New Algorithms for Predicting Conformational Polymorphism and Inferring Direct Couplings for Side Chains of Proteins

by

Laleh Soltan Ghoraie

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Doctor of Philosophy
in
Computer Science

Waterloo, Ontario, Canada, 2015

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Protein crystals populate diverse conformational ensembles. Despite much evidence that there is widespread conformational polymorphism in protein side chains, most of the x-ray crystallography data are modelled by single conformations in the Protein Data Bank. The ability to extract or to predict these conformational polymorphisms is of crucial importance, as it facilitates deeper understanding of protein dynamics and functionality. This dissertation describes a computational strategy capable of predicting side-chain polymorphisms. The applied approach extends a particular class of algorithms for side-chain prediction by modelling the side-chain dihedral angles more appropriately as continuous rather than discrete variables. Employing a new inferential technique known as particle belief propagation (PBP), we predict residue-specific distributions that encode information about side-chain polymorphisms. The predicted polymorphisms are in relatively close agreement with results from a state-of-the-art approach based on x-ray crystallography data. This approach characterizes the conformational polymorphisms of side chains using electron density information, and has successfully discovered previously unmodelled conformations.

Furthermore, it is known that coupled fluctuations and concerted motions of residues can reveal pathways of communication used for information propagation in a molecule and hence, can help in understanding the “allostery” phenomenon in proteins. In order to characterize the coupled motions, most existing methods infer structural dependencies among a protein’s residues. However, recent studies have highlighted the role of coupled side-chain fluctuations alone in the allosteric behaviour of proteins, in contrast to a common belief that the backbone motions play the main role in allostery. These studies and the aforementioned recent discoveries about prevalent alternate side-chain conformations (conformational polymorphism) accentuate the need to devise new computational approaches that acknowledge side chains’ roles. As well, these approaches must consider the polymorphic nature of the side chains, and incorporate effects of this phenomenon (polymorphism) in the study of information transmission and functional interactions of residues in a molecule. Such frameworks can provide a more accurate understanding of the allosteric behaviour.

Hence, as a topic related to the conformational polymorphism, this dissertation addresses the problem of inferring directly coupled side chains, as well. First, we present a novel approach to generate an ensemble of conformations and an efficient computational method to extract direct couplings of side chains in allosteric proteins. These direct couplings are used to provide sparse network representations of the coupled side chains. The framework is based on a fairly new statistical method, named graphical lasso (GLASSO), devised for sparse graph estimation. In the proposed GLASSO-based framework, the side-chain conformational polymorphism is taken into account. It is shown that by studying the intrinsic dynamics of an inactive structure alone, we are able to construct a network of functionally crucial residues. Second, we show that the proposed method is capable of providing a *magnified* view of the coupled and conformationally polymorphic side chains. This model reveals couplings between the alternate conformations of a coupled residue pair. To the best of our knowledge, this is the first computational method for extracting networks of side chains' alternate conformations. Such networks help in providing a detailed image of side-chain dynamics in functionally important and conformationally polymorphic sites, such as binding and/or allosteric sites. This information may assist in new drug-design alternatives.

Side-chain conformations are commonly represented by multivariate angular variables. However, the GLASSO and other existing methods that can be applied to the aforementioned inference task are not capable of handling multivariate angular data. This dissertation further proposes a novel method to infer direct couplings from this type of data, and shows that this method is useful for identifying functional regions and their interactions in allosteric proteins. The proposed framework is a novel extension of canonical correlation analysis (CCA), which we call "kernelized partial CCA" (or simply KPCCA). Using the conformational information and fluctuations of the inactive structure alone for allosteric proteins in the Ras and other Ras-like families, the KPCCA method identified allosterically important residues not only as strongly coupled ones but also in densely connected regions of the interaction graph formed by the inferred couplings. The results were in good agreement with other empirical findings and outperformed those obtained by the GLASSO-based framework. By studying distinct members of the Ras, Rho, and Rab sub-families, we show further that KPCCA is capable of inferring common allosteric characteristics in the small G protein super-family.

Acknowledgements

First, I would like to express my deepest gratitude to my supervisors, Professor Mu Zhu and Professor Forbes Burkowski, for their guidance and patience during the course of my PhD studies, their constant support and help from the initial steps of the research work performed for this dissertation to the last steps of co-authoring the published articles with me.

I would like to thank my committee members, Professor Guohui Lin, Professor Ming Li, Professor Bin Ma, and Professor Brendan McConkey for dedicating time to read my thesis and serve in my PhD committee.

I would like to thank Dr. Shuai Cheng Li for his advice and helpful discussions.

Dedication

To Soraya and Sadegh,
for their unconditional love and sacrifices that can never be described in words
To Bahador,
for his constant love, encouragement and support

Table of Contents

List of Tables

List of Figures

List of Abbreviations

BP:	Belief Propagation
CASP:	Critical Assessment of Structure Prediction
CCA:	Canonical Correlation Analysis
CRN:	Contact Rearrangement Networks
GDP:	Guanosine Diphosphate
GTP:	Guanosine Triphosphate
GLASSO:	Graphical LASSO
KCCA:	Kernelized Canonical Correlation Analysis
KPCCA:	Kernelized Partial Canonical Correlation Analysis
KS-test:	Kolmogorov-Smirnov test
MC:	Monte Carlo
MD:	Molecular Dynamics
PBP:	Particle Belief Propagation
PCCA:	Partial Canonical Correlation Analysis
PPI:	Protein-Protein Interaction
SCA:	Statistical Coupling Analysis
SCP:	Side-Chain Prediction
VM:	Von-Mises

Chapter 1

Introduction

1.1 Background and Scope

The bioinformatics field comprises three main areas, as follows: Theory and methods that include algorithms, machine learning approaches, statistical analysis, etc. Applications, such as drug discovery, sequence analysis, protein structure prediction, etc., and Data, such as gene expression profiles, structure of biological molecules (protein, DNA, RNA, ...), phylogenetic data, etc. Structural bioinformatics is a sub-discipline that has emerged due to an increase in the number of 3D macromolecular structures. Hence, it is defined as a combination of the methods and applications (first two aforementioned areas) for the biological structure data [8]. Relying on the availability of high resolution structural data, this sub-field aims to obtain a more precise understanding of the *function* of the systems and molecules and the effects of perturbations and modifications. Revelation of DNA's structure and the first protein structures created a need for computational methods that can examine and model these molecules, as well as fitting the electron densities (resulting from x-ray crystallographic data), and processing NMR-based distance constraints to produce more manageable and accessible formats for describing and storing the molecules' 3D structures.

The focus of this dissertation is solely "protein" structural data. The nature of structural data causes specific challenges in handling and computation with these data. Some of the challenges are as follows:

- Non-linearity of structural data and relationships between atoms, i.e., non-linear forces between the atoms, such as electrostatic, van der Waals, etc. This usually leads to either very expensive computations or approximations in modelling.

- Continuity of the data that leads to computationally intractable search spaces for structural problems. The structural information is represented by either atom coordinates in the 3D space or dihedral angles. Many computational methods apply simplifications, such as discritization of these search spaces. In this dissertation, attempts have been made to tackle this issue by selecting appropriate computational tools and techniques that recognize the nature of conformational data.

- Scarcity of the data. Although the number of protein structures deposited in the Protein Data Bank (PDB [5]) has grown rapidly during the recent years, the structures of many important classes of non-globular and water-insoluble proteins such as, membrane proteins are still unavailable. This lack of data prevents us from performing any statistical inference or analysis about these proteins, and leads to lack of knowledge about them.

- Noisy data. In-spite of obtaining very high resolution structural data for some proteins, flexibility and dynamics of molecules may lead to noisy data and structural disorder. Understanding flexibility of the structure assists in obtaining insights about the proteins' functions. Therefore, handling such partial information must be accommodated by appropriate computational models and tools.

Furthermore, the scientific challenges in structural bioinformatics have created the need to design computational and informatics methods to support findings of experimental structural biology, to explain biological phenomena and to obtain new insights from the data. Some of the major topics in structural biology that need to be addressed by informatics are as follows:

- Analysis of x-ray crystallographic data. Algorithms for automation of some phases of the x-ray crystallography procedure have existed in structural bioinformatics for a long time. A widely-used class of such algorithms assist in deconvolving the x-ray diffraction pattern by computing an inverse Fourier transform [8] [40]. Another useful class of algorithms assists in fitting and refinement of the structures after obtaining the electron densities so that they can be accessible as the known format (deposited in the PDB). We address a topic related to this challenge in this dissertation.

- Analysis of NMR data. NMR experiments provide us with distance constraints between atoms of a molecule. Computational and optimization methods are required to obtain 3D structures that satisfy these constraints.

- Evaluation of the 3D structures. Computational methods are required to assess the quality of the obtained structures from x-ray crystallography or NMR according to basic chemistry of the structures. Furthermore, the field is in need of automated methods to identify and annotate the functionally crucial regions such as the binding and active sites in the molecules.

- Study structure to understand function. As a fundamental principle, it is known that protein structure leads to protein function. Therefore, the promise of structural bioinformatics is to devise useful computational tools and apply them to structural data to gain insights about the function of molecules, their dynamics and the mechanisms via which they function. The main bottleneck here is the lack of sufficient structural information to create such a possibility. Therefore, simulation methods have been proposed and frequently used to generate such data. Another section of this dissertation is dedicated to addressing aspects of the function inference challenge.

According to the aforementioned aspects, one of the major challenges of structural biology for informatics is the demand for computational methods capable of "prediction", e.g., protein structure prediction (given protein sequence with(out) structural information about its backbone), predicting function from structural data, etc. As mentioned before, in-spite of recent growth in the PDB size, the number of deposited and high quality 3D structures is still far behind the available sequence data, due to the expensive and time-consuming procedure of extracting these structures. Hence, the need for accurate prediction methods that can support the current findings from experimental data becomes more significant and accentuated. Such prediction models and methods can be applied to predict future unknown 3D structures of proteins, given their sequences. The scientific studies in this regard comprise a well-established category of research endeavors in structural bioinformatics. Furthermore, world-wide competitions such as CASP (Critical Assessment of Structure Prediction) have been held since 1994 so that research groups can present capability of their structure prediction models in predicting structure of completely new proteins (3D structures cannot be found in the PDB). In addition, many attempts have been made to address an important sub-problem of the structure prediction problem. This sub-problem concerns predicting conformations of a protein's side chains, while the backbone is assumed to be fixed. Traditionally this problem is called the "side-chain prediction" (SCP) problem, and the goal is to predict the conformation of each residue so that the entire molecule achieves its lowest-energy. Hence, this problem is approached as an optimization problem. Since each side-chain conformation is parameterized by a sequence of dihedral angles (see Figure 1.1), the search space of this optimization problem is an infinite set of points representing the settings for all possible dihedral angles for all residues. Many

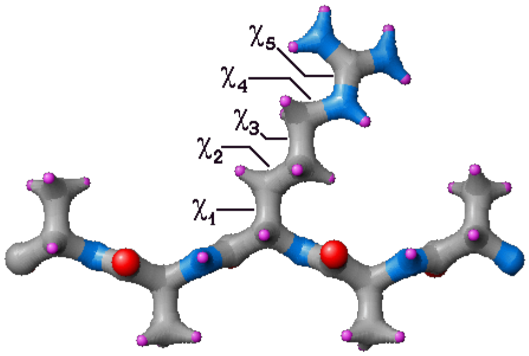


Figure 1.1: Illustration of dihedral angles, courtesy of http://www.ccp14.ac.uk/ccp/web-mirrors/garlic/garlic/ commands/dihedrals.html.

current computational approaches have reduced the problem to a combinatorial search problem by discretizing this search space and considering only the most likely conformations for the side chains. Despite all these attempts, the problem remains as a well-known NP-hard problem [2]. Therefore, a group of methods that address side-chain prediction have been concerned with finding the optimal solution (see Section 2.2). However, it is known that protein crystals populate diverse conformational ensembles. Very recent investigations have shown that further examination of x-ray crystallography data reveals more information about the residues' alternate conformations than previously thought. The phenomenon is called "conformational polymorphism". Despite much evidence that there is widespread conformational polymorphism in protein side chains, most ($\approx$ 95%) of the residues in x-ray crystallography data are modelled by single conformations in the Protein Data Bank. The ability to extract or to predict these conformational polymorphisms is of crucial importance, as it facilitates deeper understanding of protein dynamics and functionality. This is a missing aspect from all of the existing methods devised to predict side-chain conformations. This dissertation extends a group of efficient methods already devised for the SCP problem so that the final framework can both recognize continuity of the changes in the dihedral angles, i.e., avoids descretization of the dihedral angles' search space, and can predict side-chain polymorphism. This problem and its importance will be discussed in Chapter 2 of this dissertation.

Moreover, predicting function from structural information is another major goal in the structural biology. Understanding function and its correlation with structure can assist in generating new structures with desirable functionalities. This is specifically useful in pharmaceutical studies and designing new drugs. It was traditionally believed that similarity in protein sequences infers similarity in structure and function. This has been confirmed for the structure; but not for the function. In 2001, Todd et al. [104] showed that function is conserved for proteins with above 40% sequence identity (homologous proteins), and not for those below this threshold. However, there are homologous families in which function divergence has been observed [104] [1]. In contrast, there are protein families (such as the globin family) that have been subjected to mutations in their sequences; but their functions have been conserved [8]. Furthermore, some proteins share structural similarity with no sequence similarity; but converge to the same function. These proteins are called analogues. Therefore, the abundant sequence data cannot solely assist in inferring proteins' functions. This emphasizes the importance and necessity of models that can obtain such information from structural data.

In this dissertation, an important function-related topic is discussed that addresses how information is propagated in specific proteins from one site to another (distant) site. In other words, the topic concerns changes in structure or function occurring at one site of a protein as a consequence of a fluctuation at another site. This phenomenon is named "allostery". It has been proposed that conformational changes can propagate through altered fluctuations of protein, even if its average structure does not change [16] [23]. Many empirical observations, such as binding of Calmodulin to its ligand, supported this model and approved the allosteric effects of fluctuations in proteins. This led to the idea that allostery uses intrinsic long-range correlations and interactions in proteins. Hence, many attempts have been made to reveal and quantify those interactions, correlations and coupled motions through statistically correlated conformational changes of residues. These methods are mainly based on generation of a dataset of protein structures via Molecular Dynamics (MD), and application of a statistical method to extract correlations between the residues. However, the applied statistical methods such as mutual information or sample-based correlation cannot disentangle direct from indirect (transitive) dependencies or couplings between residues. In contrast, uncovering interaction pathways and networks along which information is propagated requires extracting "direct" couplings between the residues. Furthermore, traditionally backbone motions have been considered as the main source of allosteric behaviour in the proteins, recent studies have highlighted previously neglected allosteric effects of side-chain fluctuations. Furthermore, the functional roles of side-chain fluctuations become of more importance due to the discovered evidence regarding prevalent alternative conformations (conformational polymorphism) for the side chains. However, the commonly used MD simulations are solely capable of capturing very quick movements. Therefore, the slow motions of side chains are not captured by MD simulations. Hence, the recent findings and the deficiencies related to the currently existing methods emphasize the need for devising frameworks that carefully take the neglected effects of the side-chain fluctuations into account in allostery-related studies. For this purpose, this dissertation proposes two frameworks for inferring direct couplings between side chains and generating sparse networks of the coupled residues in Chapters 3 and 4.

In summary, the essential goal of this dissertation is characterizing the effects of the side-chain conformational polymorphisms and functional roles of the side-chain fluctuations in frameworks that facilitate modelling the conformational variables (dihedral angles) as continuous angular variables. To substantiate this thesis:

- Section 1.2 of this chapter briefly describes the key contributions of this dissertation.

- Chapter 2 presents a framework that extends a particular class of algorithms for SCP by modelling the side-chain dihedral angles more appropriately as continuous rather than discrete variables. The proposed method applies the Particle Belief Propagation (PBP) inferential method to predict residue-specific distributions that provide information about side-chain conformational polymorphism. By employing PBP integrated with multivariate von-Mises distributions, the framework is capable of modelling the dihedral angles as continuous angular variables.

- Chapter 3 proposes a framework to extract sparse networks of directly coupled side chains in allosteric proteins. The proposed framework, based on the application of Graphical Lasso (GLASSO), is capable of extracting couplings between alternate conformations of side chains, as well. While GLASSO is an efficient method to extract direct dependencies between variables, its application is restrictive and the conformational variables must be encoded as binary variables.

- Chapter 4 further extends the allostery-related study and presents a novel framework to extract directly coupled side chains. This framework facilitates direct modelling of the conformational variables as angular multivariate variables. It is based on integration of the following three concepts: partial correlation, Canonical Correlation Analysis (CCA), and the von-Mises kernel function. This chapter contains a comparison between this framework and the GLASSO-based (Chapter 3) framework, as well.

- Chapter 5 provides conclusion remarks, critical assessment of the key results, and discusses future research directions.

It is worthwhile to note that the material in Chapters 2, 3, and 4 have been published as journal articles [96] [97] [98].

1.2 Key Contributions

This dissertation proposes new algorithms to classic problems in structural bioinformatics with respect to major and recent state-of-the-art discoveries in the field. The following is a summary of the key contributions of this dissertation:

- The *first* computational framework capable of predicting side-chain conformational polymorphisms without requiring any discretization of the side-chain conformations in contrast to almost all of the existing methods for the SCP problem. The results are in good agreement with empirical results obtained from x-ray crystallography.

- The *first* computational framework for extracting *direct* couplings between fluctuating side chains in order to study interaction paths and sparse networks between residues in allosteric proteins. The results are in good agreement with both empirical findings and results of a model-free computational method. An interesting aspect of the framework is relying on the information of the inactive structure alone to infer critical functional residues. As well, the framework is capable of extracting couplings between alternate conformations of the side chains. This is an important aspect with respect to the recent findings about side-chain conformational polymorphism.

- Last but not least, a *novel* framework and also the *first* one that facilitates extracting direct couplings between continuous and angular multivariate variables. The framework facilitates modelling of the side chains' conformational variables directly. This is an improvement over the first proposed framework that requires discretization and encoding of the side-chain conformations. This novel framework can be considered the most significant contribution of this dissertation, since it can be adapted and applied to other problems in bioinformatics (see Section 5.3). The novel framework, as well as the previous framework, relies on the information of inactive structure alone and improved the previous framework's results for a set of small G proteins.

Chapter 2

Conformationally Polymorphic Side-Chain Prediction

2.1 Introduction

Due to the wide range of its motions, e.g., as shown by many studies using nuclear magnetic resonance (NMR) spectroscopy [47, 10, 79], a protein molecule can appear in many different conformations [30, 111]. As a result, it is insufficient to describe a protein molecule by a single model [67, 68]. One idea is to model the structure of such dynamic molecules as proteins more properly with conformational ensembles [7]. Capturing alternate conformations of a protein is of crucial importance for many applications, e.g., drug design, understanding disease mechanisms, etc; undoubtedly, doing so will bring crucial insight as well as deepen our understanding of how proteins fold, function, and bind to ligands [109, 31]. An important step in this direction is the ability to predict and describe the conformational polymorphism of each residue.

Since most residues belonging to structures in the Protein Data Bank (PDB; `http://www.rcsb.org/pdb/`) [5] are modeled by a single side-chain conformation, the majority of computational approaches for making side-chain predictions have focused on finding a single “best” conformation (more on this in Section 2.2 below). However, a few recent studies have started to reinvestigate crystallographic data, and to explore the phenomenon of side-chain polymorphism.

For example, van den Bedem et al. [108] developed a method to identify and model the conformational heterogeneity of proteins from electron density data. The output of their method is a co-called "multi-conformer model", or "an occupancy-weighted set of main-chain and side-chain conformations that collectively best represents the electron density" [108]. The word "occupancy" refers to the relative frequency of occurrence for each conformer in the crystal. The method first generates, in a sampling step, a large set of candidate conformations. In a subsequent selection step, the method fits the occupancies of this set of samples to the electron density map.

Recently, the Alber Lab at the University of California, Berkeley released a program called Ringer [56], which investigates side-chain conformational polymorphisms by sampling the electron density maps around the side-chain dihedral angles of each residue below the usual "1.0 sigma" threshold. Using Ringer, they uncovered evidence suggesting the presence of alternate, hitherto-unmodeled side-chain conformations, many of which are characterized by weak electron density features that were traditionally overlooked when building 3D models of proteins. They showed that their newly identified conformers are nonrandom and are biased towards low-energy rotational isomers. They also discovered, e.g., in Calmodulin, alternate side-chain conformations "not only on the surface but also within the structure" [73], where the protein is tightly packed and side-chain polymorphisms were rarely expected.

2.1.1 Our Contribution

We have developed a computational approach capable of predicting and describing side-chain polymorphisms — one that does *not* require experimental inputs such as electron density maps. Our approach is an extension of a particular class of algorithms for side-chain prediction that are based on belief propagation (BP) [85].

The conformation of a protein side chain is commonly represented by its dihedral angles. Each side chain may rotate flexibly about its dihedral angles, as long as there are no steric collisions. These dihedral angles are continuous in nature, but most computational approaches discretize them. Our primary extension was to model these dihedral angles more appropriately as continuous variables rather than discrete ones. Straight-forward as such an extension may sound, it would have remained difficult within the BP framework if a variation called "particle belief propagation" (PBP) [46] had not become available.

Using PBP, we were able to make inferences about residue-specific distributions in the continuous domain, and it is clear that these distributions encode information about the conformational polymorphism of each residue. We then compared the polymorphisms that we predicted with the ones extracted from crystallography data by Ringer [56]. Overall, we found the two sets of results to be in reasonably good agreement with each other. Ringer has successfully uncovered side-chain conformations that were formerly considered mere artifacts of (or noise from) electron density data. While Ringer found these alternate conformations by re-evaluating electron density maps, we can predict them with an improved — and, in fact, fundamentally different — side-chain prediction algorithm, one that works in a continuous domain.

2.1.2 Outline

The rest of this chapter is organized as follows. In Section 2.2, we quickly review computational approaches for side-chain prediction. In Section 2.3, we describe our inference method. We first give a very brief review of belief propagation (Section 2.3.1), and then describe a recent variation, called particle belief propagation, that drives our main algorithm (Section 2.3.2). In Section 2.4, we review the von-Mises (VM) distribution for angular data, and explain how we have used mixtures of VM distributions to speed up our algorithm. In Section 2.5, we report some empirical experiments and their results. Finally, in Section 2.6, we summarize our main contributions and discuss some future work.

2.2 Background and Related Work

As mentioned in the first chapter, while our goal is to address SCP problem by using strategies that recognize the continuity of the changes in the dihedral angles, many current computational approaches have reduced the problem to a combinatorial search problem by discretizing the allowed settings of the dihedral angles in the residues. This strategy capitalizes on the phenomenon of rotamericity. Even though a side chain has an infinite number of possible three dimensional conformations, it has been observed that a side chain will typically have a tendency to adopt a conformation that can be approximated by dihedral angles chosen from a small set of empirically observed settings. Each such possible conformation is called a rotamer. A rotamer library contains a discrete set of conformations for each residue type. For instance, the backbone-dependent rotamer library provided by the Dunbrack Lab [25] has been used by many researchers. Algorithms relying on these rotamer libraries essentially apply different heuristics to search for the optimal combination of rotamers, one for each residue.

A state-of-the-art heuristic is the SCWRL algorithm [9]. The main steps in SCWRL 3.0 [13] are as follows: first, a dead-end elimination procedure is applied to reduce the number of candidate rotamers for each residue; next, a graph is created by treating residues as nodes and by drawing edges between all nearby residues; then, the graph is clustered into many bi-connected components; finally, the optimization problem is solved separately on each subgraph, before the solutions are combined.

Another highly competitive heuristic is the TreePack algorithm [117]. It also models the protein molecule as a graph. However, loops are removed and the graph is decomposed into clusters and modeled as a tree. The problem of assigning an optimal rotamer to each residue is then solved efficiently by traversing the tree. TreePack is as accurate as, but significantly faster than, SCWRL 3.0.

Other standard optimization techniques such as linear programming (LP) and integer programming (IP) also have been applied to solve the side-chain prediction problem. Yanover and Weiss showed that finding the minimum energy configuration of a protein's side chains is equivalent to finding the maximum-a-posteriori (MAP) configuration of an undirected graphical model, or a Markov random field (MRF) [118, 120]. This meant the side-chain prediction problem could be formulated as a MAP estimation problem, which could be solved using belief propagation (BP). They also considered a relaxed version of the IP problem and solved the resulting convex problem with BP [119].

Besides these optimization approaches, Li et al. [61] recently showed that side-chain conformations also can be decided from backbone information without optimization.

2.3 Main Method of Inference

We have extended the class of side-chain prediction algorithms that are based on BP. To model a protein molecule with a graphical model, the backbone is regarded as being fixed and the residues $r_1, r_2, ..., r_n$ are regarded as nodes. The side chain at each node is described by a sequence of dihedral angles, collectively stored as a vector, e.g., $r_i = (\chi_{i1}, \chi_{i2}, ..., \chi_{i4})$. The exact number of dihedral angles depends on the type of amino acid. The objective is to find the minimal-energy conformation,

$$\min_{r_1, r_2, \ldots, r_n} \left[\sum_i^n E_l(r_i) + \sum_i^n \sum_{j>i} E_p(r_i, r_j) \right], \tag{2.1}$$

where E_l is the (local) intrinsic energy of a residue, E_p is the pairwise energy between two residues, and n is the total number of residues. The optimization algorithm itself is independent of the choice of the energy function. We will say more about the energy function later in Section 2.4.3.

Consider a graphical model $\mathcal{G}$, with a set of vertices $\mathcal{V}$ and a collection of edges $\mathcal{E}$. If r_i denotes the random variable associated with node i, then the joint probability distribution of $\mathbf{r} = (r_1, r_2, ..., r_n)$ can be factorized as follows:

$$P(\mathbf{r}) = \prod_{i \in \mathcal{V}} \rho_i(r_i) \prod_{(i,j) \in \mathcal{E}} \rho_{ij}(r_i, r_j). \tag{2.2}$$

The functions, $\rho_i(r_i)$ and $\rho_{ij}(r_i, r_j)$, are called node- and edge-potentials, respectively. Suppose we are given an energy function represented by E_{conf}, the Boltzmann distribution is given by

$$P_{conf}(\mathbf{r}) = \frac{1}{Z} \exp \left[\frac{-E_{conf}(\mathbf{r})}{T} \right] \tag{2.3}$$

where T is a temperature parameter and Z is a normalizing constant. Clearly, using the energy function (2.1), the distribution (2.3) can be expressed in the form of (2.2), with

$$\begin{aligned} \rho_i(r_i) &\propto \exp \left[\frac{-E_l(r_i)}{T} \right], \\ \rho_{ij}(r_i, r_j) &\propto \exp \left[\frac{-E_p(r_i, r_j)}{T} \right]. \end{aligned} \tag{2.4}$$

2.3.1 Belief Propagation (BP)

Belief propagation (BP) is an efficient local message-passing algorithm [85] for making inferences on graphical models. It performs exact inference if the graph $\mathcal{G}$ is a tree, and approximate inference for general graphs. If the graph contains cycles, the algorithm is often referred to as "loopy BP" and there is no convergence guarantee, but many groups have reported excellent results nonetheless, e.g., [33, 81, 34]. Indeed, loopy BP has been applied to the side-chain prediction problem, and the results have been comparable to such state-of-the-art software as SCWRL 3.0 [118, 120].

Given potential functions as defined by (2.4), messages (see Eq. 2.5) are computed along the edges of the graph. The sum-product BP algorithm is used to compute marginal distributions; its recursive message updating equation is as follows:

$$m_{i \to j}^{(t)}(r_j) = \sum_{r_i \in \mathcal{R}_i} \left[\rho_i(r_i) \rho_{ij}(r_i, r_j) \times \prod_{k \in \mathcal{N}(i) \backslash j} m_{k \to i}^{(t-1)}(r_i) \right], \tag{2.5}$$

where $\mathcal{R}_i$ is the (discrete) state space of r_i, and $m^{(t)}_{i\to j}$ represents the message from node i to j at iteration t. The notation, $\mathcal{N}(i)$, denotes the set of nodes that are neighbours of i. For proteins, the discrete state space $\mathcal{R}_i$ is simply the set of rotamers for residue r_i.

The max-product BP algorithm is used for finding MAP estimates; its recursive updating equation is given by

$$m^{(t)}_{i\to j}(r_j) = \max_{r_i \in \mathcal{R}_i} \left[\rho_i(r_i)\rho_{ij}(r_i, r_j) \times \prod_{k\in\mathcal{N}(i)\backslash j} m^{(t-1)}_{k\to i}(r_i) \right] \tag{2.6}$$

2.3.2 Particle Belief Propagation (PBP)

For many applications, e.g., in bioinformatics, computer vision, and other fields, the state space is continuous rather than discrete, or it can be discrete but very large so that enumerating all possible states at each iteration becomes very inefficient. Particle belief propagation (PBP) has been developed recently to address precisely such difficulties [46]. Our goal is to model the conformation of side chains more appropriately as continuous rather than discrete random variables. Hence, PBP is a crucial piece of technology for our work.

The key idea for PBP is the following: At iteration t, if we draw r_i from a certain trial distribution $W^{(t)}_i$, then (2.5) can be written as an "importance-sampling corrected expectation" [46]:

$$m^{(t)}_{i\to j}(r_j) = \mathbb{E}_{r_i\sim W^{(t)}_i} \left[\frac{\rho_i(r_i)}{W^{(t)}_i(r_i)} \rho_{ij}(r_i, r_j) \times \prod_{k\in\mathcal{N}(i)\backslash j} m^{(t-1)}_{k\to i}(r_i) \right]. \tag{2.7}$$

It is generally not possible to express the expectation $\mathbb{E}_{r_i\sim W^{(t)}_i}(\cdot)$ in analytic form, but it can be obtained using Monte Carlo techniques [22, 63]. Thus, for each node r_i, the idea is to sample a set of L particles $\{r^{(1)}_i, r^{(2)}_i, ..., r^{(L)}_i\}$ from $W^{(t)}_i$, typically using a Markov Chain Monte Carlo (MCMC) technique such as the Metropolis-Hastings algorithm [63], and then approximate (2.7) by

$$\widehat{m}^{(t)}_{i\to j}(r_j) = \frac{1}{L}\sum_{l=1}^{L} \left[\frac{\rho_i\left(r^{(l)}_i\right)}{W^{(t)}_i\left(r^{(l)}_i\right)} \rho_{ij}\left(r^{(l)}_i, r_j\right) \times \prod_{k\in\mathcal{N}(i)\backslash j} \widehat{m}^{(t-1)}_{k\to i}\left(r^{(l)}_i\right) \right]. \tag{2.8}$$

In the simplest case, the particles' locations may remain unchanged [46] but, generally, each particle's location is updated at the end of each BP iteration. This is what allows

PBP to explore a continuous state space and not be restricted to a fixed set of choices specified a priori, such as a rotamer library; and it is accomplished by re-sampling the particles at each iteration from the distribution, $W_i^{(t)}$, e.g., using the Metropolis-Hastings algorithm. At iteration t, a natural choice of $W_i^{(t)}$ is the current belief of node i,

$$W_i^{(t)}(r_i) \quad \propto \quad \rho_i(r_i) \times \prod_{k \in \mathcal{N}(i)} \widehat{m}_{k \to i}^{(t-1)}(r_i). \tag{2.9}$$

The max-product version for PBP was first given by Kothapa et al. [54]:

$$\widehat{m}_{i \to j}^{(t)}(r_j) = \max_{l=1,\ldots,L} \left[\rho_i \left(r_i^{(l)} \right) \rho_{ij} \left(r_i^{(l)}, r_j \right) \times \prod_{k \in \mathcal{N}(i) \setminus j} \widehat{m}_{k \to i}^{(t-1)} \left(r_i^{(l)} \right) \right]. \tag{2.10}$$

Their paper [54] explains in more detail why the factor $W_i^{(t)}$ does not appear in the square brackets of (2.10).

2.4 Fast Approximation of ρ_i, ρ_{ij}

We used mixtures of von-Mises distributions as a fast way to approximate the potential functions $\rho_i(r_i)$ and $\rho_{ij}(r_i, r_j)$. It is well-known that mixture densities can be used to approximate any arbitrary distribution. For example, in nonparametric belief propagation [102], mixtures of Gaussians are used to model and/or approximate $\rho_i(r_i)$ and $\rho_{ij}(r_i, r_j)$. Since we are working in the space of dihedral *angles*, the von-Mises distribution is more appropriate than the Gaussian distribution (see Section 2.4.1 below).

2.4.1 The von-Mises (VM) Distribution

The univariate von-Mises (VM) distribution is a probability distribution on a circle. The multivariate generalization was introduced by Mardia et al. [70]. In particular, $\theta \in \mathbb{R}^d$ is said to follow the multivariate von-Mises distribution, $\mathrm{MVM}(\mu, \kappa, \mathbf{\Lambda})$, if its density function is given by

$$f(\theta; \mu, \kappa, \mathbf{\Lambda}) = \frac{1}{Z(\kappa, \mathbf{\Lambda})} \times \exp \left[\kappa^T c(\theta) + \frac{s^T(\theta) \mathbf{\Lambda} s(\theta)}{2} \right] \tag{2.11}$$

where

$$c_u(\theta) \equiv \cos(\theta_u - \mu_u), \quad s_u(\theta) \equiv \sin(\theta_u - \mu_u)$$

for $u = 1, 2, ..., d$, and $Z(\kappa, \mathbf{\Lambda})$ is a normalizing constant.

The parameter $\mu \in \mathbb{R}^d$ describes the location, i.e., the mean (or center), and the parameter $\kappa \in \mathbb{R}^d > \mathbf{0}$ describes the scale, i.e., the spread (or concentration). The parameter, $\mathbf{\Lambda} = [\lambda_{uv}] \in \mathbb{R}^{d \times d}$ is a matrix whose diagonal elements are zero ($\Lambda_{uu} = 0$) and whose off-diagonal elements Λ_{uv} capture the correlation between θ_u and θ_v. It is clear from the definition above that the VM distribution is well suited for modeling angular data, and why it is sometimes referred to as the "Gaussian" distribution on the sphere.

2.4.2 Use of VM Distribution in Bioinformatics

The von-Mises distribution has been used to model dihedral angles in protein molecules. For example, Mardia et al. [69] used the EM algorithm to fit a mixture of bivariate von-Mises distributions to the two dihedral angles (ϕ, ψ) that describe protein backbones. To model higher-dimensional angular data (e.g., the dihedral angles for side chains), Mardia et al. [70] introduced the more general, multivariate von-Mises distribution (2.11) by extending the bivariate model of Singh et al. [95]. More recently, Mardia et al. [71] have extended single MVM distributions to mixtures of MVMs. For example, they fitted a 4-dimensional mixture of MVMs to model the two backbone dihedral angles (ϕ and ψ) and the first two side-chain dihedral angles (χ_1 and χ_2) of the amino acid, ILE.

In our work, we also used mixtures of MVMs (see Section 2.4.4 below). However, our work differs fundamentally from those of Mardia et al. [71]. While they fitted a *single* mixture to model the conformation of a given amino acid using data from different proteins, we used mixtures of MVMs to approximate the node- and edge-potential functions, and *different* mixture models were specified for each ρ_i and ρ_{ij} on a protein-by-protein basis.

2.4.3 Energy Functions

We now give more details about the energy function (2.1). We used a very simple energy function that essentially acted as a collision detector. This simple energy function was first popularized by SCWRL [25] and later adopted by TreePack [117] as well.

Given two atoms, a_1 and a_2, SCWRL approximates the van der Waals pairwise potential energy between them by

$$E_{apprx.vdw}(a_1, a_2) = \begin{cases} 0, & \text{if} \quad d > R_0; \\ -k_2 \frac{d}{R_0} + k_2, & \text{if} \quad k_1 R_0 \leq d \leq R_0; \\ E_{max}, & \text{if} \quad d < k_1 R_0, \end{cases} \tag{2.12}$$

where d is the distance between a_1 and a_2; R_0 is the sum of their radii; $E_{max} = 10$; $k_1 = 0.8254$; and $k_2 = E_{max}/(1 - k_1)$. For Carbon (C), Nitrogen (N), Oxygen (O), and Sulfur (S), fixed radii of 1.6, 1.3, 1.7, and 1.7 were used, respectively.

Treating each residue simply as a set of atoms, the pairwise energy function, $E_p(r_i, r_j)$ in (2.1), is merely calculated by summing over all atom-pairs:

$$E_p(r_i, r_j) = \sum_{a \in r_i, b \in r_j} E_{apprx.vdw}(a, b).$$

The intrinsic energy $E_l(r_i)$ in (2.1) is computed by

$$E_l(r_i) = -K \log \frac{p(r_i|\phi, \psi)}{p_{max}(r_i|\phi, \psi)} + \sum_{\substack{j<i-1 \\ j>i+1}} E_p(r_i, b_j), \tag{2.13}$$

where $p(r_i|\phi, \psi)$ is the rotamer probability specified by the rotamer library, which depends on the two backbone dihedral angles ϕ and ψ; $p_{max}(r_i|\phi, \psi)$ is the probability of the most probable rotamer among the rotamers listed in the library for residue i; and b_j represents the backbone part of residue j. The Dunbrack Lab [25] has suggested that the parameter K be set to 3. In order to calculate $E_p(r_i, b_j)$, r_i and b_j are again treated simply as two sets of atoms, and $E_p(r_i, b_j)$ is computed in the same fashion as (2.12) over all pairs of atoms in r_i and b_j.

2.4.4 Approximation of Potential Functions

Notice that the energy functions E_l and E_p given in the previous section — and hence the implied potential functions ρ_i and ρ_{ij}, given by (2.4) — depend on inter-atomic *distances*, whereas our state space is a set of dihedral *angles* that describe the conformation of each residue. Therefore, a conversion must take place every time the potential functions are evaluated. This is not difficult in principle, and there is existing, standard software for performing such a conversion, e.g., BALL [41]. In order to speed up our computation, however, we used a mixture of von-Mises distributions as a crude approximation to these potential functions.

For example, for residues described by four dihedral angles (e.g., the amino acid LYS), the approximation to the node potential function would be:

$$\widehat{\rho}_i(r_i) \;\; = \;\; \sum_{\tau} w_\tau f_\tau(\chi_{i1}, ..., \chi_{i4}), \tag{2.14}$$

where $f_\tau \sim \text{MVM}(\mu_\tau, \kappa_\tau, \mathbf{\Lambda}_\tau)$ is a (multivariate) von-Mises density function, given by (2.11), and w_τ denotes the weight of the mixture component τ such that $\sum w_\tau = 1$.

We used simple *spherical* or *radial basis* mixtures, that is, we set $\mathbf{\Lambda}_\tau = \mathbf{0}$. This is the same as treating the dihedral angles as being locally independent — in the future, we plan to generalize this by modeling the local correlations among the χ-angles. We chose $\kappa_\tau = (10, 10, ..., 10)$ for all τ. For each residue i, the set of discrete rotamers from the (backbone-dependent) rotamer library were used as mixture centers, μ_τ. We specified the weight of each component τ to be

$$w_\tau = \alpha\ p(\mu_\tau|\phi, \psi) + (1-\alpha)\frac{1}{\#(\text{mixture components})},$$

where α was chosen to be 0.1, and $p(\mu_\tau|\phi, \psi)$ is the rotamer probability for rotamer μ_τ from the rotamer library.

For the edge potential, our approximation was:

$$\widehat{\rho}_{ij}(r_i, r_j) = \sum_{\tau}\sum_{\tau'} w_{\tau,\tau'} \times f_{\tau,\tau'}(\chi_{i1}, ..., \chi_{id_i}, \chi_{j1}, ..., \chi_{jd_j}), \tag{2.15}$$

where d_i and d_j are the number of dihedral angles for r_i and r_j, respectively; $f_{\tau,\tau'}$ is, again, a spherical MVM density function, with $\kappa_{\tau,\tau'} = (10, 10, ..., 10)$ and $\mathbf{\Lambda}_{\tau,\tau'} = \mathbf{0}$ for all τ, τ' as before. If the pairwise edge potential between two rotamers — one for residue r_i and another for r_j — exceeded 0.05, their dihedral angles were concatenated together and used as a mixture center, $\mu_{\tau,\tau'}$, with $w_{\tau,\tau'} \propto \rho(\mu_\tau, \mu_{\tau'})$.

The expressions (2.14) and (2.15) can be viewed as approximations of ρ_i and ρ_{ij} using a crudely specified single-layer radial basis function network (RBFnet) in angular space.

2.5 Experiments and Results

We used a data set containing 362 diverse proteins that were previously analyzed by the Alber group [56] with their Ringer program (http://ucxray.berkeley.edu/ringer.htm). The protein data files were retrieved from the PDB. We used functions and modules from the Biochemistry Algorithms Library (BALL) [41] to read and process the PDB files. The electron density maps for the proteins, which were required to run the Ringer program, were downloaded from the Electron Density Server [53]. All experiments were performed on the Sharcnet system (http://www.sharcnet.ca/).

2.5.1 PBPMixVM

To reflect the fact that we used PBP for inference and mixtures of (multivariate) VM distributions for approximating the potential functions, from this point on we shall refer to our algorithm as PBPMixVM. We implemented it in `C++`, using the overall architecture provided by GraphLab (`http://select.cs.cmu.edu/code/graphlab/`) [65].

2.5.2 The Kolmogorov-Smirnov (KS) Test

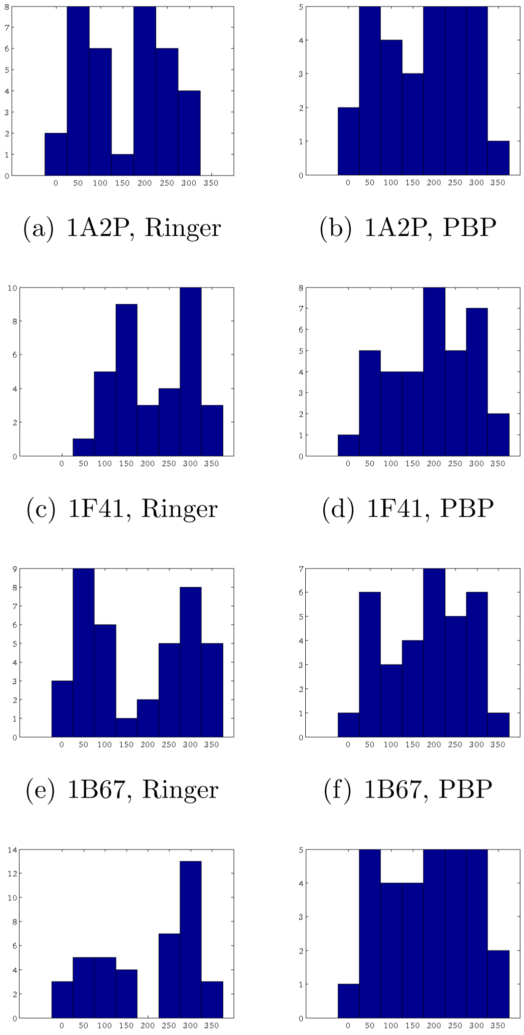
For each residue, we used a two-sample Kolmogorov-Smirnov (KS) test [15] to compare the results from Ringer with those from PBPMixVM. The KS-test is a widely used non-parametric test for determining whether two distributions are significantly different from each other. The significance level for the KS-test was set to be 0.05.

Fig. 2.1 provides some visual illustrations of what the KS-test does. Based on the p-values from individual KS-tests, we selected four residues, whose corresponding p-values from the aforementioned KS-tests were 0.99, 0.77, 0.53, and 0.25, respectively. Such a selection is meant to illustrate varying levels of agreement between the PBPMixVM results and the Ringer results — larger p-values indicate higher levels of agreement.

The four selected residues are: residue ASN23, 1A2P, Chain C [72], p-value=0.99; residue HIS88, 1F41, Chain B [43], p-value=0.77; residue GLU134, 1B67 [19], p-value=0.53; and residue LYS87, 1F4P [88], p-value=0.25. For these four residues, the polymorphisms in their respective χ_1-angles as characterized by Ringer and predicted by PBPMixVM are displayed next to each other in Fig. 2.1, ordered by p-values from top to bottom.

2.5.3 Comparative Results

From the individual KS-tests, we computed two summary statistics to evaluate the overall agreement between our results from PBPMixVM and those given by Ringer: (i) *percent in agreement* — the fraction of residues for which the KS-test failed to show a statistically significant difference; and (ii) *mean p-value* — the average p-value from individual KS-tests.



(a) 1A2P, Ringer

(b) 1A2P, PBP

(c) 1F41, Ringer

(d) 1F41, PBP

(e) 1B67, Ringer

(f) 1B67, PBP

(g) 1F4P, Ringer

(h) 1F4P, PBP

Figure 2.1: Illustration of results from Ringer (left) and those from PBPMixVM (right; simply labeled as “PBP” in the plots): four specific residues, whose polymorphisms in χ_1 differed to varying degrees as characterized by Ringer and predicted by PBPMixVM. (a) & (b) Residue ASN23, 1A2P, Chain C; KS-test p-value=0.99. (c) & (d) Residue HIS88, 1F41, Chain B; KS-test p-value=0.77. (e) & (f) Residue GLU134, 1B67; KS-test p-value=0.53. (g) & (h) Residue LYS87, 1F4P; KS-test p-value=0.25. Larger p-values indicate better agreement between Ringer and PBP.

Table 2.1: Comparison of PBPMixVM and Ringer results: χ_1 and χ_2, average results over all residues, across all proteins in the data set.

Dihedral angle	Percent in agreement (%)	Mean p-value
χ_1	57	0.19
χ_2	56	0.17

Table 2.2: Comparison of PBPMixVM and Ringer results: χ_1 only, average results over all residues of the same amino acid type, across all proteins in the data set.

Amino acid	Total No.	Percent in agreement (%)	Mean p-value
ARG	4067	57	0.18
ASN	3613	67	0.24
ASP	5072	63	0.21
CYS	1342	62	0.21
GLN	3180	67	0.24
GLU	5466	71	0.26
HIS	2080	52	0.16
ILE	4894	52	0.17
LEU	8250	64	0.22
LYS	4890	53	0.16
MET	1348	69	0.23
PHE	3675	48	0.14
PRO	4097	42	0.09
SER	4776	77	0.29
THR	4749	77	0.29
TRP	1259	54	0.16
TYR	3044	50	0.15
VAL	6422	66	0.23

Table 2.3: Comparison of PBPMixVM and Ringer results: χ_1 only, average results over all residues of the same amino acid type *and* having the same secondary structure, across all proteins in the data set.

Amino acid	Secondary structure	Total No.	Percent in agreement (%)	Mean p-value
ARG	Helix	1774	60	0.19
	Strand	815	56	0.17
	Loop	1478	56	0.17
ASN	Helix	1070	64	0.21
	Strand	541	64	0.23
	Loop	2002	69	0.25
ASP	Helix	1764	62	0.21
	Strand	571	59	0.19
	Loop	2737	65	0.22
CYS	Helix	423	63	0.22
	Strand	409	57	0.20
	Loop	510	66	0.20
GLN	Helix	1573	71	0.26
	Strand	546	66	0.22
	Loop	1060	61	0.20
GLU	Helix	2854	73	0.27
	Strand	861	66	0.24
	Loop	1751	70	0.25
HIS	Helix	679	53	0.16
	Strand	456	52	0.17
	Loop	945	51	0.16
ILE	Helix	1804	51	0.16
	Strand	1885	52	0.17
	Loop	1205	52	0.18
LEU	Helix	3947	71	0.26
	Strand	2107	57	0.18
	Loop	2195	60	0.20
LYS	Helix	2203	54	0.17
	Strand	870	54	0.17
	Loop	1817	50	0.15
MET	Helix	581	68	0.24

	Strand	318	67	0.22
	Loop	449	70	0.24
PHE	Helix	1383	51	0.14
	Strand	1179	42	0.13
	Loop	1112	50	0.15
PRO	Helix	841	44	0.09
	Strand	363	44	0.10
	Loop	2893	42	0.09
SER	Helix	1462	78	0.30
	Strand	884	74	0.28
	Loop	2430	77	0.29
THR	Helix	1450	78	0.29
	Strand	1255	75	0.27
	Loop	2043	76	0.29
TRP	Helix	487	58	0.18
	Strand	392	48	0.14
	Loop	380	55	0.17
TYR	Helix	1122	51	0.15
	Strand	1009	46	0.13
	Loop	913	52	0.16
VAL	Helix	2087	72	0.26
	Strand	2736	63	0.21
	Loop	1599	65	0.24

Table 2.1 shows the overall results for the first two dihedral angles, χ_1 and χ_2, averaged over all residues across all proteins in our data set. Table 2.2 shows results for χ_1 only, averaged over residues of the same amino acid type across all proteins in our data set. Based on the KS-tests, these proteins' side-chain polymorphisms as predicted by PBPMixVM and as described by Ringer agreed for well over 50% of all the residues, and the average p-value from these KS-tests was about 0.20, much higher than the typical cutoff value of 0.05.

We can observe from Table 2.2 that, for some residue types, including not only those having just one χ-angle, e.g., Serine (SER), Threonine (THR), Valine (VAL), but also those having relatively large structures and hence, multiple χ-angles, e.g., Asparagine (ASN), Glutamine (GLN), Glutamic Acid (GLU), Methionine (MET), the agreement between PBPMixVM and Ringer can be noticeably higher than the overall average (Table 2.1)

— the *percent in agreement* for some residue types was close to 70% and 80%, and the corresponding *mean p-value* was close to 0.25 and 0.30. To the extent that the Ringer program can discover alternate side-chain conformations, PBPMixVM can be seen to have the ability to predict alternate side-chain conformations for these residues as well. The agreement with Ringer for large residues having multiple χ-angles, in particular, are indications that using mixtures of *locally independent* VM distributions in our approximation of the potential functions (see Section 2.4.4) has not had a significant impact on our algorithm.

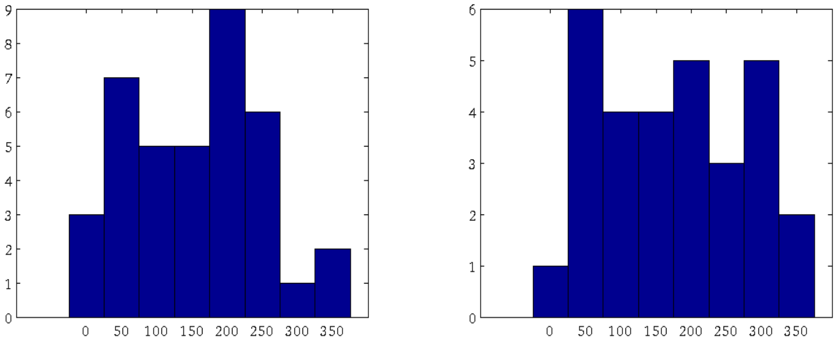
Table 2.3 further groups the results by the secondary structures of the residues, which we obtained using the DSSP software [50, 49]. Here we easily can see that the agreement between PBPMixVM and Ringer was generally better for residues whose secondary structures are helices. This is not surprising, since helices typically are more stable.

Earlier, we mentioned in Section 2.1 that, using Ringer, Lang et al. [56] had uncovered interesting polymorphisms in the protein, Calmodulin (CaM) [114]. An interesting residue in that protein is SER38. The currently modeled χ_1-angle for SER38 changes conformation from 80° in the unbound form of CaM (PDB ID: 1EXR) to 295° in the complex form (PDB ID: 2O5G). By analyzing crystallography data for 1EXR (the unbound form of CaM), Ringer successfully recognized the 295° conformation — often detectable only from 2O5G (the complex form of CaM) — as a secondary peak. This result was scientifically significant because, previously, such conformational changes could not have been easily identifiable without a complete structural refinement of both the bound and the unbound proteins, but Ringer was able to detect this conformational polymorphism from the unbound molecule alone. We also analyzed 1EXR with PBPMixVM. Our predicted polymorphism for the χ_1-angle of SER38 agreed well with the result from Ringer (Fig. 2.2; KS-test p-value = 0.64). In particular, PBPMixVM also predicted the secondary conformation near 295° from the unbound form of CaM (1EXR).

2.5.4 Some Computational Details

The PBPMixVM algorithm was deemed to have converged when the Kullback-Leibler (KL) divergence between $W_i^{(t)}(r_i)$ and $W_i^{(t-1)}(r_i)$ fell below 10^{-8} for each residue i, where $W_i^{(t)}$ denotes the belief function of node i at iteration t, given previously in (2.9). For all proteins in our data set, PBPMixVM converged in < 50 iterations.

At the moment, PBPMixVM is relatively slow, compared with some other side-chain prediction algorithms such as SCWRL. Whereas the running time of SCWRL is on the order of seconds or minutes, that of PBPMixVM is on the order of hours. This is mainly because, within *each* PBP iteration, we must run a *separate* MCMC to update the particles



(a) Ringer

(b) PBP

Figure 2.2: Calmodulin (1EXR), residue SER38: polymorphism in χ_1 as extracted by Ringer from crystallography data (left) and predicted by PBPMixVM (right).

for *each* residue! To speed up the computation, we used a relatively small number of particles and relatively short MCMC chains to analyze all the proteins.

However, we did examine, using a small subset of 20 proteins, how much the performance of PBPMixVM could be affected by these computational parameters. On this small subset at least, increasing the length of the MCMC chains and the number of particles per residue had little effect on the overall level of agreement between PBPMixVM and Ringer.

2.6 Chapter Summary

Proteins in crystals undergo a lot of large- and small-scale motions. Hence, studying a protein molecule with a single conformational model is not adequate. We have developed a computational approach capable of predicting residue-specific conformational polymorphisms. We modeled side-chain dihedral angles as continuous random variables in an MRF, and used PBP as our main inference technique. To speed up the computation, we approximated the (continuous) node- and edge-potential functions by mixtures of VM distributions. For each node in the MRF, a set of particles were sampled at each iteration to represent its distribution. After convergence, these node-specific marginal distributions could be seen to encode information about alternate side-chain conformations. To the best of our knowledge, this work is the first to address the prediction of side-chain polymorphisms from a purely computational point of view, without relying on additional experimental inputs such as electron density data.

A distinct feature of our method is the treatment of side-chain dihedral angles as continuous variables. We think it constitutes an important (and necessary) step toward being able to provide an accurate description of side-chain ensembles, and to discover low-occupancy conformers.

We used SCWRL energy function that is a linear approximation of the van der Waals (see Section 2.4.3) energy function. An advantage of our method is that it is independent from the choice of energy function. Therefore, more complex energy functions can be integrated into the framework easily.

As mentioned earlier (Section 2.5.4), PBPMixVM is relatively slow at the moment, due to the need to update the particles for *each* residue by a *separate* MCMC within *each* PBP iteration. Although we haven't yet done so, the running time of our algorithm could be improved significantly by parallelizing some of these updates. For local message passing, we are currently using a synchronous schedule, but there have been suggestions that using an asynchronous schedule could further accelerate BP-types of algorithms [27].

We are also considering some other refinements to our algorithm, for example, improving our approximation of the potential functions by modeling the local correlations among the dihedral angles. We also think that combining the results from PBPMixVM with those from state-of-the-art side-chain prediction algorithms, such as SCWRL [55] and TreePack [117], can further enhance the accuracy and reliability of the predicted side-chain polymorphisms.

Chapter 3

Inferring Direct Couplings between Polymorphic and Functional Side Chains via Graphical Lasso

3.1 Introduction

A key component of information conveyance and communication in a protein molecule is its *allosteric* behaviour. Allostery is defined as any change in a protein's function or structure at one site in response to a modification or perturbation at another site [23]. Ligand-binding information or other modifications can be propagated in a protein from one site to another through the protein's altered conformational fluctuations. The observed role of correlated motions in allosteric behaviour of proteins has led to the idea that allostery occurs through pre-defined correlations and interaction networks in a protein [23]. Hence, studying correlated conformational changes of residues is of significant importance for understanding allostery.

Common methods to examine a protein's allosteric behaviour include NMR spectroscopy techniques or simulation-based methods such as molecular dynamics (MD) [74], Monte Carlo (MC) simulations [23] and Rosetta-based analysis [52]. Recently, using Monte Carlo simulations, DuBay et al. [23] demonstrated and emphasized the contribution of correlated side-chain fluctuations alone in transmitting intra-protein communications, whereas backbone motions used to be considered the key element of these communications. DuBay et al. [23] performed the experiments on proteins previously shown to have a relatively rigid backbone but significantly variable side-chain conformations, specifically in the core area [115]. The authors verified their results by comparing to previous NMR experiments that had shown mutation in a residue can influence conformation of residues distant from the mutation site. Such observation had previously suggested presence of virtual networks of correlated residues [62].

Studying correlated motions of side chains in order to understand a protein's allosteric behaviour becomes more crucial considering recent investigations by Lang et al. [56] and Van den Bedem et al. [108] on alternate conformations of side chains. These studies suggest that despite the fact that about 90% of the side chains are modelled uniquely, more careful examination of x-ray crystallography data reveals undiscovered conformations, and hence, alternate side-chain conformations are more prevalent than previously assumed. This phenomenon is known as side-chain *conformational polymorphism*. Furthermore, very recently, Van den Bedem et al. [107], have extended their study and proposed an application, named CONTACT, to determine correlated displacements of conformationally polymorphic residues. Applying CONTACT to the results of their polymorphic/multi-conformer model (qFit [108]), they identified contact networks of conformationally polymorphic residues that collectively respond to perturbations. Their results were in agreement with the findings of NMR experiments, and thus, highlighted the importance of considering conformational polymorphism in studying residue couplings.

The observed role of side-chain motions in information propagation and the newly obtained polymorphism data from x-ray crystallography suggest that studying interaction networks of conformationally polymorphic residues may reveal key insights to the information transmission within the molecule and the molecule's dynamics and functionality. It may provide clearer explanations of some unknown characteristics and mechanisms of a protein's allosteric behaviour. To the best of our knowledge, CONTACT is the only application that provides interaction networks based on information about side-chain conformational polymorphism. The field still lacks methods that take such information into account. Furthermore, currently, the proposed approaches to study or predict side-chain conformational polymorphism [56] [108] [96] lack a component capable of extracting the dependencies among alternate conformations of polymorphic side chains. Such dependency information can provide a complete understanding of side-chain dynamics during an allosteric and/or a ligand-binding event.

In this chapter, we propose an efficient computational approach to the aforementioned problems. While simulation-based techniques can be time-consuming, and not completely capable of capturing the slower rearrangements of protein side chains, developing a fast and efficient computational technique to extract correlations can be very helpful. In this study, we first predict directly correlated side chains by applying a sparse graph estimation technique, and generate intuitive network representations from the couplings (please see

Section 3.2). As done in CONTACT [107], we take side-chain conformational polymorphism into account, and characterize dependencies between each pair of residues based on the computed couplings for the alternate conformations of the two (for more information, see Section 3.2.3). By applying our method to the inactive structure of an allosteric protein, we reveal directly coupled side chains in allosterically important regions. Hence, we show that our method is capable of inferring interaction networks of functional residues. Moreover, our method is capable of revealing possible co-occurrences of alternate conformations of a side-chain pair as well, via the computed couplings of alternate conformations. We can provide such information as networks and/or lists of coupled alternate conformations for any coupled residue pair of a protein.

This chapter is organized as follows. First, we describe the proposed computational methods in Section 3.2, including the generation of a conformational ensemble and the application of a sparse graph estimation technique (GLASSO) to quantify the strengths of pairwise side-chain couplings. Then, we describe our experiments and main results in Section 3.3. Finally, we discuss some observations from our experiments, summarize our contributions and discuss some future work in Sections 3.4 and 3.5, respectively.

3.2 Methods

Our method consists of two phases. First, for a given protein, we generate a diverse conformational ensemble. Then, we estimate *direct* residue-residue dependence from the ensemble with the graphical LASSO (or simply GLASSO) algorithm [32].

3.2.1 Generation of a Conformational Ensemble

As mentioned in the Section 3.1, we are interested in a newly extended perspective concerning allostery, which states that the absence of backbone conformational changes does not imply the absence of allosteric behaviour [16] [105]. Tsai et al. [105] categorized allosteric proteins into three types: I, II, and III, according to the types of backbone conformational changes that a protein undergoes during an allosteric event — type I proteins undergo no change or subtle changes; type II proteins undergo minor changes; and type III proteins undergo larger domain changes.

A recent study has highlighted the role of side-chain fluctuations alone in information propagation within proteins, with subtle or no backbone conformational changes [23]. To conduct such a study, one would focus on type I proteins and require a heterogeneous dataset (or ensemble) of protein structures, in which the source of diversity among the different structures comes from alternate side-chain conformations alone, since the backbone is held fixed. Commonly used methods for generating such datasets include Monte Carlo (MC) simulations and molecular dynamics (MD). The common practice is to introduce a fluctuation or structural change that in a real environment may be caused by heat or other environmental factors. The introduced change can be an amino acid mutation or a small change in dihedral angles. The fluctuation stimulates the response of the system (molecule) accordingly. The simulation techniques approximate the final stabilized structure that would be a candidate member for the ensemble of conformations.

We apply a different approach to generate the required protein ensembles for our study, by using two state-of-the-art and fast side-chain prediction (SCP) algorithms, namely, SCWRL 4.0 [55] and TreePack [116] [117]. Although this approach is different from the commonly used simulation methods, it follows the same principles. For a protein with m residues, we repeatedly generate an ensemble consisting of $n \equiv [m(m-1)]/2$ structures, as follows. First, we randomly select 20% of the side chains and set each of their conformations to a randomly chosen rotamer from the backbone-dependent rotamer library provided by the Dunbrack Lab [25]. In this step, we basically fluctuate the system. Then, the rest of the side chains are packed by SCWRL or TreePack, and the final structure is added to the ensemble. Essentially, this amounts to solving the side-chain packing problem (a complex optimization problem known to have many local solutions) with many different initial values in order to create a diverse ensemble. In this step, we simulate the response of the system to the introduced fluctuations. We choose $n = [m(m-1)]/2$ because the parameters we are trying to estimate are pairwise coupling strengths, and there are a total of $[m(m-1)]/2$ such parameters for a protein with m residues. We want our ensemble to contain at least the same number of structures as the number of parameters we are estimating.

We think using SCP methods is a reasonable and efficient alternative for data (ensemble) generation, as long as we focus on allosteric effects caused mainly by side-chain fluctuations alone (as opposed to backbone movements), since these SCP methods assume that the backbone is fixed and predict energy-optimal side-chain conformations given the backbone.

3.2.2 The GLASSO Algorithm

The GLASSO algorithm [32] is concerned with the following problem: given a dataset of n observations in $\mathbb{R}^d$ from a multivariate Gaussian distribution with mean μ and covariance matrix Σ, we would like to estimate the inverse covariance matrix, $\Theta \equiv \Sigma^{-1}$, also referred to as the *concentration* or *precision* matrix. The reason why Θ is more interesting than Σ for Gaussian data is that $\Theta_{ij} = 0$ implies the variables i and j are conditionally independent given the other variables, whereas $\Sigma_{ij} = 0$ does not. Even though the "conditional independence" interpretation holds only for Gaussian data, in practice the GLASSO algorithm is often applied for various exploratory purposes, even if the data are not Gaussian.

Let S be the empirical (sample) covariance matrix computed from the observations. The GLASSO algorithm estimates Θ under extra sparsity constraints by solving the convex optimization problem,

$$\underset{\Theta}{max} \quad \underbrace{\log\det(\Theta) - \mathrm{tr}(S\Theta)}_{\text{(I)}} \underbrace{-\lambda \sum_{i,j=1}^{d} |\Theta_{ij}|,}_{\text{(II)}} \tag{3.1}$$

where $\lambda > 0$ and $\mathrm{tr}(A)$ denotes the trace of matrix A. The term (I) in Eq. (3.1) is the Gaussian log-likelihood function, and the term (II) is an L_1 penalty function which forces the solution to be sparse (i.e., containing many zeros) [77][122][4][32]. The positive parameter λ controls the "degree" of sparsity. Generally speaking, sparsity increases (more zero entries) as the value of λ is increased. Please see Section 3.3 for more information on how we select λ.

Direct Coupling via GLASSO

Recently, Jones et al. [48] have applied GLASSO to infer residue-residue *contacts* using available multiple sequence alignment information within a *family* of proteins. Two residues are defined to be in contact if the distance between their β-carbons (or α-carbons, in case of glycine) is less than a certain pre-specified value [28] [48]. A common approach for predicting contacts is to identify correlated mutations from multiply aligned sequences by calculating the mutual information (MI) between pairs of residues [48]. This is considered a *local* approach [80], in the sense that each residue pair is considered independent from the rest of the residues — that is, the impact of other residues is not considered. But a few recent studies have shown that global approaches, such as maximum entropy techniques that consider effects of all other residues, performed more accurately than purely local statistical analyses for predicting direct contacts [11][80]. Local methods cannot disentangle "causal" (or direct) correlations from "transitive" (or indirect) ones [80]. If variable A is correlated with variable B, and B with C, a transitive correlation requires A to be correlated with C. While studying residue contacts in protein-protein interactions, Weight et al. [112] have emphasized the importance of being able to disentangle direct couplings from indirect ones. Indirect couplings do not require large conformational changes, and may occur as a result of having accumulated small effects from a number of direct interactions [112]. Due to the "conditional independence" interpretation of zeros in the inverse covariance matrix, Jones et al. [48] saw the GLASSO as an effective tool for extracting direct coupling information.

We use the GLASSO for the same reason, that is, to avoid unwanted transitive correlations and to focus on side chains that have direct correlations with each other. But, in our study, the information about direct correlations is obtained by tracking how conformational changes at one side-chain location affect conformational changes at another location within *a single protein*, whereas, for Jones et al. [48], the information is obtained by tracking correlated mutations of amino acids at different side-chain locations within *a protein family.*

3.2.3 Applying the GLASSO to the Conformational Ensemble

We now explain how we use the GLASSO algorithm to extract direct coupling information from our conformational ensemble (see Section 3.2.1).

Binary Encoding of Side-chain Conformations

Given a protein, let m denote its total number of residues. For each residue $1 \leq i \leq m$, the SCWRL backbone-dependent rotamer library [9][25] provides a set of discretized candidate conformations, called *rotamers*, ranked according to their occurrence frequencies. For each protein, we first generate an ensemble of n heterogeneous structures (see Section 3.2.1). For each structure $l \in \{1, 2, ..., n\}$ in the ensemble, we use a set of binary random variables to encode its conformation:

$$b_{ik}^{(l)} = \begin{cases} 1, & \text{if the conformation of residue } i \text{ in structure } l \text{ is ``closest'' to rotamer } k; \\ 0, & \text{if not,} \end{cases}$$

where "closeness" is measured in terms of the root mean squared deviation (RMSD). It should be noted, however, that since the rotamer library does not define rotamers for alanine and glycine, we exclude these two types of residues from our analysis.

Calculation of the Sample Covariance Matrix

For each residue i, let $|R_i|$ denote the total number of different rotamers that appear in our ensemble. Entries of the sample covariance matrix S (to be used as input to the GLASSO algorithm) are computed directly from the definition of covariance.

For each residue pair (i, j), we compute an $|R_i|\times|R_j|$ sub-covariance matrix, whose entries are

$$\begin{aligned} S_{ik,jt} &\equiv \mathbb{C}\mathrm{ov}(b_{ik}, b_{jt}) \\ &= \mathbb{E}(b_{ik}b_{jt}) - \mathbb{E}(b_{ik})\mathbb{E}(b_{jt}), \end{aligned} \tag{3.2}$$

where each expectation $\mathbb{E}(\cdot)$ is estimated empirically by the corresponding weighted sample averages,

$$\widehat{\mathbb{E}}(b_{ik}) = \sum_{l=1}^{n} w_l b_{ik}^{(l)}, \quad \widehat{\mathbb{E}}(b_{jt}) = \sum_{l=1}^{n} w_l b_{jt}^{(l)}, \quad \text{and} \quad \widehat{\mathbb{E}}(b_{ik}b_{jt}) = \sum_{l=1}^{n} w_l b_{ik}^{(l)} b_{jt}^{(l)}, \tag{3.3}$$

over the heterogeneous structures $l = 1, 2, ..., n$ within the ensemble. Respectively, these quantities measure how often rotamer k occurs at residue i, how often rotamer t occurs at residue j, and how often rotamers (k, t) co-occur at residues (i, j) — over the entire conformational ensemble. The weight w_l in Equation 3.3 above is chosen to be inversely proportional to the total energy of each structure l so that the more stable structures in our ensemble will contribute the most information to our covariance calculation and overall procedure.

For $i = j$, the expression (3.2) can be simplified further. Within such a sub-matrix, the non-diagonal entries correspond to two different rotamers $k \neq t$ for the same residue i. Since two different rotamers can never occur simultaneously at a single residue site, we have $b_{ik}b_{it} \equiv 0$. For the diagonal entries $(k = t)$, we have $\mathbb{E}(b_{ik}b_{ik}) = \mathbb{E}(b_{ik})$, since $b_{ik}^2 = b_{ik}$ when b_{ik} is either zero or one. Therefore, for $i = j$, the expression (3.2) simply becomes

$$S_{ik,it} = \begin{cases} -\mathbb{E}(b_{ik})\mathbb{E}(b_{it}), & \text{if} \quad k \neq t; \\ \mathbb{E}(b_{ik}) - [\mathbb{E}(b_{ik})]^2, & \text{if} \quad k = t, \end{cases} \tag{3.4}$$

where each expectation $\mathbb{E}(\cdot)$ is again estimated by the corresponding (weighted) sample averages, given above (Eq. 3.3).

Thus, the dimension of our sample covariance matrix S is $d \times d$, where

$$d = \sum_{i=1}^{m} |R_i|.$$

Notice that our R_i is not always equal to the total number of rotamers available for residue i in the rotamer library, because some (low-probability) rotamers may never appear in our (finite) ensemble. Fig. 3.1 contains a schematic illustration of the matrix S and its block structure.

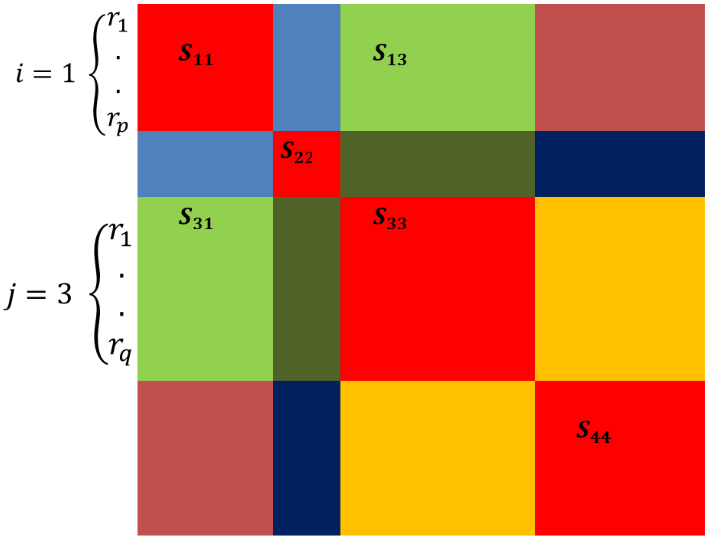


Figure 3.1: Schematic illustration of the sample covariance matrix, with four residues ($m = 4$). Each coloured block is a covariance sub-matrix. Each diagonal block (S_{ii}, $1 \leq i \leq m$) (red) is a size $|R_i|\times|R_i|$ covariance sub-matrices for residue i. Each non-diagonal block (S_{ij}, $1 \leq i, j \leq m$) corresponds to an $|R_i|\times|R_j|$ covariance sub-matrices for the residue pair (i, j). Within a block, each row (and column) corresponds to a rotamer of the given residue. In the figure, residues $i = 1$ and $j = 3$ are shown with $|R_i|= p$ and $|R_j|= q$. The covariance sub-matrices corresponding to the pair of residues have been coloured in light green.

Calculation of the Coupling Score

Using the $d \times d$ matrix S as input, we apply the GLASSO to obtain a sparse estimate of Θ by solving (3.1). Of course, the matrix Θ has the same block structure as S (Fig. 3.1), consisting of sub-matrices for each residue pair (i, j). The final coupling score for each residue-pair is obtained by conducting a two-step post-processing on the estimated Θ matrix, $\widehat{\Theta}$.

First, we aggregate over each sub-matrix:

$$\widehat{\Theta}^{A}_{ij} = \sum_{k=1}^{|R_i|} \sum_{t=1}^{|R_j|} |\widehat{\Theta}_{ik,jt}| \quad \text{for each} \quad 1 \leq i, j \leq m. \tag{3.5}$$

This gives us an $m \times m$ matrix, $\widehat{\Theta}^{A}$, where the superscript "A" stands for "aggregated". The entry $\widehat{\Theta}^{A}_{ij}$ can be seen to contain information about the overall strength of direct coupling between residues i and j.

Next, we correct for the so-called "entropic bias" [48][24][60], using the "average product correction" formula proposed by Dunn et al. [24]:

$$\widehat{\Theta}^{AC}_{ij} \equiv \widehat{\Theta}^{A}_{ij} - \frac{\widehat{\Theta}^{A}_{i\cdot} \times \widehat{\Theta}^{A}_{\cdot j}}{\widehat{\Theta}^{A}_{\cdot\cdot}}, \tag{3.6}$$

where

$$\widehat{\Theta}^{A}_{i\cdot} = \frac{1}{m} \sum_{j=1}^{m} \widehat{\Theta}^{A}_{ij}, \quad \widehat{\Theta}^{A}_{\cdot j} = \frac{1}{m} \sum_{i=1}^{m} \widehat{\Theta}^{A}_{ij}, \quad \text{and} \quad \widehat{\Theta}^{A}_{\cdot\cdot} = \frac{1}{m^2} \sum_{i=1}^{m} \sum_{j=1}^{m} \widehat{\Theta}^{A}_{ij}.$$

The superscript "AC" stands for "aggregated and corrected". The basic idea behind the "entropic bias" is as follows. There is an intrinsic upward bias in the estimate $\widehat{\Theta}^{A}_{ij}$ for larger residues, simply because they have more "choices" in their conformations — a reality also reflected in the fact that the rotamer library contains a larger number of available rotamers for such residues as well. As such, these residues' conformations have higher variability, or entropy [60].

We use these corrected estimates $\widehat{\Theta}^{AC}_{ij}$ as the final coupling scores to rank all residue-residue pairs in terms of their direct coupling strengths.

Information about Alternate Side-chain Conformations

Finally, we notice that the matrix $\widehat{\Theta}$ itself also contains valuable information about the couplings between pairs of *alternate* side-chain conformations (rotamers) for each residue pair. Such information is contained in the corresponding sub-matrices of $\widehat{\Theta}$ — for example, the entry $\widehat{\Theta}_{ik,jt}$ contains coupling information between the k-th (rotamer) conformation at side chain i and the t-th (rotamer) conformation at side chain j. This information can be used to obtain a more detailed view of coupled conformational changes in various critical regions of an allosteric protein. We will show an example below in the Section, 3.3.2.

3.3 Experiments and Results

We implemented the dataset generation phase (Section 3.2.1) and the covariance matrix calculation (Section 3.2.3) in C++ using APIs from the Biochemical Algorithms Library (BALL[41]), SCWRL 4.0 [55] and TreePack [116] [117]. This phase was run on Sharcnet (`http://www.sharcnet.ca/`). Furthermore, we adapted MATLAB code (`http://statweb.stanford.edu/~tibs/glasso/`) to run the GLASSO. The parameter λ (see Eq. 4.3.2) was set to an initial value of 0.001, and it was adjusted iteratively to obtain a target density of 3% for the precision matrix. This target density was suggested by Jones et al. [48].

We selected a few previously studied allosteric proteins of Type I (for more information, please see Section 3.2.1), and performed our experiments on the inactive or unbound molecules (Table 3.1) in order to test the capability of our method in identifying functionally crucial regions by studying the correlated motions of the side chains alone.

3.3.1 Coupled Side-Chain Pairs

For each test case, our results (Table 3.1) basically formed a list of side-chain couplings ranked according to the coupling strength. We examined the lists to determine whether our computed couplings corresponded to known correlated motions of the functionally important residues for each protein. For each case, we generated the corresponding 'Contact Rearrangement Networks' (CRN) [17] to compare with our results. Generating CRN requires both inactive and active structures for each test case [18] (Table 3.1). We used our own implementation of the method and applied the suggested parameters used by Daily et al. [17].

We performed a quantitative comparison of our results against the CRN's lists of couplings using the receiver-operating characteristic (ROC) curve. Treating the list of CRN results as "ground truths", the Area Under the ROC Curve (AUC) is a numeric summary of how well our ranked list matched the CRN findings (see Table 1). These AUC values show quite conclusively that our detection of allosteric couplings is significantly better than

Table 3.1: Set of Proteins Used for Experiments: the 2nd and 3rd columns show the PDB [5] IDs of inactive and active structures of each protein, respectively. The 4th and 5th columns contain the area under the ROC curve (AUC) for comparing the performance of our method (using only the inactive structure) against that of the CRN, and against that of our method using only the active structure, respectively.

Protein	Inactive/Unbound PDB ID	Active/Bound PDB ID	Inactive (AUC against CRN)	Inactive (AUC against Active)
Ras	4Q21	6Q21	0.776	0.792
UPRTase	1XTU	1XTT	0.774	0.874
CheY	3CHY	1F4V (& 1FQW)	0.675	0.874
FixJ	1DBW	1D5W	0.723	0.768
rheb	1XTQ	1XTS	0.711	0.788
hemoglobin	4HHB	1HHO	0.724	0.823
arg repressor	1XXC	1XXA	0.675	0.690
AraC	2ARA	2ARC	0.725	0.756
rhoA	1FTN	1A2B	0.719	0.789
YsxC	1SVI	1SVW	0.755	0.764
rap2a	1KAO	2RAP	0.677	0.847
tet repressor	2TRT	1QPI	0.788	0.796

random, and that there is a good deal of agreement between our results and those from the CRN.

Furthermore, we constructed 3D networks by superimposing the couplings on the 3D structure of the proteins. Network nodes are commonly located at the α carbons of the residues, and edges are drawn between the nodes corresponding to the coupled residue pairs[1]. To visualize our results as networks, we needed more than just a *ranked* list of couplings; we needed to make an active decision as to how many couplings to include in the network. To this effect, we generated a few different networks, rather than committing to any particular one of them. First, we generated a network including only the top K couplings, where K is the number of couplings identified by the CRN. We then created a scree plot of our coupling scores — ranked from the highest to the lowest, chose a few more cut-off thresholds along the scree plot, and generated a network graph for each of the chosen thresholds. These sets of networks for four selected proteins (see Sections Ras, UPRTase, CheY, and FixJ below) are displayed in Figures 3.2 (page 43), 3.5 (page 46), 3.8

[1] 3D molecular visualizations and graphs in this chapter have been produced using the StructBio package [12] (in Python) for Chimera applications [86]

(page 49) and 3.11 (page 52), as well as the aforementioned scree plots and the ROC curves, that are displayed in Figures 3.3 (page 44), 3.6 (page 47), 3.9 (page 50) and 3.12 (page 53).

Examining the networks revealed subregions formed by strongly correlated residues. The subregions usually contain residues that are close in the 3D structure. Relying on GLASSO's feature in extracting direct couplings (see Section 3.2.2), we think that we can obtain the underlying pathways formed by direct interactions between the key residues responsible for allosteric behaviour in proteins. We picked four cases (Ras (PDB ID: 4Q21 [78]), UPRTase (PDB ID: 1XTU [3]), CheY (PDB ID: 3CHY [110]) and FixJ (PDB ID: 1DBW [37])) from the test set to analyse further in the subsequent Subsections (3.3.1 to 3.3.1), respectively. A summary of comparisons is given in Table 3.2.

Ras

Milburn et al. [78] identified two regions named switches I and II (residues 30-38 and 60-73, respectively) to exhibit the major motions in Ras. Switch II is known to be directly involved in switching the protein from inactive to active status [52]. Examining the 3D structure of this protein, we determined that the residues located in the binding site are as follows: 28-35 (has overlap with switch I), 12-19, 145-147, 116-120 (Fig. 3.2). Kidd et al. [52] reported strong correlations in switch II, and the area of the hydrophobic core which is conserved in the Ras family. The authors did not report connectivity between the switches.

The generated CRN revealed a connected component of 75 correlations in and between the two switches (Fig. 3.2 (a)). The component was also connected to a few binding site residues: GLY 12, GLY 13, SER 17 and LYS 16. A few correlations were also seen in the β-sheet close to the switches.

The top 75 correlations of our results revealed strong correlations in the switch II region (Fig. 3.2 (b), and the coupling (GLU 37, ARG 68) connected switches II and I. We captured the couplings (PHE 28, ASP 30) and (ASP 30, GLU 31) in switch I as well. Furthermore, couplings among residues of certain secondary structures (especially α-helices) and the β-sheet close to the switch II were observed.

Considering the top 150 couplings, connectivity within the switch regions improved. Moreover, interesting couplings in the binding site region emerged; the involved residues were 147, 116, 117, 119, 120 and 19 (Fig. 3.2 (c)). Interestingly, couplings among these residues (the binding site) were not observed in the CRN.

Except for the correlations in and among the α-helices, the noticeable difference between our results and those of the CRN was that the CRN captured correlations in the binding site only partially: in the area close to the switches, i.e., among residues in switch I and the α1-β1 loop. In contrast, we obtained correlations among other residues in the binding site as well. The α1-β1 loop contains alanine and glycine residues (GLY 10, ALA 11, GLY 12 and GLY 13), which have been excluded from our model, due to lack of rotameric conformations (more information in Section 3.2.3). Therefore, we did not retrieve correlations for this segment. A 2D network[2] is presented in Figure 3.4 (b) (page 45). The discovered couplings in the binding site may not be directly related to the allosteric behaviour of Ras; but the binding site is categorized as an important functional region of the molecule. Therefore, extracting couplings between residues within and close to this site can be very helpful, especially if the method can extract the couplings between alternate conformations of these residues, as well (see Section 3.3.2 for more information).

[2] The 2D networks in this chapter are produced by the software, Cytoscape [92]

Uracil Phosphoribosyltransferase (UPRTase)

UPRTase is categorized as an enzyme and, as a drug target, it has attracted attention. Sequence identities in this family range from 20% to 45% [3] [91]. Hence, the sequences are rather dissimilar; but the residues at the active site are strongly conserved. The regulatory behaviour of this enzyme differs in different organisms [3]. We studied UPRTase from *Sulfolobus solfataricus*. The inactive UPRTase (PDB ID = 1XTU) is in complex with uridine 5′-monophosphate (UMP) and has the allosteric effector cytidine 5′-triphosphate (CTP). The presence of both UMP and CTP improves the inhibition in UPRTase, because they are not strong inhibitors individually [3].

Arent et al. [3] designated functionally important regions in UPRTase that play an important role in the binding of 5-phosphate-α-1-diphosphate (PRPP): first, a flexible loop between $\beta5$ and $\beta6$ (residues 106-118); second, a loop region between $\beta4$ and $\alpha3$ (residues 78-80); third, a highly conserved region of the PRPP recognition motif which includes $\beta7$ and residues 135-150 of $\alpha4$; finally, another conserved characteristic in the protein family, named the *hood* region, which contains $\beta10$, $\beta11$ and $\alpha6$, i.e., residues 196-216. Furthermore, the allosteric binding site comprises a few residues on the surface of $\alpha2$ and $\alpha3$.

The CRN of this protein consists of 40 couplings (Fig. 3.5 (a)- page 46) that are concentrated in the flexible loop region, $\alpha6$, and the loop between $\beta4$ and $\alpha3$. Almost no connection between the allosteric effector's binding site and the main binding site can be recognized in the CRN. By looking at the top 40 couplings in our result, a noticeable difference between our results and the CRN was that we obtained strong correlations in the binding site of the allosteric effector, i.e., $\alpha2$ and $\alpha3$ (Fig. 3.5 - page 46). We identified strong couplings in the important regions such as the flexible loop, the PRPP recognition motif and the hood region; but these couplings did not form a connected component. Connectivity within each of the allosteric and main binding sites increased when we considered the top 100 couplings, and interesting couplings emerged in the aforementioned crucial regions (Fig. 3.5(c)- page 46). By going further down our ranked list to include the top 175 couplings, the coupling (GLN 25, ARG 80) emerged and we discovered that the two subnetworks of allosteric site and main binding site (consisting of the four segments mentioned above) are connected. Observing this connected component in the set of 175 top-ranked couplings of the protein is significant for two reasons. First, the connected component can explain how the communication between these two sites occurs during allostery through side-chain interactions. Second, the set of 175 couplings approximately contains 0.8% of the total possible couplings for this protein. Therefore, the set consists of very highly ranked and strong couplings extracted for UPRTase. A 2D network of these 175 couplings is shown in Figure 3.7 (page 48) to clearly demonstrate the connection between the two sites.

CheY

Several studies have noted the allosteric behaviour of CheY [57] [14] [26] [58] [52]. In the presence of Mg^{2+}, a phospohrylation happens at ASP 57. Upon this phosphorylation event, CheY undergoes conformational changes that allows the molecule to bind to the flagellar motor switch protein FliM. This conformational changes starts with a hydrogen bond formation and displacement of THR 87 which allows isomerization of TYR 106 [29] [57] [14]. The main conformational changes between the active and inactive structures of CheY are known to be: motions of the $\beta4$-$\alpha4$ loop (ALA 88 - LYS 91) with isomerization of TYR 106, and small changes in the $\beta5$-$\alpha5$ loop (LYS 109 - THR 112) with side-chain motions of LYS 109 and PHE 14 [58] [66] [76].

We generated the CRN for CheY using two different active structures (PDB IDs: 1F4V and 1FQW). The networks contained 33 and 21 correlations, respectively. Both networks contained a connected component that connected ASP 57, ASP 13, and residues of the $\alpha4$-$\beta4$ loop. We picked the first 33 correlations from our list, and produced a network representation (Fig. 3.8 (b)- page 49). Our first observation was that, in addition to the common correlations between the two CRNs, our list contained correlations that were exclusively found in each CRN, e.g., couplings (ARG 19, ARG 22) and (LEU 46, MET 78) were only found in the CRN for 1FQW, whereas (ARG 18, GLU 37), (LYS 122, LYS 126) and (ASP 57, MET 60) were found only in the CRN for 1F4V. This showed that our method can capture a wider range of correlations, in contrast to the CRN which captured correlations by studying two different structures of a molecule.

Furthermore, we noticed that, in higher thresholds (100-120), which still contain a very small portion (less than 1.5%) of all possible couplings, interesting connections started to emerge, such as a connected component of residues ASP 13, MET 60, ASP 57, ASP 12, PHE 14, MET 17, LYS 109, and VAL 86, at a threshold of 100. These residues reside in the Mg^{+2} binding site. With the same threshold, another connected component containing residues TYR 106, MET 85, TRP 58, ASN 59, ASN 94, LYS 91, and GLU 89 is observed. As mentioned above, this set contains residues known to be involved in the CheY's allostery. By increasing the threshold to 120, coupling between 57 and 58 (at rank 118) emerged and connected the two sub-networks. In addition to this coupling, (THR 87, TYR 106) emerged at rank 106 (Fig. 3.8 (d)- page 49). These connections are in agreement with the new findings of McDonald et al. [76] who proposed an extended and previously undescribed allosteric network for CheY consisting of residues ALA 88 (excluded from our study; see Section 3.2.3), THY 87, TRP 58, MET 85, GLU 89, and TYR 106. We think the network connecting the binding and allosteric sites through top-ranked direct couplings is a significant finding of our method, which can explain the side-chain interactions of important residues in these two sites. A 2D network representation of top 120 couplings is presented in Fig. 3.10 (page 51).

Furthermore, examining the list of couplings, we made an interesting observation about interactions between LYS 109 and PHE 14. The couplings (MET 17, LYS 109) and (PHE 14, MET 17) were ranked 4th and 51st, respectively, while (PHE 14 , LYS 109) was in a lower position in the list. The capability of GLASSO in extracting direct couplings suggests that the correlated motions between LYS 109 and PHE 14 occurs indirectly through MET 17.

FixJ

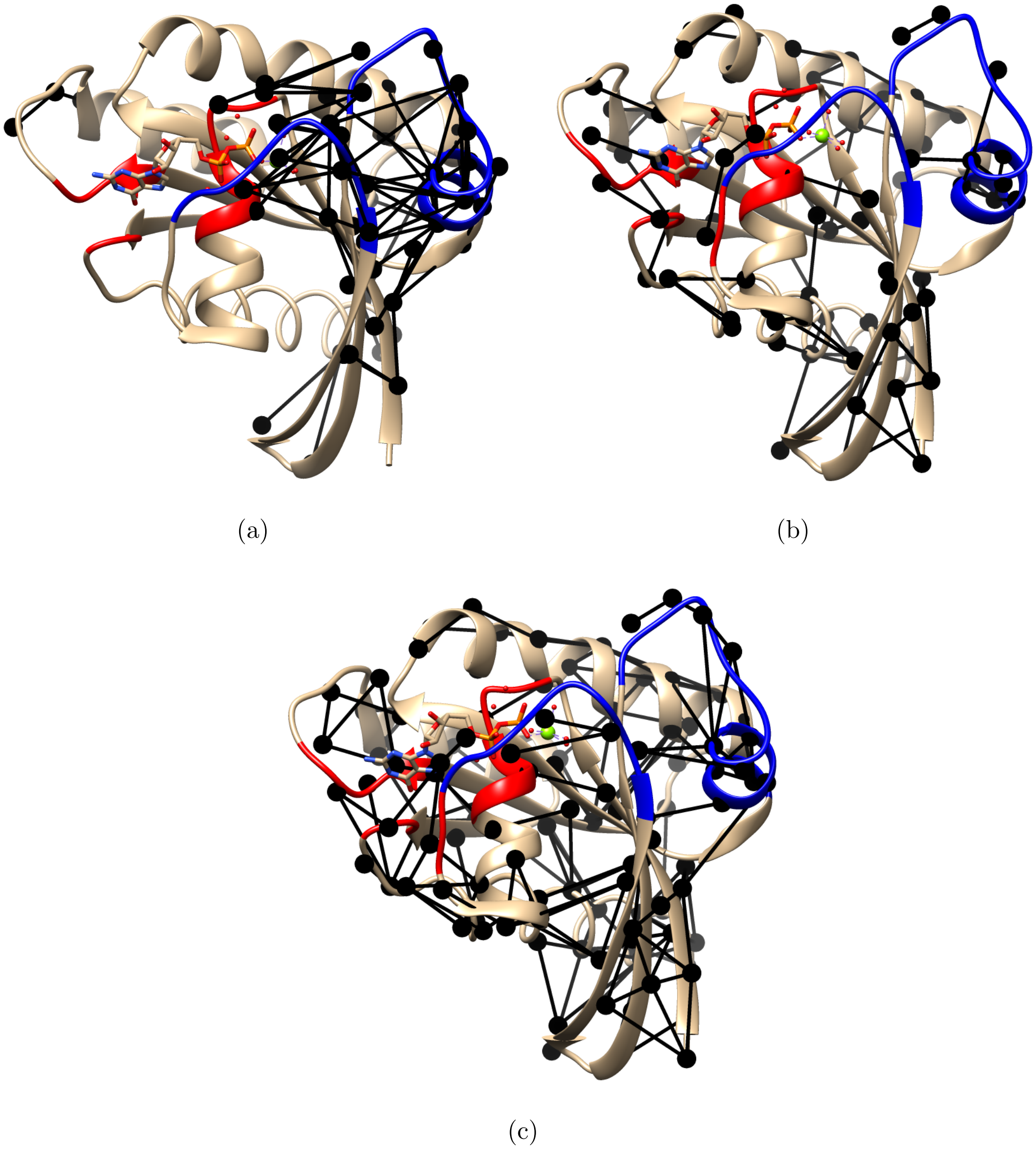
FixJ is an example of response regulator proteins that control different cellular processes. Birck et al. [6] analysed the structural displacements of the receiver domain of FixJ in its unphosphorylated and phosphorylated states. Phosphorylation happens in the active site at ASP 54. Major motions are the phosphorylation-induced conformational changes that have been noticed in the $\beta4$-$\alpha4$-$\beta5$ region, especially in the $\beta4$-$\alpha4$ loop (GLY 83- GLY 85) whose motion leads to a reorientation of $\alpha4$, and in the region between the C-terminal of $\beta1$ (VAL 9) and the first turn of $\alpha1$ (VAL 15). The hydrophobic core of the protein is formed by strands $\beta1$, $\beta3$ and $\beta4$. The C-terminal of the β strands and the $\beta3$-$\alpha3$ (also known as γ loop: LEU 55- MET 60), $\beta1$-$\alpha1$ and $\beta4$-$\alpha4$ loop regions constitute the active site of the protein [37].

The generated CRN for this protein contained 63 couplings and showed a connected component concentrated in $\beta4$, $\alpha4$, $\beta5$, and the intermediate loop regions. The connectivity expanded to the $\beta3$-$\alpha3$ and $\beta1$-$\alpha1$ loop regions of the active site (Fig. 3.11- page 52). Among the top 63 couplings from our results, we recognized strong correlations among the active site residues, and some correlations within $\alpha7$. By increasing the cut-off threshold, we noticed that connection between the $\beta4$-$\alpha4$ loop and the active site was established through the coupling (HIS 84, LYS 104), which emerged at rank 71. The results are also presented as a 2D network in Figure 3.13 (page 54). The revealed interaction path can explain how the phospohrylation at ASP 54 induces a cascade of conformational changes through the side chains' interactions.

3.3.2 Coupled Alternate Side-Chain Conformations

Emphasizing the prevalence of side-chain polymorphism that has recently been revealed from x-ray crystallography data [56] [108], we modelled alternate conformations by various rotamers obtained from the rotamer library at each residue site (Section 3.2.3). As mentioned in Section 3.2.3, we computed the inverse covariance matrix (Θ) which contains correlations between the alternate conformations of a protein's residues. This information can facilitate the prediction of alternate conformations of residues that tend to co-occur during an event that requires conformational changes, e.g., ligand-binding. This capability is another contribution of our method, i.e., providing a *magnified* view of a coupled residue pair and revealing details about the couplings between the alternate conformations of the pair.

The couplings between alternate conformations can be represented as a network and/or by a ranked list. For example, we selected two residues ASP 119 and LYS 147 at the binding site of Ras (Section 3.3.1). These residues are highlighted in magenta in Fig. 3.14 (a) (page 55). In order to retrieve the correlations between the alternate conformations, we extracted the sub-matrix corresponding to the residue pair from the matrix (Θ) (for more information, please see Section 3.2.3). A network (hierarchical layout) of correlated alternate conformations for this residue pair is displayed in Fig. 3.14 (b) (page 55). The rotamer library provides 9 and 81 rotamers for ASP 119 and LYS 147, respectively. Our approach identifies 89 non-zero interactions between the rotamers of this pair. The strongest coupling occurs between the first rotamers of residue 119 and residue 147. We denote this coupling as (1,1) in which the first and second number correspond to the rotamer number of residues 119 and 147, respectively. The next 9 correlations (sorted in descending order of coupling strength) are: (7,30), (5,52), (7,64), (2,1), (5,12), (5,22), (7,51), (7,80), and (7,68). We can extract useful information about the dynamics of the residues from the graph or the list, e.g., all of the strong couplings involve rotamers 1, 5, 7 and 2 of residue 119 (see Fig. 3.14- page 55). Rotamer 3 of residue 119 is the highest-degree node in the graph, i.e., this rotamer of residue 119 may co-occur with many rotamers of residue 147, although none of the corresponding couplings are strong. Furthermore, the graph shows that rotamer 8 of residue 119 only co-occurs with rotamer 11 of residue 147.

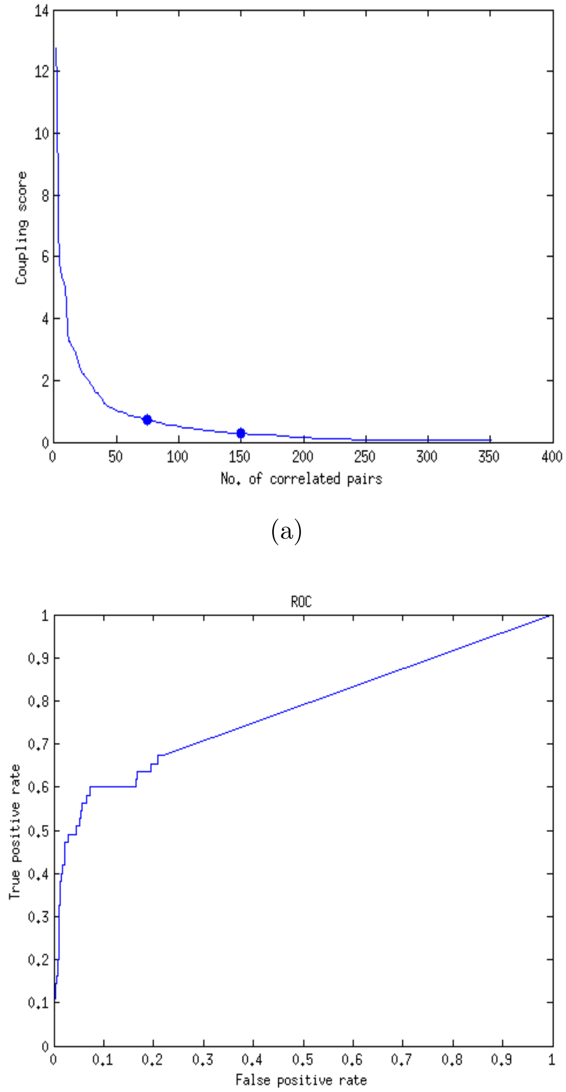


(a)

(b)

(c)

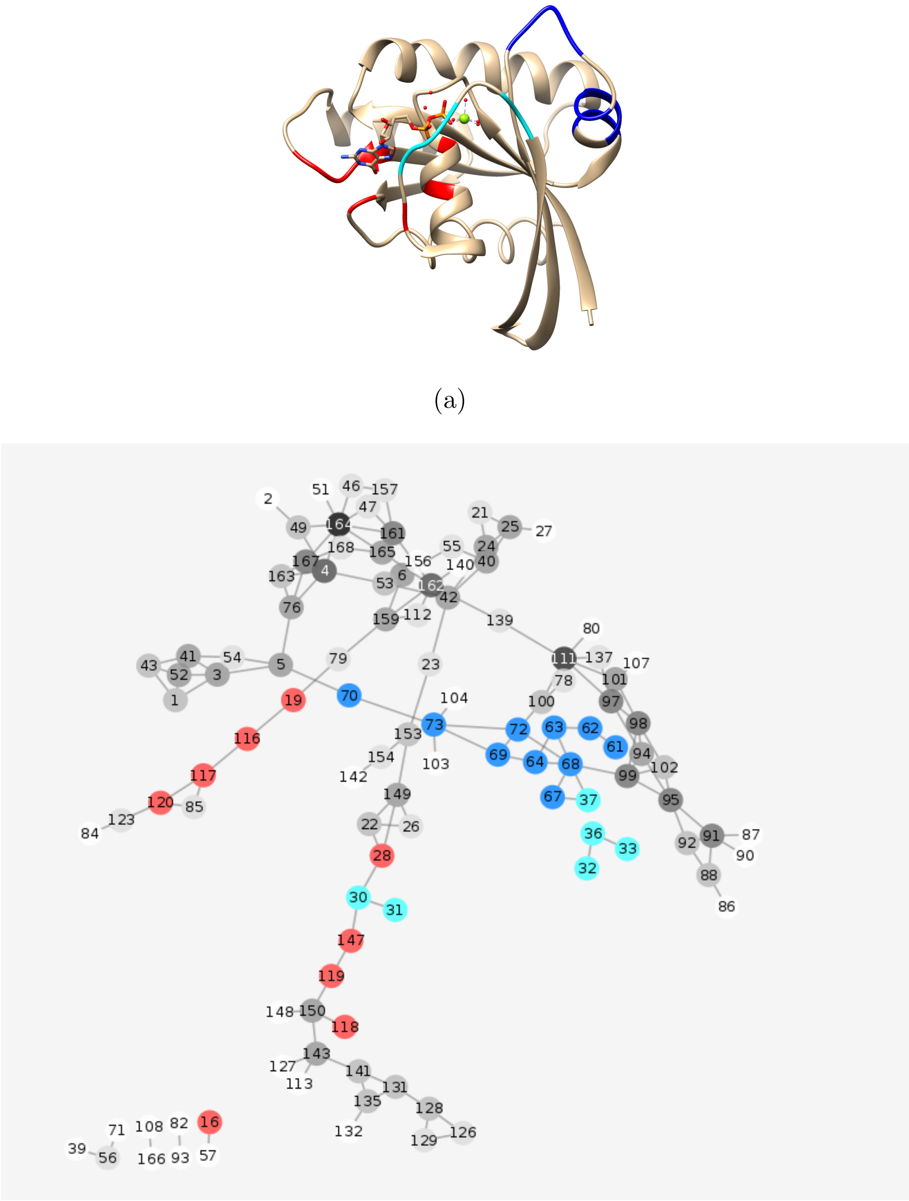
Figure 3.2: 3D network representations of the couplings for Ras. The switch I and II regions are coloured in blue. The binding site residues are coloured in red.(a) CRN results: 75 couplings that are observed in and between the switches. The cluster is connected to a couple of the binding site residues. (b) & (c) 3D network representation of our method's results using the 75, and 150 strongest couplings, respectively.



(a)

(b)

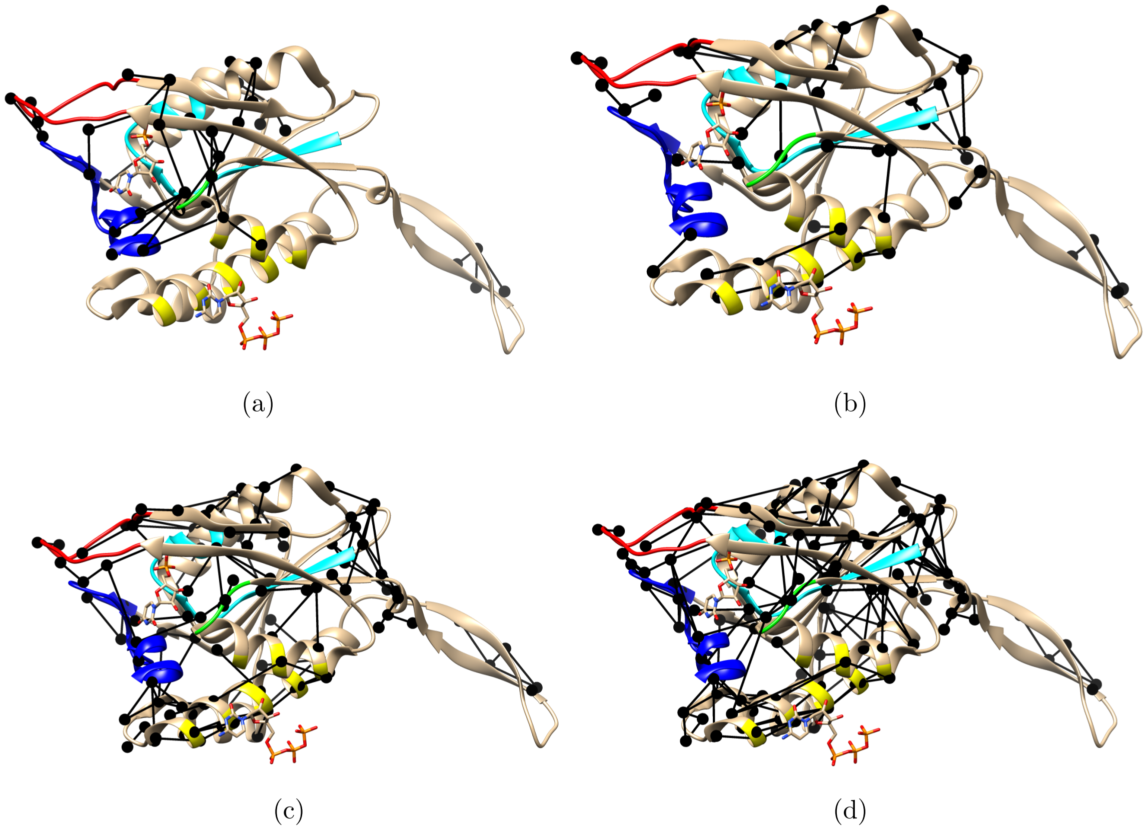
Figure 3.3: Ras: (a) Plot of the coupling score computed by our method vs. number of couplings, cut-offs 75 and 150 have been marked. (b) ROC plot: comparison with CRN's list of couplings. AUC = 0.776.



(a)

(b)

Figure 3.4: Ras: Important residues that appear in the top 150 couplings of Ras have been annotated in (a) a 3D model and (b) a 2D network. The coloured/annotated regions in dark and light blue correspond to switch II and I, respectively. The residues coloured/annotated in red are located in the binding site that overlaps with switch I residues. Nodes degrees are coded by different shades of gray (black: largest, white: smallest).

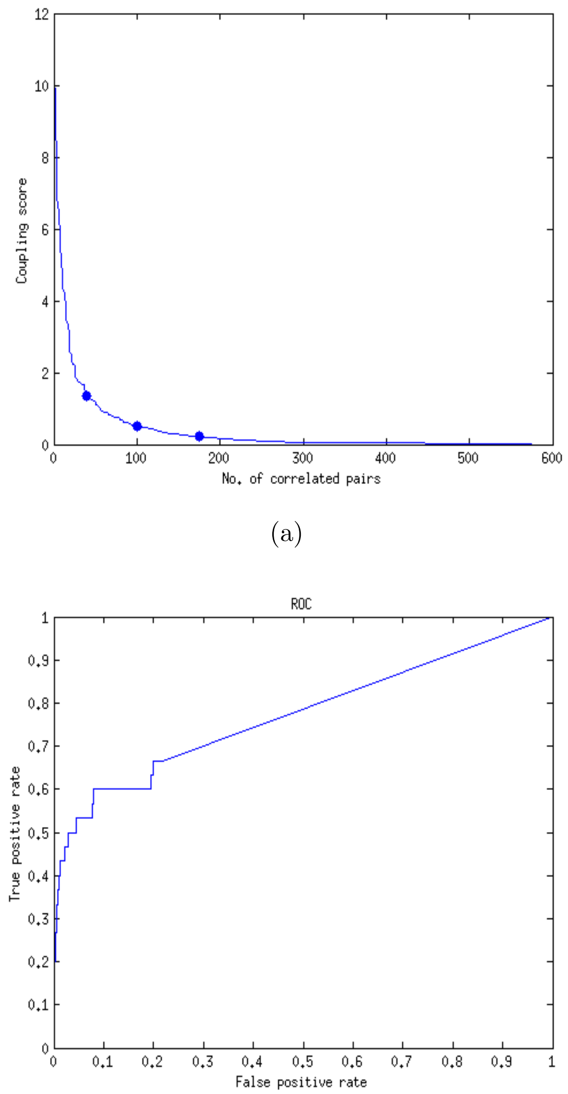

(a)

(b)

(c)

(d)

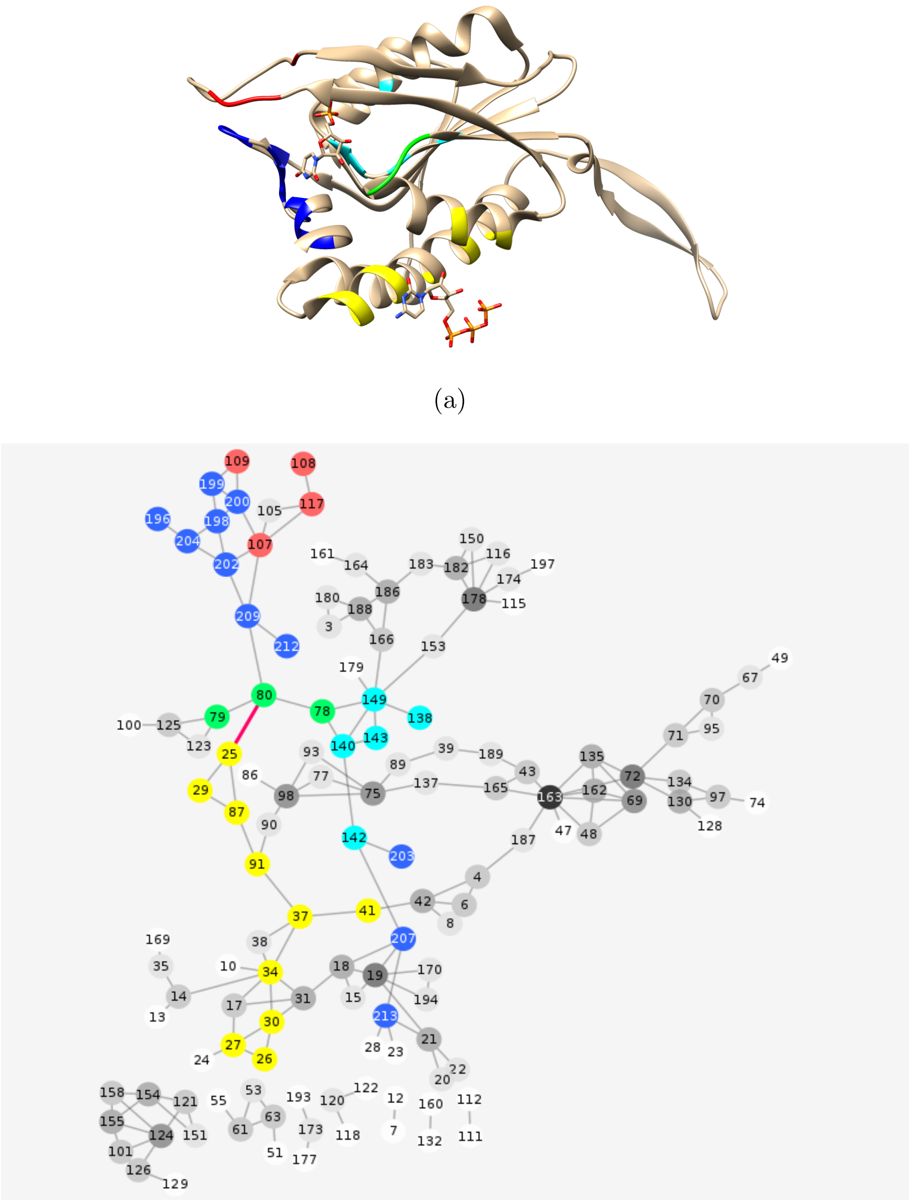
Figure 3.5: 3D network representation of the couplings for UPRTase. The main binding site contains: the flexible loop between $\beta5$ and $\beta6$ (red), the loop region between $\beta4$ and $\alpha3$ (green), $\beta7$ and a part of $\alpha4$ (cyan), and the 'hood' region formed by $\beta10$, $\beta11$ and $\alpha6$ (blue). The allosteric effector's binding site is coloured in yellow. (a) CRN contains 40 couplings, mostly concentrated around the main binding site. (b)-(d) 3D network presentation of our method's results using the 40, 100, and 175 strongest couplings, respectively.



(a)

(b)

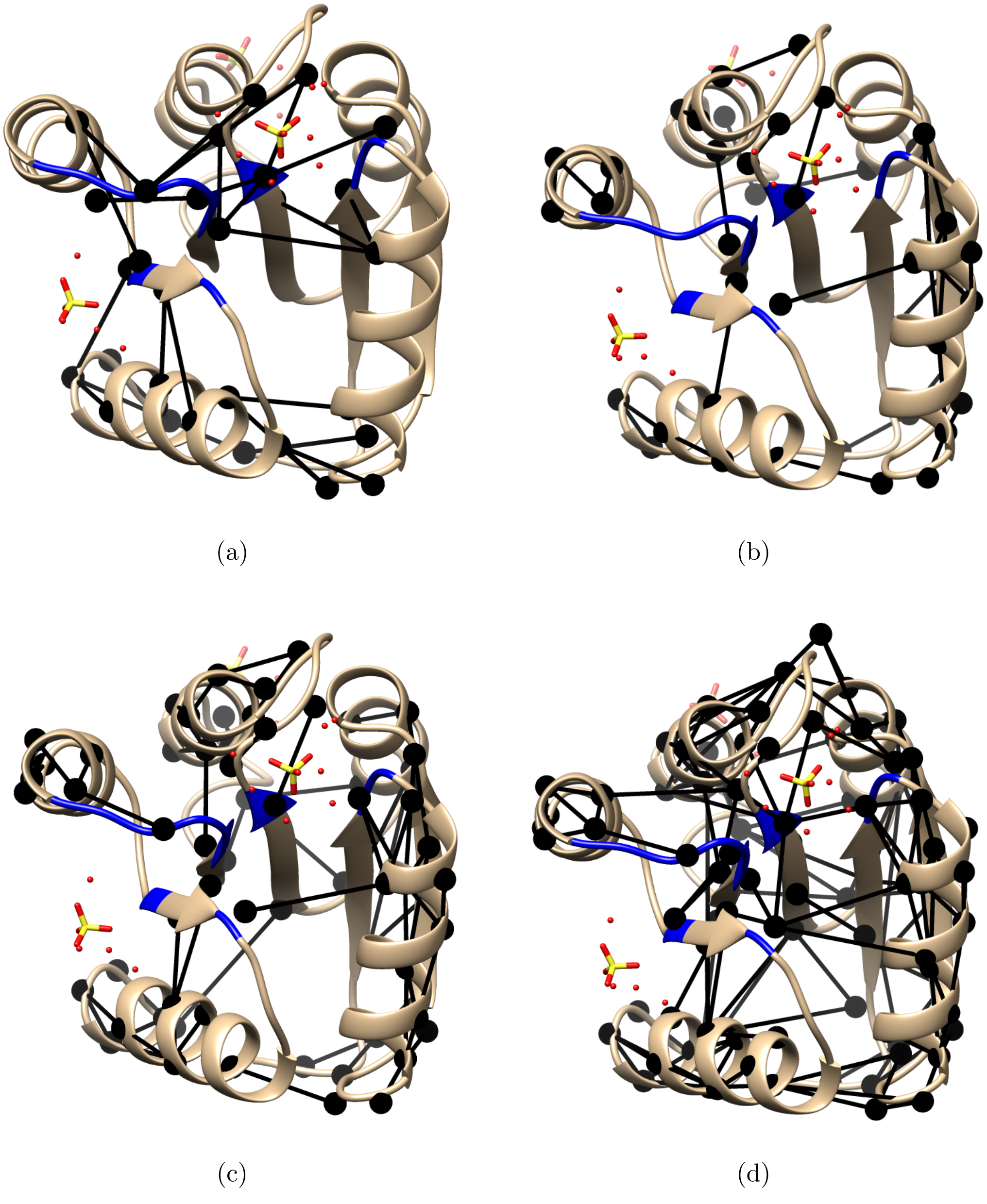
Figure 3.6: UPRTase: (a) Plot of the coupling scores computed by our method vs. number of couplings. Chosen cut-off thresholds used in Fig. 3.5 (b)-(d) are designated on the plot. (b) ROC plot: comparison against CRN's list of couplings. AUC = 0.774.



(a)

(b)

Figure 3.7: UPRTase: Important residues that were observed in the top 175 couplings have been annotated in (a) a 3D model and (b) a 2D network. The two binding sites can be recognized from the computed couplings: the coloured regions in yellow correspond to the allosteric effector's binding site. The residues coloured in light green, cyan, red, and blue correspond to residues in the main binding site. The edge in red connects the two sub-networks of main and allosteric binding sites.

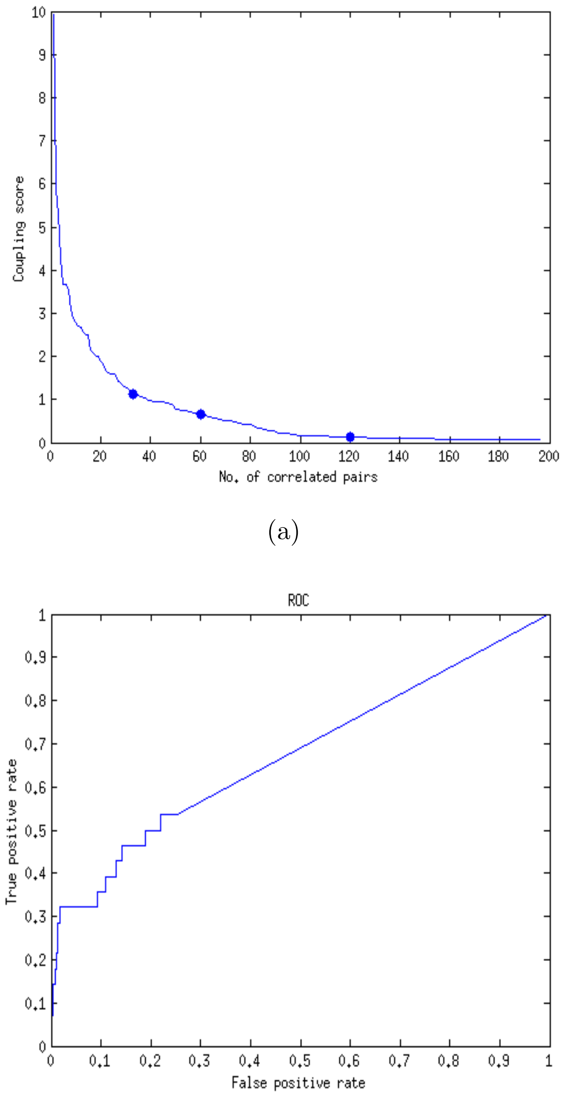

(a)

(b)

(c)

(d)

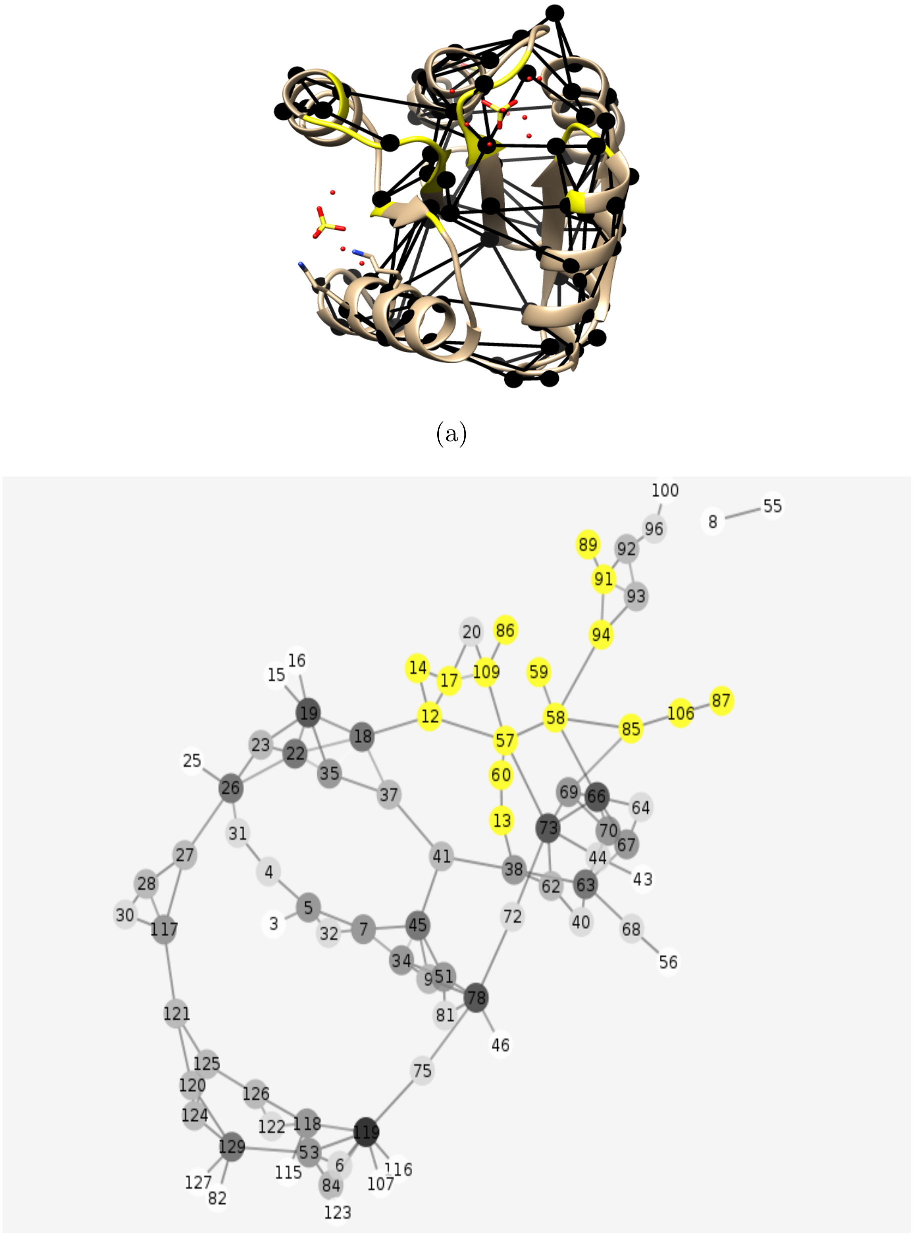
Figure 3.8: 3D network representation of the couplings for CheY. Residues in blue are crucial to allostery. (a) CRN applied to the active structure 1F4V. (b)-(d) Our results — top 33, 60 and 120 couplings, respectively.



(a)

(b)

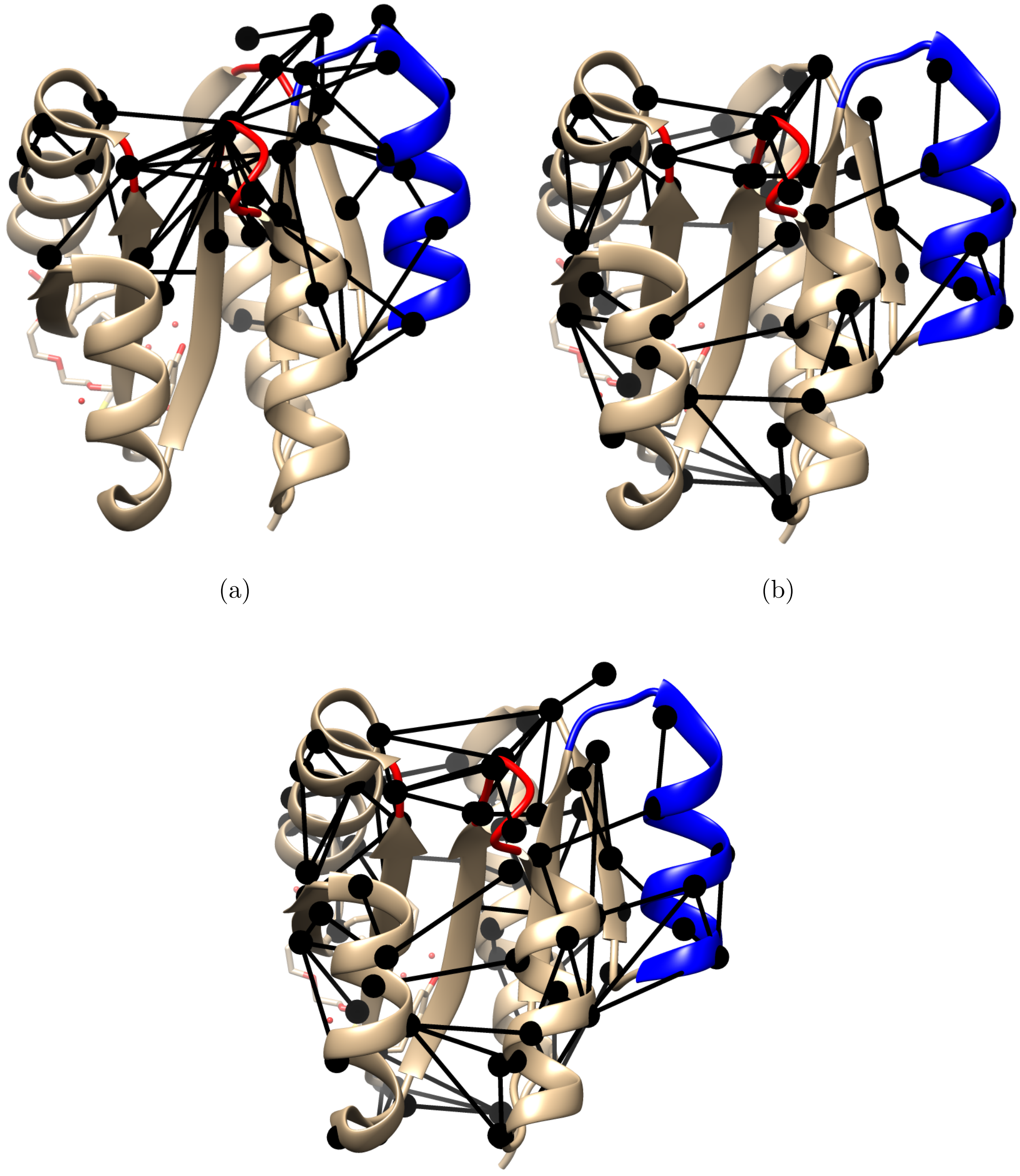
Figure 3.9: CheY: (a) Plot of the coupling score computed by our method vs. number of couplings, with cut-offs 33, 60 and 120 marked. (b) ROC plot: comparison against CRN's list of couplings. AUC = 0.675.



(a)

(b)

Figure 3.10: CheY: Important residues that were observed in the top 120 couplings have been annotated in (a) a 3D model and (b) a 2D network. The coloured regions in yellow correspond to the allosteric site and the Mg^{2+} binding site. Node degrees are coded by different shades of gray (black: largest, white: smallest)

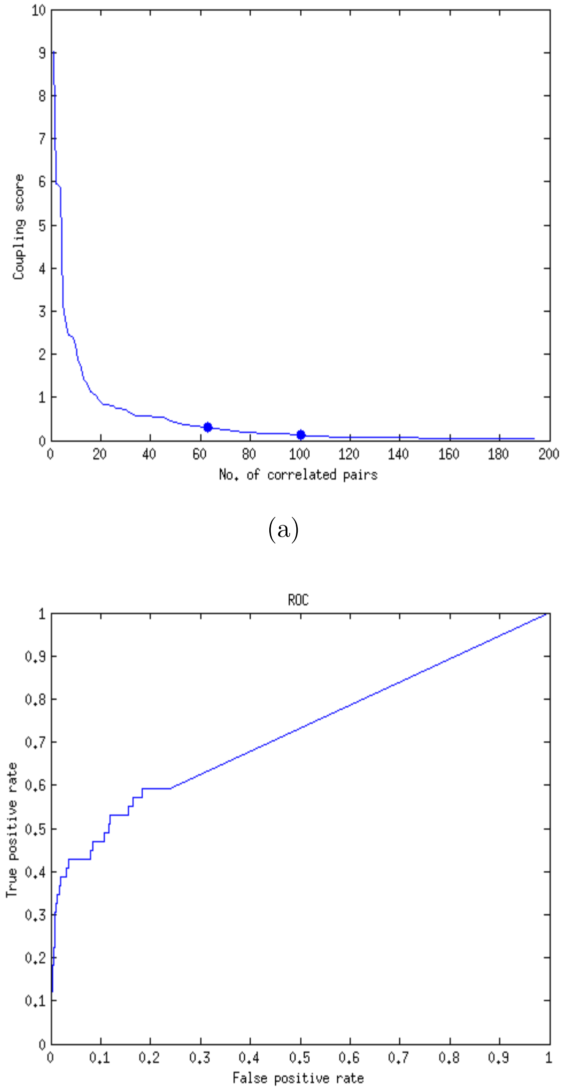


(a)

(b)

(c)

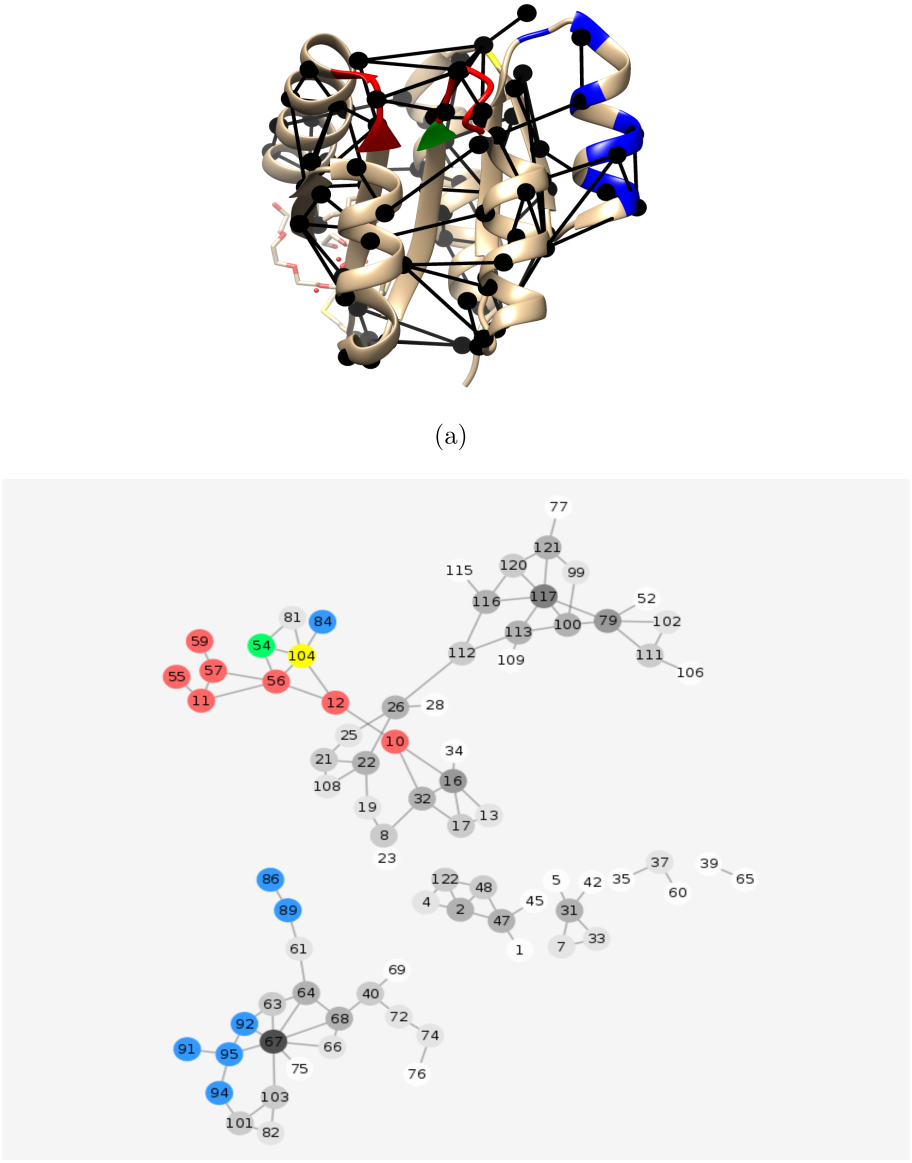
Figure 3.11: 3D network representation of the couplings for FixJ. The active-site residues (except for $\beta4$-$\alpha4$ segment) and the residues of $\beta4$-$\alpha4$-$\alpha4$ region are coloured in red and blue, respectively. (a) CRN contain 63 correlations; crucial regions are connected.(b)&(c) Our results — top 63 and 100 couplings, respectively.



(a)

(b)

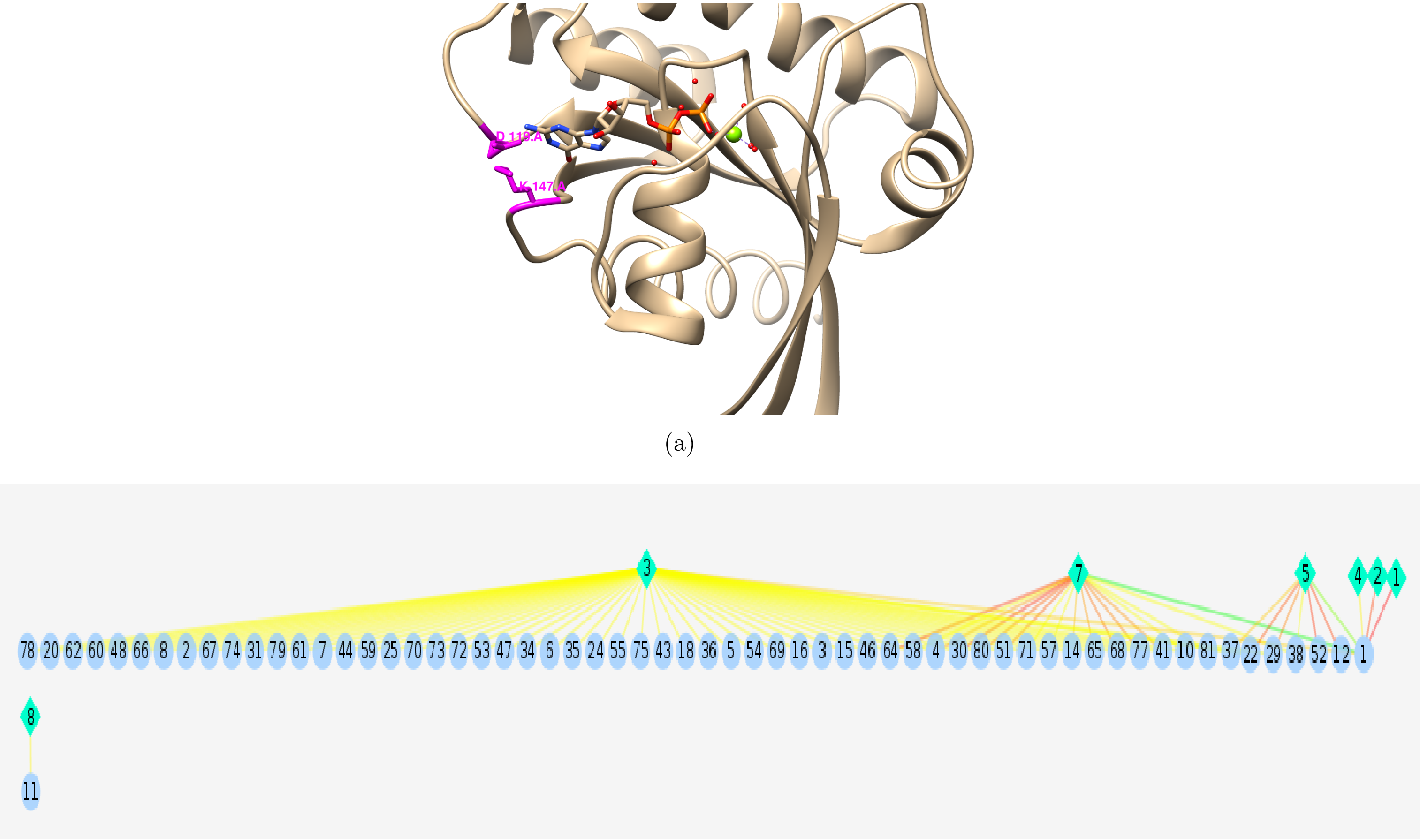
Figure 3.12: FixJ: (a) Plot of the coupling score computed by our method vs. number of couplings, with cut-offs 63 and 100 marked. (b) ROC plot: comparison against CRN's list of couplings. AUC = 0.723.



(a)

(b)

Figure 3.13: FixJ: Important residues in the top 100 couplings have been annotated in (a) a 3D model and (b) a 2D network. The region coloured in blue corresponds to the residues of $\alpha4$ and HIS 84 from the $\beta4$-$\alpha4$ loop. The active site (γ loop and the conserved residues 10, 11 and 12) are coloured in red. The phosphorylation site (ASP 54) and the conserved residue 104 are coloured in green and yellow, respectively. The active site and the $\beta4$-$\alpha4$ loop are connected through the coupling between LYS 104 and HIS 84. Node degrees are coded by different shades of gray (black: largest, white: smallest)



(a)

(b)

Figure 3.14: (a) Molecular visualization of Ras (4Q21). Residues ASP 119 and LYS 147 at the binding site are highlighted. (b) A hierarchical representation of coupled alternate conformations for the residue pair (119, 147). Rotamers for residues 119 and 147 are shown as cyan diamonds and blue circles, respectively. The labels of the nodes denote the rotamer number for each residue. The edge colours show coupling strengths (red: strongest, green: weakest).

3.3.3 Comparison with Statistical Coupling Analysis (SCA)

SCA was initially proposed by Lockless and Ranganathan [64] and used to extract groups of co-evolved residues, called a *sector* in a family of proteins. The method relies on sequence data from the family. We compared our results to the results of the SCA method for the Ras and PDZ families [39] [101] using the ROC. We noticed that neither our results nor those of the CRN are in good agreement with the SCA results, with AUCs around 0.50. Similar results have been reported by Demerdash et al. [20]. This lack of agreement might be due to the different sources of data used by our method and by the SCA, i.e., structural data vs. sequence data. Furthermore, initially the SCA was developed to distinguish conserved and co-evolved residues in a protein family. It was not specifically devised to extract allosteric sites. Additionally, we use data obtained from only one (inactive) structure to infer our results, while the SCA uses a large amount of data from all sequences in a family to infer the results.

3.4 Discussion

Quantitative and qualitative comparison to the computational and empirical methods (Table 3.2) indicated that our proposed method can obtain networks of functional residues and possible interaction paths among them in allosteric proteins. It should be noted that a meaningful comparison with the existing computational methods was difficult, since these methods mostly rely on information from both active and inactive allosteric structures, whereas we assumed that only the inactive structure was accessible.

Another notable point is that, while our results showed good agreement with the corresponding CRNs, our results covered a larger range of functional residues. This became clear by visualizing and comparing the obtained networks. We often obtained more couplings between residues in the main binding site of the protein than those obtained by CRN, e.g. in the Ras case. Although the binding-site residues are implicated as important functional residues of a protein, some of them may not necessarily be directly involved in the allosteric event. Furthermore, we noticed that in the case of CheY which had two active structures, our obtained network captured the specific couplings that were found exclusively by each of the generated CRNs for this protein. The two sets of results demonstrated by our method and by the CRN are not contradicting, and we think that both CRNs and our networks can be used as complementary tools.

3.4.1 Residue Couplings in the Active Structure Are Obtainable from the Inactive Structure

We ran further experiments (data not shown) on the active structures (Table 3.1) to compute the sets of directly coupled residues for these structures as well as the inactive ones, so that we can study and compare the two sets of couplings in the active and inactive structures and possibly achieve more insights about the transition of the molecule from inactive to active state. We applied the same cut-off thresholds, used for the inactive structures (Section 3.3), to pick the top-ranked couplings for the active structures. Comparison of the two sets (inactive and active), at the same cut-offs, revealed that we can categorize each set into two subsets of *common* and *specific* couplings. The former subset refers to the couplings that were common between the two sets, and the latter subset refers to the couplings specific to either the inactive or the active set. Our calculations showed that, on average, about 70% of the couplings detected within the active structure were also detected within the inactive structure (using the same cut-off threshold). Furthermore, we noticed that the remaining 30% of the couplings from the active compound often could be detected within the inactive compound at slightly lower thresholds. Despite appearing at lower ranks, most of these couplings were still among the top 0.5% to 16% of the ones from the inactive structure. The last column of Table 3.1 contains AUC values measuring how well our ranked list of couplings (detected within the *inactive* structure alone) matched couplings that could be identified in a similar fashion from the corresponding *active* structure. These relatively high AUCs ($\sim 0.70 - 0.90$) seem to suggest that transitions from inactive to active structures can be characterized by changes in the relative strengths of couplings detected within the inactive structure alone. That is, during the transition from inactive to active structure, some highly-ranked couplings in the inactive structure become even higher-ranked couplings in the active structure. Therefore, it seems that the set of couplings for the inactive structure contains significant information about couplings in the other (active) structure. This may open a door to predicting functional regions and residue couplings for the active structure by studying the inactive structure only.

Table 3.2: Summary of Results: the critical regions/residues revealed by different methods for each test case have been provided.

Protein	Methods:		
		type of method	observations
Ras			
	Milburn [78]	Empirical	Major motions in switches I (residues 30-38) and II (residues 60-73). Switch II is directly involved in switching the protein from inactive to active status.
	Kidd et al. [52]	Computational	Strong correlations in switch II, and the area of the hydrophobic core (conserved in the Ras family). No connectivity was reported between the switches.
	Daily et al. [17]	Computational	A connected component was formed by residues in the two switches and a few binding site residues (12, 13, 17 and 16).
	Our Approach	Computational	Strong correlations in the switch II region and a couple of correlations connecting the switches II and I (cut-off threshold 75). Couplings were noticed between the binding site residues (residues 19, 116, 117, 120 and 28, 30, 31, 147, and 119).
UPRTase			

	Arent et al. [3]	Empirical	Four critical regions: flexible loop region between $\beta5$ and $\beta6$ (residues 106-118), loop region between $\beta4$ and $\alpha3$ (residues 78-80); highly conserved region including $\beta7$ and residues 135-150 of $\alpha4$, and the $\beta10$, $\beta11$ and $\alpha6$ (residues 196-216)
	Daily et al. [17]	Computational	Interactions mostly in the flexible loop region, $\alpha6$, and the loop between $\beta4$ and $\alpha3$. Connectivity between the allosteric effector's binding site and the main binding site could barely be recognized.
	Our Approach	Computational	Couplings were found in all the critical regions (identified by the empirical approach) and in the the binding site of the allosteric effector ($\alpha2$ and $\alpha3$). A sparse network connects the two binding and allosteric sites.
CheY			
	Lee et al. [57]	Empirical	Conserved residues (57, 12, 13, 87, and 109) involved in a phosphorylation event. Consecutive major conformational changes in $\alpha4$-$\beta4$ loop (residues 88-91). Isomerization of TYR 106 is caused by phosphorylation and coupling between residues 87 and 106. The isomerization is correlated with the binding affinity to FliM.

	McDonald et al. [76]	Empirical	An extended and previously undescribed signalling network was recognized that involves residues 88, 87, 58, 85, 89, and 106.
	Daily et al. [17]	Computational	Residues 57, 13, and residues in the $\alpha4$-$\beta4$ loop were connected.
	Our Approach	Computational	Our set of couplings contains couplings from either CRNs generated from two different active structures of CheY. Using the top 1.5% of all couplings, our graph revealed a sparse network between the allosteric and Mg^{2+} binding sites.
FixJ			
	Birck et al. [6]	Empirical	Upon phosphorylation at residue 54, major motions occur in $\beta4$-$\alpha4$-$\beta5$ and residues 9-15. The active site is formed by the C-terminal of the β strands and the $\beta3$-$\alpha3$, $\beta1$-$\alpha1$ and $\beta4$-$\alpha4$ loop regions.
	Daily et al. [17]	Computational	A component connected residues in $\beta4$, $\alpha4$, $\beta5$, and the intermediate loop regions, and the $\beta3$-$\alpha3$ and $\beta1$-$\alpha1$ of the active site.
	Our Approach	Computational	Residues within the active site region and within $\alpha7$ were connected through strong couplings. Increasing the threshold to 100 resulted in connectivity between $\beta4$-$\alpha4$ loop and the active site via HIS 84.

3.5 Chapter Summary

In this chapter we studied the previously undermined role of side-chain fluctuations in allostery via a computational approach. We focused on the recently extended view of allostery emphasizing that the presence of allostery does not require significant backbone conformational changes [105] [23].

We have presented an efficient method for extracting direct couplings of protein side chains applying a sparse graph estimation technique named GLASSO. We modelled the problem according to another important recent finding which places emphasis on the prevalence of alternate side-chain conformations (conformational polymorphism) [56] [108]. We explicitly took polymorphism into account in our modelling and correlation computations by using rotamers to represent different conformations.

The predicted residue pair couplings indicated correlated motions in critical regions of previously studied allosteric proteins. This suggests that our method is capable of extracting useful correlation information by studying the inactive structures alone. Sparse network representations of the couplings show the underlying correlation pathways that may be used to propagate information in a protein during an allosteric event.

Another contribution of this work is the capability of zooming into the residue-residue couplings to extract the couplings of alternate conformations. We think this step helps to provide a clearer and more detailed image of conformational changes during an allosteric event such as ligand-binding.

We are currently developing a direct coupling detection algorithm, that can be applied in the continuous space of alternate conformations, rather than being restricted to the discretized (rotamer) conformations. We think that avoiding discretization will lead to higher accuracy and more flexibility of the model, and will facilitate the integration of continuous and conformationally polymorphic models, such as Ringer [56], qFit [108], or PBPMixVM [96], when studying allosteric effects of side chains.

Chapter 4

A Novel Algorithm to Infer Direct Couplings between Continuous Multivariate Angular Variables

4.1 Introduction

Predicting allosteric regions in proteins and understanding their interaction mechanisms are challenging problems in bioinformatics. It is common to mainly identify backbone motions responsible for the allosteric behaviour of proteins. However, recent studies have not only highlighted the commonly neglected role of side-chain fluctuations in information transmission within a molecule [23], but also emphasized the presence of allostery in proteins with minimal backbone motions [105]. Moreover, recent discoveries by x-ray crystallography reveal that alternate side-chain conformations are more prevalent than previously thought [56] [108]. These findings further accentuate the importance of a thorough study of the role played by side-chains in allostery [107] [23].

Intrinsic networks of correlated residues are known to play an important role in the propagation of information during an allosteric event. Identifying networks of *directly* coupled allosteric residues is thus of crucial importance for understanding the allosteric mechanism and interaction paths in a molecule. Common correlation-based analyses, however, cannot disentangle *causal* (or direct) correlations from *transitive* (or indirect) ones [80]. To this effect, the statistical concept of *partial* correlation, a conditional dependence measure between two variables given all other variables, is more appropriate for inferring direct couplings. Existing methods based on the partial correlation such as the GLASSO [32]

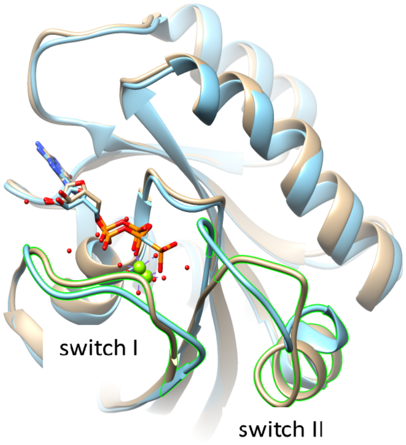


Figure 4.1: Superimposed 3D structures of active and inactive H-Ras. Major conformational changes are known to occur in the Switch I and Switch II regions during an allosteric event.

(see Section 3.2.2), on the other hand, often assume that the data are generated from a multivariate normal distribution — although Jones et al. [48] also applied it to binary variables. This is a restrictive assumption for many applications such as the one discussed in this chapter and Chapter 3. In particular, side-chain conformations are commonly modelled by multidimensional *angular* variables, for which the normal distribution is not a good fit. Another challenge is the quantification of correlations between two multidimensional random variables. Here, a useful statistical tool is canonical correlation analysis (CCA) [44]. We developed a novel extension of CCA by incorporating *partial* correlation and by using a multivariate von-Mises kernel function [70] to capture similarities between two multidimensional *angular* variables.

We tested our method on a number of well-studied allosteric proteins from the Ras, Rho, and Rab sub-families of the small G protein super-family. While the sequence similarity within a sub-family may be relatively high (50-55%), members of two different sub-families tend to share low (∼30%) sequence identity [103]. Despite the low similarity and having distinct functions, 3D structural analysis of these proteins has revealed common characteristics. For example, they cycle between two inter-convertible forms [103] [87] [113] – the inactive form [bound to guanosine diphosphate (GDP)] and the active form [bound to guanosine triphosphate (GTP)]. Furthermore, during this cycle, all of the small G proteins undergo major conformational changes in two common regions, referred to as Switch I and Switch II (see Figure 4.1) in the literature [78] [38] [93].

Our method successfully identified the aforementioned allosteric regions in these test cases. In each case, residues belonging to these regions are specifically involved in the strongest couplings and are among the highest-degree nodes in the *interaction graph* formed by the inferred couplings. Furthermore, allosteric and binding sites in these test cases are connected in the interaction graphs as well. This means that, by studying side-chain fluctuations, our method can infer pathways between these sites and shed light on how information propagates between these functionally important residues.

It is worthwhile to note that we obtained our results by using information from only the inactive (GDP-bound) structure of each allosteric protein. Most methods for studying allostery use both the active and the inactive structures. However, in many situations, both structures are not readily available. We think our method can provide especially valuable information in these types of situations.

4.2 Methods

Our method for inferring direct couplings comprises a few fundamental components. First, we rely on the mathematical notion of partial correlation to measure *direct* couplings (Section 4.2.1). Second, we use canonical correlation analysis (CCA) to quantify the notion of correlation (more specifically, partial correlation) for *multivariate* data (Section 4.2.2). Third, we use a specific kernel function — the von-Mises kernel — to measure similarity between two sets of conformational variables expressed in terms of dihedral angles (Section 4.2.4, Section 4.2.5).

4.2.1 Direct Coupling and Partial Correlation

If variable x is correlated with a set of variables $\boldsymbol{z} = (z_1, z_2, ..., z_d)^{\mathrm{T}}$ and so is y, a transitive correlation requires that x be correlated with y. For direct couplings between residues, we are interested in the *direct* correlation between x and y, not the kind of transitive correlations between them. In many applications, computing direct couplings between residues is crucial [48] [80].

The “partial correlation” between x and y is a measure of their dependence after having removed the effect of $\boldsymbol{z}$. It can be computed as follows. First, we respectively regress both x and y onto $\boldsymbol{z}$, that is, we fit the following models to x and y:

$$\begin{aligned} x &= \beta_0 + \beta_1 z_1 + ... + \beta_d z_d + \varepsilon_x, \\ y &= \gamma_0 + \gamma_1 z_1 + ... + \gamma_d z_d + \varepsilon_y. \end{aligned}$$

Let $\widehat{\beta}_j$ and $\widehat{\gamma}_j$ denote the estimated regression coefficients, for $j = 0, 1, 2, ..., d$. Let r_x and r_y denote the residuals from these regression models, i.e.,

$$\begin{aligned} r_x &= x - (\widehat{\beta}_0 + \widehat{\beta}_1 z_1 + ... + \widehat{\beta}_d z_d), \\ r_y &= y - (\widehat{\gamma}_0 + \widehat{\gamma}_1 z_1 + ... + \widehat{\gamma}_d z_d). \end{aligned}$$

The partial correlation between x and y is the usual Pearson correlation between r_x and r_y.

4.2.2 Canonical Correlation Analysis (CCA)

As indicated above, if both x and y are *univariate* random variables, we can use their usual Pearson correlation to measure their marginal association, or their partial correlation to measure their direct association. But what if both of them are *multivariate* random variables? Moreover, what if they have different dimensions, e.g., $\boldsymbol{x} = (x_1, ..., x_p)^{\mathrm{T}}$ and $\boldsymbol{y} = (y_1, ..., y_q)^{\mathrm{T}}$ for some $p \neq q$?

One way to come up with a *single* numeric measure of the association between two *multivariate* variables $\boldsymbol{x} \in \mathbb{R}^p$ and $\boldsymbol{y} \in \mathbb{R}^q$ is to compute the quantity,

$$\rho(\boldsymbol{x}, \boldsymbol{y}) \equiv \underset{\boldsymbol{u} \in R^p, \boldsymbol{v} \in R^q}{max} \quad \mathbb{C}\mathrm{orr}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{v}^{\mathrm{T}}\boldsymbol{y}), \tag{4.1}$$

sometimes referred to as the *canonical correlation coefficient* between $\boldsymbol{x}$ and $\boldsymbol{y}$. More specifically, since

$$\begin{aligned} \mathbb{C}\mathrm{orr}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{v}^{\mathrm{T}}\boldsymbol{y}) &= \frac{\mathbb{C}\mathrm{ov}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x}, \boldsymbol{v}^{\mathrm{T}}\boldsymbol{y})}{\sqrt{\mathbb{V}\mathrm{ar}(\boldsymbol{u}^{\mathrm{T}}\boldsymbol{x})\mathbb{V}\mathrm{ar}(\boldsymbol{v}^{\mathrm{T}}\boldsymbol{y})}} \\ &= \frac{\boldsymbol{u}^{\mathrm{T}}\mathbb{C}\mathrm{ov}(\boldsymbol{x}, \boldsymbol{y})\boldsymbol{v}}{\sqrt{[\boldsymbol{u}^{\mathrm{T}}\mathbb{V}\mathrm{ar}(\boldsymbol{x})\boldsymbol{u}][\boldsymbol{v}^{\mathrm{T}}\mathbb{V}\mathrm{ar}(\boldsymbol{y})\boldsymbol{v}]}}, \end{aligned}$$

the maximization problem in Eq. (4.1) is equivalent to

$$\underset{\boldsymbol{u} \in R^p, \boldsymbol{v} \in R^q}{max} \quad \boldsymbol{u}^{\mathrm{T}}[\mathbb{C}\mathrm{ov}(\boldsymbol{x}, \boldsymbol{y})]\boldsymbol{v}, \tag{4.2}$$

subject to the constraints

$$\boldsymbol{u}^{\mathrm{T}}[\mathbb{V}\mathrm{ar}(\boldsymbol{x})]\boldsymbol{u} = 1 \quad \text{and} \quad \boldsymbol{v}^{\mathrm{T}}[\mathbb{V}\mathrm{ar}(\boldsymbol{y})]\boldsymbol{v} = 1. \tag{4.3}$$

Given a data set, $\{(\boldsymbol{x}_i, \boldsymbol{y}_i) : i = 1, 2, ..., n\}$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $\boldsymbol{y}_i \in \mathbb{R}^q$, let

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix}_{n\times p} \quad \text{and} \quad \boldsymbol{Y} = \begin{bmatrix} \boldsymbol{y}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{y}_n^{\mathrm{T}} \end{bmatrix}_{n\times q}$$

be the usual data matrices, respectively stacking n samples of $\boldsymbol{x}$ and $\boldsymbol{y}$ as row vectors. If both $\boldsymbol{X}$ and $\boldsymbol{Y}$ are centered so that each column has mean zero, then the sample estimates of $\mathbb{V}\mathrm{ar}(\boldsymbol{x})$, $\mathbb{V}\mathrm{ar}(\boldsymbol{y})$ and $\mathbb{C}\mathrm{ov}(\boldsymbol{x}, \boldsymbol{y})$ are simply

$$\widehat{\mathbb{V}\mathrm{ar}(\boldsymbol{x})} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X},$$

$$\widehat{\mathbb{V}\mathrm{ar}(\boldsymbol{y})} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{y}_i\boldsymbol{y}_i^{\mathrm{T}} = \frac{1}{n}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y},$$

$$\widehat{\mathbb{C}\mathrm{ov}(\boldsymbol{x}, \boldsymbol{y})} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i\boldsymbol{y}_i^{\mathrm{T}} = \frac{1}{n}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}.$$

Hence, the empirical estimate of the canonical correlation coefficient given in (4.1) can be obtained by solving the maximization problem (4.2)-(4.3) using the three sample estimates above, i.e.,

$$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}) \;=\; \underset{\substack{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}=1 \\ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}=1}}{max} \; \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}. \tag{4.4}$$

The maximization problem in Eq. (4.4) is well-known to be a generalized eigenvalue problem (see, e.g. Shawe-Taylor and Cristianini [94]), and can be solved in many scientific computing platforms, including MATLAB.

4.2.3 Partial CCA

If, instead, we are interested in a single numeric measure of the *direct* association between $\boldsymbol{x}$ and $\boldsymbol{y}$, we can use the same idea as that of the partial correlation (Section 4.2.1). That is, we can first remove the effect of $\boldsymbol{z}$ from both of them, before computing their canonical correlation coefficient. More specifically, let

$$\boldsymbol{Z} = \begin{bmatrix} \boldsymbol{z}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{z}_n^{\mathrm{T}} \end{bmatrix}_{n\times d}.$$

We simply compute (4.4) using

$$\begin{aligned} \check{\boldsymbol{X}} &= \boldsymbol{X} - \boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{X} \quad \text{and} \\ \check{\boldsymbol{Y}} &= \boldsymbol{Y} - \boldsymbol{Z}(\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Z})^{-1}\boldsymbol{Z}^{\mathrm{T}}\boldsymbol{Y} \end{aligned}$$

instead of the original data matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$. We refer to the resulting estimate,

$$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = \max_{\substack{\boldsymbol{u}^{\mathrm{T}}\check{\boldsymbol{X}}^{\mathrm{T}}\check{\boldsymbol{X}}\boldsymbol{u}=1 \\ \boldsymbol{v}^{\mathrm{T}}\check{\boldsymbol{Y}}^{\mathrm{T}}\check{\boldsymbol{Y}}\boldsymbol{v}=1}} \boldsymbol{u}^{\mathrm{T}}\check{\boldsymbol{X}}^{\mathrm{T}}\check{\boldsymbol{Y}}\boldsymbol{v}, \tag{4.5}$$

as the *partial canonical correlation coefficient* between $\boldsymbol{x}$ and $\boldsymbol{y}$.

4.2.4 Kernelization of CCA and Partical CCA

It is easy to see that, if we reparameterize $\boldsymbol{u} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}$ and $\boldsymbol{v} = \boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\theta}$ for some $\boldsymbol{\alpha}, \boldsymbol{\theta} \in \mathbb{R}^n$, the sample canonical correlation coefficient [Eq. (4.4)] can be computed as

$$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}) = \max_{\substack{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}=1 \\ \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\theta}=1}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\theta}. \tag{4.6}$$

Let $\boldsymbol{K}_X = \boldsymbol{X}\boldsymbol{X}^{\mathrm{T}} =$

$$\begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_n \\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{x}_n \end{bmatrix}$$

be an $n \times n$ matrix whose (i, j)-th entry is equal to $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$, the inner-product between observations $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and likewise for $\boldsymbol{K}_Y$. Then, Eq. (4.6) can be written as

$$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}) = \max_{\substack{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_X^2\boldsymbol{\alpha}=1 \\ \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{K}_Y^2\boldsymbol{\theta}=1}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_X\boldsymbol{K}_Y\boldsymbol{\theta}. \tag{4.7}$$

This shows that CCA can easily be "kernelized" (see, e.g. Shawe-Taylor and Cristianini [94]) — simply replace the inner-products, $\boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x}_j$ and $\boldsymbol{y}_i^{\mathrm{T}}\boldsymbol{y}_j$, with $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and $K(\boldsymbol{y}_i, \boldsymbol{y}_j)$, for some kernel function $K(\cdot, \cdot)$.

When a different kernel function is used in Eq. (4.7) other than the usual inner-product, we will use the notation, $\widehat{\rho}_K(\boldsymbol{x}, \boldsymbol{y})$, to refer to the quantity in Eq. (4.7). Clearly, the

Table 4.1: Summary of Notations

Notation	Meaning	Section
$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y})$	CCA	4.2.2
$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}\|\boldsymbol{z})$	Partial CCA	4.2.3
$\widehat{\rho}_K(\boldsymbol{x}, \boldsymbol{y})$	Kernelized CCA	4.2.4
$\widehat{\rho}_K(\boldsymbol{x}, \boldsymbol{y}\|\boldsymbol{z})$	Kernelized Partial CCA	4.2.4

same argument applies to sample estimate of the partial canonical correlation coefficient [Eq. (4.5)] as well, that is, the quantity $\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$ can be obtained from Eq. (4.7), too, by simply letting $\boldsymbol{K}_X = \check{\boldsymbol{X}}\check{\boldsymbol{X}}^{\mathrm{T}}$ and $\boldsymbol{K}_Y = \check{\boldsymbol{Y}}\check{\boldsymbol{Y}}^{\mathrm{T}}$. Similarly, when a different kernel function is used other than the usual inner-product, we will use the notation, $\widehat{\rho}_K(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$, to distinguish it from $\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$. Table 4.1 summarizes our notations.

A technical detail, which we largely have suppressed in our exposition here, is that, in Eq. (4.7), it is necessary to add a regularization term such as $\lambda\boldsymbol{I}$ to both $\boldsymbol{K}_X^2$ and $\boldsymbol{K}_Y^2$ in the constraints to avoid an otherwise degenerate solution.

4.2.5 Application of KPCCA to the Study of Allostery

In this study, we use Kernelized Partial CCA (or simply KPCCA) to quantify the direct coupling between pairs of residues and study the allosteric behaviour of proteins. Let m denote the number of residues in a given protein. For any given pair of residues $1 \leq a, b \leq m$, we let

- $\boldsymbol{x}$ be the vector of p dihedral angles describing the side-chain conformation of residue a;
- $\boldsymbol{y}$ be the vector of q dihedral angles describing the side-chain conformation of residue b; and
- $\boldsymbol{z}$ be the vector of d dihedral angles describing the side-chain conformations of all other residues.

In general, $0 \leq p, q \leq 4$, depending on the type of amino acids for the two residues, whereas d is much larger. When we say that we use KPCCA, we mean that we use the quantity, $\widehat{\rho}_K(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$ (see Table 4.1), to quantify how strongly the two residues a and b are directly coupled. We do this for all $m(m-1)/2$ pairs of residues. In order to compute $\widehat{\rho}_K(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$, we need

- multiple observations for $\boldsymbol{x}$, $\boldsymbol{y}$ and $\boldsymbol{z}$, that is, different conformations of the same protein; and
- an appropriate kernel function $K(\cdot, \cdot)$ for measuring the similarity of two different conformations (expressed in terms of dihedral angles).

In the next three subsections, we explain in more detail how we addressed these specific issues.

Generation of a Conformational Ensemble

As explained previously in Chapter 3, Section 3.2.1, inferring the couplings requires a heterogeneous dataset. We applied the same method for generating a conformational ensemble as used in Section 3.2.1.

Weighted Von-Mises Kernel Function

For $\boldsymbol{x}_i, \boldsymbol{x}_j \in \mathbb{R}^p$, we used the following kernel function to perform KPCCA (see Section 4.2.4),

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) \quad = \quad w_i w_j \prod_{t=1}^{p} \exp\left[\kappa_t \cos(x_{it} - x_{jt})\right], \tag{4.8}$$

and likewise for $\boldsymbol{y}_i, \boldsymbol{y}_j \in \mathbb{R}^q$. This is based on the multivariate von-Mises distribution [70] and treating the dihedral angles as if they were independent.

An angular random variable $\boldsymbol{x} \in \mathbb{R}^p$ is said to follow the multivariate von-Mises distribution if it has density function,

$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\kappa}, \boldsymbol{\Lambda}) = \frac{1}{Z(\boldsymbol{\kappa}, \boldsymbol{\Lambda})} \exp\left[\boldsymbol{\kappa}^{\mathrm{T}} \boldsymbol{c}(\boldsymbol{x}) + \frac{\boldsymbol{s}^{\mathrm{T}}(\boldsymbol{x}) \boldsymbol{\Lambda} \boldsymbol{s}(\boldsymbol{x})}{2}\right],$$

where

$$c_t(\boldsymbol{x}) \equiv \cos(x_t - \mu_t) \quad \text{and} \quad s_t(\boldsymbol{x}) \equiv \sin(x_t - \mu_t)$$

for $t = 1, 2, ..., p$, and $Z(\boldsymbol{\kappa}, \boldsymbol{\Lambda})$ is a normalizing constant. The parameter $\boldsymbol{\mu} \in \mathbb{R}^p$ describes the location, i.e., the mean (or center), and the parameter $\boldsymbol{\kappa} \in \mathbb{R}^p$ ($\boldsymbol{\kappa} > \mathbf{0}$) describes the scale, i.e., the spread (or concentration). The parameter, $\boldsymbol{\Lambda} = [\lambda_{st}] \in \mathbb{R}^{p \times p}$ is a matrix whose diagonal elements are zero ($\Lambda_{ss} = 0$) and whose off-diagonal elements Λ_{st} capture the correlation between x_s and x_t. Setting $\Lambda_{st} = 0$ ignores the correlation between x_s and x_t. The multivariate von-Mises distribution frequently has been used as an appropriate tool for modelling angular variables that describe residue conformations in proteins [69] [71].

We also introduced weights w_i, w_j in our kernel function [Eq. (4.8)]. These weights were set to be inversely proportional to the energies of the two structures, i and j, in our ensemble (Sections 4.2.5 and 3.2.1). This allows structures with lower energies — i.e., the ones that are more stable in our ensemble — to contribute more information to our overall procedure.

Choice of κ_t

The kernel function (4.8) contains p concentration parameters, $\kappa_1, ..., \kappa_p$, one for each dihedral angle. An advantage of the von Mises kernel is that these concentration parameters can be set to reflect the intrinsic nature of side-chain dihedral angles. For example, the first two dihedral angles are known to undergo more restricted motions, while the 3rd and 4th have more freedom of movement. Hence, we assigned higher concentration parameters to the first two angles ($\kappa_1 = \kappa_2 = 8$) to allow less freedom in motion, and lower concentration parameters to the 3rd and 4th angles ($\kappa_3 = 4; \kappa_4 = 2$) to allow more freedom of movement. The reason to select these values for the concentration parameters will be explained later in this subsection.

The von-Mises kernel can be thought of as the Gaussian kernel (or radial basis kernel) for angular data. To see this, notice that, using the Taylor approximation, $\cos(x) \approx 1 - x^2/2$, we can write

$$\exp\left[\kappa_t \cos(x_{it} - x_{jt})\right] \approx \exp\left[\kappa_t - \frac{\kappa_t(x_{it} - x_{jt})^2}{2}\right] = (e^{\kappa_t}) \exp\left[-\frac{(x_{it} - x_{jt})^2}{2/\kappa_t}\right]. \tag{4.9}$$

On the other hand, the corresponding Gaussian (or radial basis) kernel is given by

$$K(x_{it}, x_{jt}) = \exp\left[-\frac{(x_{it} - x_{jt})^2}{2\sigma_t^2}\right].$$

Since e^{κ_t} is a constant not depending on either input to the kernel function, we can see that Eq. (4.9) is equivalent to a Gaussian (or radial basis) kernel with "standard deviation unit"

$$\sigma_t = \sqrt{\frac{1}{\kappa_t}} \quad \text{or} \quad \sigma_t = \sqrt{\frac{1}{\kappa_t}} \times \frac{360^\circ}{2\pi}. \tag{4.10}$$

Table 4.2: Conversion Between κ_t And σ_t [Eq. (4.10)]

Dihedral Angle		σ_t (degrees)	
(t)	κ_t	One Decimal	Nearest 10-th
1st	8	20.3°	20°
2nd	8	20.3°	20°
3rd	4	28.6°	30°
4th	2	40.5°	40°

Therefore, our choices of $\kappa_1 = \kappa_2 = 8$, $\kappa_3 = 4$ and $\kappa_4 = 2$ roughly correspond to using "standard deviation units" of 20°, 30° and 40° in the corresponding Gaussian kernel (see Table 4.2). A side-chain prediction is often deemed successful if the predicted dihedral angle is within 40° of the true angle [55]. Thus, our choice of κ_4 agreed with this convention, and we used larger values for $\kappa_3, \kappa_2, \kappa_1$ to permit less movement for the lower dihedral angles.

4.3 Results and Discussion

We tested our method on a number of well-studied Ras and Ras-like proteins (see Table 4.3). They have been of special interest due to their diverse range of functions. The inactive and active structures of many family members have been crystallized and are known.

We performed both quantitative and qualitative comparisons of our results with those obtained by the Contact Rearrangement Network (CRN) [17] and by the GLASSO [97]. For an allosteric protein, the CRN method generates networks of allosteric pathways by calculating significant differences in the residue-residue contact network derived from the inactive structure and that derived from the active structure. Therefore, it provides a direct and model-free analysis of both structures [17]. The GLASSO is a relatively new statistical method, which we have used in an earlier study (see Chapter 3) to extract direct couplings between residues, but its application required that we work with discrete conformation variables rather than angular variables that describe the conformations more directly (more details in Section 4.3.2). We implemented the KPCCA in MATLAB using the Kernel Methods Toolbox [106].

All three methods' outputs consisted of a list of coupled residues, each ranked by a score indicating their coupling strength. The quantitative comparison was performed using the receiver-operating characteristic (ROC) curve. Treating the list of CRN results as "ground truths", the Area Under the ROC Curve (AUC) is a numeric summary of how well the ranked list produced by the KPCCA or by the GLASSO matched the CRN findings (see

Table 4.3). These AUC values show quite conclusively that KPCCA's detection of allosteric couplings is significantly better than random, and that there is a good deal of agreement between our results and those from the CRN. This is a significant finding considering that the CRN relies on structural information of both the inactive and the active structure of an allosteric protein, while we have analysed the dynamics of the side chains in the inactive structure alone.

Furthermore, we evaluated our results qualitatively (see Section 4.3.1 below) by visualizing them as *interaction graphs*, and comparing them to the interaction graphs generated by the CRNs and by the GLASSO. The inferred couplings for each test case were visualized as a 3D network graph superimposed onto the 3D structure of the protein itself. All 3D molecular visualizations and graphs were produced using the StructBio package [12] for the software, Chimera [86]. For each coupling, nodes were placed at the α-carbon for each of the involved residues and edges were drawn between them. We used two different cut-offs to threshold the top-ranked couplings when generating the interaction graphs, and studied a small subset of these couplings in more detail. The first threshold was equal to the number of couplings identified by the CRN [17] for each individual test case, so that we could make a fair comparison. The second threshold was 100 for all test cases, and used for generating 2D graphs (Figures 4.3, 4.4 and 4.5), so that connections between residues in important regions could be shown more clearly. It should be noted that both types of cut-offs allowed only a small subset of all the couplings ($\approx$ 0.6%-1%) to be shown. From these graphs, we noticed that the top-ranked couplings often involved allosterically crucial residues (more details in Section 4.3.1). Moreover, these allosterically important residues often appeared as high-degree nodes in the graphs; sometimes, they could be seen to act as *hubs* connecting the allosteric region to other functionally important parts of the protein, such as the binding site.

Both our quantitative and qualitative results indicated that the KPCCA outperformed the GLASSO in that it was able to capture couplings that correspond to more significant connections in the crucial regions of the test cases. This confirms that, to infer direct residue-residue couplings from the same conformational data, the KPCCA – which facilitates data modelling by continuous, multivariate angular variables — is more accurate than the GLASSO. In some cases, such as Rheb (Table 4.3), although we noticed a smaller AUC value for the KPCCA (indicating that the GLASSO had slightly better agreement with the CRN), the interaction graphs still showed that the KPCCA identified the crucial residues more effectively (see, e.g., Figures 4.6-4.7 and more discussions in Section 4.3.2).

Table 4.3: Allosteric Proteins from Three Sub-families of the Small G Protein Super-family with PDB IDs of Active and Inactive Structures. Active structure: bound to GTP. Inactive structure: bound to GDP.

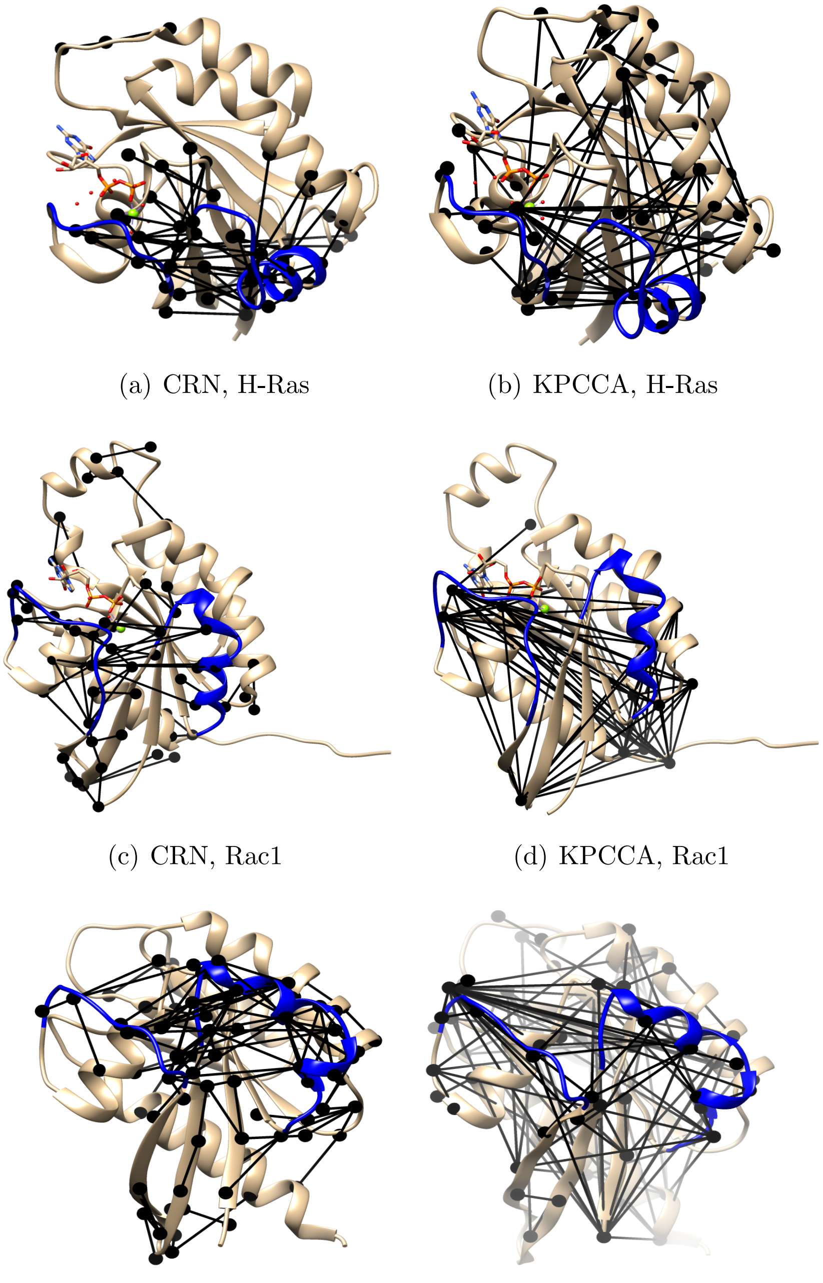
Sub		PDB ID		AUC against CRN	
Family	Protein	Inactive	Active	KPCCA	GLASSO
Ras	H-Ras	4Q21	6Q21	0.796	0.776
	Rap2A	1KAO	2RAP	0.693	0.677
	Rheb	1XTQ	1XTS	0.699	0.711
Rho	RhoA	1FTN	1A2B	0.750	0.719
	Rac1	1HH4(A)	1MH1	0.672	0.594
	Cdc42	1AN0	1NF3	0.681	0.675
Rab	Sec4	1G16	1G17	0.676	0.683
	Ypt7p	1KY3	1KY2	0.717	0.666

4.3.1 Small G Proteins

The members of this super-family are structurally categorized into five sub-families: Ras (Section 4.3.1), Rho (Section 4.3.1), Rab (Section 4.3.1), Sar1/Arf and Ran. Both NMR and crystallographic analyses have shown that members of different sub-families act as molecular switches that cycle between on (active) and off (inactive) states [87], and that they share a common topology in the GDP/GTP binding domain [103]. In this section, we highlight our findings for a few representative and well-studied members of these sub-families.

Ras Sub-Family

Members from the Ras subfamily are the primary members of the super-family; they play a critical role in human oncogenesis [113]. When activated, they regulate cell proliferation and survival through gene expression [113] [103]. We experimented with three members of this sub-family: H-Ras, Rap2A, and Rheb (see Table 4.3). We used the software, Blastp (protein-protein BLAST; `http://blast.ncbi.nlm.nih.gov/Blast.cgi`), to perform pairwise sequence alignment between the test cases. The sequences of Rap2A and Rheb respectively shared 49% and 36% amino-acid identity with H-Ras. The regions that undergo major conformational changes in H-Ras have been identified [78] as Switch I (residues 30-38) and Switch II (residues 60-73); see Figure 4.1. Switch II is known to be directly involved in switching the protein from inactive to active status [52]. Residues residing in the binding site are residues 28-35, 12-19, 145-147, and 116-120. Using the Rosetta software for structural prediction [89] [52] obtained strong correlations in Switch II and the hydrophobic core, which is conserved in the Ras family, though they did not report connectivity between the two Switches. The CRN for H-Ras generated using both active and inactive structures contained 75 couplings [Figure 4.2(a)]. The interaction graph based on the top 75 couplings inferred by the KPCCA is shown in Figure 4.2(b); it clearly shows that strong couplings connect the two Switches to each other and to the binding site. In addition, the residues in these two regions are among the highest-degree nodes in the interaction graph — e.g., the node with maximum degree of 21 (see Table 4.4) in the 2D interaction graph [Figure 4.3(a), based on the top 100 couplings] is associated with residue 34 in the Switch I region.



(a) CRN, H-Ras

(b) KPCCA, H-Ras

(c) CRN, Rac1

(d) KPCCA, Rac1

(e) CRN, Sec4

(f) KPCCA, Sec4

Figure 4.2: 3D interaction graphs of H-Ras (top), Rac1 (middle), and Sec4 (bottom), showing the top 75, 63, and 112 couplings, respectively. The couplings are mostly seen in the Switch I and Switch II regions (blue) and the nearby β-sheets.

Rho Sub-Family

The best-studied members of this sub-family are RhoA, Rac1 [38], and Cdc42 (Table 4.3). Sequence alignment results from Blastp showed that RhoA, Rac1 and Cdc42 shared 30%, 29% and 32% amino-acid identity with H-Ras, respectively, whereas the amino-acid identity is higher within the sub-family, e.g., 58% between RhoA and Rac1, and 69% between Rac1 and Cdc42. Like the Ras family, the Rho proteins are involved as regulators in cell cycle progression (cell polarity, movement, shape, and so on) and gene expression [113]. The three aforementioned members are known to be involved in very diverse cellular processes [113] [103]. The CRN for Rac1 consisted of 63 couplings [Figure 4.2(c)]. The edges were mostly concentrated in the classic Switch regions of this super-family. Figure 4.2(d) shows a 3D interaction graph based on the top 63 couplings inferred from the KPCCA; this network included residues 29, 32, and 34 from Switch I as well as residues 69 and 73 from Switch II. The set of top-ranked couplings also included residues in the C-terminus (residues 179, 180 and 182) and those from a loop segment (residues 106-110). These regions were also identified by the CRN. Figure 4.4(a) shows a 2D interaction graph formed by the top 100 couplings from the KPCCA; we can see that it is a highly connected and concentrated network consisting of a single connected component with only 15 residues (see also Table 4.4).

Rab Sub-Family

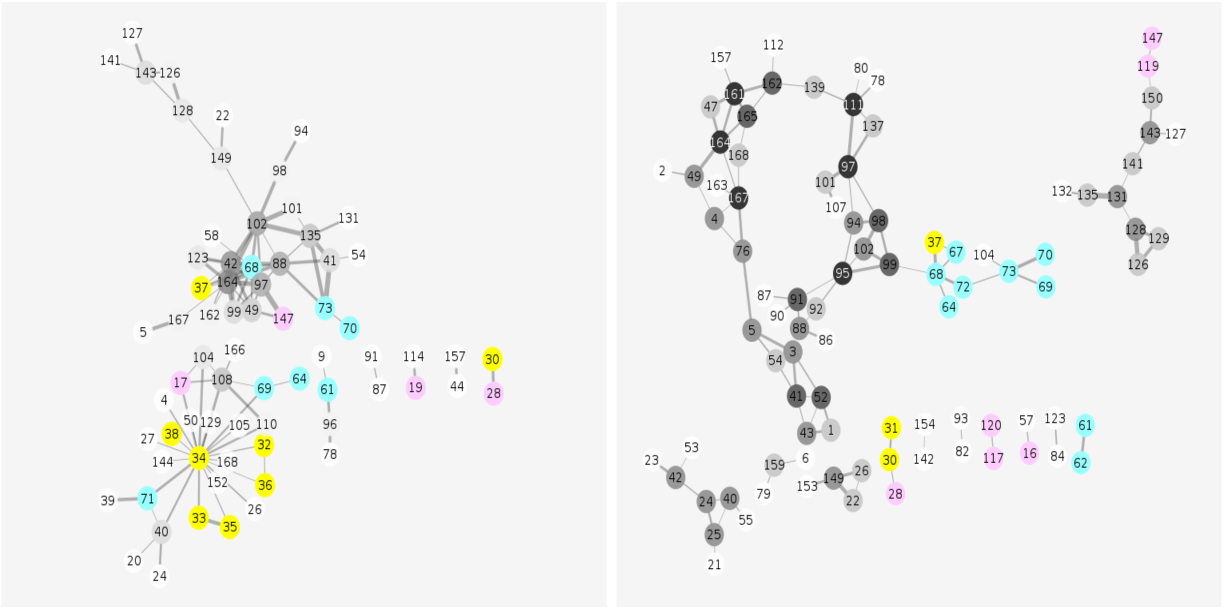
The Rab proteins constitute the largest sub-family in the small G protein super-family [35]. They are involved in the regulation of intracellular vesicular trafficking, vesicle formation, budding and fusion [113] [99] [75]. For our experiments, we selected the inactive structures of a few well-known members, such as Rab7, Sec4, Ypt7p, Rab11b, Rab11a, and Rab6b (PDB IDs: 1VG1, 1G16, 1KY3, 2F9L, 4LWZ, and 2FE4). The PDB structures for many members of this family are incomplete in the critical and functional regions, i.e., the two Switches. To perform our experiments, we used the software, MODELLER (`http://toolkit.tuebingen.mpg.de/modeller`), to complete their structures [90]. By comparing the active and inactive structures, we noticed that, except for Sec4 and Ypt7p, the others underwent a secondary structural change (from loop to helix, or vice versa) in the Switch II region during the transition from inactive to active form. We excluded them from the current study, which focuses on proteins with minor backbone motions [105]. The common amino acids shared between H-Ras and (Sec4, Ypt7p) are about (35%, 32%), respectively. The CRN for Sec4 consisted of 112 couplings [Figure 4.2(e)]. The interaction graph based on top-ranked couplings by the KPCCA [Figures 4.2(e) and 4.5(a)] showed connections between the two Switches through the edges, (47,87) and (83,56).

Ran and Arf/Sar1 Sub-Family

The Ran proteins [93] [84] [100] are best known for their involvement in nucleocytoplasmic transport of macromolecules (e.g., RNAs, proteins), whereas members of the Arf family function as regulators of vesicular transport [113] [103], like the Rab proteins. Comparing the active and inactive structures of the best studied members from these families, we noticed that they underwent drastic conformational changes during the activation procedure. Calculated RMSDs between the GDP- and GTP-bound pairs for Ran (PDB IDs: 1BYU-1RRP, 1BYU-1IBR, 3GJ0-1WA5), Arf1 (PDB ID: 1HUR-1O3Y) and Arf6 (PDB ID: 1E0S-1HFV) were in the range of approximately 4-14Å. Hence, these proteins do not belong to the category characterized by "minor backbone motions" [105] and we excluded them from the current study.

4.3.2 KPCCA and GLASSO

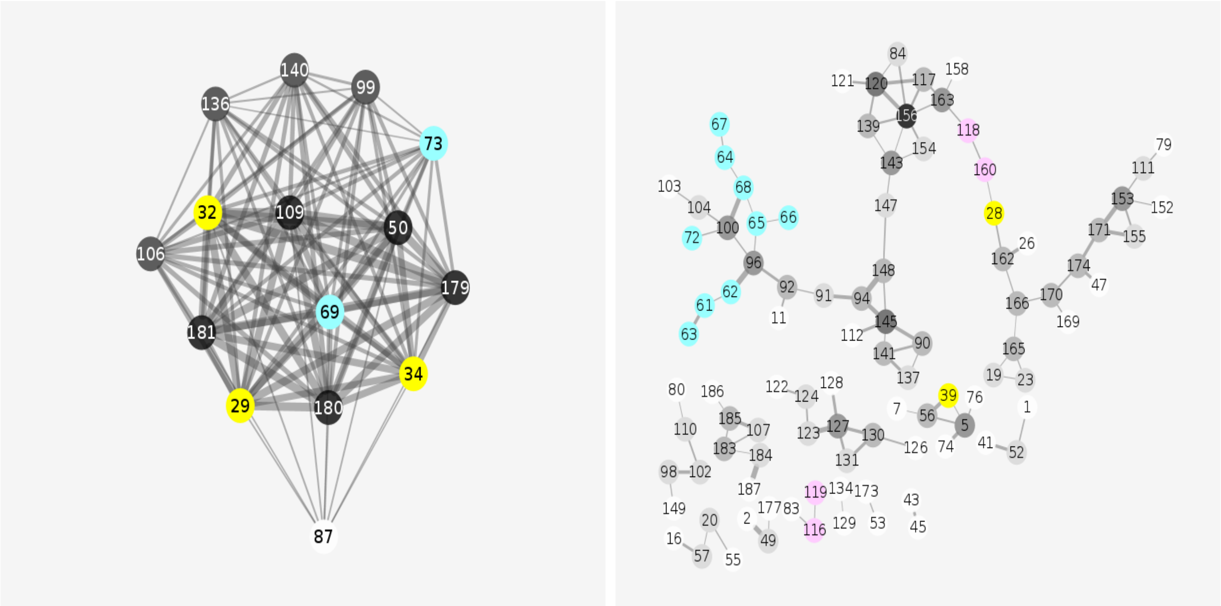
In the Chapter 3, application of the GLASSO to infer direct couplings between side chains was explained [96]. The GLASSO is incapable of handling multivariate angular variables. The GLASSO-based method was also applied to the same set of proteins and a thorough comparison between the overall features and final results of the two methods is provided.



(a) KPCCA, H-Ras

(b) Glasso, H-Ras

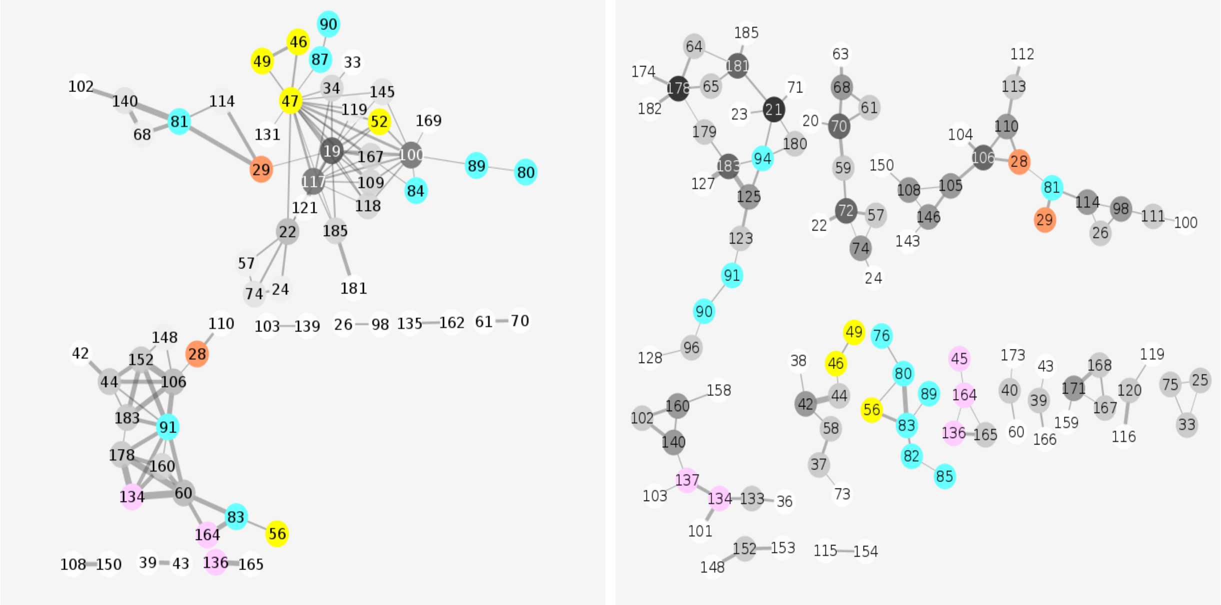
Figure 4.3: 2D interaction graphs of H-Ras, showing the top 100 couplings. For nodes shaded in grey, darkness is proportional to node degree. For edges, thickness is proportional to coupling strength. Residues in Switch I are coloured yellow and those in Switch II, blue. The binding site residues (which do not overlap with Switch I) are highlighted in pink and the phosphate-binding loop, orange. The KPCCA produced more connected networks whereas, for the GLASSO, the inferred couplings are more “spread out” within the molecule.



(a) KPCCA, Rac1

(b) Glasso, Rac1

Figure 4.4: 2D interaction graphs of Rac1, all showing the top 100 couplings. For nodes shaded in grey, darkness is proportional to node degree. For edges, thickness is proportional to coupling strength. Residues in Switch I are coloured yellow and those in Switch II, blue. The binding site residues (which do not overlap with Switch I) are highlighted in pink and the phosphate-binding loop, orange. The KPCCA produced more connected networks whereas, for the GLASSO, the inferred couplings are more “spread out” within the molecule.



(a) KPCCA, Sec4

(b) Glasso, Sec4

Figure 4.5: 2D interaction graphs of Sec4, all showing the top 100 couplings. For nodes shaded in grey, darkness is proportional to node degree. For edges, thickness is proportional to coupling strength. Residues in Switch I are coloured yellow and those in Switch II, blue. The binding site residues (which do not overlap with Switch I) are highlighted in pink and the phosphate-binding loop, orange. The KPCCA produced more connected networks whereas, for the GLASSO, the inferred couplings are more "spread out" within the molecule.

Table 4.4: Statistical Features of Interaction Graphs (based on graphs formed by the top 100 inferred couplings)

Protein	Method	No. of Nodes	No. of Connected Components	Max. Node Degree	Avg. Node Degree
H-Ras	GLASSO	91	12	5	1.176
	KPCCA	70	7	21	2.829
Rac1	GLASSO	97	12	7	2.062
	KPCCA	15	1	14	13.333
Sec4	GLASSO	100	14	5	2.000
	KPCCA	64	9	17	3.125

Comparison

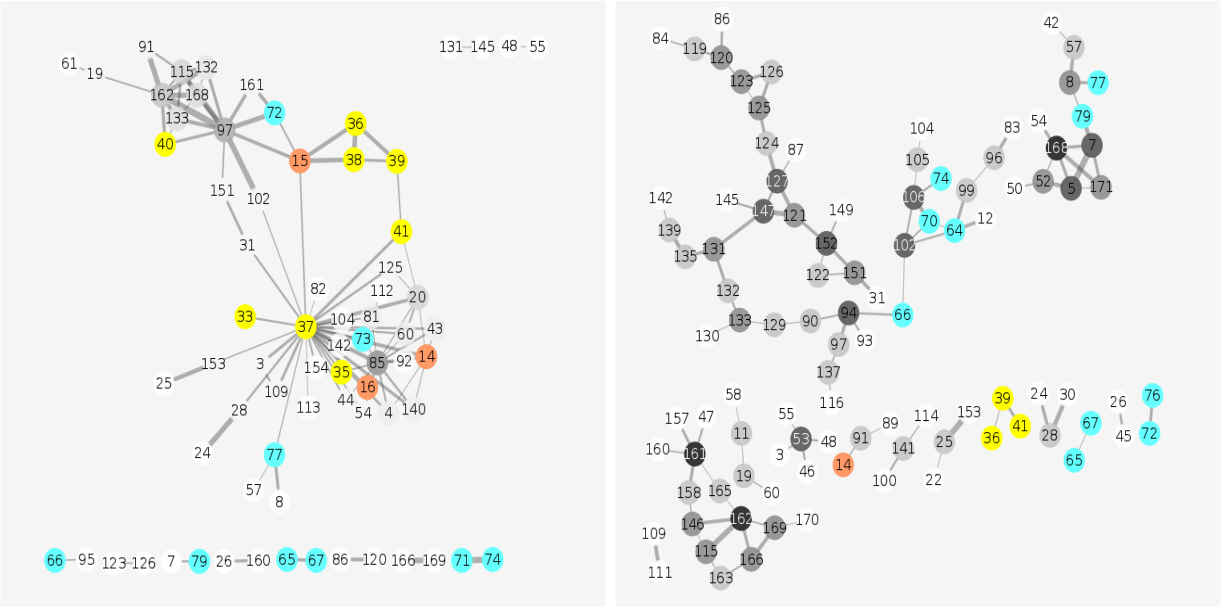
Based on our observations, the KPCCA has the following advantages over the GLASSO:

(i) The KPCCA models side-chain conformations more appropriately with continuous rather than discrete variables. As we stated above, one main disadvantage of the GLASSO algorithm was the need to encode side-chain conformation information using a discrete rotamer library. By contrast, the KPCCA algorithm facilitates a more realistic modelling approach, consistent with the intrinsic nature of conformational data, by allowing us to use continuous, multidimensional angular variables to characterize side-chain conformations.

(ii) The KPCCA identifies allosteric regions more effectively. Quantitatively, Table 4.3 showed that the KPCCA results agreed well with those of the CRN, and that, in this respect, it either compared favourably to the GLASSO or obtained similar results. The superiority of the KPCCA becomes more evident from the qualitative comparisons based on interaction graphs. In particular, the strongest couplings inferred by the KPCCA are concentrated in the allosteric regions, whereas couplings inferred by the GLASSO are more "spread out" within the entire molecule. The GLASSO often identified couplings between residues that may undergo concerted motions in the inactive structure but do not necessarily reside in allosteric regions. Some of these residues are located in semi-rigid secondary structures such as helices; they may reside in or close to the binding site but do not necessarily participate directly in allosteric events. This can be seen more clearly in Figures 4.3, 4.4, and 4.5 which contain 2D interaction graphs formed by the top 100 couplings identified by each method for a few representative test cases. For Rheb and Sec4, although it is the GLASSO that appeared to be in better agreement with the CRN (see Table 4.3, the AUC values), their respective interaction graphs lead to the same qualitative conclusion as that in other cases. Even for these two cases, the KPCCA can be seen to have identified more dependencies specific to coupled motions during allosteric events (see Figure 4.6). In addition, couplings in important functional regions also tend to emerge earlier (i.e., at *higher* positions) in the ranked list of the KPCCA than in that of the GLASSO, another indication that the KPCCA is better at identifying regions crucial to function. For example, Figure 4.7 contains 2D interaction graphs for Rheb using a cut-off threshold < 100, and shows that the KPCCA has identified more residues in the functionally crucial regions at the top of its ranked list than has the GLASSO.

(iii) The KPCCA's interaction graphs tend to show better connectivity among functionally important regions. Another important observation was that the GLASSO obtained significantly sparser clusters of residues (see Figures 4.3, 4.4 and 4.5). The increased connectivity among couplings inferred by the KPCCA was a notable advantage; these connections can potentially explain the mechanism of information propagation within the molecule. If interaction pathways between the allosteric and/or binding sites can be discerned using an interaction graph based on K top-ranked couplings from the GLASSO, using top-ranked couplings from the KPCCA it often can be done with much fewer than K couplings. Table 4.4 contains various statistical features showing the overall connectivity of the interaction graphs produced by the GLASSO and by the KPCCA for H-Ras, Rac1 and Sec4.

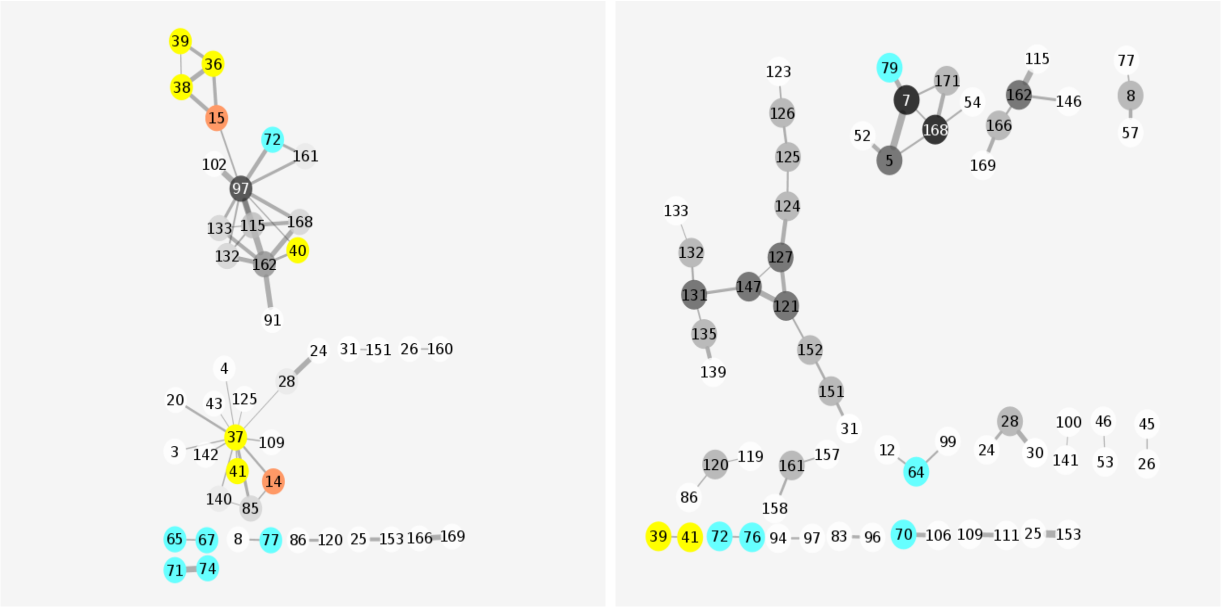
(iv) The KPCCA is less prone to entropic bias. One drawback of using discrete rotamers to encode conformations is that the results produced by the GLASSO were biased towards larger amino acids that naturally have more diverse rotamer conformations. This is referred to as the "entropic bias" in the literature, e.g., Jones et al. [48], Dun et al. [24] who also suggested techniques for its correction. By contrast, the KPCCA does not appear to suffer from such biases. Figure 4.8 shows the average number of available rotamer conformations for residues involved in the top 1-5, 2-6, ..., up to the top 300-305 paired couplings, as computed by the KPCCA and by the GLASSO for H-Ras, Rac1 and Sec4. For the GLASSO, the rankings were based on scores after bias correction. Although no correction was introduced for the KPCCA, its results do not show significant bias.



(a) KPCCA, Rheb

(b) Glasso, Rheb

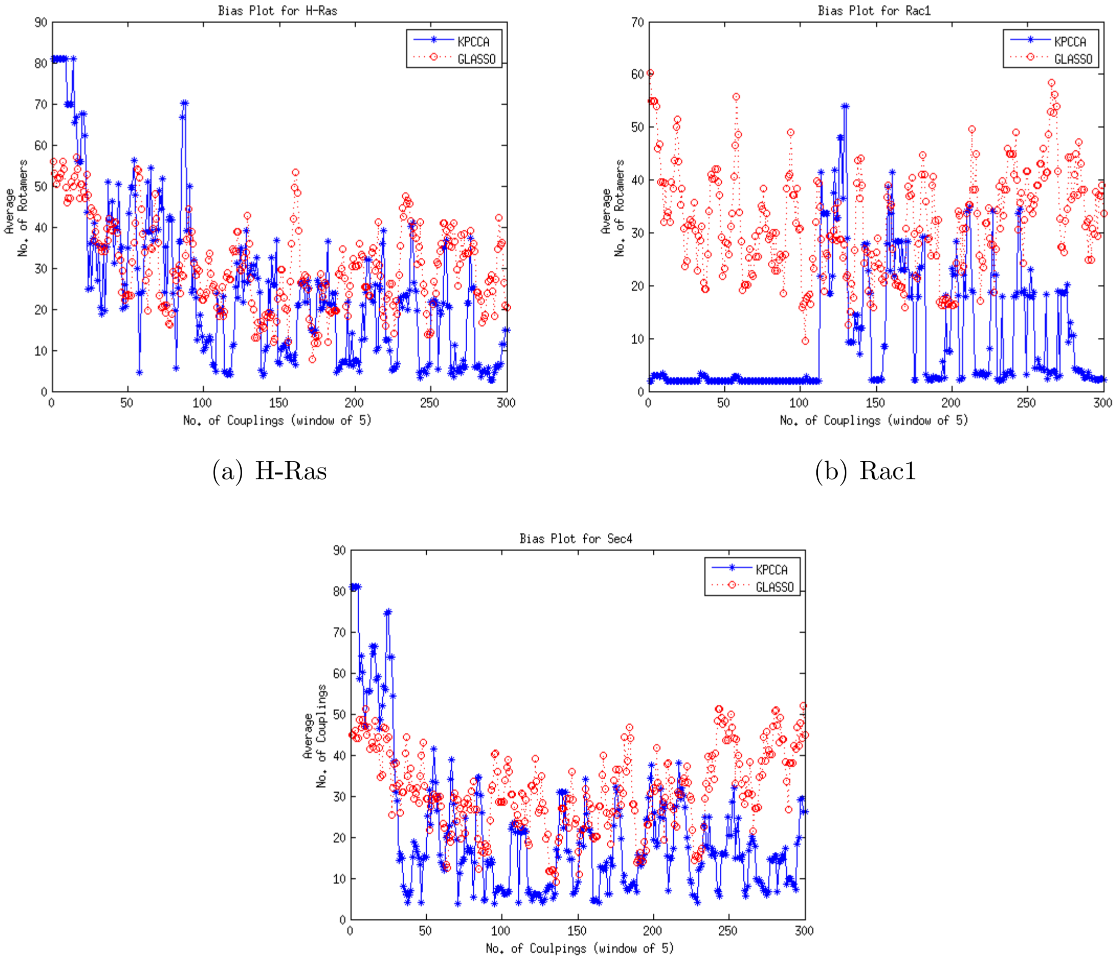
Figure 4.6: 2D interaction graphs for Rheb, using the top 100 couplings. Residues in Switch I (II) are coloured yellow (blue). (a) KPCCA: The two Switch regions are directly connected (residue 37 from Switch I with residues 73 and 77 from Switch II); moreover, Switch I is indirectly connected to Switch II by residue 15 in the phosphate-binding loop [121]. (b) GLASSO: No connection between the Switches is identified.



(a) KPCCA, Rheb

(b) Glasso, Rheb

Figure 4.7: 2D interaction graphs for Rheb, using the top 47 couplings (the same number as identified by the CRN). Residues in Switch I (33-41), Switch II (63-79) and the phosphate-binding loop (p-loop) are coloured yellow, blue and orange, respectively. The p-loop residues are connected to the Switches in the CRN. Crucial couplings emerge at higher positions in the ranked list of the KPCCA than in that of the GLASSO.



(a) H-Ras

(b) Rac1

(c) Sec4

Figure 4.8: X-axis: Rank order of the inferred couplings ($x = 1, 2, ..., 300$). Y-axis: Average number of available rotamers for residues involved in couplings ranked at positions x, $x + 1$, ..., $x + 4$. Red: GLASSO (with bias correction). Blue: KPCCA (without bias correction). The KPCCA does not show significant bias towards residues with more rotamer alternatives; in fact, the average number of rotamers is lower for the KPCCA than for the GLASSO in general.

4.4 Chapter Summary

We have proposed a novel extension of CCA, namely KPCCA, to quantify direct correlations between multidimensional angular data. Existing methods for inferring direct correlations do not handle data of this type, which are common in structural bioinformatics, where side-chain conformations of proteins are characterized by a number of dihedral angles. Using information about side-chain fluctuations in the inactive structure alone, we are able to identify common, allosterically crucial regions (e.g., Switch I and Switch II) in the Ras, Rho, and Rab sub-families of small G proteins. Residues in these allosteric regions appear in the strongest couplings inferred by our method and in the densest regions of the corresponding interaction graph. Furthermore, allosteric sites and binding sites are connected in these graphs, which may explain the mechanism with which allostery occurs in these proteins.

Our analytic framework is modular. In principle, ensembles generated by other techniques such as MC simulations and/or MD can be used as well. But currently they are much less efficient. For instance, in one of our test cases (Rap2A; PDB ID: 1KAO, 167 residues), SCWRL took about 1 second to generate a structure whereas an MC method in Rosetta, like that described by [51], took as much as 40 seconds. Hence, for an ensemble of size $[167 \times (167-1)]/2 \approx 14,000$, our current method took about 4 hours but an MC method would have taken 160 hours, almost a full week, for a single protein!

In future studies, our proposed analytic framework can be extended to include backbone dihedral angles as well. This will allow us to study allosteric behaviours of all protein types, even those that may undergo drastic backbone motions. The method also can be applied to other problems in bioinformatics, e.g., for revealing the “hot spot” residues in protein-protein interactions by using only the fluctuation information of the “unbound” protein [82].

Chapter 5

Concluding Remarks and Future Work

This chapter presents closing arguments for this dissertation. A summary of contributions, conclusive remarks, and potential directions for future research work are presented.

5.1 Highlights of the Thesis

This dissertation has addressed two recent crucial findings about side-chain conformations in new computational frameworks. The findings are as follows:

First, the prevalence of alternative conformations for the side chains in the x-ray crystallography data (more than previously thought). The phenomenon is referred to as *conformational polymorphism*. Despite this new discovery, the fact that still more than $\approx 95\%$ of the side chains in the PDB are uniquely modelled shows that the conformational polymorphism topic requires more attention from the scientific community in order to devise computational methods capable of precisely characterizing this phenomenon for protein residues. This dissertation's proposal can be considered as the first step towards achieving such goal.

Second, significant effects of concerted side-chain fluctuations in information transmission within protein molecules. Backbone conformational changes were traditionally considered to play the main role for allosteric behaviour of molecules. By contrast, several studies have shown the necessity of re-assessing the side chains' roles in allostery. For this purpose, this dissertation proposes two different computational frameworks that avoid transitive dependencies between side-chain conformational variables and extract direct couplings. One of the frameworks is novel and can contribute to other applications in bioinformatics (See 5.3).

A common characteristic of the proposed algorithms in this dissertation is recognizing the nature of side-chain conformational data and handling them as continuous angular multivariate data, i.e., avoiding any discretization or encoding.

5.2 Concluding Remarks

We successfully applied the particle-based inferential technique, PBP, to design the first computational method that obtains residue-specific conformational distributions and predicts conformational polymorphisms for the side chains. The final results were in good agreement with results of direct analysis of the x-ray crystallography data. However, the proposed framework is computationally expensive due to the use of MC sampling from the conformations.

Furthermore, the successful application of the direct-coupling concept in the identification of functional residues, for the first time in this dissertation, demonstrated that:

- In proteins with minor backbone changes, strongly coupled side chains alone can reveal functional regions known to be involved in allosteric events.
- Disentangling direct from indirect couplings can help in finding meaningful interaction paths between crucial regions of a protein, and the obtained paths or networks can reveal how information is propagated from one site to another.
- Relying on information from inactive structure's fluctuations alone, we can infer a lot about the functional regions of the molecule. This finding implies that these functional regions exist intrinsically in the inactive structure's dynamics. This is a significant observation since in many cases we do not have access to the active structure of an allosteric protein. Interestingly, similar conclusions were recently drawn by studying dynamics of unbound structures involved in protein-protein interactions (PPI) in order to identify the functional residues named as "hot-spots" [82].

Moreover, as explained in Chapters 3 and 4, the task of direct coupling inference requires a conformational dataset or ensemble that can represent many fluctuated conformational states of a protein's side chains. The common methods to generate a conformational ensemble are MD and MC simulations. However, the MD-based techniques cannot capture slow re-arrangements of the side chains, and the MC techniques are computationally very expensive. The dissertation proposes and successfully applies a new and efficient method based on SCP applications.

The coupling results obtained for the G family proteins (Chapter 4) showed an overlap (not a complete agreement) with the results of the popular SCA (see Section 3.3.3) that relies on extracting couplings from sequence information of an entire protein family. The method has been applied to the G family to identify allosteric sites/residues [101]. However, a lack of agreement has been reported by other groups that rely on protein's structural information rather than the sequence information in their allostery-related studies. Possible reasons for the lack of full agreement between the results of SCA and those reported in this dissertation may be the following:

- Methodological differences: SCA is based on calculating sample covariance (or marginal dependence) for the amino acid positions in a given sequence alignment, while the proposed methods in this dissertation extract direct couplings or conditional dependence between the side-chain pairs. As mentioned before, the covariance calculation does not remove transitive relationships between variables (amino acid positions); but direct coupling does.

- Different sources of data: SCA uses sequence data, while we rely on structural data.

- Amount of data: SCA uses sequences of a whole family; we only use information of side chains' fluctuations obtained from the inactive structure alone.

- Purpose of method: Initially, SCA was not developed for extracting allosteric sites; but it was devised to discover groups of co-evolved residues in a protein family, while the allosteric characteristics might not necessarily be conserved in all sub-families of a super-family. "In principle, SCA is also limited to detecting correlated changes in residues during evolution, presumably highlighting only correlation networks with a selected function and can therefore say little about presence or absence of other correlation"[23]. Furthermore, as pointed by Ranganathan and co-workers [101], the co-evolution might occur due to different reasons, such as, maintaining the stability of a protein fold and not solely due to its allosteric functionalities.

The proposed methods in the allostery-related study section of this dissertation obtain more protein-specific results than what SCA obtains, because they are inferred from each individual protein's structural fluctuations. However, the remarkable findings of SCA suggest that an integration between the two types of methods (structure-based and sequence-based) may be the key to reveal and characterize allosteric behaviour in proteins.

5.3 Future Research

In the very recent years, after initiation of this thesis, several interesting studies related to the topics discussed in this dissertation have appeared. These studies may help in improving the reported results in this dissertation or assist in extending the proposed frameworks and algorithms to be applicable to other problems in the bioinformatics field.

In the following, we present several interesting directions for future work moving beyond the current scope of the research efforts of this thesis:

- The proposed PBP-based framework for conformational polymorphism (Chapter 2) is currently computationally expensive. Furthermore, unless we use a very large number of particles, the stochastic re-sampling applied in the algorithm may result in instabilities and degeneracies [83]. Recently, Pacheco et al. [83] have proposed an improved version of Max-PBP, named *diverse particle max-product* (D-PMP). In the new algorithm, the particle set is kept diverse and within a computationally tractable size by avoiding the original stochastic re-sampling and employing an optimization perspective to the particle generation and message approximation. As reported, the new modifications lead to computation speed-ups and prevent common degeneracies. Hence, applying the D-PMP to the polymorphism prediction may improve the current results by revealing more diverse modes in the final residue-specific conformational distributions.

- The novel framework (KPCCA) proposed in Chapter 4 was originally devised to extract the directly coupled side chains; but the framework is extendable to infer coupled fluctuations of both side chain and backbone. The extended framework can obtain more comprehensive insights about allosteric behaviour in different categories of proteins (not only those with minor conformational changes in the backbone). However, this inference task will require the following: first, a comprehensive dataset containing conformational fluctuations for both side chains and backbone. This may be achieved by applying MC sampling techniques integrated to the protein design frameworks such as Rosetta [51], second, a modification in the kernel function to include the backbone dihedral angles, as well as the side chains'.

- Recently, a study by Ozbek et al. [82] showed for the first time that relying on the fluctuation information of “unbound” structure of proteins that are involved in protein-protein interactions, can reveal functional residues or the “hot spots”. Since PKCCA is also a framework to extract functional residues based on the conformational fluctuations of the side chains, one future direction can be to investigate application of this framework to the PPI problem.

- This dissertation has addressed the side-chain conformational polymorphism in Chapter 2. However, calculating the residue-specific conformational distributions does not completely characterize the polymorphism, and an important sub-problem still remains unresolved. The field is still in need of efficient methods that can compute pairwise or joint conformational distributions between the conformationally polymorphic side chains. This topic has roughly been touched upon in Chapter 3 (Section 3.3.2) by calculating and discussing the coupling scores between alternate conformations of a selected pair of coupled side chains; but computing joint distributions for a pair of coupled side chains remains a challenging task that can be a direction for significant future research.

References

[1] K. R. Acharya, J. Ren, D. I. Stuart, D. C. Phillips, & R. E. Fenna. Crystal structure of human α-lactalbumin at 1.7Å resolution. *Journal of molecular biology*, 221(2), 571-581, 1991.

[2] T. Akutsu, NP-hardness results for protein side-chain packing, in Miyano, S., Takagi, T., eds., *Genome Informatics*, 8, 180-186, 1997.

[3] S. Arent, P. Harris, K. F. Jensen, and S. Larsen. Allosteric regulation and communication between subunits in Uracil Phosphoribosyltransferase from Sulfolobus Solfataricus. *Biochemistry*, 44:883-892, 2005.

[4] O. Banerjee, L. E. Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *J. Machine Learning Research*, 9:485-516, 2008.

[5] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The Protein Data Bank, *Nucleic Acids Research*, vol. 28, pp. 235-242, 2000.

[6] C. Birck, L. Mourey, P. Gouet, B. Fabry, J. Schumacher, P. Rousseau, D. Kahn, J. P. Samama. Conformational changes induced by Phosphorylation of the FixJ receiver domain. *Structure*, 7:1505-1515, 1999.

[7] D. D. Boehr, R. Nussinov, and P. E. Wright, The role of dynamic conformational ensembles in biomolecular recognition, *Nat. Chem. Biol.*, vol. 5, no. 11, pp. 789-796, 2009.

[8] Bourne, Philip E., and Helge Weissig. ”Structural bioinformatics.” (2003).

[9] M. Bower, F. Cohen, and R. Dunbrack, Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool,*J. Mol. Biol.*, vol. 267, pp. 1268-1282, 1997.

[10] R. Bruschweiler, New approaches to the dynamic interpretation and prediction of NMR relaxation data from proteins,*Curr. Opin. Struct. Biol.*, vol. 13, pp. 175-183, 2003.

[11] L. Burger and E. Van Nimwegen. Disentangling direct from indirect co-evolution of residues in protein alignments. *PLoS computational biology*, 6(1), e1000633, 2010.

[12] F. Burkowski. Computational and Visualization Techniques for Structural Bioinformatics Using Chimera. Chapman and Hall/CRC. 461 p., 2014.

[13] A. Canutescu, A. Shelenkov, and R. Dunbrack, A graph-theory algorithm for rapid protein side-chain prediction, *Protein Sci.*, vol. 12, no. 9, pp. 2001-2014, 2003.

[14] H. S. Cho, S. Y. Lee, D. Yan, X. Pan, and J. Parkinson, S. Kustu, D. E. Wemmer, J. G. Pelton. NMR structure of activated CheY. *J. Mol. Biol.*, 297(3):543551, 2000.

[15] W. J. Conover, Practical Nonparametric Statistics, New York: John Wiley & Sons, 1971.

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

[17] M. D. Daily, T. Upadhyaya, and J. J. Gray. Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins: Structure, Function, and Bioinformatics*, 71(1):455-466, 2008.

[18] M. D. Daily and J. J. Gray. Local motions in a benchmark of allosteric proteins. *Proteins: Structure, Function, and Bioinformatics*, 67:385-399, 2007.

[19] K. Decanniere, A. M. Babu, K. Sandman, J. N. Reeve, and U. Heinemann, Crystal structures of recombinant histones HMfA and HMfB from the hyperthermophilic archaeon Methanothermus fervidus, *J. Mol. Biol.*, 303(1), 35-47, 2000.

[20] O. N. Demerdash, M. D. Daily, and J. C. Mitchell. Structure-based predictive models for allosteric hot spots. *PLoS computational biology*, 5(10), e1000531. 2009.

[21] J. Desmet, M. De Maeyer, and I. Lasters, The dead-end elimination theorem and its use in protein side-chain positioning, *Nature*, 356:539-542, 1992.

[22] A. Doucet, N. de Freitas, and N. J. Gordon, Sequential Monte Carlo Methods in Practice, New York: Springer-Verlag, 2001.

[23] K. H. DuBay, J. P. Bothma, and P. L. Geissler. Long-range intra-protein communication can be transmitted by correlated side-chain fluctuations alone. *PLoS Comput. Biol.*, 7, e1002168, 2011.

[24] S. D. Dunn, L. M. Wahl LM, and G. B. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333-340, 2008.

[25] R. Dunbrack and M. Kurplus, Backbone-dependent rotamer library for proteins: application to side-chain prediction, *J. Mol. Biol.*, 230:543-574, 1993.

[26] C. Dyer, M. Quillin, A. Campos, J. Lu, M. McEvoy, A. Hausrath, E. Westbrook, P. Matsumura, B. Matthews, and F. Dahlquist. Structure of the constitutively active double mutant CheY(D13K) Y-106W alone and in complex with a Flim peptide. *J. Mol. Biol.*, 2004. 342:1325-1335.

[27] G. Elidan, I. McGraw, and D. Koller, Residual belief oropagation: informed scheduling for asynchronous message passing, *Proc. 22nd Conf. Uncertainty in Artificial Intelligence (UAI '06)*, 2006.

[28] I. Ezkurdia, O. Graña, J. M. Izarzugaza, and M. L. Tress. Assessment of domain boundary predictions and the prediction of intramolecular contacts in CASP8. *Proteins: Structure, Function, and Bioinformatics*, 77(S9), 196-209, 2009.

[29] M. S. Formaneck, L. Ma, and Q. Cui. Reconciling the "old" and "new" views of protein allostery: A molecular simulation study of Chemotaxis Y protein (CheY). *Proteins: Structure, Function, and Bioinformatics*, 63:846-867, 2006.

[30] H. Frauenfelder, G. A. Petsko, and D. Tsernoglou, Temperature-dependent x-ray-diffraction as a probe of protein structural dynamics, *Nature*, 280:558-563, 1979.

[31] K. K. Frederick, M. S. Marlow, K. G. Valentine, and A. J. Wand, Conformational entropy in molecular recognition by proteins, *Nature*, 448:325-329, 2007.

[32] J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9:432-441, 2008.

[33] W. Freeman and E. Pasztor, Learning to estimate scenes from images, *Advances in Neural Information Processing Systems 11 (NIPS '99)*, MIT Press, 1999.

[34] B. Frey, R. Koetter, and N. Petrovic, Very loopy belief propagation for unwrapping phase images, *Advances in Neural Information Processing Systems 14 (NIPS '04)*, MIT Press, 2004.

[35] I. Garcia-Saez, S. Tcherniuk, and F. Kozielski. The structure of human neuronal Rab6B in the active and inactive form. *Acta Crystallographica Section D: Biological Crystallography*, 62(7), 725-733, 2006.

[36] R. F. Goldstein, Efficient rotamer elimination applied to protein side chains and related spin glasses, *Biophys. J.*, 66:1335-1340, 1994.

[37] P. Gouet, B. Fabry, V. Guillet, C. Birck, L. Mourey, D. Kahn, and J. P. Samama. Structural transitions in the FixJ receiver domain. *Structure*, 7:15171526, 1999.

[38] S. Grizot, J. Faure, F. Fieschi, P. V. Vignais, M. C. Dagher, and E. Pebay-Peyroula. Crystal structure of the Rac1-RhoGDI complex involved in nadph oxidase activation. *Biochemistry*, 40(34), 10007-10013, 2001.

[39] M. E. Hatley, S. W. Lockless, S. K. Gibson, A. G. Gilman, and R. Ranganathan. Allosteric determinants in guanine nucleotide-binding proteins. *Proc Natl Acad Sci*, USA, 100(24):14445-50, 2003.

[40] W. A. Hendrickson. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science*, 254(5028), 51-58, 1991.

[41] A. Hildebrandt, A. K. Dehof, A. Rurainski, A. Bertsch, M. Schumann, N. C. Toussaint, A. Moll, D. Stöckel, S. Nickels, S. C. Mueller, H.-P. Lenhof, and O. Kohlbacher, BALL - biochemical algorithms library 1.3, *BMC Bioinformatics*, 11, article 531, 2010.

[42] L. Holm and C. Sander, Fast and simple Monte Carlo algorithm for side-chain optimization in proteins: application to model, *Proteins: Structure, Function and Bioinformatics*, 14(2), 213-223, 1992.

[43] A. Hörnberg, T. Eneqvist, A. Olofsson, E. Lundgren, and A. E. Sauer-Eriksson, A comparative analysis of 23 structures of the Amyloidogenic Protein Transthyretin, *J. Mol. Biol.*, 302(3), 649-669, 2000.

[44] H. Hotelling. Relations between two sets of variates. *Biometrika* 28(3-4), 321-377, 1936.

[45] J. Hwang and W. Liao, Side-chain prediction by neural networks and simulated annealing optimization, *Protein Eng.*, 8(4), 363-370, 1995.

[46] A. Ihler and D. McAllester, Particle belief propagation, *Proc. 12th Int. Conf. on Artificial Intelligence and Statistics (AISTATS '09)*, 256-263, 2007.

[47] R. Ishima and D. A. Torchia, Protein dynamics from NMR, *Nat. Struct. Biol.*, 7:740-743, 2000.

[48] D. T. Jones, D. W. Buchan, D. Cozzetto, and M. Pontil. PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, 28:184-190. 2012.

[49] R. P. Joosten, T. A. H. te Beek, E. Krieger, M. L. Hekkelman, R. W. W. Hooft, R. Schneider, C. Sander, and G. Vriend, A series of PDB related databases for everyday needs, *Nucleic Acids Research*, 39:D411-D419, doi: 10.1093/nar/gkq1105, 2010.

[50] W. Kabsch and C. Sander, Dictionary of protein secondary structure: pattern recognition of Hydrogen-bonded and geometrical features, *Biopolymers*, 22:2577-2637, 1983.

[51] K. W. Kaufmann, G. H. Lemmon, S. L. DeLuca, J. H. Sheehan, and J. Meiler. Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*, 49(14), 2987-2998, 2010.

[52] B. A. Kidd BA, D. Baker, W. E. Thomas. Computation of conformational coupling in allosteric proteins. *PLoS Comput. Biol.* 5, e1000484, 2009.

[53] G. J. Kleywegt, M. R. Harris, J. Y. Zou, T. C. Taylor, A. Wählby, and T. A. Jones, The Uppsala Electron-Density Server, *Acta Cryst.*, D60:2240-2249, 2004.

[54] R. Kothapa, J. Pacheco, and E. Sudderth, Max-product particle belief propagation, Technical Report, Department of Computer Science, Brown University, 2011.

[55] G. G. Krivov, M. V. Shapovalov, and R. Dunbrack, Improved prediction of protein side-chain conformations with SCWRL4, *Proteins: Structure, Function, and Bioinformatics*, 77(4), 778-795, 2009.

[56] P. T. Lang, H.-L. Ng, J. S. Fraser, J. E. Corn, N. Echols, M. Sales, J. M. Holton, and T. Alber, Automated electron-density sampling reveals widespread conformational polymorphism in proteins, *Protein Sci.*, 19(1420-1431), 2010.

[57] S. Y. Lee, H. S. Cho, J. G. Pelton, D. Yan, E. A. Berry, and D. E. Wemmer. Crystal structure of activated CheY. *J. Biol. Chem.*, 276:16425-16431, 2001.

[58] S. Y. Lee, H. S. Cho, J. G. Pelton, D. Yan, R K. Henderson, D. S. King, L. Shar Huang, S. Kustu, E. A. Berry, D. E. Wemmer. Crystal atructure of an activated response regulator bound to its target. *Nat. Struct. Biol.*, 1:5256. 2001.

[59] C. Lee and S. Subbiah, Prediction of protein side-chain conformation by packing optimization, *J. Mol. Biol.*, 213:373-388, 1991.

[60] T. Lenaerts, J. Ferkinghoff-Borg, F. Stricher, L. Serrano, J. W. H. Schymkowitz, and F. Rousseau. Quantifying information transfer by protein domains: analysis of the Fyn SH2 domain structure. *BMC structural biology*, 8:43. doi: 10.1186/1472-6807-8-43, 2008.

[61] S. C. Li, D. Bu, and M. Li, Residues with similar hexagon neighborhoods share similar side-chain conformations, *IEEE/ACM Trans. Comput. Biology Bioinform.*, 9(1), 240-248, 2012.

[62] H. Li, and C. Frieden. Comparison of c40/82a and p27a c40/82a barstar mutants using 19f NMR. *Biochemistry*, 46:4337-4347, 2007.

[63] J. S. Liu, Monte Carlo strategies in scientific computing, New York: Springer-Verlag, 2001.

[64] S. W. Lockless, and R. Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. Science, 286(5438):295-9, 1999.

[65] Y. Low, J. Gonzalez, A. Kyrola, D. Bickson, C. Guestrin, and J. Hellerstein, Graphlab: A new framework for parallel machine learning, *Proc. 26th Conf. Uncertainty in Artificial Intelligence (UAI '10)*, 2010.

[66] L. Ma, and Q. Cui. Activation mechanism of a signaling protein at atomic resolution from advanced computations. *Journal of the American Chemical Society*, 129(33), 10261-10268, 2007.

[67] B. Ma, S. Kumar, C. J. Tsai, and R. Nussinov, Folding funnels and binding mechanisms, *Protein Eng.*, 12(713-720), 1999.

[68] B. Ma, S. Kumar, C. J. Tsai, H. Wolfson, N. Sinha, and R. Nussinov, Protein-ligand interactions: Induced fit, in *Encyclopedia of Life Sciences*, Chichester: John Wiley, 2002, doi:10.1038/npg.els.0003140.

[69] K. V. Mardia, C. C. Taylor, and G.K. Subramaniam, Protein bioinformatics and mixtures of bivariate von-Mises distributions for angular data, *Biometrics*, 63:505-512, 2007.

[70] K. V. Mardia, G. Hughes, C. C. Taylor, and H. Singh, A multivariate von-Mises distribution with applications to bioinformatics, *Can. J. Stat.*, 36(1), 99-109, 2008.

[71] K. V. Mardia, J. T. Kent, Z. Zhang, C. C. Taylor, and T. Hamelryck, Mixtures of concentrated multivariate sine distributions with application to bioinformatics, *J. Appl. Stat.*, 39(11), 2475-2492, 2012.

[72] C. Martin, V. Richard, M. Salem, R. Hartley, and Y. Mauguen, Refinement and structural analysis of Barnase at 1.5Å resolution, *Acta Crystallogr. D. Biol. Crystallogr.*, 55(2), 386-398, 1999.

[73] B. W. Matthews, Peripatetic Proteins, *Protein Sci.*, 19, 1279-1280, 2010.

[74] C. L. McClendon, G. Friedland, D. L. Mobley, H. Amirkhani, M. P. Jacobson. Quantifying correlations between allosteric sites in thermodynamic ensembles. *J. Chem. Theory Comput*, 5:24862502, 2009.

[75] B. A. McCray, E. Skordalakes, and J. P. Taylor. Disease mutations in Rab7 result in unregulated nucleotide exchange and inappropriate activation. *Human molecular genetics*, ddp567, 2009.

[76] L. R. McDonald, J. A. Boyer, and A. L. Lee. Segmental motions, not a two-state concerted switch, underlie allostery in CheY. *Structure*, 20(8), 1363-1373, 2012.

[77] N. Meinshausen and P. Bühlmann P. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34:1436-1462, 2006.

[78] M. V. Milburn MV, L. Tong, A. M. deVos, Z. Yamaizumi, S. Nishimura, and S. H. Kim. Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic Ras proteins. *Science*, 247:939945, 1990.

[79] A. Mittermaier and L. E. Kay, New tools provide new insights in NMR studies of protein dynamics, *Science*, 312:224-228, 2006.

[80] F. Morcos, A. Pagnani, L. Bryan, A. Bertolino, S. Marks, C. Sander, R. Zecchina, J. N. Onuchic, T. Hwa, and M. Weigt. Direct-coupling analysis of residue co-evolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. U.S.A.*, 108, E1293, 2011.

[81] K. Murphy, Y. Weiss, and M. Jordan, Loopy belief propagation for approximate inference: An empirical study, *Proc. 15th Conf. Uncertainty in Artificial Intelligence (UAI '99)*, 467-475, 1999.

[82] P. Ozbek, S. Soner, and T. Haliloglu. Hot spots in a network of functional sites. *PloS One*, 8(9), e74320.

[83] J. Pacheco, S. Zuffi, M. J. Black and E. B. Sudderth, Preserving modes and messages via diverse particle selection, *International Conference on Machine Learning (ICML)*, Jun. 2014.

[84] J. R. Partridge, and T. U. Schwartz, Crystallographic and biochemical analysis of the Ran-binding zinc finger domain. *Journal of molecular biology*, 391(2), 375-389, 2009.

[85] J. Pearl, Probabilistic reasoning in intelligent systems: Networks of plausible inference, San Francisco: Morgan Kaufmann, 1988.

[86] E. F. Pettersen, T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, T. E. Ferrin. UCSF Chimera–a visualization system for exploratory research and analysis. *J. Comput. Chem.*, 25(13):1605-12, 2004.

[87] F. Raimondi, G. Portella, M. Orozco, and F. Fanelli. Nucleotide binding switches the information flow in ras GTPases. *PLoS computational biology*, 7(3), e1001098, 2011.

[88] R. A. Reynolds, W. Watt, and K. D. Watenpaugh, Structures and comparison of the Y98H (2.0 Å) and Y98W (1.5 Å) mutants of Flavodoxin (Desulfovibrio Vulgaris), *Acta Crystallogr. D. Biol. Crystallogr.*, 57(4), 527-535, 2001.

[89] C. A. Rohl, C. E. Strauss, K. M. Misura, and D. Baker. Protein structure prediction using Rosetta. *Methods in enzymology*, 383, 66-93, 2004.

[90] A. Sali, L. Potterton, F. Yuan, H. van Vlijmen, M. and Karplus. Evaluation of comparative protein modelling by MODELLER. Proteins, 23, 318-326, 1995.

[91] M. A. Schumacher, D. Carter, D. M. Scott, D. S. Roos, B. Ullman, and R. G. Brennan. Crystal structures of Toxoplasmagondii Uracil Phosphoribosyltransferase reveal the atomic basis of pyrimidine discrimination and prodrug binding. *EMBO J.*, 17:3219-3232, 1998.

[92] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13(11):2498-504, 2003.

[93] K. Scheffzek, C. Klebe, K. Fritz-Wolf, W. Kabsch, and A. Wittinghofer. Crystal structure of the nuclear Ras-related protein Ran in its GDP-bound form, 1995.

[94] J. Shawe-Taylor, and N. Cristianini. Kernel methods for pattern analysis. Cambridge university press, 2004.

[95] H. Singh, V. Hnizdo, and E. Demchuk, Probabilistic model for two dependent circular variables, *Biometrika*, 89:719-723, 2002.

[96] L. Soltan Ghoraie, F. Burkowski, S. C. Li, and M. Zhu. Residue-specific side-chain polymorphisms via particle belief propagation. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 11(1), 33-41, 2014.

[97] L. Soltan Ghoraie, F. Burkowski, and M. Zhu. Sparse networks of directly coupled, polymorphic and functional side chains in allosteric proteins. *Proteins: Structure, Function, and Bioinformatics*, 2014.

[98] L. Soltan Ghoraie, F. Burkowski, and M. Zhu. Using kernelized partial canonical correlation analysis to study directly coupled side chains and allostery in small G proteins. Accepted to ISMB 2015 (to be published in Bioinformatics)

[99] M. Stein, M. Pilli, S. Bernauer, B. H. Habermann, M. Zerial, and R. C. Wade. The interaction properties of the human Rab GTPase family: A comparative analysis reveals determinants of molecular binding selectivity. *PloS One*, 7(4), e34870, 2012.

[100] M. Stewart, H. M. Kent, and A. J. McCoy. The structure of the Q69L mutant of GDP-Ran shows a major conformational change in the switch II loop that accounts for its failure to bind nuclear transport factor 2 (NTF2). *Journal of molecular biology*, 284(5), 1517-1527, 1998.

[101] G. M. Süel, S. W. Lockless, M. A. Wall, and R. Ranganathan. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. *Nat Struct Biol*, 10(1):59-69, 2003.

[102] E. Sudderth, A. Ihler, W. Freeman, and A. Wilsky, Nonparametric belief propagation, *Proc. 2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03)*, 605-612, 2003.

[103] Y. Takai, T. Sasaki, and T. Matozaki. Small GTP-binding proteins. *Physiological reviews*, 81(1), 153-208, 2001.

[104] A. E. Todd, C. A. Orengo, and J. M. Thornton. Evolution of function in protein superfamilies, from a structural perspective. *Journal of molecular biology*, 307(4), 1113-1143, 2001.

[105] C. Tsai, A. Del Sol, and R. Nussinov. Allostery: absence of a change in shape does not imply that allostery is not at play. *Journal of molecular biology*, 378(1):1-11, 2008.

[106] S. V. Vaerenbergh. Kernel Methods for Nonlinear Identification, Equalization and Separation of Signals. Thesis, Universidad de Cantabria, 2010.

[107] H. van den Bedem, G. Bhabha, K. Yang, P. E. Wright, and J. S. Fraser. Automated identification of functional dynamic contact networks from x-ray crystallography. *Nat. Methods*, 10:896-902, 2013.

[108] H. van den Bedem, A. Dhanik, J. C. Latombe, and A. M. Deacon, Modeling discrete heterogeneity in x-ray diffraction data by fitting multi-conformers, *Acta Cryst.*, D65:1107-1117, 2009.

[109] M. Vendruscolo, Determination of conformationally heterogeneous states of proteins, *Curr. Opin. Struct. Biol.*, 17:15-20, 2007.

[110] K. Volz and P. Matsumura. Crystal structure of Escherichia coli CheY refined at 1.7 resolution, *J Biol Chem*, 266:1551115519, 1991.

[111] J. Wang, M. Dauter, R. Alkire, A. Joachimiak, and Z. Dauter, Triclinic Lysozyme at 0.65 Angstrom resolution, *Acta Crystallogr. D. Biol. Crystallogr.*, 63(12), 1254-1268, 2007.

[112] M. Weigt, R. A. White, H. Szurmant, J. A. Hoch, and T. Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci*, 106:67-72, 2009.

[113] K. Wennerberg, K. L. Rossman, and C. J. Der. The Ras superfamily at a glance. *Journal of cell science*, 118(5), 843-846, 2005.

[114] M. A. Wilson and A. T. Brunger, The 1.0Å crystal structure of Ca^{2+}-bound Calmodulin: An analysis of disorder and implications for functionally relevant plasticity, *J. Mol. Biol.*, 301(5), 1237-1256, 2000.

[115] K. B. Wong and V. Daggett. Barstar has a highly dynamic hydrophobic core: evidence from molecular dynamics simulations and nuclear magnetic resonance relaxation data. *Biochemistry*, 37:11182-11192, 1998.

[116] J. Xu. Rapid protein side-chain packing via tree decomposition. *In Research in computational molecular biology*, 423-439. Springer Berlin Heidelberg, 2005.

[117] J. Xu and B. Berger, Fast and accurate algorithms for protein side-chain packing, *J. ACM.*, 53(4), 533-557, 2006.

[118] C. Yanover and Y. Weiss, Approximate inference and protein folding, *Advances in Neural Information Processing Systems (NIPS '02)*, MIT Press, 2002.

[119] C. Yanover, T. Meltzer, and Y. Weiss, Linear programming relaxations and belief propagation: An empirical study, *J. Mach. Learn. Res.*, 7:1887-1907, 2006.

[120] C. Yanover and Y. Weiss, Approximate inference and side-chain prediction, Technical Report, School of Computer Science and Engineering, Hebrew University of Jerusalen, 2007.

[121] Y. Yu, S. Li, X. Xu, Y. Li, K. Guan, E. Arnold, and J. Ding. Structural basis for the unique biological function of small GTPase RHEB. *Journal of Biological Chemistry*, 280(17), 17093-17100, 2005.

[122] M. Yuan and Y. Lin. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19-35, 2007.