymbol{M}_{\text{contact}}$ element-wise. The entries of $\boldsymbol{W}_{\text{simp}}$ that equal zero represent node-node interactions that are subsequently ignored. Alternatively, users can provide their own $\boldsymbol{M}_{\text{contact}}$ if desired.

Second, to further reduce the complexity of the functionalized correlation matrix $\boldsymbol{W}$, a pruning algorithm identifies nodes that only participate in pathways having lengths in network space that are greater than another user-defined cutoff (d_{cutoff}). As the ultimate goal is to identify suboptimal paths with lengths less than d_{cutoff}, these nodes can be effectively discarded as well. To identify these nodes, we first generate the set of all forced-node paths (FNPs). An FNP is the optimal pathway between two user specified nodes na and nb that is forced to pass through a given third node ni. For any two fixed nodes na and nb, each third node ni is associated with a single FNP. The set of all FNPs can therefore be generated by iterating over all the nodes, ni, of the system.

To calculate an FNP, Dijkstra's algorithm, included in NetworkX[98]21 is first used to identify the optimal paths between $n_a \rightarrow n_i$, and $n_b \rightarrow n_i$, respectively. The FNP has a length equal to the sum of these two constituent paths. Any path between n_a and n_b that passes through n_i must have a length equal to or greater than that of the associated FNP. Consequently, if the length of the FNP is greater than d_{cutoff}, all entries in $\boldsymbol{W}_{\text{simp}}$ associated with n_i are set to zero, so that n_i is effectively ignored.

3.3.4 Calculating suboptimal pathways

Having generated $\boldsymbol{W}_{\text{simp}}$, we are now ready to search for both the single optimal and multiple suboptimal paths between n_a and n_b. Fortunately, the optimal path is fairly easy to identify using Dijkstra's algorithm, mentioned above. In contrast, identifying all suboptimal paths is difficult because the number of possible pathways between n_a and n_b grows rapidly as the total number of nodes increases.

To identify suboptimal paths, a recursive, bidirectional approach is employed. Simultaneous searches start from n_a and n_b (Fig. 3.1, in blue and red,

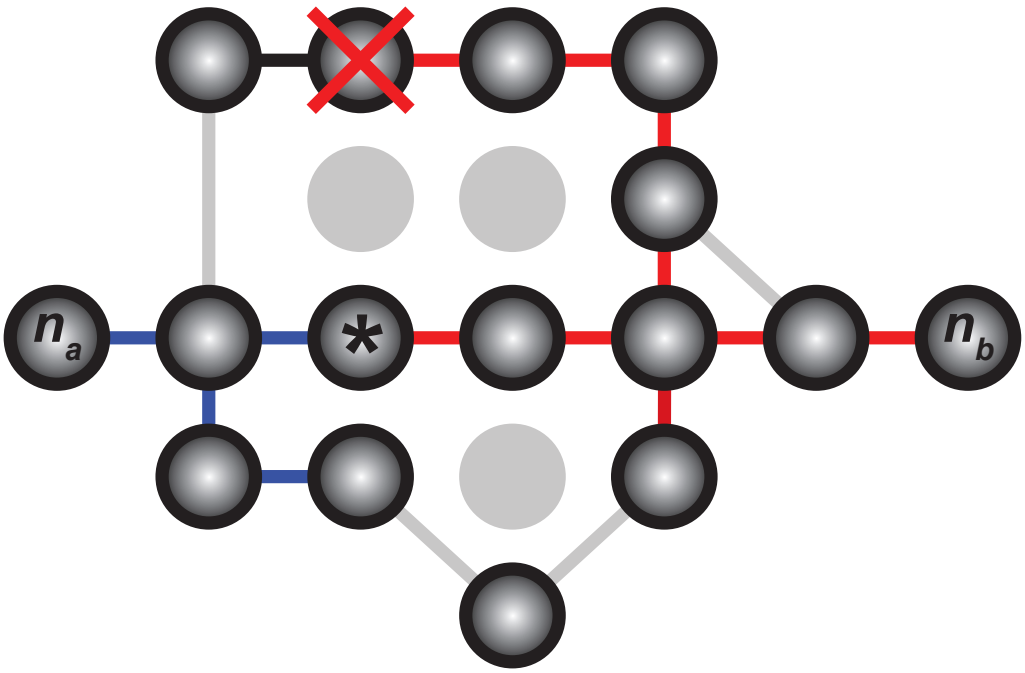


Figure 3.1: Simultaneous searches start from n_a and n_b (blue and red, respectively) and recursively traverse the nodes of the dynamical network. Connections/correlations between nodes that are physically distant are ignored (grey lines). Nodes eliminated using the FNP technique are also ignored (grey circles). As soon as any of the lengthening paths grows too long, that branch of the recursion is killed (red X). At each recursive step, all branches originating from n_a and n_b are compared for common nodes (asterisk). If a common node exists, the two paths are joined. If the length of this composite path is sufficiently short, a suboptimal path has been identified.

respectively) and recursively traverse the nodes of the dynamical network. The recursive algorithm ignores the connections/correlations between nodes that are physically distant (Fig. 3.1, grey lines). Additionally, nodes eliminated using the FNP technique described above are likewise ignored (Fig. 3.1, grey circles). As soon as any of the lengthening paths grows longer than d_{cutoff}, that branch of the recursion is killed (Fig. 3.1, red X). At each recursive step, all branches originating from n_a and n_b are compared for common nodes (Fig. 3.1, the node marked with an asterisk). If a common node exists, the two paths are joined at this node. If the length of this composite path is less than d_{cutoff}, a suboptimal path has been identified. As WISP has been developed to take advantage of multiple processors, running the program on a multi-core system can lead to further speedups beyond the software optimizations described above.

3.3.5 Program output

The program output is a directory containing multiple files, including the specific $\boldsymbol{W}$ and $\boldsymbol{M}_{\text{contact}}$ matrices used. The primary output file is a TCL script that, when loaded into VMD, draws three-dimensional splines representative of the optimal and suboptimal paths. User defined parameters control the relationship between spline thickness, color, opacity and path length. Useful information is also given as comments in the TCL file, including path lengths and participating protein residues.

3.3.6 Graphical user interface

In addition to the command-line program, we have also developed a Visual Molecular Dynamics[97] (VMD) plugin and Tcl-based GUI for easy preparation and visualization of WISP results. The plugin can be accessed through the VMD Extensions menu. The main window of the WISP GUI (Fig. 3.2) allows the user to specify the molecular trajectory and the allosteric-signal source and sink residues. Several additional window interfaces allow the user to modify more advanced program options if needed. All options available through the WISP command-line

interface are available to users of the GUI.

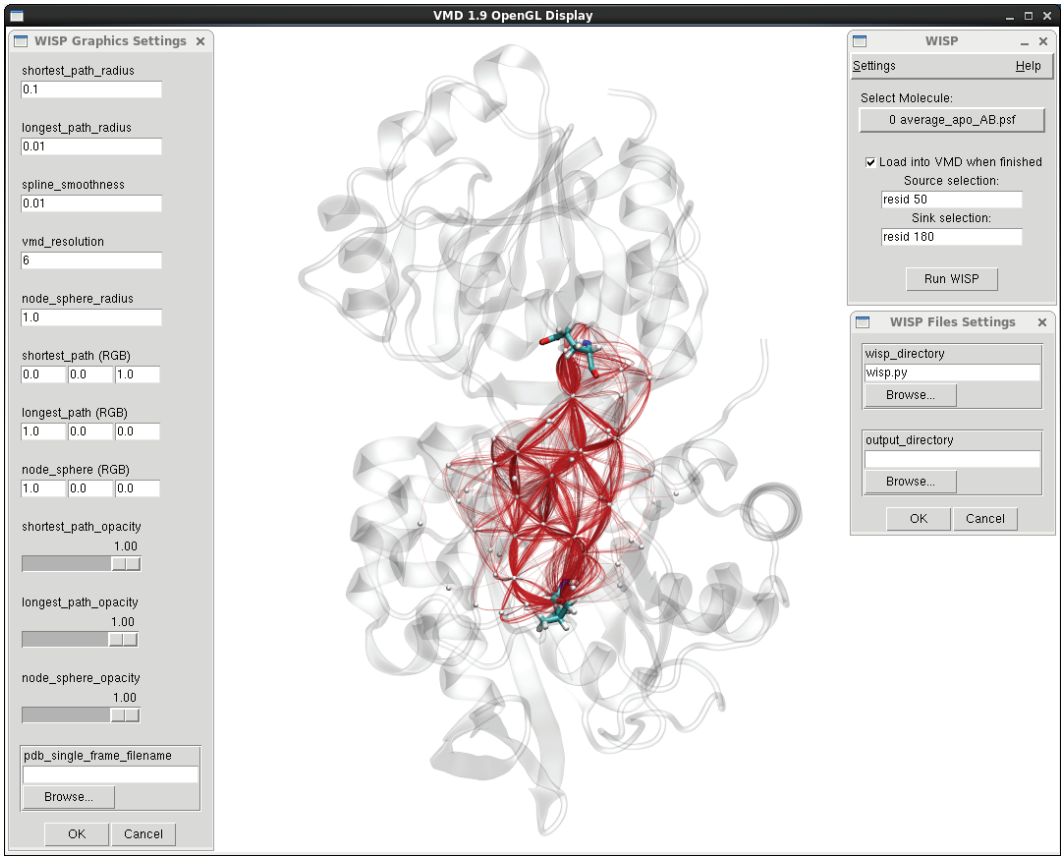


Figure 3.2: In this demonstration, the GUI is used to visualize the allosteric pathways between Leu50:HisF and Glu180:HisH. In the main window (top left), the user selects the relevant molecule, and which residues to use as the source and sink. The user may also select to load the visualization into VMD upon job completion. The setting option windows (left and bottom right) allow the user to specify additional WISP arguments.

Once satisfied with the run specifications, the user may click the Run WISP button at the bottom of the WISP main window to execute the job. The plugin loads the visualization of the allosteric pathways into the main VMD window, where the appearance can be further modified according to the users preferences.

3.3.7 HisH-HisF Details

The molecular dynamics simulations of HisH-HisF used in the current study have been described previously.[90] In brief, a model of the HisH-HisF apo dimer was prepared from the 1GPW[80] crystal structure (Thermotoga maritima). To generate the corresponding holo structure, the 1OX5[64] crystal structure (Saccharomyces cerevisiae), which contains a co-crystallized PRFAR allosteric effector molecule, was aligned to the apo model, effectively positioning PRFAR within the 1GPW:HisF allosteric site. The aligned 1OX5 PRFAR was then merged with the 1GPW-based apo model to yield the corresponding holo structure. Following solvation and equilibration, 20 ns of production dynamics were run for both the apo and holo systems using NAMD[14], the CHARMM27 force field[20], and the same PRFAR parameterization used previously.[73]

3.4 Results/Discussion

Allosteric regulation is crucial to many biological processes. Consequently, one natural strategy for rational drug design is to impede or agonize protein function via allosteric modulation. Classic views of allostery suggest that the binding of an effector molecule at an allosteric site induces large conformational shifts that alter the activity of the primary site. However, as allostery is not necessarily limited to large shifts, this reasoning does not explain some examples of regulation at a distance. For instance, Chung-Jung Tsai et al.[27] recently showed that significant backbone deformations are not required for an allosteric effect; rather, in the absence of large conformational changes, subtle shifts in local dynamics driven by entropic effects[42] govern certain types of allostery.

Quasi-harmonic analysis (e.g., like that used by software packages such as CARMA[76, 75] to calculate entropy) is commonly used to build dynamical network models that quantify signaling pathways among protein constituents. Optimal and suboptimal pathways are calculated that connect protein constituents believed to be important for allostery (i.e., sources and sinks). An optimal pathway is the shortest distance traversed between source and sink along weighted

edges (e.g., as determined by correlated motions), and suboptimal pathways are those closest in length to, but not including, the optimal path. Existing tools can compute optimal and suboptimal pathways between residues [55]36; however, these programs lack the speed required to compute more than fifty suboptimal pathways within a reasonable amount of time (several hours or days). As statistics related to suboptimal pathways may provide important insights that cannot be gleaned from the single optimal pathway, faster algorithmic advances must be made.

WISP was designed to facilitate the calculation of hundreds of suboptimal pathways in minutes, thereby permitting fast and robust statistical analysis of biological systems modeled as dynamical networks. For example, using a modern workstation with 24 cores, we recently used a 20,000-frame trajectory to identify 750 pathways. WISP loaded and analyzed the trajectory, generated the functionalized correlation matrix, and identified the 750 pathways in 21 min and 52 seconds. When the calculation was repeated using a copy of the functionalized correlation matrix saved from the first run, the 750 pathways were identified in only 5 minutes and 44 seconds.

To demonstrate the utility of the WISP algorithm, we used it to study HisH-HisF, a multidomain globular protein known to exhibit allostery. The activity of HisH-HisF, which regulates the fifth step of the histidine biosynthetic pathway in plants, fungi, and microbes, is substantially altered by the allosteric effector N1-[(5'-phosphoribulosyl)-formimino]-5-aminoimidazole-4-carboxamide ribonucleotide (PRFAR).[63] Guided by previous work,[90] we investigated the suboptimal pathways between residues Leu50:HisF and Glu180:HisH using 20-ns molecular dynamics simulations of both apo and holo HisH-HisF. A total of 700 pathways (Fig. 3.3) between Leu50:HisF and Glu180:HisH were calculated using WISPs default correlation (eqs. 1-3) and contact-map matrices, described in the Materials and Methods. Had only the two optimal pathways (apo vs. holo) been considered, we would have concluded that communication between the allosteric and primary site is fundamentally different in the presence and absence of the PRFAR effector molecule (Fig. 3,4). The optimal pathway between Leu50:HisF and Glu180:HisH in the apo state was LEU50:HisF $\rightarrow$ PHE49:HisF $\rightarrow$ PHE77:HisF $\rightarrow$ PRO76:HisF

→ LYS181:HisH → GLU180:HisH. In contrast, the optimal pathway with PRFAR bound was LEU50:HisF → GLY80:HisF → VAL79:HisF → LYS99:HisF → ASP98:HisF → LYS181:HisH → GLU180:HisH.

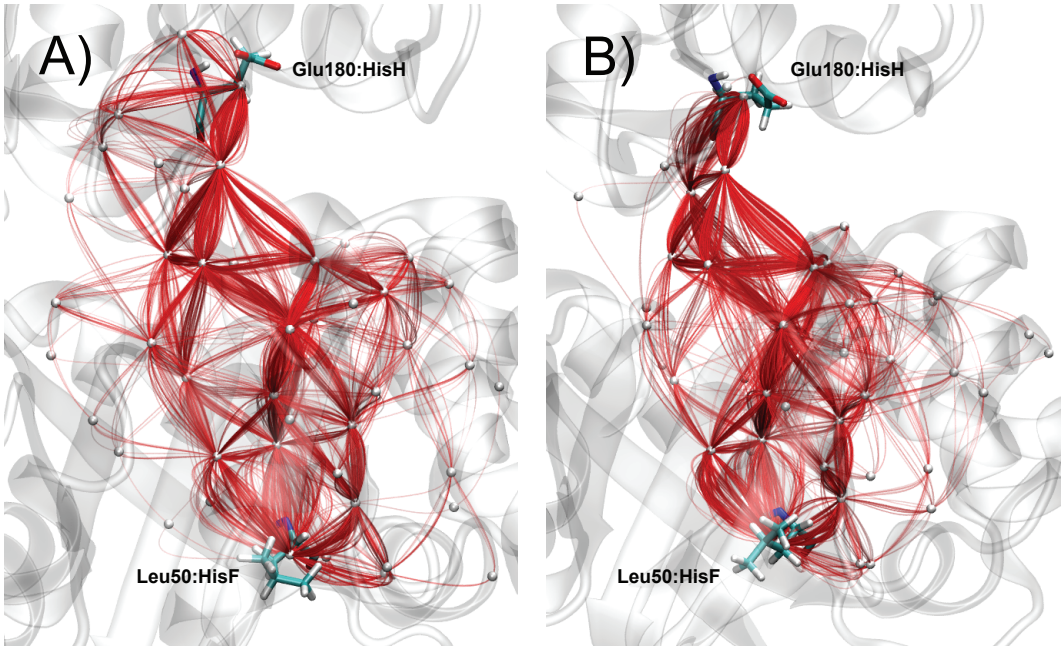


Figure 3.3: The 700 shortest paths between Leu50:HisF and Glu180:HisH, shown as red splines, derived from A) the apo trajectory, and B) the holo trajectory. Wisp allows the user to choose between a number of graphical settings to better visualize signaling among nodes.

However, when multiple suboptimal paths were considered, it became apparent that the allosteric mechanism is in fact far more intricate. The optimal path in the apo simulation is the shortest suboptimal path in the holo simulation (top 0.3%), and the optimal path in the holo simulation is the 13th shortest suboptimal path in the apo simulation (top 2.0%). In light of this multi-pathway analysis, the idea that PRFAR binding fundamentally alters a solitary line of communication between the allosteric and primary site becomes less tenable. Rather, the binding of the effector molecule likely has small effects on multiple pathways, both optimal and suboptimal, that when taken together yield a substantial allosteric effect.

We next sought to characterize the strength of the allosteric effect. The lengths of the two optimal pathways did not differ substantially (apo: 2.97; holo:

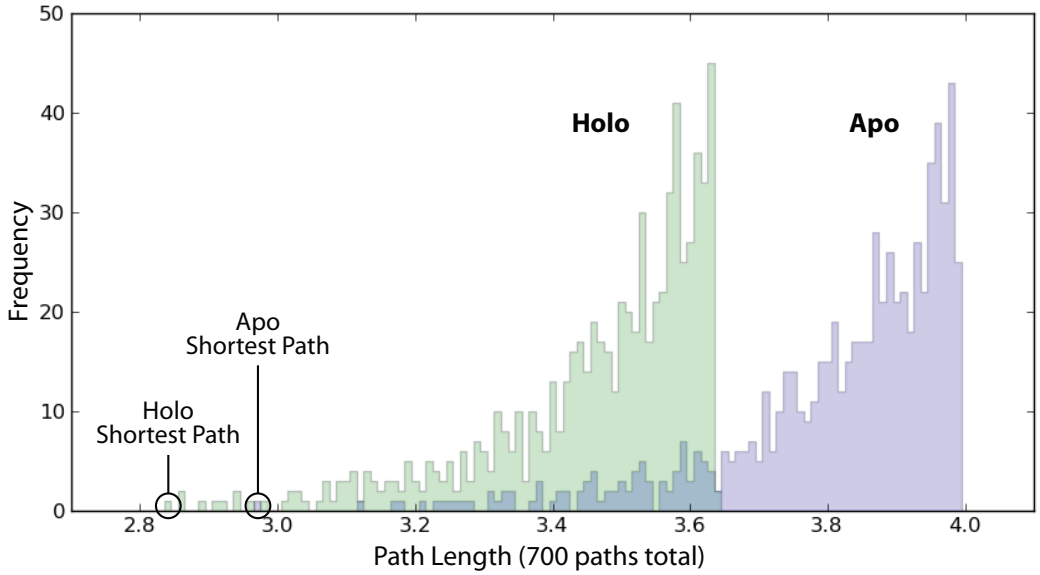


Figure 3.4: A histogram of the 700 path lengths associated with the apo and holo trajectories are shown. The optimal paths are denoted "Shortest Path". The path distribution is largely shifted to the left for the holo (allosteric) state. This shift likely results from a more coherent signal in the holo simulation, indicating a possible decrease in the entropy along the pathways due to PRFAR binding.

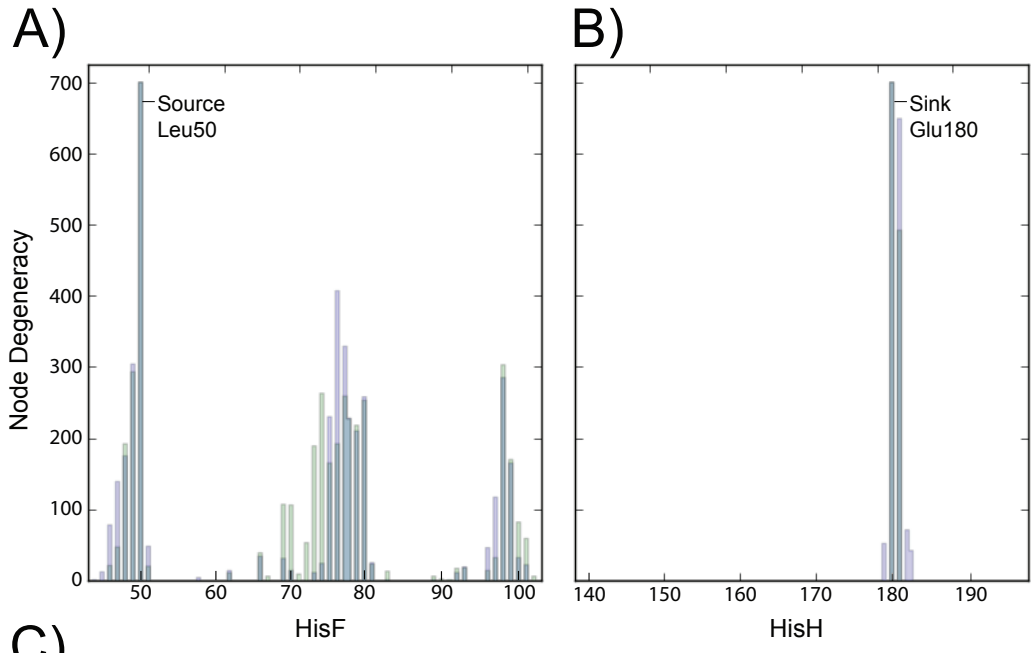
2.84). Consequently, had only these two pathways been considered, some might have mistakenly concluded that the allosteric consequences of PRFAR binding are minor. In contrast, when hundreds of suboptimal paths were also considered, a large PRFAR-dependent shift in communication between the allosteric and primary site became apparent. To demonstrate this shift, we generated a histogram of all path lengths for both the holo and apo simulations (Fig. 3.4). The distribution derived from the holo trajectory is substantially skewed towards shorter path lengths, suggesting that the motions of the residues connecting the allosteric and primary sites are more tightly correlated when PRFAR is bound. A loss of entropy along the pathways may therefore explain the allosteric signal.

To identify protein residues critical for allosteric transmission, we counted the number of times each residue appeared in any of the 700 paths associated with the apo and holo trajectories, respectively (i.e., the degeneracy of each node, Fig. 3.5). Importantly, a number of residues had large effector-molecule dependent shifts in degeneracy (i.e., HisF LEU47, VAL48, ILE75, THR78, and ALA97 from the apo trajectory; and HisF VAL48, ILE73, ILE75, and THR78 from the holo trajectory; see Fig. 5C). Importantly, these residues, which may be crucial for the regulation of protein activity, did not all appear in the optimal apo and holo paths and so would not have been identified had the suboptimal paths been ignored. Compounds that target these residues may serve as useful precursors to future allosteric modulating drugs.

3.5 Conclusion

We present WISP, a program that rapidly calculates both optimal and suboptimal communication pathways between distinct protein residues. The program is available as a VMD plugin or a standalone command-line script. WISP outputs path members and lengths that can be subsequently used in the analysis of path distributions, node degeneracy, etc.

To demonstrate the utility of our program, we presented a dynamical analysis of the HisH-HisF protein. Allosteric modulation in our test case was likely



C)

700 Paths																		
HisF Nodes	**LEU** 47	**VAL** 48	**PHE** 49	**LEU** 50	**VAL** 69	**ALA** 70	**ILE** 73	**ASP** 74	**ILE** 75	**PRO** 76	**PHE** 77	**THR** 78	**VAL** 79	**GLY** 80	**ALA** 97	**ASP** 98	**LYS** 99
Degeneracy Apo	139	175	304	700	31	14	11	24	230	407	329	228	210	258	117	285	165
Degeneracy Holo	47	192	293	700	107	106	189	263	165	192	259	227	218	253	32	303	170

HisH Nodes	**GLU** 180	**LYS** 181
Degeneracy Apo	700	649
Degeneracy Holo	700	492

Figure 3.5: The total number of times a given residue participates in any of the 700 paths (i.e., node degeneracy) is shown in A) HisF and B) HisH. Green indicates the holo state, blue indicates the apo state, and cyan indicates an overlap. Note that Leu50:HisF and Glu180:HisH are in all 700 paths. C) A numerical representation of the same data. The comparison between the apo and holo states shows that certain residues are more sensitive to the allosteric effector PRFAR than others (shaded columns).

the result of subtle changes in multiple suboptimal pathways, rather than large changes in a single optimal path. Additionally, we showed that PRFAR binding causes a large shift towards shorter path lengths (i.e., more correlated motions) in 700 communication pathways between residues HisF:Leu50 and HisH:Glu180, explaining the strong allosteric effects of this modulator (Fig. 4.4). The multiple suboptimal pathways are dominated by a few select residues, as indicated by the shift in node degeneracy between the apo and holo states (Fig. 4.5).

Table 3.1: WISP has been tested on a number of operating systems, using various versions of NumPy, SciPy, and NetworkX. We note that installation under Windows was difficult; however, the command-line version of the program was successfully executed after installing the appropriate dependencies using the ActivePython software package.

Operating System	Python	NumPy	SciPy	NetworkX
Scientific Linux 6.4	2.6	1.7	0.9.0	1.7
Mac OSX 10.6	2.7.2	1.6.1	0.9.0	1.8.1
Ubuntu 12.04	2.7.5	1.7.1	0.12.0	1.8.1
Windows XP	2.7.3	1.7.0rc1	0.11.0	1.8.1

WISP has been successfully tested on a number of platforms (Table 3.1). We are hopeful that the program will be a useful tool for the computational-biology community.

3.6 Acknowledgements

This work was funded by the National Institutes of Health through the NIH Directors New Innovator Award Program DP2-OD007237, and the NSF TeraGrid Supercomputer resources grant RAC CHE060073N to R.E.A. Funding and support from the National Biomedical Computation Resource is provided through NIH P41 GM103426. We would also like to thank Dr. Robert Swift and Dr. John Eargle for their helpful discussions.

Chapter 3 is a minimally modified reprint of the material as it appears in

Van Wart A. T., Durrant J., Votapka L., and Amaro R. E., "Weighted Implementation of Suboptimal Paths (WISP): An optimized algorithm and tool for dynamical network analysis", *J. Chem. Theory Comput.*-(in revision 2013)

Chapter 4

Solving evolution equations using interacting trajectory ensembles

4.1 Abstract

In this paper, we describe a general approach to solving evolution equations for probability densities using interacting trajectory ensembles. Assuming the existence of a positive definite (probabilistic) description of the state of the system, we derive general equations of motion for the trajectories in the kinematic space (e.g., configuration or phase space). The vector field describing the time rate of change of the trajectory ensemble members depends, in general, on both external forces and on the probability density itself. The dependence of the equations of motion on the probability density lead to interactions between the ensemble members and a loss of their statistical independence. The formalism is illustrated by a number of numerical examples. For multidimensional systems, a gauge-like freedom exists in the choice of the underlying vector field, which leaves the evolution of the probability density invariant.

4.2 introduction

The trajectory is a central object in the theoretical description of the time evolution of particles. In classical mechanics [105], the position x of a particle of mass m evolves with time under the forces defining the dynamical system. The function $x(t)$ then traces out the trajectory of the particle in the configuration space in the Newtonian or Lagrangian formulations of dynamics, while the time derivative $v(t) = \dot{x}(t)$ defines the instantaneous velocity. In Hamiltonian dynamics, the corresponding trajectory $x(t) = (q(t), p(t))$ moves in the systems phase space, with the functions $q(t)$ and p(t) solving Hamiltons equations [105],

$$\dot{q} = \frac{\partial H}{\partial p}, \tag{4.1}$$

$$\dot{p} = -\frac{\partial H}{\partial q}, \tag{4.2}$$

where $H(q, p)$ is the Hamiltonian of the system.

The classical mechanics of deterministic systems can be generalized to incorporate the effect of many external bath degrees of freedom in the form of friction and random forces; an example is the Langevin equation [106, 107, 108]

$$m\ddot{x} = -U'(x) - m\gamma_0\dot{x} + f(t), \tag{4.3}$$

where $U'(x)$ is the derivative of the potential energy, γ_0 is the friction constant, and $f(t)$ is a random force. The force and friction are related by the second fluctuation-dissipation theorem:

$$\langle f(0)f(t)\rangle = 2k_BT\gamma_0\delta(t), \tag{4.4}$$

where k_B is Boltzmanns constant, and the average on the left is over realizations of the random force. In this case, the correlation function of the force is a δ-function, indicating Markovian (memoryless) stochastic dynamics; this can be generalized to include finite memory effects. The trajectory $x(t)$ that solves Eq. (3) is itself a random function of time. In Ref. [109], Pollak describes the connection between infinite dimensional Hamiltonian systems and reduced stochastic descriptions, such as the generalized Langevin equation (see also [110]).

In classical mechanics, an alternative view of the dynamics of a system is given by considering statistical distributions in the systems configuration or phase space, consistent with constraints on the system. These distributions connect with the trajectory description as representations of ensembles of trajectories evolving in time with their initial conditions sampled from the initial distribution. This viewpoint is important for describing the equilibrium or non-equilibrium statistical mechanics of many-body systems, where the number of degrees of freedom is much higher than the measurable and controllable macroscopic parameters, such as temperature and volume [106, 107, 108]. A single trajectory in such cases contains too much microscopic detail and is statistically insignificant, so the distribution itself is the central quantity.

For classical deterministic systems, the dynamics are determined by the classical Liouville equation [106]

$$\frac{\partial \rho}{\partial t} = \{H, \rho\}, \tag{4.5}$$

where $\rho(q, p, t)$ is the probability density in phase space, $H(q, p)$ is the system Hamiltonian, and $\{H, \rho\}$ is the Poisson bracket

$$\{H, \rho\} = \frac{\partial H}{\partial q}\frac{\partial \rho}{\partial p} - \frac{\partial \rho}{\partial q}\frac{\partial H}{\partial p}. \tag{4.6}$$

For a classical system in contact with a thermal environment, the generalization analogous to the Langevin equation (3) is the Kramers equation [106, 107, 108],

$$\frac{\partial \rho}{\partial t} = \{H, \rho\} + \gamma_0 \frac{\partial}{\partial p}\left(p + mk_B T \frac{\partial}{\partial \rho}\right)\rho. \tag{4.7}$$

The stochastic force appearing in the trajectory representation, Eq. (3), is here replaced by a diffusive but deterministic term.

The state of a quantum mechanical system is described by its wave function $\psi(x, t)$ whose absolute square $|\psi(x, t)|^2$ gives the configuration space probability density, or by the density operator $\widehat{\rho}(\mathrm{t})$, whose diagonal element $\langle x|\widehat{\rho}|x\rangle$ gives the same quantity, even for statistical mixed states [111]. The density operator obeys the quantum Liouville equation,

$$i\hbar \frac{d\widehat{\rho}}{dt} = [\widehat{H}, \widehat{\rho}], \tag{4.8}$$

where $\widehat{H}$ is the Hamiltonian operator, $[\widehat{H}, \widehat{\rho}]$ is the commutator of the two operators, and $\hbar$ is Plancks constant divided by 2π. In phase space, quantum systems can be described by the Wigner function [112, 113, 114, 115],

$$\rho_w(q,p,t) = \frac{1}{2\pi\hbar}\int_{-\infty}^{\infty}\langle q - \frac{y}{2}|\widehat{\rho}(t)|q + \frac{y}{2}\rangle e^{ipy/\hbar}dy. \tag{4.9}$$

This function is always real, and is analogous to the classical phase space density that solves the classical Liouville equation. The Wigner function is not a true probability density, however, as it can assume negative values.

For quantum mechanical systems, the uncertainty principle complicates the concept of trajectory, as a system cannot have both a definite position $x(t)$ and velocity $\dot{x}(t)$. In phase space, the state of a system cannot be localized beyond the limit $\Delta q \Delta p = \hbar/2$. Despite this, formulations of quantum mechanics in terms of trajectories can be constructed. The path integral formulation expresses the propagator $\langle x,t|exp(-i\widehat{H}(t-t_0)/\hbar)|x_0,t_0\rangle$ in terms of a sum over all paths connecting the points (x,t) and (x_0,t_0) in spacetime [116]. Classical Lagrangian mechanics then emerges from the asymptotic stationary phase limit of the path integral expression. A quantum formulation built explicitly on configuration space trajectories is given by so-called Bohmian dynamics [117, 118, 119, 120], where classical-like trajectories evolve under the influence of a force derived from the sum of the classical potential $U(x)$ and a quantum potential $U_Q(x,t)$ that depends on the wave function of the system:

$$U_Q(x,t) = -\frac{\hbar^2}{2m}\frac{\nabla^2\psi(x,t)}{\psi(x,t)}. \tag{4.10}$$

Treating quantum systems with classical or classical-like trajectories has been pursued since the early days of quantum theory as both a method for solving problems and as a tool for interpretation. In chemical physics, this remains an active area of research, where quantum and semiclassical methods are developed that are based on using classical molecular dynamics to solve the corresponding quantum equations of motion [117, 121, 122, 123, 124, 125].

In a recent series of papers, we proposed and developed a trajectory- based method for solving the quantum mechanical equations of motion for the Wigner

function [126, 126, 127, 128, 129]. The quantum Liouville equation is solved numerically by representing iton averageas a positive continuous distribution function approximated by an ensemble of phase space trajectories. Equations of motion for these trajectories are derived from the exact equation of motion of the Wigner function. The nonclassical nonlocality of quantum mechanics leads to a breakdown of the statistical independence of the trajectories in the ensemble. The motion of the quantum trajectories are entangled with each other through interactions that model quantum nonlocality, and the ensemble must be propagated as a unified whole. The method can be used to simulate the dynamics of model quantum systems with good accuracy. In addition, it provides insight into the nonlocal nature of quantum dynamics and a classical mechanical perspective on the nonclassical phenomena underlying quantum processes.

The nonlocality of quantum mechanics leads to interactions between the trajectories in the method, and this provides an appealing perspective on the uncertainty principle and quantum entanglement as manifested in phase space representations. However, this effect is more general than one due to the nonclassical components of quantum mechanics. Any evolution equations in phase space that incorporate additional terms beyond the Poisson bracket structure of Eq. (5) will lead to nonlocal interactions between ensemble members. Indeed, we illustrated these effects and proposed a corresponding numerical methodology in the context of diffusion equations in phase space, such as Eq. (7), in Ref. [127].

In this paper, we describe a general approach to solving evolution equations for probability densities using interacting trajectory ensembles. Assuming the existence of a positive definite (probabilistic) description of the state of the system, we derive general equations of motion for the trajectories in the kinematic space (e.g., configuration or phase space). The equations of motion for the ensemble members depend, in general, on both external forces and on the probability density itself. This second part then leads to interaction terms between the ensemble members.

The organization of the rest of the paper is as follows. In Section 4.3, we describe the general formalism for the case of distributions in one dimension.

The equations of motion for a continuous distribution are derived, and a variational approach is implemented to derive the equations for a finite sampling of the underlying distribution by a trajectory ensemble. In Section 4.4, the approach is applied to modeling a solution of the one- and two-dimensional Smoluchowski equations describing diffusion under the influence of time-independent and time-dependent potentials. The results are compared with the results of conventional Brownian dynamics simulations. In Section 4.5, we discuss trajectory representation of evolution equations in higher dimensions related to a gauge-like freedom in the choice of the resulting vector field. The general principles are illustrated with several examples. Finally, a summary and discussion are given in Section 4.6.

4.3 Formalism

We start with a general evolution equation in one space coordinate x and time t (the generalization to higher dimensions is given below):

$$\frac{\partial \rho}{\partial t} = \widehat{L}\rho. \tag{4.11}$$

Here, $\rho(x,t)$ is a positive probability distribution and $\widehat{L}$ is a linear operator. The solution $\rho(x,t)$ can be expressed trivially as the convolution of itself with a delta function:

$$\rho(x,t) = \int \delta(x-\xi)\rho(\xi,t)d\xi. \tag{4.12}$$

We now assume that the solution at time t can be written in the form

$$\rho(x,t) = \int \delta(x - X(\xi,t))\rho(\xi,0)d\xi, \tag{4.13}$$

where $\delta(x - X(\xi,t))$ is a propagator that evolves initial density from point ξ at $t = 0$ to point x at time t. $X(\xi,t)$ satisfies the initial condition $X(\xi,0) = \xi$. This form preserves the normalization of the density:

$$\int \rho(x,t)dx = \int \rho x, 0dx. \tag{4.14}$$

We take Eq. (13) as our ansatz, recognizing that the equations of motion for the function $X(\xi,t)$ will be complicated, depending in general on the entire history of the evolution of $\rho(x,t)$.

We now derive these equations of motion. From Eq. (13), the time derivative of the density becomes

$$\frac{\partial \rho(x,t)}{\partial t} = -\int \dot{X}(\xi,t)\delta'(x - X(\xi,t))\rho(\xi,0)d\xi, \tag{4.15}$$

where $\delta'(x - X) = (\partial/\partial x)\delta(x - X)$. The derivative with respect to x acting on the delta function can come out from under the integral, giving

$$\frac{\partial \rho(x,t)}{\partial t} = -\frac{\partial}{\partial x}\int \dot{X}(\xi,t)\delta(x - X(\xi,t))\rho(\xi,0)d\xi, \tag{4.16}$$

The quantity $\dot{X}(\xi,t)$ is the velocity of the point $X(\xi,t)$, which is the point that evolves from ξ as time increases from 0 to t. We can thus write $\dot{X}(\xi,t) = \dot{X}(X(\xi,t)) \equiv \dot{X}(X,t)$. (Note that with this definition of notation we do not mean to imply that $X = \xi$. Rather, we express the vector field as a function of the current position X and time t, rather than as a function of the initial position ξ and t.) The delta function allows us to write $\dot{X}(X,t) = \dot{X}(x,t)$ and the integral then yields

$$\frac{\partial \rho(x,t)}{\partial t} = -\frac{\partial}{\partial x}[\dot{X}(x,t)\rho(x,t)], \tag{4.17}$$

or in terms of the current $j(x,t) = \dot{X}(x,t)\rho(x,t)$,

$$\frac{\partial \rho(x,t)}{\partial t} + \frac{\partial j(x,t)}{\partial x} = 0, \tag{4.18}$$

This is just the equation of continuity. It allows us to cast the evolution equation in the form

$$-\frac{\partial}{\partial x}[\dot{X}(x,t)\rho(x,t)] = \widehat{L}\rho(x,t). \tag{4.19}$$

The vector field as a function of x at time t then becomes, by integration

$$\dot{X}(x,t) \equiv V(x,t) = -\frac{1}{\rho(x,t)}\int^{x} \widehat{L}\rho(y,t)dy, \tag{4.20}$$

These equations of motion are then integrated to propagate the density using Eq. (13).

As a check on our results, consider an infinitesimal time propagation of the density. Under the evolution equation, an infinitesimal propagation is

$$\rho(x,t+\epsilon) \simeq \rho(x,t) + \epsilon\widehat{L}\rho(x,t), \tag{4.21}$$

which becomes, in terms of our ansatz,

$$\rho(x, t+\epsilon) \simeq \int \delta(x - \xi - \epsilon \dot{X}(\xi, 0))\rho(\xi, t)d\xi, \tag{4.22}$$

where we have used the short time expression

$$X(\xi, t+\epsilon) \simeq \xi + \epsilon \dot{X}(\xi, t) \tag{4.23}$$

for the trajectory evolution. Expanding the delta function gives

$$\delta(x - \xi - \epsilon \dot{X}(\xi, t)) \simeq \delta(x - \xi) - \epsilon \dot{X}(\epsilon, t)\delta'(x - \xi), \tag{4.24}$$

where $\delta'(x - \xi) = (\partial/\partial x)\delta(x - \xi)$. Inserting this gives

$$\rho(x, t+\epsilon) \simeq \rho(x, t) - \epsilon \int \dot{X}(\xi, t)\delta'(x - \xi)\rho(\xi, t)d\xi. \tag{4.25}$$

Then, using

$$\dot{X}(\xi, t) = -\frac{1}{\rho(\xi, t)} \int^{\xi} \widehat{L}\rho(y, t)dy, \tag{4.26}$$

we obtain

$$\rho(x, t+\epsilon) \simeq \rho(x, t) + \epsilon \int \left[\int^{\xi} \widehat{L}\rho(y, t)dy \right] \delta'(x - \xi)d\xi. \tag{4.27}$$

We now note that $(\partial/\partial x)\delta(x - \xi) = -(\partial/\partial \xi)\delta(x - \xi)$, and integrate by parts with respect to ξ. This gives

$$\rho(x, t+\epsilon) \simeq \rho(x, t) + \epsilon \int \widehat{L}\rho(\xi, t)\delta(x - \xi)d\xi. \tag{4.28}$$

Finally, performing the integral reproduces Eq. (21), as it should. Numerical implementation involves sampling the density with a finite number of points,

$$\rho(x, t) = \frac{1}{N} \sum_{j=1}^{N} \delta(x - x_j(t)), \tag{4.29}$$

0 and then propagating the points using

$$\dot{x}_j(t) = V(x_j(t), t), \tag{4.30}$$

where $V(x, t)$ is given by Eq. (20). The function ρ and its derivatives must be estimated at each point x_j from the evolving ensemble.

In practice, a finite ensemble of N trajectories is employed to represent the evolving probability distribution. In addition, smoothing functions $\phi(x)$ must be used for interpolating the solution between ensemble members, allowing derivatives to be computed. Here we assume for simplicity the same smoothing function for each member of the ensemble; this can be generalized in numerical implementation.

We now derive the equations of motion for a finite smoothed ensemble using a variational approach, and investigate its equivalence to the analysis given above in the continuum limit. We take an ansatz for the density in terms of an ensemble of trajectories:

$$\rho(x,t) = \frac{1}{N}\sum_{j=1}^{N}\phi(x - x_j(t)) \tag{4.31}$$

where $\phi(x)$ is the smoothing function (in practice, a Gaussian). Inserting Eq. (31) into the evolution equation, Eq. (11), gives

$$\frac{\partial\rho}{\partial t} = -\frac{1}{N}\sum_{j=1}^{N}\dot{x}_j\phi'(x - x_j(t)), \tag{4.32}$$

where $\phi'(x) = d\phi(x)/dx$. We now define an error function $\epsilon(\dot{x}_1, \dot{x}_2, ..., \dot{x}_N)$ as

$$\epsilon = \int\left(\frac{\partial\rho}{\partial t} - \widehat{L}\rho\right)^2 dx. \tag{4.33}$$

This becomes

$$\epsilon(\dot{x}_1, \dot{x}_2, ..., \dot{x}_N) = \int\left[\sum_{j=1}^{N}\dot{x}_j\phi'(x - x_j) + \widehat{L}\phi(x - x_j)\right]^2 dx. \tag{4.34}$$

This error is to be minimized with respect to the velocities $\dot{x}_k$. This gives

$$\frac{\partial\epsilon}{\partial\dot{x}_k} = 2\int\phi'(x - x_j)\left[\sum_{j=1}^{N}\dot{x}_j\phi'(x - x_j) + \widehat{L}\phi(x - x_j)\right]dx = 0. \tag{4.35}$$

Thus, a system of linear equations for the velocities $\dot{x}_k$ results:

$$\sum_{j=1}^{N}(\Delta_{kj}\dot{x}_j + \Lambda_{kj}) = 0, \tag{4.36}$$

where

$$\Delta_{kj} = \int\phi'(x - x_k)\phi'(x - x_j)dx, \tag{4.37}$$

$$\Lambda_{kj} = \int \phi'(x - x_k)\widehat{L}\phi(x - x_j)dx. \tag{4.38}$$

Defining the vector λ with components

$$\lambda_k = \sum_{j=1}^{N} \Lambda_{kj} \tag{4.39}$$

we can write Eq. (36) as

$$\Delta\dot{x} = -\lambda, \tag{4.40}$$

which gives the formal solution for $\dot{x}$ as

$$\dot{x} = -\Delta^{-1}\lambda. \tag{4.41}$$

It is instructive to derive these quantities explicitly for a simple case: the diffusion equation in one dimension. Here, the operator $\widehat{L}$ is given by

$$\widehat{L} = D\frac{\partial^2}{\partial x^2}. \tag{4.42}$$

We take the smoothing function to be a Gaussian: $\phi(x) = exp(-ax^2)$. Then, the elements of the matrices Δ and Λ can be computed analytically. We find

$$\Delta_{kj} = a\sqrt{\frac{\pi}{2a}}\left[1 - a(x_k - x_j)^2\right] e^{-\frac{1}{2}a(x_k - x_j)^2}, \tag{4.43}$$

$$\Lambda_{kj} = -Da^2\sqrt{\frac{\pi}{2a}}\left[3(x_k - x_j) - a(x_k - x_j)^3\right] e^{-\frac{1}{2}a(x_k - x_j)^2}. \tag{4.44}$$

We can compare the predictions of this variational approach with the results obtained previously by taking the continuum limit of the former. The first term of Eq. (36) becomes

$$\sum_j \Delta_{kj}\dot{x}_j \to \int \rho(\xi, t)\left[\int \phi'(y - x)\phi'(y - \xi)dy\right]\dot{x}(\xi, t)d\xi, \tag{4.45}$$

where $x_j \to x$ and the sum over j becomes an integral over ξ with the weighting factor of the instantaneous density $\rho(\xi, t)$. We can further let the smoothing functions $\phi(x - y) \to \delta(x - y)$, yielding

$$\sum_j \Delta_{kj}\dot{x}_j \to \int \rho(\xi, t)\left[\int \delta'(y - x)\delta'(y - \xi)dy\right]\dot{x}(\xi, t)d\xi. \tag{4.46}$$

Noting that

$$\delta'(x-\xi) = \frac{\partial}{\partial x}\delta(x-\xi) = -\frac{\partial}{\partial x}\delta(x-\xi), \tag{4.47}$$

we can integrate by parts to find

$$\int \rho(\xi,t)\left[\int \delta'(y-x)\delta'(y-\xi)dy\right]\dot{x}(\xi,t)d\xi \\ = \int\int \left(\frac{\partial}{\partial x}\left[\dot{x}(\xi,t)\rho(\xi,t)\right]\right)\delta'(y-x)\delta(y-\xi)dyd\xi \tag{4.48}$$

and evaluating the integral over ξ gives

$$\sum_j \Delta_{kj}\dot{x}_j \to \int \left(\frac{\partial}{\partial y}[\dot{x}(y,t)\rho(y,t)]\right)\delta'(y-x)dy \tag{4.49}$$

or, following a second integration by parts, we obtain finally,

$$\sum_j \Delta_{kj}\dot{x}_j \to -\frac{\partial^2}{\partial x^2}\left[\dot{x}(x,t)\rho(x,t)\right]. \tag{4.50}$$

For self-adjoint operators $\widehat{L}$, a similar analysis yields

$$\sum_j \Lambda_{kj} \to -\frac{\partial}{\partial x}\left[\widehat{L}\rho(x,t)\right]. \tag{4.51}$$

Putting the pieces together yields

$$\frac{\partial^2}{\partial x^2}[\dot{x}(x)\rho(x)] = -\frac{\partial}{\partial x}\left[\widehat{L}\rho(x)\right]. \tag{4.52}$$

This is the derivative with respect to x of the continuity equation form of the evolution equation.

$$-\frac{\partial j(x,t)}{\partial x} = -\frac{\partial}{\partial x}\left[\dot{x}(x,t)\rho(x,t)\right] = \widehat{L}\rho(x,t). \tag{4.53}$$

Integrating Eq. (52) with respect to x twice (and setting integration constants equal to zero) then gives the same result as the equations of motion derivation:

$$\dot{x}(x,t) = -\frac{1}{\rho(x,t)}\int^x \widehat{L}\rho(y,t)dy. \tag{4.54}$$

The one-dimensional analysis given above can be generalized to multidimensional systems. Considering an N dimensional space spanned by Cartesian coordinates $= (x_1, x_2, ..., x_N)$, we have the continuity equation

$$\frac{\partial\rho(\boldsymbol{x},t)}{\partial t} + \nabla\cdot\boldsymbol{j}(\boldsymbol{x},t) = 0 \tag{4.55}$$

and the relation between the vector field and the evolution equation

$$-\nabla \cdot \left[\dot{\boldsymbol{X}}(\boldsymbol{x},t)\rho(\boldsymbol{x},t)\right] = \widehat{L}\rho(\boldsymbol{x},t). \tag{4.56}$$

In the multidimensional case, there is considerable freedom in the integration of this partial differential equation. In general, we can decompose the operator $\widehat{L}$ into N terms, as

$$\widehat{L} = \sum_{j=1}^{N} \widehat{\lambda}_j \tag{4.57}$$

(some of the terms can be chosen to equal zero). Each of these terms can be paired with one term of the divergence $\nabla \cdot \left[\dot{\boldsymbol{X}}(\boldsymbol{x},t)\rho(\boldsymbol{x},t)\right]$ on the left side of Eq. (56), yielding a set of N partial differential equations:

$$\frac{\partial}{\partial x_j}(\dot{X}_j(\boldsymbol{x},t))\rho(\boldsymbol{x},t) = -\widehat{\lambda}_j\rho(\boldsymbol{x},t) \tag{4.58}$$

for $j = 1, 2, ..., N$. The components of the vector field as functions of x at time t then become, by integration,

$$\dot{X}_j(\boldsymbol{x},t) \equiv V_j(\boldsymbol{x},t) = -\frac{1}{\rho(\boldsymbol{x},t)} \int^{x_j} \widehat{\lambda}_j \rho(x_a, x_2, ..., y_j, ..., x_n, t) dy_j \tag{4.59}$$

$(j = 1, 2, ..., N)$.

There is great freedom in the partition given in Eq. (57). For a separable operator $\widehat{L} = \sum_j \widehat{L}_j(x_j)$, it is natural (but not necessary) to associate each $\widehat{\lambda}_j = \widehat{L}_j(x_j)$. Alternatively, each $\widehat{\lambda}_j$ can be assigned to a fraction of $\widehat{L}$ itself:

$$\widehat{\lambda}_j = \alpha_j \widehat{L}, \tag{4.60}$$

where the sum $\sum_j \alpha_j = 1$. Many other decompositions can be contemplated.

4.4 Numerical examples

We now illustrate the general formalism described above by showing the results of numerical solution of evolution equations using the interacting trajectory

method, and compare the results with the standard method of Brownian dynamics. We treat evolution equations of the form

$$\frac{\partial \rho(x,t)}{\partial t} = D\frac{\partial}{\partial x}\left(\frac{1}{k_b T}\frac{\partial U(x,t)}{\partial x} + \frac{\partial}{\partial x}\right)\rho(x,t) \tag{4.61}$$

in one dimension, or

$$\frac{\partial \rho(\boldsymbol{x},t)}{\partial t} = D\nabla\cdot\left(\frac{1}{k_b T}\nabla U(\boldsymbol{x},t) + \nabla\right)\rho(\boldsymbol{x},t) \tag{4.62}$$

in multiple dimensions, where $D = k_B T/m\gamma$. These correspond to diffusion in overdamped systems, as described by the Smoluchowski equation [107].

For the one-dimensional case, the interacting trajectory vector field becomes

$$\dot{X}(x,t) = -\frac{D}{k_B T}\frac{\partial U(x,t)}{\partial x} - \frac{D}{\rho(x,t)}\frac{\partial \rho(x,t)}{\partial x}. \tag{4.63}$$

In higher dimensions this generalizes to

$$\dot{\boldsymbol{X}}(\boldsymbol{x},t) = -\frac{D}{k_B T}\nabla U(\boldsymbol{x},t) - \frac{D}{\rho(\boldsymbol{x},t)}\nabla\rho(\boldsymbol{x},t). \tag{4.64}$$

The corresponding Langevin equation describing the Brownian motion of the system is

$$\dot{x}(t) = -\frac{D}{k_B T}\frac{\partial U(x,t)}{\partial x} + \frac{1}{m\gamma}R(t), \tag{4.65}$$

where $R(t)$ is a random force. The generalization to higher dimensions is

$$\dot{\boldsymbol{x}}(t) = -\frac{D}{k_B T}\nabla U(\boldsymbol{x},t) + \frac{1}{m\gamma}\boldsymbol{R}(t). \tag{4.66}$$

We compare numerical solutions of the equations of motion for the interacting trajectory ensembles with an ensemble of trajectories evolving under the stochastic Langevin equation. We consider two one-dimensional potentials and a two-dimensional system.

The first one-dimensional model consists of a linear plus cosine potential, given by

$$U(x) = 1 - 0.15x - cos(0.6x). \tag{4.67}$$

The evolving density $\rho(x,t)$ is approximated by an ensemble of N trajectories $x_j(t)(j = 1, 2, ..., N)$ and smoothed using

$$\rho(x,t) = \frac{1}{N}\sum_{j=1}^{N}\phi(x - x_j(t)) \tag{4.68}$$

with the smoothing functions $\phi(x) = (1/\sqrt{2\pi h})exp(-x^2/2h^2)$. The widths h of the smoothing functions are given by $h = [4/N(d+2)]^{1/(d+4)}$, where $d = 1$ is the dimension of the system [130]. We take $m = 1$, $\gamma = 1$, and $k_BT = 3.0$ in our simulations. Both the interacting trajectory and Brownian dynamics simulations are performed using 1000 trajectories with initial conditions sampled from a Gaussian distribution centered at $x = 0$ with a standard deviation $\sigma = 1$. The interacting trajectory equations of motion are solved using a RungeKutta RK4 algorithm [131] with timestep of $\Delta t = 0.02$. The stochastic Langevin equation is integrated using the method of Ermak, as described in Ref. [132].

The results of the simulations are shown in Fig. 4.1. Four snapshots of the evolving density $\rho(t)$ for the interacting trajectory method are shown as solid curves. These are compared with the Brownian dynamics simulation, shown as dashed curves. The two results are nearly superimposable. Also depicted in the figure is the potential $U(x)$. The system starts initially ($t = 0$) in one well of the metastable system. The system undergoes diffusion and directed transport from left to right, down the energy stairway, as time increases.

The second one-dimensional model is a time-dependent sawtooth potential $U(x,t) = u(x)g(t)$ modeling the operation of a Brownian ratchet [133, 134]. The spatial part is periodic in space with period $L : u(x+L) = u(x)$. Between $x = 0$ and $x = L$ it is given by

$$u(x) = \begin{cases} a(x/x_s), & 0 < x < x_s, \\ a(1-(x-x_s)/x_L), & x_s < x < L, \end{cases} \tag{4.69}$$

where $x_s + x_L = L$. In our calculations $x_s = 2$ and $x_L = 10$. The function $g(t+\tau) = g(t)$ switches periodically between 1 and 0. The period τ is taken to be 13.5, and during $(0,\tau)$ the function is given by

$$g(t) = \begin{cases} 1, & 0 < t < 10.0, \\ 0, & 10.0 < t < 13.5. \end{cases} \tag{4.70}$$

The barrier height $a = 8k_BT$, and we take $k_BT = 2.0$. Other parameters are the same as for Fig. 4.1.

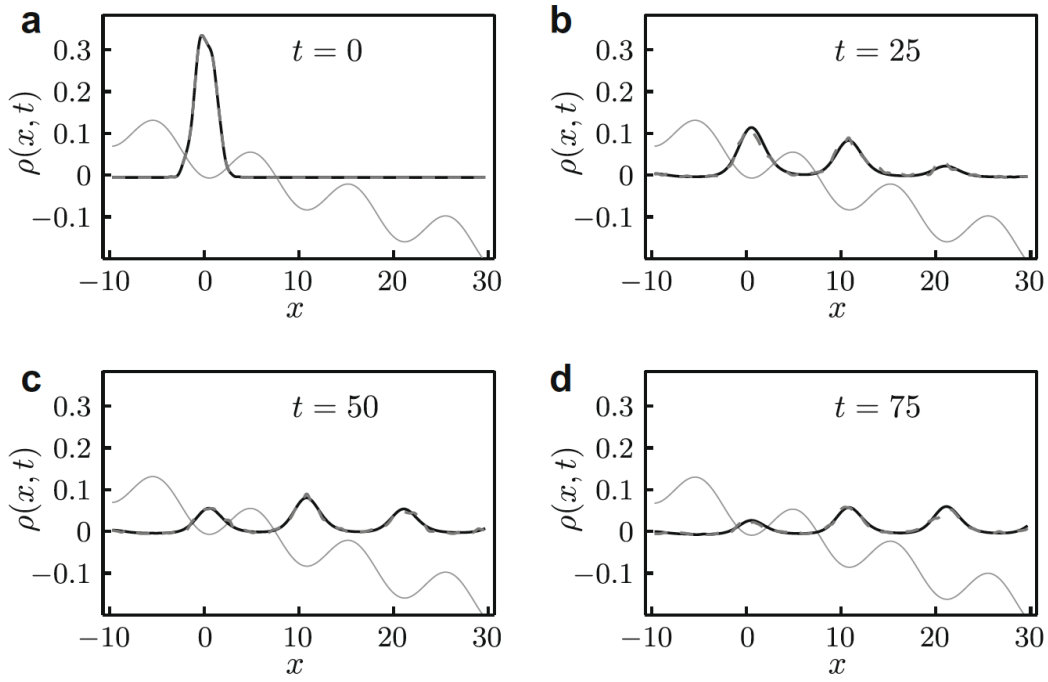


Figure 4.1: Comparison of interacting trajectory ensemble simulation and Brownian dynamics simulation for the linear plus cosine potential, Eq. (4.67). Four snapshots of the evolving probability distributions are shown. The interacting trajectory method is shown as solid curves. These are compared with the Brownian dynamics simulation, shown as dashed curves. (a) $t = 0$. (b) $t = 25$. (c) $t = 50$. (d) $t = 75$.

The results of this model are shown in Fig. 4.2. Initially, the system is localized in one minimum of the sawtooth potential. At $t = 10.0$, the potential switches to zero, and the system undergoes diffusion. At $t = 13.5$, the potential turns back on, and its asymmetric shape causes a net transport of the ensemble from left to right. The convention for the plot is the same as in Fig. 4.1, and, again, close agreement between interacting trajectory and Brownian dynamics results is observed.

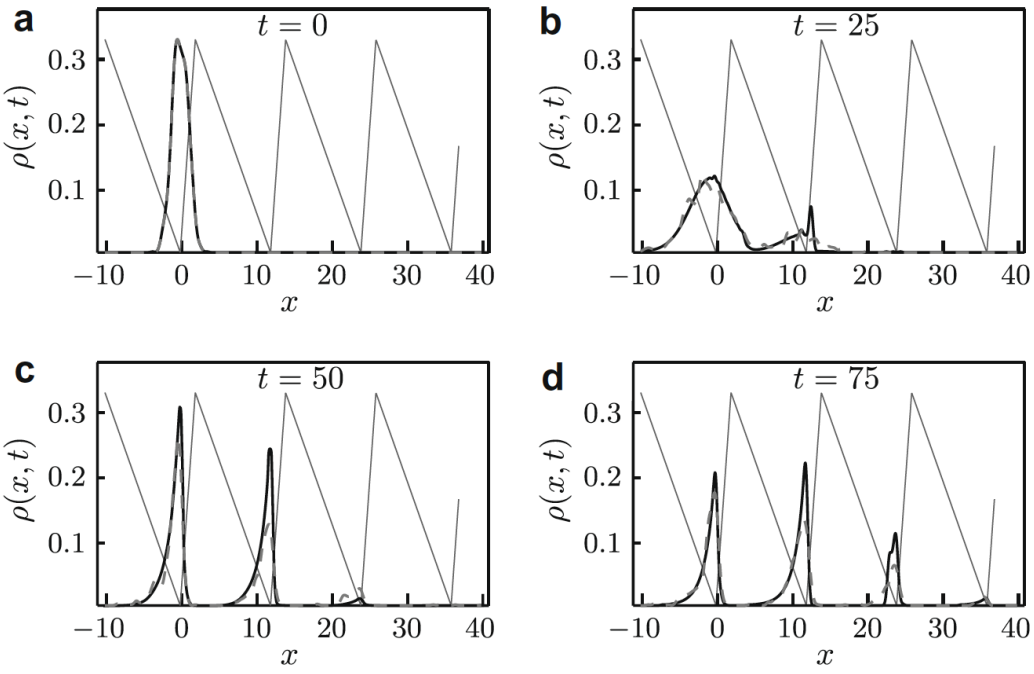


Figure 4.2: Comparison of interacting trajectory ensemble simulation and Brownian dynamics simulation for the time-dependent Brownian ratchet potential. Four snapshots of the evolving probability distributions are shown. The interacting trajectory method is shown as a solid curve, and compared to the Brownian dynamics simulation, shown as a dashed curve. (a) $t = 0$. (b) $t = 25$. (c) $t = 50$. (d) $t = 75$.

We now consider a two-dimensional model, constructed by adding a harmonic term in y to the one-dimensional linear plus cosine system. The frequency of the harmonic oscillator depends on x.

$$U(x,y) = \frac{1}{2}(1.15 - 0.03x)y^2 + (1 - 0.15x - cos(0.6x)). \tag{4.71}$$

The same parameters and method are employed to solve the equations, but with $N = 2000$ trajectories for the interacting ensemble method. We find that to get smooth results, 5000 trajectories are required for the Brownian simulations.

The two-dimensional results are shown in Fig. 4.3. Net motion down the energy stairway, from left to right, is observed, as in the one-dimensional case of Fig. 4.1. In addition, spreading in the y direction is observed, consistent with the weakening of the harmonic restoring force in y as x increases. Excellent agreement between the two methods is again obtained.

4.5 Gauge freedom in multidimensional systems

We now return to the issue of ambiguity in defining a trajectory representation for evolution equations in higher dimensions. The freedom associated with decomposing the equations of motion is reminiscent of the freedom associated with the choice of gauge in electromagnetic theory [135]. In particular, we can transform the vector field $\boldsymbol{V}(\boldsymbol{x}, t)$ governing the trajectory representation by a gauge-like transformation

$$\boldsymbol{v}(\boldsymbol{x}, t) \rightarrow \boldsymbol{V}'(\boldsymbol{x}, t) = \boldsymbol{V}(\boldsymbol{x}, t) + \boldsymbol{W}(\boldsymbol{x}, t) \tag{4.72}$$

where $\boldsymbol{W}(\boldsymbol{x}, t)$ can be any vector field that satisfies the condition

$$\nabla \cdot \boldsymbol{W}(\boldsymbol{x}, t)\rho(\boldsymbol{x}, t)) = 0 \tag{4.73}$$

Thus, there are infinitely many trajectory representations of a given evolution equation, all related by transformations of the form Eqs. (72) and (73).

We illustrate this general principle with a concrete example, consisting of isotropic diffusion in two dimensions. The operator $\widehat{L}$ is then given by

$$\widehat{L} = D\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2}\right) \tag{4.74}$$

We decompose the operator using its separability: $\widehat{\lambda_x} = D\partial^2/\partial x^2$ and $\widehat{\lambda}_y = D\partial^2/\partial y^2$. Eq. (20) then gives the vector field

$$\dot{X}(x, y, t) = -\frac{D}{\rho(x, y, t)}\frac{\partial \rho(x, y, t)}{\partial x}, \tag{4.75}$$

$$\dot{Y}(x,y,t) = -\frac{D}{\rho(x,y,t)}\frac{\partial\rho(x,y,t)}{\partial y}, \tag{4.76}$$

A gauge-like transformation can be performed on this vector field by adding a two-dimensional vector $(W_x.W_y)$ that obeys $\nabla\cdot(\boldsymbol{W}\rho) = 0$. Many such vectors are possible; here we consider a family of vectors given by

$$W_x(x,y,t) = \frac{\delta}{\rho(x,y,t)}\frac{\partial\rho(x,y,t)}{\partial y} \tag{4.77}$$

$$W_y(x,y,t) = -\frac{\delta}{\rho(x,y,t)}\frac{\partial\rho(x,y,t)}{\partial x} \tag{4.78}$$

where δ is a variable parameter with the same units as the diffusion constant D. It is easy to check that the divergence of $\boldsymbol{W}\rho$ is indeed zero.

In Fig. 4.4, we show a comparison of two interacting ensemble simulations. The first, represented by solid lines in the contour plots, is identical to the results of Fig. 4.3, while the dashed contours show the results of including gauge terms in the vector field. Vector fields of the form given in Eqs. (77) and (78) are employed, with $\delta = 1$. All other parameters are as before. It is apparent that the evolving functions are virtually identical, although analysis of individual trajectories in the ensembles show their motion is quite different between the two simulations.

As a further example of gauge freedom in trajectory methods, we consider standard Hamiltonian dynamics. Here, the vector field in phase space is simply that given by Hamiltons equations, and is independent of the density $\rho(\boldsymbol{q},\boldsymbol{p},t)$. Still, there are infinite possible vector fields which yield the same evolution of q.

The classical Liouville equation is

$$\frac{\partial\rho}{\partial t} = \widehat{L}\rho = \{H,\rho\}, \tag{4.79}$$

where $\rho(\boldsymbol{q},\boldsymbol{p},t)$ is a density in phase space $\boldsymbol{x} = (\boldsymbol{q},\boldsymbol{p}), H(\boldsymbol{q},\boldsymbol{p})$ is the Hamiltonian, and $\{A,B\}$ is the Poisson bracket of A and B:

$$\{A,B\} = \sum_{k=1}^{N}\frac{\partial A}{\partial q_k}\frac{\partial B}{\partial p_k} - \frac{\partial B}{\partial q_k}\frac{\partial A}{\partial p_k} = \nabla A\cdot\boldsymbol{J}\cdot\nabla B. \tag{4.80}$$

where $\boldsymbol{J}$ is the symplectic matrix:

$$\boldsymbol{J} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \tag{4.81}$$

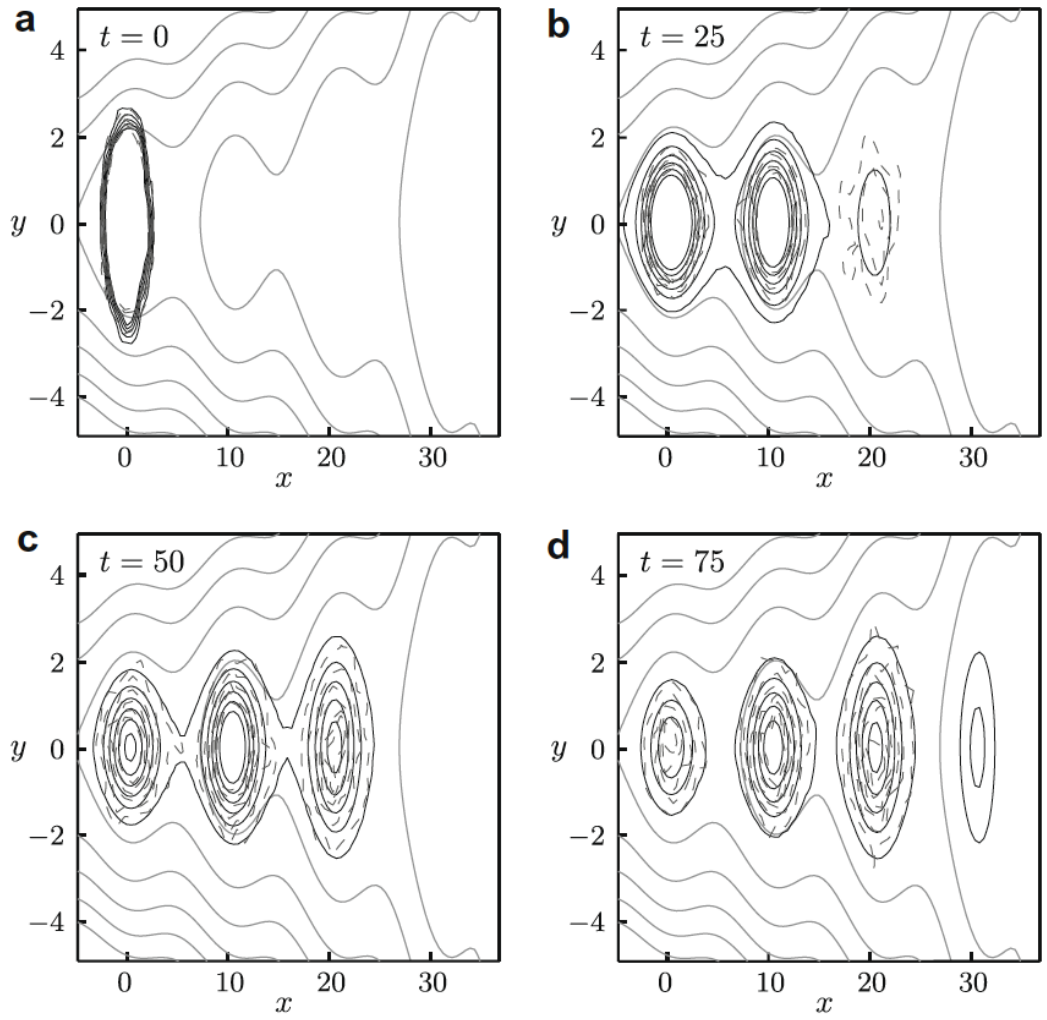


Figure 4.3: Comparison of interacting trajectory ensemble simulation and Brownian dynamics simulation for the two-dimensional potential, Eq. (71). Four snapshots of the evolving probability distributions are shown. The interacting trajectory method is shown as solid contour lines, and compared to the Brownian dynamics simulation, shown as dashed contours. (a) $t = 0$. (b) $t = 25$. (c) $t = 50$. (d) $t = 75$.

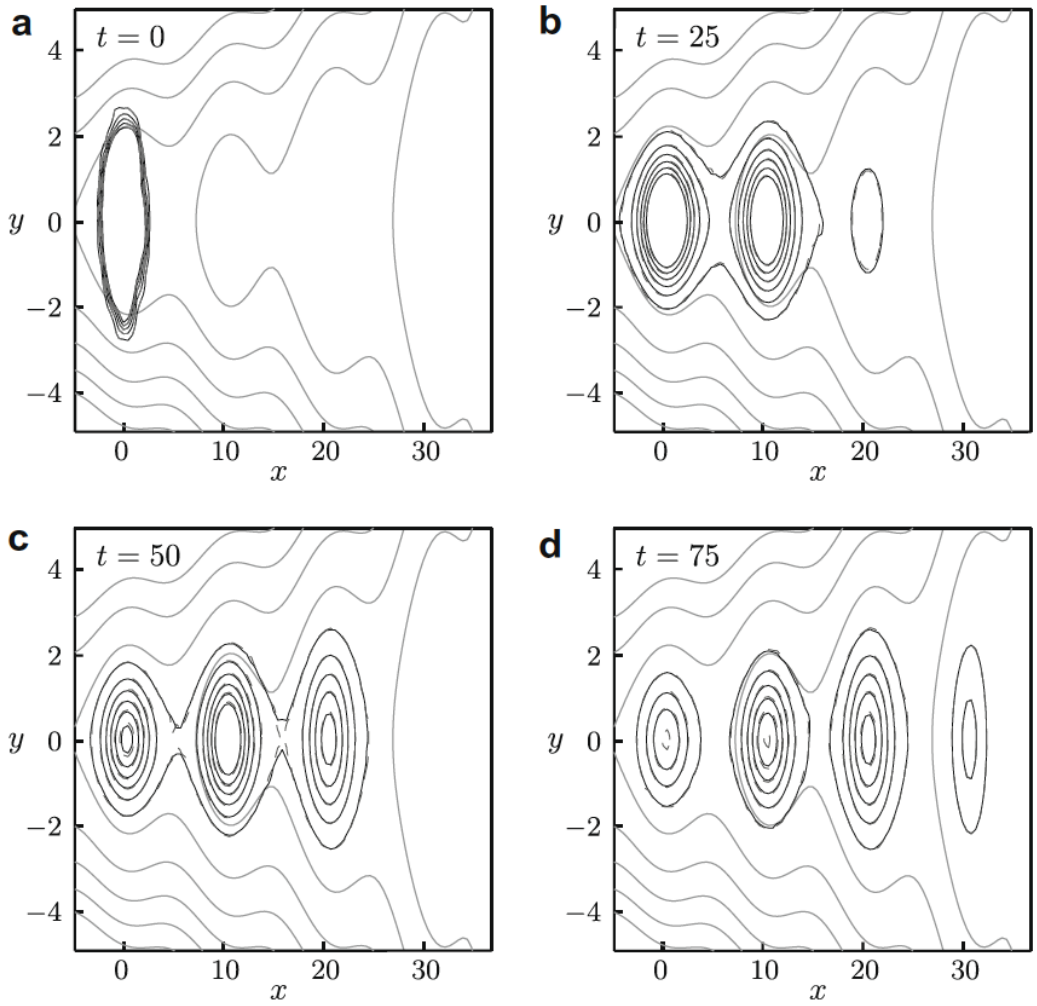


Figure 4.4: Comparison of two interacting trajectory ensemble simulations for the two-dimensional system with potential $U(x, y)$ given by Eq. (4.71). Four snapshots of the evolving probability distributions are shown. The interacting trajectory method is shown as a solid contour lines, and compared to an interacting trajectory simulation modified by addition of the gauge-like terms to the vector field, given by Eqs. (4.77) and (4.78) with $\delta = 1$, shown as dashed contours. (a) $t = 0$. (b) $t = 25$. (c) $t = 50$. (d) $t = 75$.

and ∇ is the gradient in phase space:

$$\nabla = \begin{pmatrix} \frac{\partial}{\partial \boldsymbol{q}} \\ \frac{\partial}{\partial \boldsymbol{p}} \end{pmatrix}. \tag{4.82}$$

The Liouville equation then becomes

$$\frac{\partial \rho}{\partial t} = \nabla H \cdot \boldsymbol{J} \cdot \nabla \rho. \tag{4.83}$$

We can write the Liouville equation as a continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot \boldsymbol{j} = 0, \tag{4.84}$$

where $\boldsymbol{j} = \dot{\boldsymbol{x}}\rho$. For classical Hamiltonian dynamics the flow in phase space is incompressible: $(\nabla \cdot \dot{\boldsymbol{x}}) = 0$. In this case the Liouville equation becomes

$$\frac{\partial \rho}{\partial t} = -(\dot{\boldsymbol{x}} \cdot \nabla)\rho. \tag{4.85}$$

By noting that $\boldsymbol{J}^T = -\boldsymbol{J}$, we can then identify the equation for the vector field $\dot{\boldsymbol{x}}$ as

$$\dot{\boldsymbol{x}} = \boldsymbol{J} \cdot \nabla H, \tag{4.86}$$

which are just the usual Hamiltons equations:

$$\dot{\boldsymbol{q}} = \frac{\partial H}{\partial \boldsymbol{p}},$$

$$\dot{\boldsymbol{p}} = -\frac{\partial H}{\partial \boldsymbol{q}}. \tag{4.87}$$

We now consider a gauge-like transformation, of the form

$$\dot{\boldsymbol{x}} \to \dot{\boldsymbol{x}} + \boldsymbol{w}, \tag{4.88}$$

where the invariance of the evolution equation requires

$$\nabla \cdot (\boldsymbol{w}\rho) = 0. \tag{4.89}$$

If we further require the vector field to itself have a zero divergence, we have the condition

$$\nabla \cdot \boldsymbol{w} = 0. \tag{4.90}$$

Together, these give the gauge condition

$$\boldsymbol{w} \cdot \nabla \rho. \tag{4.91}$$

The result makes sense, physically: the modification to the vector field must be everywhere orthogonal to the direction of change of ρ if the transformation is to leave the evolution of ρ invariant.

We now consider a simple but important example: an invariant distribution given as a function of the Hamiltonian, $\rho = f(H)$. Then,

$$\boldsymbol{w} \cdot \nabla \rho = (\boldsymbol{w} \cdot \nabla H) f'(H) = 0 \tag{4.92}$$

or

$$\boldsymbol{w} \cdot \nabla H = 0. \tag{4.93}$$

One solution to this is

$$\boldsymbol{w} = -\boldsymbol{J} \cdot \nabla H. \tag{4.94}$$

This transforms the vector field as follows:

$$\dot{\boldsymbol{x}} \to 0. \tag{4.95}$$

The result again makes sense: if we sample an invariant distribution with trajectories and then do not let them move, the distribution does not change. In general, we can consider the family of transformations $\boldsymbol{w}(\alpha) = \alpha \boldsymbol{j} \cdot \nabla H$, where α is an adjustable parameter. This transformation yields Hamilton-like equations that differ from the usual ones in that the rate of change of the phase space coordinates can be tuned by α. For the special case of $\alpha = -1$, we have the case of time standing still, as described above, although all values of α yield the correct invariant behavior for ρ.

We now briefly consider the implications for trajectory-based approaches to quantum mechanics. The quantum Liouville equation in the Wigner representation is given by [112, 113, 114, 115]

$$\frac{\partial \rho_w}{\partial t} = -\frac{p}{m}\frac{\partial \rho_w}{\partial q} + V'(q)\frac{\partial \rho_w}{\partial p} - \frac{\hbar^2}{24} V'''(q)\frac{\partial^3 \rho_w}{\partial p^3} + \cdots . \tag{4.96}$$

For simplicity, we will ignore the higher order terms in what follows. Following our previous work [126, 126, 127, 128, 129], we identify the right hand side as (the negative of) the divergence of a quantum flux written in a classical form $\boldsymbol{j}_w = \dot{\boldsymbol{x}}\rho_w$, so

$$\frac{\partial \rho_w}{\partial t} + \nabla \cdot \boldsymbol{j}_w = 0 \tag{4.97}$$

giving

$$\begin{aligned} \frac{\partial}{\partial q}(\dot{q}\rho_w) &= \frac{\partial}{\partial q}\left(\frac{p}{m}\rho_w + \theta_q \rho_w\right), \\ \frac{\partial}{\partial p}(\dot{p}\rho_w) &= \frac{\partial}{\partial p}\left(-V'(q)\rho_w + \theta_p \rho_w\right), \end{aligned} \tag{4.98}$$

which defines the quantum vector field $\theta = (\theta_q, \theta_p)$. We then consider the equation of motion condition

$$\nabla \cdot (\theta \rho_w) = \frac{\hbar^2}{24} V'''(q) \frac{\partial^3 \rho_w}{\partial p^3}. \tag{4.99}$$

If we want incompressible phase space flow, we should also impose the divergence condition

$$\nabla \cdot \theta = 0. \tag{4.100}$$

If the incompressibility condition is ignored and the choice $\theta_q = 0$, the equation of motion condition can be integrated to yield

$$\theta_p = \frac{\hbar^2}{24} V'''(q) \frac{1}{\rho_w} \frac{\partial^2 \rho_w}{\partial p^2}. \tag{4.101}$$

This leads to the method described in Refs. [126, 126, 127, 128, 129]. Other choices are possible, however. For instance, one can choose $\theta_p = 0$, which then yields

$$\theta_q = \frac{\hbar^2}{24} \frac{1}{\rho_w} \int^q V'''(q') \frac{\partial^3 \rho_w(q', p)}{\partial p^3} dq'. \tag{4.102}$$

This defines an alternative (and untested) quantum trajectory method. Many other divisions of the quantum vector field between its q and p components are possible, which all lead to the same quantum Liouville equation.

It would be of interest to consider general solutions to this problem which also incorporate the incompressibility condition $\nabla \cdot \theta = 0$.

4.6 Summary

In this paper,we have described a general formal method for solving evolution equations for probability distributions using trajectory ensembles. For systems described by a positive definite distribution function, we derived the general equations of motion for the trajectories. We showed that the vector field describing the time rate of change of the trajectory ensemble members depends, in general, on both external forces and on the probability density itself. The dependence of the equations of motion on the probability density lead to interactions between the ensemble members and a loss of their statistical independence. The formalism was illustrated by a number of numerical examples consisting of diffusion on potential energy surfaces. Excellent agreement between the numerical implementation of the interacting trajectory formalism and conventional Brownian dynamics was obtained. For multidimensional systems, a gauge-like freedom exists in the choice of the underlying vector field, which leaves the evolution of the probability density invariant. We illustrated this point with several examples.

When taken as a mathematical approach to solving partial differential equations using trajectory ensembles, we have shown that no unique set of trajectories result. Rather, there are infinite possibilities that are related by families of gauge transformations. Thus, trajectories lose their realistic interpretation, and become just constructs in the solution of the underlying partial differential equation for the probability distribution. This observation has relevance to more philosophical analyses of the foundations and interpretation of quantum mechanics, the correspondence principle, hidden variable theories, and other contexts where particle trajectories are postulated in quantum mechanical systems.

From a practical standpoint, such transformations may be useful in molecular dynamics or stochastic sampling of statistical mechanical ensembles. There may be particular gauge choices corresponding to modified dynamics that are more efficient in practice than standard classical mechanics or Langevin dynamics. This will be investigated in future publications.

4.7 Acknowledgements

The authors gratefully recognize the many contributions made by Eli Pollak to the field of theoretical chemistry, and dedicate this paper to him. This work was supported by the National Science Foundation.

Chapter 4 is a minimally modified reprint of the material as it appears in Hogan P., **Van Wart A. T.**, Donoso A., Martens C. C., "Solving evolution equations using interacting trajectory ensembles" *Chemical Physics* (2010), 370, 2028. I was the primary investigator and author of the computational aspects of this paper.

Bibliography

[1] T. Darden, D. York, and L. Pedersen. Particle mesh ewald - an n.log(n) method for ewald sums in large systems. *J. Chem. Phys.*, 98:10089–10092, 1993.

[2] Grumuller H., Heller H., Windemuth A., and Schulten K. *Mol. Simulat.*, 6:121, 1991.

[3] Loncharich R. J. and Brooks B. R. *Macromolecules*, 6:32, 1989.

[4] Feller S. E., Pastor R. W., Rojnuckarin A., Bogusz S., and Brooks B. R. *J. Phys. Chem.*, 6:32, 1989.

[5] Hairer E., Lubich C., Rojnuckarin A., and Wanner G. *Geometric Numerical Integration vol. 31.* Springer-Verlag, Berlin, 2002.

[6] Reich S. *SIAM J Numer Anal*, 36:1549, 1999.

[7] Frenkel D. and Smit B. *Understanding Molecular Simulation from Algorithms to Applications.* Academic Press, San Diego CA, 2002.

[8] Bishop T. C., Skeel T. D., and Schulten K. *J. Comp. Chem.*, 18:1785, 1997.

[9] Feller S. E., Zhang Y., Pastor R. W., and Brooks B. R. *J. Chem. Phys.*, 103:4613, 1995.

[10] Hoover W. G. *Phys. Rev. A*, 31:1695, 1985.

[11] Hoover W. G. *Phys. Rev. A*, 34:2499, 1986.

[12] Hoover W. G. *Computational Statistical Mechanics.* Elsevier, Amsterdam, 1991.

[13] Hoover W. G. *Phys. Rev. A*, 120:11432, 2004.

[14] J. C. Phillips, R. Braun, W. Wang, J. Gumbart, E. Tajkhorshid, E. Villa, C. Chipot, R. D. Skeel, L. Kale, and K. Schulten. Scalable molecular dynamics with namd. *J. Comput. Chem.*, 26:1781–802, 2005.

[15] Bormuth A., Hofmann M., Henritzi P., Vogel M., and Rssler E. A. Chain-length dependence of polymer dynamics: A comparison of results from molecular dynamics simulations and field-cycling ^{1}H NMR. *Macromolecules*, 46:7805–7811, 2013.

[16] Showalter S. A. and Bruschweiler R. Validation of molecular dynamics simulations of biomolecules using nmr spin relaxation as benchmarks:? application to the amber99sb force field. *J. Chem. Theory Comput.*, 3:961–975, 2007.

[17] Kuwata K., Matumoto T., Cheng H., Nagayama K., James T. L., and Roder H. Nmr-detected hydrogen exchange and molecular dynamics simulations provide structural insight into fibril formation of prion protein fragment 106126. *Proc. Natl. Acad. Sci.*, 100:14790–14795, 2003.

[18] Salomon-Ferrer R., Case D. A., and Walker R. C. An overview of the amber biomolecular simulation package. *Comp. Mol. Sci.*, 3:198–210, 2013.

[19] Par Bjelkmar, Per Larsson, Michel A. Cuendet, Berk Hess, and Erik Lindahl. Implementation of the charmm force field in gromacs: Analysis of protein stability effects from correction maps, virtual interaction sites, and water models. *J. Chem. Theory Comput.*, 6:459–466, 2010.

[20] B. R. Brooks, 3rd Brooks, C. L., Jr. Mackerell, A. D., L. Nilsson, R. J. Petrella, B. Roux, Y. Won, G. Archontis, C. Bartels, S. Boresch, A. Caflisch, L. Caves, Q. Cui, A. R. Dinner, M. Feig, S. Fischer, J. Gao, M. Hodoscek, W. Im, K. Kuczera, T. Lazaridis, J. Ma, V. Ovchinnikov, E. Paci, R. W. Pastor, C. B. Post, J. Z. Pu, M. Schaefer, B. Tidor, R. M. Venable, H. L. Woodcock, X. Wu, W. Yang, D. M. York, and M. Karplus. Charmm: the biomolecular simulation program. *J. Comput. Chem.*, 30:1545–614, 2009.

[21] McCammon J. A., Gelin B. R., and Karplus M. Dynamics of folded proteins. *Nature*, 267:585–590, 1977.

[22] Salomon-Ferrer R., Gotz A. W., Poole D., Grand S. L., and Walker R. C. Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald. *J. Chem. Theory Comput.*, 9:38783888, 2013.

[23] Bowman G. R., Voelz V. A., and Pande V. S. Atomistic folding simulations of the five-helix bundle protein $\lambda(685)$. *J. Am. Chem. Soc.*, 133:664–667, 2011.

[24] Shaw D. E., Maragakis P., Lindorff-Larsen K., Piana S., Dror R. O., Eastwood M. P., Bank J. A., Jumper J. M., Salmon J. K., Shan Y., and et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330:341–346, 2010.

[25] J. Monod, J. Wyman, and J. P. Changeux. On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.*, 12:88–118, 1965.

[26] Jr. Koshland, D. E., G. Nemethy, and D. Filmer. Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry*, 5:365–85, 1966.

[27] C. J. Tsai, A. del Sol, and R. Nussinov. Allostery: absence of a change in shape does not imply that allostery is not at play. *J. Mol. Biol.*, 378:1–11, 2008.

[28] G. M. Suel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat. Struct. Biol.*, 10:59–69, 2003.

[29] B. N. Chaudhuri, S. C. Lange, R. S. Myers, S. V. Chittur, V. J. Davisson, and J. L. Smith. Crystal structure of imidazole glycerol phosphate synthase: a tunnel through a (/)8 barrel joins two active sites. *Structure*, 9:987–97, 2001.

[30] R. E. Amaro, A. Sethi, R. S. Myers, V. J. Davisson, and Z. A. Luthey-Schulten. A network of conserved interactions regulates the allosteric signal in a glutamine amidotransferase. *Biochemistry*, 46:2156–73, 2007.

[31] Jeong H., Mason S. P., Barabasi A. L., and Oltvai Z. N. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

[32] Zoraghi R. and Reiner N. E. Protein interaction networks as starting points to identify novel antimicrobial drug targets. *Curr. Opinion Micro.*, 16, 2013.

[33] J. P. Changeux. 50th anniversary of the word "allosteric". *Protein Sci.*, 20:1119–24, 2011.

[34] J. Monod, J. P. Changeux, and F. Jacob. Allosteric proteins and cellular control systems. *J. Mol. Biol.*, 6:306–29, 1963.

[35] M. M. Rubin and J. P. Changeux. On the nature of allosteric transitions: implications of non-exclusive ligand binding. *J. Mol. Biol.*, 21:265–74, 1966.

[36] Y. Chen, K. Reilly, and Y. Chang. Evolutionarily conserved allosteric network in the cys loop family of ligand-gated ion channels revealed by statistical covariance analyses. *J. Biol. Chem.*, 281:18184–92, 2006.

[37] A. del Sol, H. Fujihashi, D. Amoros, and R. Nussinov. Residues crucial for maintaining short paths in network communication mediate signaling in proteins. *Mol. Syst. Biol.*, 2:1–12, 2006.

[38] S. W. Lockless and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286:295–9, 1999.

[39] J. P. Changeux and S. J. Edelstein. Allosteric mechanisms of signal transduction. *Science*, 308:1424–8, 2005.

[40] N. M. Goodey and S. J. Benkovic. Allosteric regulation and catalysis emerge via a common route. *Nat. Chem. Biol.*, 4:474–82, 2008.

[41] C. Chennubhotla, Z. Yang, and I. Bahar. Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin groel. *Mol. Biosyst.*, 4:287–92, 2008.

[42] A. Cooper and D. T. F. Dryden. Allostery without conformational change - a plausible model. *Eur. Biophys. J. Biophys. Lett.*, 11:103–109, 1984.

[43] M. R. Arkin and J. A. Wells. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nat. Rev. Drug Discovery*, 3:301–17, 2004.

[44] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules .2. analysis of experimental results. *J. Am. Chem. Soc.*, 104:4559–4570, 1982.

[45] G. Lipari and A. Szabo. Model-free approach to the interpretation of nuclear magnetic-resonance relaxation in macromolecules .1. theory and range of validity. *J. Am. Chem. Soc.*, 104:4546–4559, 1982.

[46] P. J. Sapienza and A. L. Lee. Using nmr to study fast dynamics in proteins: methods and applications. *Curr. Opin. Pharmacol.*, 10:723–30, 2010.

[47] D. Yang and L. E. Kay. Contributions to conformational entropy arising from bond vector fluctuations measured from nmr-derived order parameters: application to protein folding. *J. Mol. Biol.*, 263:369–82, 1996.

[48] L. M. Horstink, R. Abseher, M. Nilges, and C. W. Hilbers. Functionally important correlated motions in the single-stranded dna-binding protein encoded by filamentous phage pf3. *J. Mol. Biol.*, 287:569–77, 1999.

[49] J. B. Watney and S. Hammes-Schiffer. Comparison of coupled motions in escherichia coli and bacillus subtilis dihydrofolate reductase. *J. Phys. Chem. B*, 110:10130–8, 2006.

[50] M. J. Bradley, P. T. Chivers, and N. A. Baker. Molecular dynamics simulation of the escherichia coli nikr protein: equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *J. Mol. Biol.*, 378:1155–73, 2008.

[51] O. F. Lange and H. Grubmuller. Generalized correlation for biomolecular dynamics. *Proteins*, 62:1053–61, 2006.

[52] B. J. Killian, J. Yundenfreund Kravitz, and M. K. Gilson. Extraction of configurational entropy from molecular simulations via an expansion approximation. *J. Chem. Phys.*, 127:024107, 2007.

[53] C. L. McClendon, G. Friedland, D. L. Mobley, H. Amirkhani, and M. P. Jacobson. Quantifying correlations between allosteric sites in thermodynamic ensembles. *J. Chem. Theory Comput.*, 5:2486–2502, 2009.

[54] A. Sethi, J. Eargle, A. A. Black, and Z. Luthey-Schulten. Dynamical networks in trna:protein complexes. *Proc. Natl. Acad. Sci. U. S. A.*, 106:6620–5, 2009.

[55] J. Eargle and Z. Luthey-Schulten. Networkview: 3d display and analysis of protein.rna interaction networks. *Bioinformatics*, 28(22):3000–1, 2012.

[56] M. G. Chaparian and D. R. Evans. The catalytic mechanism of the amidotransferase domain of the syrian hamster multifunctional protein cad. evidence for a cad-glutamyl covalent intermediate in the formation of carbamyl phosphate. *J. Biol. Chem.*, 266:3387–95, 1991.

[57] B. Roux and C. T. Walsh. p-aminobenzoate synthesis in escherichia coli: kinetic and mechanistic characterization of the amidotransferase paba. *Biochemistry*, 31:6904–10, 1992.

[58] J. Nakamura, K. Straub, J. Wu, and L. Lou. The glutamine hydrolysis function of human gmp synthetase. identification of an essential active site cysteine. *J. Biol. Chem.*, 270:23450–5, 1995.

[59] M. Willemoes. Thr-431 and arg-433 are part of a conserved sequence motif of the glutamine amidotransferase domain of ctp synthases and are involved in gtp activation of the lactococcus lactis enzyme. *J. Biol. Chem.*, 278:9407–11, 2003.

[60] D. Lang, R. Thoma, M. Henn-Sax, R. Sterner, and M. Wilmanns. Structural evidence for evolution of the beta/alpha barrel scaffold by gene duplication and fusion. *Science*, 289:1546–50, 2000.

[61] R. S. Myers, J. R. Jensen, I. L. Deras, J. L. Smith, and V. J. Davisson. Substrate-induced changes in the ammonia channel for imidazole glycerol phosphate synthase. *Biochemistry*, 42:7013–22, 2003.

[62] R. Amaro, E. Tajkhorshid, and Z. Luthey-Schulten. Developing an energy landscape for the novel function of a (beta/alpha)8 barrel: ammonia conduction through hisf. *Proc. Natl. Acad. Sci. U. S. A.*, 100:7599–604, 2003.

[63] P. Alifano, R. Fani, P. Lio, A. Lazcano, M. Bazzicalupo, M. S. Carlomagno, and C. B. Bruni. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol. Rev.*, 60:44–69, 1996.

[64] B. N. Chaudhuri, S. C. Lange, R. S. Myers, V. J. Davisson, and J. L. Smith. Toward understanding the mechanism of the complex cyclization reaction catalyzed by imidazole glycerolphosphate synthase: crystal structures of a ternary complex and the free enzyme. *Biochemistry*, 42:7003–12, 2003.

[65] S. Warshall. A theorem on boolean matrices. *J. Assoc. Comput. Mach.*, 9:11–13, 1962.

[66] R. W. Floyd. Algorithm-97 - shortest path. *Commun. ACM*, 5:345–345, 1962.

[67] M. Girvan and M. E. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. U. S. A.*, 99:7821–6, 2002.

[68] M. Rostkowski, M. H. Olsson, C. R. Sondergaard, and J. H. Jensen. Graphical analysis of ph-dependent properties of proteins predicted using propka. *BMC Struct Biol*, 11:6, 2011.

[69] E. Roberts, J. Eargle, D. Wright, and Z. Luthey-Schulten. Multiseq: unifying sequence and structure data for evolutionary analysis. *BMC Bioinf.*, 7:382, 2006.

[70] R. B. Russell and G. J. Barton. Multiple protein-sequence alignment from tertiary structure comparison - assignment of global and residue confidence levels. *Proteins: Struct., Funct., Genet.*, 14:309–323, 1992.

[71] R. B. Russell and G. J. Barton. The limits of protein secondary structure prediction accuracy from multiple sequence alignment. *J. Mol. Biol.*, 234:951–957, 1993.

[72] R. E. Amaro, R. S. Myers, V. J. Davisson, and Z. A. Luthey-Schulten. Structural elements in igp synthase exclude water to optimize ammonia transfer. *Biophys. J.*, 89:475–87, 2005.

[73] Rommie Amaro and Zaida Luthey-Schulten. Molecular dynamics simulations of substrate channeling through an a b barrel protein. *Chem. Phys.*, 307:147–155, 2004.

[74] H. Grubmuller. Solvate v 1.0. *Theoretical Biophysics Group, Institute for Medical Optics, Ludwig-Maximilians University, Munich*, 1996.

[75] N. M. Glykos. Software news and updates. carma: a molecular dynamics analysis program. *J. Comput. Chem.*, 27:1765–8, 2006.

[76] I. Andricioaei and M. Karplus. On the calculation of entropy from covariance matrices of the atomic fluctuations. *J. Chem. Phys.*, 115:6289–6292, 2001.

[77] R. S. Myers, R. E. Amaro, Z. A. Luthey-Schulten, and V. J. Davisson. Reaction coupling through interdomain contacts in imidazole glycerol phosphate synthase. *Biochemistry*, 44:11974–85, 2005.

[78] J. M. Lipchock and J. P. Loria. Nanometer propagation of millisecond motions in v-type allostery. *Structure*, 18:1596–1607, 2010.

[79] S. Beismann-Driemeyer and R. Sterner. Imidazole glycerol phosphate synthase from thermotoga maritima. quaternary structure, steady-state kinetics, and reaction mechanism of the bienzyme complex. *J. Biol. Chem.*, 276:20387–96, 2001.

[80] A. Douangamath, M. Walker, S. Beismann-Driemeyer, M. C. Vega-Fernandez, R. Sterner, and M. Wilmanns. Structural evidence for ammonia tunneling across the (beta alpha)(8) barrel of the imidazole glycerol phosphate synthase bienzyme complex. *Structure*, 10:185–193, 2002.

[81] R. Omi, H. Mizuguchi, M. Goto, I. Miyahara, H. Hayashi, H. Kagamiyama, and K. Hirotsu. Structure of imidazole glycerol phosphate synthase from thermus thermophilus hb8: Open-closed conformational change and ammonia tunneling. *J. Biochem.*, 132:759–765, 2002.

[82] S. Korolev, T. Skarina, E. Evdokimova, S. Beasley, A. Edwards, A. Joachimiak, and A. Savchenko. Crystal structure of glutamine amidotransferase from thermotoga maratama. *Proteins: Struct., Funct., Genet.*, 49:420–422, 2002.

[83] R. Nussinov, C. J. Tsai, and B. Y. Ma. The underappreciated role of allostery in the cellular network. *Annual Review of Biophysics, Vol 42*, 42:169–189, 2013.

[84] Q. Cui and M. Karplus. Allostery and cooperativity revisited. *Protein Science*, 17:1295–1307, 2008.

[85] I. N. Berezovsky. Thermodynamics of allostery paves a way to allosteric drugs. *Biochimica Et Biophysica Acta-Proteins and Proteomics*, 1834:830–835, 2013.

[86] R. Nussinov and C. J. Tsai. Allostery in disease and in drug discovery. *Cell*, 153:293–305, 2013.

[87] E. Laine, L. Martinez, D. Ladant, T. Malliavin, and A. Blondel. Molecular motions as a drug target: Mechanistic simulations of anthrax toxin edema

factor function led to the discovery of novel allosteric inhibitors. *Toxins*, 4:580–604, 2012.

[88] A. R. Atilgan and C. Atilgan. Local motifs in proteins combine to generate global functional moves. *Brief Funct Genomics*, 11:479–88, 2012.

[89] P. M. Gasper, B. Fuglestad, E. A. Komives, P. R. Markwick, and J. A. McCammon. Allosteric networks in thrombin distinguish procoagulant vs. anticoagulant activities. *Proc Natl Acad Sci U S A*, 109:21216–22, 2012.

[90] A. T. VanWart, J. Eargle, Z. Luthey-Schulten, and R. E. Amaro. Exploring residue component contributions to dynamical network models of allostery. *Journal of Chemical Theory and Computation*, 8:2949–2961, 2012.

[91] Y. Miao, S. E. Nichols, P. M. Gasper, V. T. Metzger, and J. A. McCammon. Activation and dynamic network of the m2 muscarinic receptor. *Proc Natl Acad Sci U S A*, 110:10982–7, 2013.

[92] I. Rivalta, M. M. Sultan, N. S. Lee, G. A. Manley, J. P. Loria, and V. S. Batista. Allosteric pathways in imidazole glycerol phosphate synthase. *Proc Natl Acad Sci U S A*, 109:1428–36, 2012.

[93] A. Ghosh, R. Sakaguchi, C. Liu, S. Vishveshwara, and Y. M. Hou. Allosteric communication in cysteinyl trna synthetase: a network of direct and indirect readout. *J Biol Chem*, 286:37721–31, 2011.

[94] D. D. Boehr, J. R. Schnell, D. McElheny, S. H. Bae, B. M. Duggan, S. J. Benkovic, H. J. Dyson, and P. E. Wright. A distal mutation perturbs dynamic amino acid networks in dihydrofolate reductase. *Biochemistry*, 52:4605–4619, 2013.

[95] D. Long and R. Bruschweiler. Structural and entropic allosteric signal transduction strength via correlated motions. *Journal of Physical Chemistry Letters*, 3:1722–1726, 2012.

[96] M. S. Liu, B. D. Todd, and R. J. Sadus. Allosteric conformational transition in adenylate kinase: Dynamic correlations and implication for allostery. *Australian Journal of Chemistry*, 63:405–412, 2010.

[97] W. Humphrey, A. Dalke, and K. Schulten. Vmd: Visual molecular dynamics. *Journal of Molecular Graphics and Modelling*, 14:33–38, 1996.

[98] A. A. Hagberg, D. A. Schult, and P. J. Swart. Exploring network structure, dynamics, and function using networkx. pages 11–15.

[99] P. F. Dubois and T. Y. Yang. Extending python with fortran. *Comput. Sci. Eng.*, 1:66–73, 1999.

[100] P. Peterson. F2py: a tool for connecting fortran and python programs. *International J. Comput. Sci. and Eng.*, 4:296–305, 2009.

[101] Eric Jones, Travis Oliphant, Pearu Peterson, and Others. Scipy: Open source scientific tools for python, 2001.

[102] David Ascher, Paul F. Dubois, Konrad Hinsen, James J. Hugunin, and Travis Oliphant. *Numerical Python*. Lawrence Livermore National Laboratory, Livermore, CA, 1999.

[103] Travis E. Oliphant. *Guide to NumPy*. Brigham Young University, Provo, UT, 2006.

[104] G. S. Couch, E. F. Pettersen, C. C. Huang, and T. E. Ferrin. Annotating pdb files with scene information. *Journal of Molecular Graphics*, 13:153–158, 1995.

[105] H. Goldstein. *Classical Mechanics, 2nd ed.* Addison-Wesley, Reading, 1980.

[106] D. A. McQuarrie. *Statistical Mechanics*. HarperCollins, New York, 1980.

[107] R. Kubo, M. Toda, and N. Hashitsume. *Statistical Physics, vol. II*. Springer-Verlag, Berlin, 1985.

[108] D. Chandler. *Introduction to Modern Statistical Mechanics*. Oxford University Press, Oxford, 1987.

[109] E. Pollack. *J. Chem. Phys.*, 85:865, 1986.

[110] C. C. Martens. *J. Chem. Phys.*, 116:2516, 2002.

[111] C. Chohen-Tannoudji, B. Diu, and F. Laloe. *Quantum Mechanics*. Wiley, New York, 1977.

[112] E. P. Wigner. *Phys. Rev.*, 40:749, 2002.

[113] K. Takahashi. *Prog. Theor. Phys. Suppl.*, 98:109, 1989.

[114] H. W. Lee. *Phys. Rep.*, 259:147, 1995.

[115] S. Mukamel. *Principles of Nonlinear Optical Spectroscopy*. Oxford University Press, Oxford, 1995.

[116] R. P. Feynmann and A. R. Hibbs. *Quantum Mechanics and Path Integrals*. McGraw Hill, New York, 1965.

[117] R. Wyatt. *Quantum Dynamics with Trajectories: Introduction to Quantum Hydrodynamics*. Springer, New York, 2005.

[118] D. Bohm. *Phys. Rev.*, 85:166, 1952.

[119] D. Bohm. *Phys. Rev.*, 85:180, 1952.

[120] P.R. Holland. *The Quantum Theory of Motion*. Cambridge University Press, Cambridge, 1995.

[121] C. I. Lopermore and R.E. Wyatt. *Phys. Rev. Lett.*, 82:5190, 1999.

[122] R.E. Wyatt. *J. Chem. Phys.*, 11:4406, 1999.

[123] B. K. Dey, A. Askar, and H. Rabitz. *J. Chem. Phys.*, 109:8770, 1998.

[124] E. R. Bittner and R. E. Wyatt. *J. Chem. Phys.*, 113:8888, 2000.

[125] E. Gindensperger, C. Meier, and J. A. Beswick. *J. Chem. Phys.*, 113:9369, 2000.

[126] A. Donoso and C. C. Martens. *Int. J. Quantm Chem.*, 87:1348, 2002.

[127] A. Donoso and C. C. Martens. *J. Chem. Phys.*, 116:10598, 2002.

[128] A. Donoso, Y. Zheng, and C. C. Martens. *J. Chem. Phys.*, 119:5010, 2003.

[129] H. Lopez, C. C. Martens, and A. Donoso. *J. Chem. Phys.*, 1125:154111, 2006.

[130] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.

[131] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vettering. *Numerical Recipes in C: The Art of Scientific Computing*. Cambridge University Press, Cambrindge, 1992.

[132] M. P. Allen and D. J. Tildesley. *Compter Simulations of Liquids*. Claredon Press, Oxford, 1987.

[133] R. D. Astumian. *Science*, 276:917, 1997.

[134] P. Reimann. *Phys. Rep.*, 361:57265, 2002.

[135] J. D. Jackson. *Classical Electrodynamics*. Wiley, New York, 1998.