Bioinformatics, 31, 2015, i124–i132
doi: 10.1093/bioinformatics/btv241
ISMB/ECCB 2015

Using kernelized partial canonical correlation analysis to study directly coupled side chains and allostery in small G proteins

Laleh Soltan Ghoraie[1,*], Forbes Burkowski[1] and Mu Zhu[2]

[1]Department of Computer Science and [2]Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON, Canada

*To whom correspondence should be addressed.

Abstract

Motivation: Inferring structural dependencies among a protein's side chains helps us understand their coupled motions. It is known that coupled fluctuations can reveal pathways of communication used for information propagation in a molecule. Side-chain conformations are commonly represented by multivariate angular variables, but existing partial correlation methods that can be applied to this inference task are not capable of handling multivariate angular data. We propose a novel method to infer direct couplings from this type of data, and show that this method is useful for identifying functional regions and their interactions in allosteric proteins.

Results: We developed a novel extension of canonical correlation analysis (CCA), which we call 'kernelized partial CCA' (or simply KPCCA), and used it to infer direct couplings between side chains, while disentangling these couplings from indirect ones. Using the conformational information and fluctuations of the inactive structure alone for allosteric proteins in the Ras and other Ras-like families, our 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. Our results were in good agreement with other empirical findings. By studying distinct members of the Ras, Rho and Rab sub-families, we show further that KPCCA was capable of inferring common allosteric characteristics in the small G protein super-family.

Availability and implementation: https://github.com/lsgh/ismb15

Contact: lsoltang@uwaterloo.ca

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 (DuBay *et al.*, 2011), but also emphasized the presence of allostery in proteins with minimal backbone motions (Tsai *et al.*, 2008). Moreover, recent discoveries by X-ray crystallography reveal that alternate side-chain conformations are more prevalent than previously thought (Lang *et al.*, 2010; van den Bedem *et al.*, 2009). These findings further accentuate the importance of a thorough study of the role played by side-chains in allostery (DuBay *et al.*, 2011; van den Bedem *et al.*, 2013).

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 (Morcos *et al.*, 2011). 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 graphical LASSO (GLASSO; Friedman *et al.*, 2008), on the other hand, often assume that the data are generated from a multivariate normal distribution—although Jones *et al.* (2012) also applied it to binary variables. This is a restrictive assumption for many applications such as the one discussed in this paper. 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 how to quantify correlations between two multidimensional random variables. Here, a useful statistical tool is canonical correlation analysis (CCA; Hotelling, 1936). We developed a novel extension of CCA by incorporating *partial* correlation and by using a multivariate von-Mises kernel function (Mardia *et al.*, 2008) 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 (Takai *et al.*, 2001). 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 (Raimondi *et al.*, 2011; Takai *et al.*, 2001; Wennerberg *et al.*, 2005)—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 Fig. 1) in the literature (Grizot *et al.*, 2001; Milburn *et al.*, 1990; Scheffzek *et al.*, 1995).

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, not both structures are readily available. We think our method can provide especially valuable information in these types of situations.

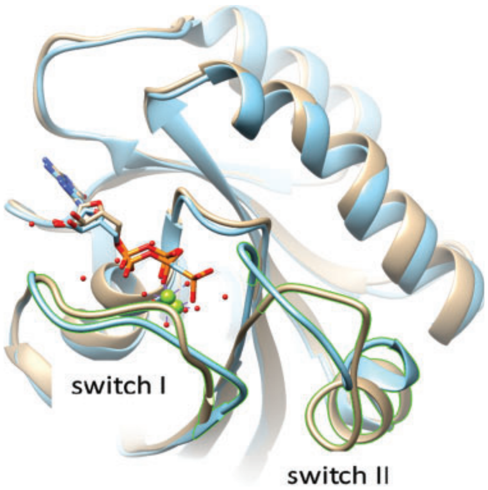

Fig. 1. Superimposed 3D structures of active and inactive H-Ras. Major conformational changes are known to occur in the Switch I and Switch II regions during an allosteric event

2 Methods

Our method for inferring direct couplings comprises a few fundamental components. First, we rely on the mathematical notion of the partial correlation to measure *direct* couplings (Section 2.1). Second, we use CCA to quantify the notion of correlation (more specifically, partial correlation) for *multivariate* data (Section 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 (Sections 2.4, 2.5.2).

2.1 Direct coupling and partial correlation

If variable x is correlated with a set of variables $\mathbf{z} = (z_1, z_2, \ldots, z_d)^{\mathrm{T}}$ and so is y, a transitive correlation requires that x be also 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 (Jones *et al.*, 2012; Morcos *et al.*, 2011).

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

$$x = \beta_0 + \beta_1 z_1 + \ldots + \beta_d z_d + \varepsilon_x,$$
$$y = \gamma_0 + \gamma_1 z_1 + \ldots + \gamma_d z_d + \varepsilon_y.$$

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

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

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

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. $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ and $\mathbf{y} = (y_1, \ldots, y_q)^{\mathrm{T}}$ for some $p \neq q$?

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

$$\rho(\mathbf{x}, \mathbf{y}) \equiv \max_{\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q} \mathbb{C}\mathrm{orr}(\mathbf{u}^{\mathrm{T}}\mathbf{x}, \mathbf{v}^{\mathrm{T}}\mathbf{y}), \quad (1)$$

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

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

the maximization problem in Equation (1) is equivalent to

$$\max_{\mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q} \mathbf{u}^{\mathrm{T}}[\mathbb{C}\mathrm{ov}(\mathbf{x}, \mathbf{y})]\mathbf{v}, \quad (2)$$

subject to the constraints

$$\boldsymbol{u}^{\mathrm{T}}[\mathbb{V}\mathrm{ar}(\boldsymbol{x})]\boldsymbol{u} = 1 \qquad \text{and} \qquad \boldsymbol{v}^{\mathrm{T}}[\mathbb{V}\mathrm{ar}(\boldsymbol{y})]\boldsymbol{v} = 1. \tag{3}$$

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

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

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

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

Hence, the empirical estimate of the canonical correlation coefficient given in Equation (1) can be obtained by solving the maximization problem (2)–(3) using the three sample estimates above, i.e.

$$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}) = \max_{\substack{\boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{u}=1 \\ \boldsymbol{v}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}=1}} \boldsymbol{u}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{v}. \tag{4}$$

The maximization problem in Equation (4) is well-known to be a generalized eigenvalue problem (see, e.g. Shawe-Taylor *et al.*, 2004), and can be solved in many scientific computing platforms, including MATLAB.

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 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) using

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

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

$$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z}) = \max_{\substack{\boldsymbol{u}^{\mathrm{T}}\check{\boldsymbol{X}}^{\mathrm{T}}\check{\boldsymbol{X}}\boldsymbol{u}=1 \\ \boldsymbol{v}^{\mathrm{T}}\check{\boldsymbol{Y}}^{\mathrm{T}}\check{\boldsymbol{Y}}\boldsymbol{v}=1}} \boldsymbol{u}^{\mathrm{T}}\check{\boldsymbol{X}}^{\mathrm{T}}\check{\boldsymbol{Y}}\boldsymbol{v}, \tag{5}$$

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

2.4 Kernelization of CCA and partial 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 $\boldsymbol{\alpha}, \boldsymbol{\theta} \in \mathbb{R}^{\mathrm{n}}$, the sample canonical correlation coefficient [Equation (4)] can be computed as

$$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}) = \max_{\substack{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\alpha}=1 \\ \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\theta}=1}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{\theta}. \tag{6}$$

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

$$\begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}} \\ \vdots \\ \boldsymbol{x}_n^{\mathrm{T}} \end{bmatrix} [\boldsymbol{x}_1 \quad \ldots \quad \boldsymbol{x}_n] = \begin{bmatrix} \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_2 & \ldots & \boldsymbol{x}_1^{\mathrm{T}}\boldsymbol{x}_n \\ \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_2 & \ldots & \boldsymbol{x}_2^{\mathrm{T}}\boldsymbol{x}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{x}_1 & \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{x}_2 & \ldots & \boldsymbol{x}_n^{\mathrm{T}}\boldsymbol{x}_n \end{bmatrix}$$

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, Equation (6) can be written as

$$\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}) = \max_{\substack{\boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_X^2\boldsymbol{\alpha}=1 \\ \boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{K}_Y^2\boldsymbol{\theta}=1}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{K}_X\boldsymbol{K}_Y\boldsymbol{\theta}. \tag{7}$$

This shows that CCA can easily be ‘kernelized’ (see, e.g. Shawe-Taylor *et al.*, 2004)—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 Equation (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 Equation (7). Clearly, the same argument applies to sample estimate of the partial canonical correlation coefficient [Equation (5)] as well, that is, the quantity $\widehat{\rho}(\boldsymbol{x}, \boldsymbol{y}|\boldsymbol{z})$ can be obtained from Equation (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 1 summarizes our notations.

A technical detail, which we largely have suppressed in our exposition here, is that, in Equation (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.

2.5 Application of KPCCA to the study of allostery

In this article, 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 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

Table 1. Summary of notations

Notation	Meaning	Section
$\hat{\rho}(\boldsymbol{x},\boldsymbol{y})$	CCA	2.2
$\hat{\rho}(\boldsymbol{x},\boldsymbol{y}\vert\boldsymbol{z})$	Partial CCA	2.3
$\hat{\rho}_K(\boldsymbol{x},\boldsymbol{y})$	Kernelized CCA	2.4
$\hat{\rho}_K(\boldsymbol{x},\boldsymbol{y}\vert\boldsymbol{z})$	Kernelized Partial CCA	2.4

- 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.

2.5.1 Generation of a conformational ensemble

As mentioned earlier (Section 1), a recent study has highlighted the role of side-chain fluctuations alone in information propagation within 'natively-folded' proteins (DuBay *et al.*, 2011). To conduct such a study, one requires 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, while the backbone is held fixed. Commonly used methods for generating such datasets include Monte Carlo (MC) 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 a response from the system (molecule) accordingly. The simulation techniques approximate the final stabilized structure that would be a candidate member for the ensemble of conformations.

We applied a different (and much more efficient) 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 (Krivov *et al.*, 2009) and TreePack (Xu, 2005). Although this approach is different from the commonly used simulation methods, it follows the same principles. For each given protein of m residues, we produced an ensemble consisting of $n \equiv [m(m-1)]/2$ structures, as follows. First, we randomly selected 20% of the side chains and set each of their conformations to a randomly chosen rotamer from the backbone-dependent rotamer library provided by Dunbrack *et al.* (1993). This step introduced fluctuations to the protein's conformation. Then, the rest of the side chains were packed by SCWRL or TreePack, and the final structure was added to the ensemble. Essentially, this amounted 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. This step simulated the response of the system to the fluctuations introduced in the first step. We believe 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).

2.5.2 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 2.4),

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j) = w_i w_j \prod_{t=1}^{p} \exp\left[\kappa_t \cos(x_{it}-x_{jt})\right], \tag{8}$$

and likewise for $\boldsymbol{y}_i, \boldsymbol{y}_j \in \mathbb{R}^q$. This is based on the multivariate von-Mises distribution (Mardia *et al.*, 2008) 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

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

for $t = 1,2,\ldots,p$, and $Z(\boldsymbol{\kappa},\Lambda)$ is a normalizing constant. The parameter $\boldsymbol{\mu} \in \mathbb{R}^p$ describes the location, i.e. the mean (or center), and the parameter $\boldsymbol{\kappa} \in \mathbb{R}^p$ ($\boldsymbol{\kappa} > \boldsymbol{0}$) describes the scale, i.e. the spread (or concentration). The parameter, $\Lambda = [\lambda_{st}] \in \mathbb{R}^{p\times p}$ 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 (e.g. Mardia *et al.*, 2007, 2012) as an appropriate tool for modelling angular variables that describe residue conformations in proteins.

We also introduced weights w_i, w_j in our kernel function [Equation (8)]. These weights were set to be inversely proportional to the energies of the two structures, i and j, in our ensemble (Section 2.5.1). 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.

2.5.3 Choice of $\boldsymbol{\kappa}_\mathbf{t}$

The kernel function (8) contains p concentration parameters, $\kappa_1,\ldots,\kappa_p$, 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-chains dihedral angles. For example, the first two dihedral angles are known to undergo more restricted motions, while the third and fourth 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 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

$$\begin{aligned}\exp\left[\kappa_t\cos(x_{it}-x_{jt})\right] &\approx \exp\left[\kappa_t - \frac{\kappa_t(x_{it}-x_{jt})^2}{2}\right] \\ &= (e^{\kappa_t})\exp\left[-\frac{(x_{it}-x_{jt})^2}{2/\kappa_t}\right].\end{aligned} \tag{9}$$

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 Equation (9) is equivalent to a Gaussian (or radial basis) kernel with 'standard deviation unit'

$$\sigma_t = \sqrt{\frac{1}{\kappa_t}} \qquad \text{or} \qquad \sigma_t = \sqrt{\frac{1}{\kappa_t}} \times \frac{360^\circ}{2\pi}. \tag{10}$$

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°

Table 2. Conversion Between κ_t And σ_t [Equation (10)]

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

and 40° in the corresponding Gaussian kernel (see Table 2). A side-chain prediction is often deemed successful if the predicted dihedral angle is within 40° of the true angle (Krivov *et al*., 2009). 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.

Table 3. Allosteric proteins from three sub-families of the small G protein super-family with PDB (Berman *et al.*, 2000) IDs of active and inactive structures*

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

*Active structure: bound to GTP. Inactive structure: bound to GDP.

3 Results and discussion

We tested our method on a number of well-studied Ras and Ras-like proteins (see Table 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; Daily *et al*., 2008) and by the GLASSO (Soltan Ghoraie *et al*., 2015). 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 (Daily *et al*., 2008). The GLASSO is a relatively new statistical method, which we have used in an earlier study 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 3.2). We implemented the KPCCA in MATLAB using the Kernel Methods Toolbox (Vaerenbergh, 2010).

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 against the CRN findings (see Table 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 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 (Burkowski, 2014) for the software, Chimera (Pettersen *et al*., 2004). 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 (Daily *et al*., 2008) 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 (Fig. 3), 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 (≈ 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 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 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. Figs 4, 5 and more discussions in Section 3.2).

3.1 Small G proteins

The members of this super-family are structurally categorized into five sub-families: Ras (Section 3.1.1), Rho (Section 3.1.2), Rab (Section 3.1.3), 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 (Raimondi *et al*., 2011), and that they share a common topology in the GDP/GTP binding domain (Takai *et al*., 2001). In this section, we highlight our findings for a few representative and well-studied members of these sub-families.

3.1.1 Ras sub-family

Members from the Ras subfamily are the primary members of the super-family; they play a critical role in human oncogenesis (Wennerberg *et al*., 2005). When activated, they regulate cell proliferation and survival through gene expression (Takai *et al*., 2001;

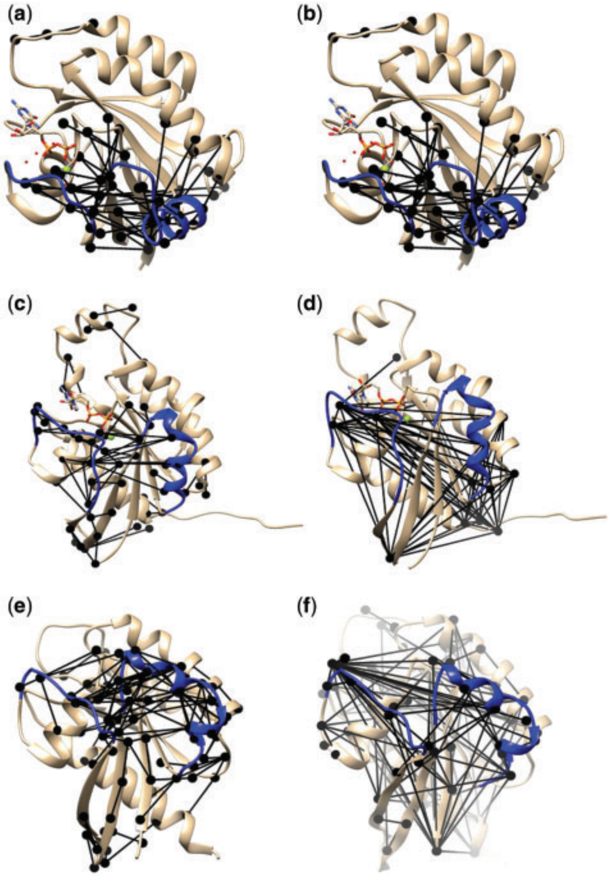

Fig. 2. 3D interaction graphs of H-Ras (top), Rac1 (middle) and Sec4 (bottom), showing the top 75, 63 and 112 couplings, respectively. These cut-offs are chosen to be equal to the number of couplings identified by the CRN. The couplings are mostly seen in the Switch I and Switch II regions (blue) and the β-sheets close by Left: CRN; Right: KPCCA

Wennerberg *et al.*, 2005). We experimented with three members of this sub-family: H-Ras, Rap2A and Rheb (see Table 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 (Milburn *et al.*, 1990) as Switch I (residues 30–38) and Switch II (residues 60–73); see Figure 1. Switch II is known to be directly involved in switching the protein from inactive to active status (Kidd *et al.*, 2009). Residues residing in the binding site are residues 28–35, 12–19, 145–147 and 116–120. Using the Rosetta software for structural prediction (Rohl *et al.*, 2004), Kidd *et al.* (2009) 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 (Fig. 2a). The interaction graph based on the top 75 couplings inferred by the KPCCA is shown in Figure 2(b); it clearly shows that strong couplings connected 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) in the 2D interaction graph [Fig. 3(a), based on the top 100 couplings] is associated with residue 34 in the Switch I region.

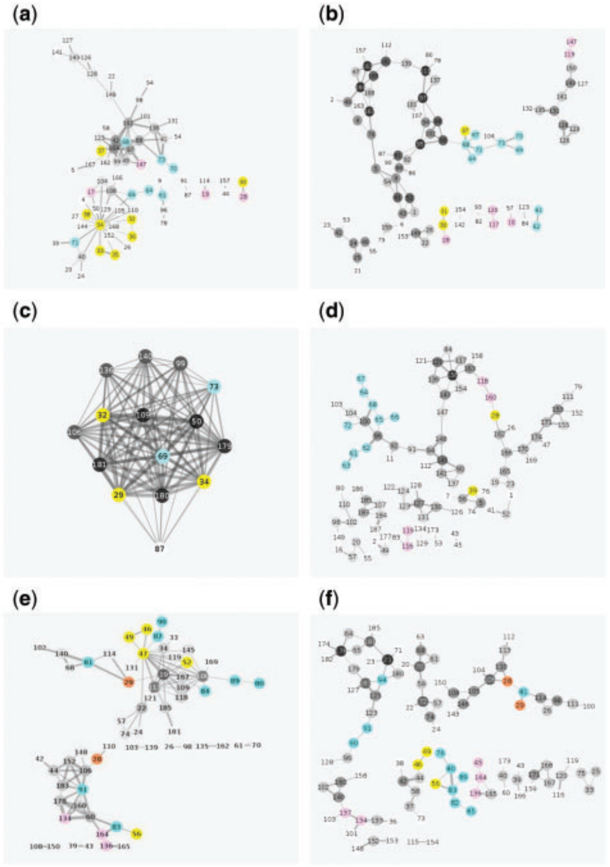

Fig. 3. 2D interaction graphs of H-Ras (top), Rac1 (middle) and Sec4 (bottom), all showing the top 100 couplings. For nodes shaded in grey, darkness is proportional to node degree. For edges, thickness is proportional to coupling strength. Residues in Switch I are coloured yellow and those in Switch 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 (left) produced more connected networks whereas, for the GLASSO (right), the inferred couplings are more 'spread out' within the molecule

3.1.2 Rho sub-family

The best-studied members of this sub-family are RhoA, Rac1 (Grizot *et al.*, 2001) and Cdc42 (Table 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 also are involved as regulators in cell cycle progression (cell polarity, movement, shape, and so on) and gene expression (Wennerberg *et al.*, 2005). The three aforementioned members are known to be involved in very diverse cellular processes (Takai *et al.*, 2001; Wennerberg *et al.*, 2005). The CRN for Rac1 consisted of 63 couplings (Fig. 2c). The edges were mostly concentrated in the classic Switch regions of this super-family. Figure 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 3(c) shows a 2D interaction graph formed by the top 100 couplings from the KPCCA; we can see that it is a highly connected

Table 4. Statistical features of interaction graphs*

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

*Based on graphs formed by the top 100 inferred couplings.

and concentrated network consisting of a single connected component with only 15 residues (see also Table 4).

3.1.3 Rab sub-family

The Rab proteins constitute the largest sub-family in the small G protein super-family (Garcia-Saez *et al.*, 2006). They are involved in the regulation of intracellular vesicular trafficking, vesicle formation, budding and fusion (McCray *et al.*, 2009; Stein *et al.*, 2012; Wennerberg *et al.*, 2005). 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 (Sali *et al.*, 1995). 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 (Tsai *et al.*, 2008). The common amino acids shared between H-Ras and (Sec4, Ypt7p) are about (35, 32%), respectively. The CRN for Sec4 consisted of 112 couplings (Fig. 2e). The interaction graph based on top-ranked couplings by the KPCCA (Figs 2e and 3e) showed connections between the two Switches through the edges, (47,87) and (83,56).

3.1.4 Ran and Arf/Sar1 sub-family

The Ran proteins (Partridge *et al.*, 2009; Scheffzek *et al.*, 1995; Stewart *et al.*, 1998) 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 (Takai *et al.*, 2001; Wennerberg *et al.*, 2005), 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 4–14 Å. Hence, these proteins do not belong to the category characterized by 'minor backbone motions' (Tsai *et al.*, 2008) and we excluded them from the current study.

3.2 KPCCA and GLASSO

We have recently applied the GLASSO to infer direct couplings between side chains (Soltan Ghoraie *et al.*, 2015). The GLASSO is incapable of handling multivariate angular variables. Thus, for each structure in the ensemble (which we generated with the same procedure as what we explained in Section 2.5.1), the conformation of each side chain (i) is matched against a set of candidate *rotamers* (R_i) from the standard rotamer library (Dunbrack *et al.*, 1993), and encoded using a set of binary random variables, b_{ik}. More specifically, for each structure ℓ in the ensemble,

$$b_{ik}^{(\ell)} = \begin{cases} 1, & \text{if the conformation of residue } i \text{ in structure } \ell \\ & \text{is ``closest'' to rotamer } k; \\ 0, & \text{otherwise,} \end{cases}$$

for $k = 1, 2, \ldots, |R_i|$. Using these binary variables, a sample covariance matrix S can be computed as follows. For each residue pair, (i, j), we compute an $|R_i| \times |R_j|$ sub-covariance matrix, $S_{i,j}$, whose entries are

$$S_{ik,jt} \equiv \mathbb{C}\text{ov}(b_{ik}, b_{jt}) = \mathbb{E}(b_{ik} b_{jt}) - \mathbb{E}(b_{ik})\mathbb{E}(b_{jt}),$$

where each expectation $\mathbb{E}(\cdot)$ is estimated empirically by a weighted sample average, using weights inversely proportional to the energy of each structure (see also Section 2.5.2). Using S as the input, the GLASSO estimates a sparse inverse covariance matrix, Θ. After removing the entropic bias (more on this below), entries in each submatrix $\Theta_{i,j}$ are aggregated to form a coupling score for the residue pair, (i, j).

3.2.1 Comparison

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

i. It 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. It identifies allosteric regions more effectively. Quantitatively, Table 3 showed that the KPCCA results agreed well with the CRN ones, 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 Figure 3, which contains 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 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 Fig. 4). 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 5 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.

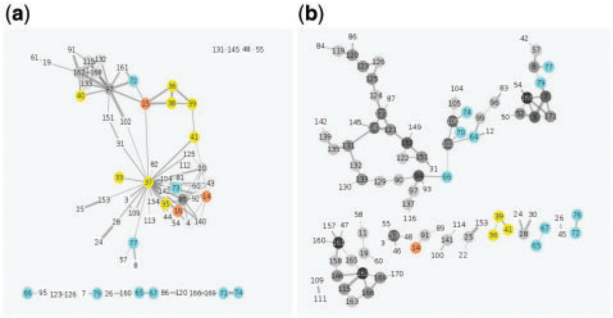

Fig. 4. 2D interaction graphs for Rheb, using the top 100 couplings. Residues in Switch I (II) are coloured yellow (blue). (**a**) KPCCA: The two Switch regions are directly connected (residue 37 from Switch I with residues 73 and 77 from Switch II); moreover, Switch I is indirectly connected to Switch II by residue 15 in the phosphate-binding loop (Yu *et al.*, 2005). (**b**) GLASSO: No connection between the Switches is identified

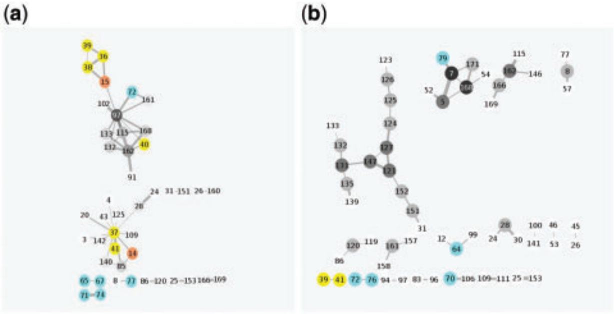

Fig. 5. 2D interaction graphs for Rheb, using the top 47 couplings (the same number as identified by the CRN). Residues in Switch I (33–41), Switch II (63–79) and the phosphate-binding loop (p-loop) are coloured yellow, blue and orange, respectively. The p-loop residues are connected to the Switches in the CRN. Crucial couplings emerge at higher positions in the ranked list of the KPCCA (**a**) than in that of the GLASSO (**b**)

iii. Its 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 Fig. 3). 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 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. It 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.*, 2012; Dunn *et al.*, 2008, who also suggested techniques for its correction). By contrast, the KPCCA does not appear to suffer from such biases. Figure 6 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.

4 Conclusion

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

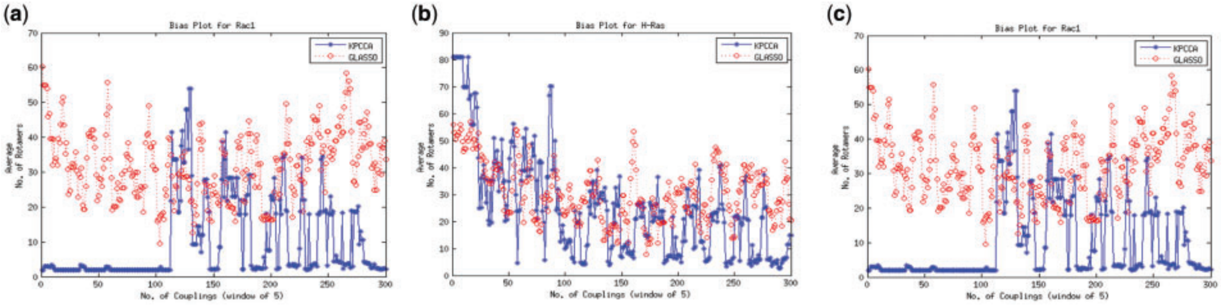

Fig. 6. X-axis: Rank order of the inferred couplings ($x = 1, 2, \ldots, 300$). Y-axis: Average number of available rotamers for residues involved in couplings ranked at positions $x, x+1, \ldots, x+4$ for (**a**) H-Ras (**b**) Rac1 (**c**) Sec4. Red: GLASSO (with bias correction). Blue: KPCCA (without 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

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 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 Kaufman *et al.* (2010), 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 the bioinformatics, e.g. for revealing the 'hot spot' residues in protein–protein interactions by using only the fluctuation information of the 'unbound' protein (Ozbek *et al.*, 2013).

Funding

This research is partially supported by the Natural Sciences and Engineering Research Council (NSERC) of Canada.

Conflict of Interest: none declared.

References

Berman,H.M. *et al.* (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.

Burkowski,F. (2014) *Computational and Visualization Techniques for Structural Bioinformatics Using Chimera*. Chapman and Hall/CRC, UK.

Daily,M.D. *et al.* (2008) Contact rearrangements form coupled networks from local motions in allosteric proteins. *Proteins Struct. Funct. Bioinf.*, **71**, 455–466.

DuBay,K.H. *et al.* (2011) Long-range intra-protein communication can be transmitted by correlated side-chain fluctuations alone. *PLoS Comput. Biol.*, **7**, e1002168.

Dunbrack,R.L. and Karplus,M. (1993) Backbone-dependent rotamer library for proteins application to side-chain prediction. *J. Mol. Biol.*, **230**, 543–574.

Dunn,S.D. *et al.* (2008) Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, **24**, 333–340.

Friedman,J. *et al.* (2008) Sparse inverse covariance estimation with the graphical Lasso. *Biostatistics*, **9**, 432–441.

Garcia-Saez,I. *et al.* (2006) The structure of human neuronal Rab6B in the active and inactive form. *Acta Crystall. Sect. D*, **62**, 725–733.

Grizot,S. *et al.* (2001) Crystal structure of the Rac1-RhoGDI complex involved in nadph oxidase activation. *Biochemistry*, **40**, 10007–10013.

Hotelling,H. (1936) Relations between two sets of variates. *Biometrika* **28**, 321–377.

Jones,D.T. *et al.* (2012) PSICOV: precise structural contact prediction using sparse inverse covariance estimation on large multiple sequence alignments. *Bioinformatics*, **28**, 184–190.

Kidd,B.A. *et al.* (2009) Computation of conformational coupling in allosteric proteins *PLoS Comput. Biol.*, **5**, e1000484.

Krivov,G.G. *et al.* (2009) Improved prediction of protein side-chain conformations with SCWRL4. *Proteins Struct. Funct. Bioinf.*, **77**, 778–795.

Kaufmann,K.W. *et al.* (2010) Practically useful: What the Rosetta protein modeling suite can do for you. *Biochemistry*, **49**, 2987–2998.

Lang,P.T. *et al.* (2010) Automated electron-density sampling reveals widespread conformational polymorphism in proteins. *Protein Sci.*, **19**, 1420–1431.

Mardia,K.V. *et al.* (2007) Protein bioinformatics and mixtures of bivariate von Mises distributions for angular data. *Biometrics*, **63**, 505–512.

Mardia,K.V. *et al.* (2008) A multivariate von Mises distribution with applications to bioinformatics. *Can. J. Stat.*, **36**, 99–109.

Mardia,K.V. *et al.* (2012) Mixtures of concentrated multivariate sine distributions with applications to bioinformatics. *J. Appl. Stat.*, **39**, 2475–2492.

McCray,B.A. *et al.* (2009) Disease mutations in Rab7 result in unregulated nucleotide exchange and inappropriate activation. *Hum. Mol. Genet.*, **19**, 1033–1047.

Milburn,M.V. *et al.* (1990) Molecular switch for signal transduction: structural differences between active and inactive forms of protooncogenic ras proteins. *Science*, **247**, 939–945.

Morcos,F. *et al.* (2011) Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci.*, **108**, E1293–E1301.

Ozbek,P. *et al.* (2013) Hot spots in a network of functional sites. *PloS One*, **8**, e74320.

Partridge,J.R. and Schwartz,T.U. (2009) Crystallographic and biochemical analysis of the Ran-binding zinc finger domain. *J. Mol. Biol.*, **391**, 375–389.

Pettersen,E.F. *et al.* (2004) UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.*, **25**, 1605–1612.

Raimondi,F. *et al.* (2011) Nucleotide binding switches the information flow in ras GTPases. *PLoS Comput. Biol.*, **7**, e1001098.

Rohl,C.A. *et al.* (2004) Protein structure prediction using Rosetta. *Methods Enzymol.*, **383**, 66–93.

Sali,A. *et al.* (1995) Evaluation of comparative protein modelling by MODELLER. *Proteins*, **23**, 318–326.

Shannon,P. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.

Scheffzek,K. *et al.* (1995) Crystal structure of the nuclear Ras-related protein Ran in its GDP-bound form. *Nature*, **374**, 378–381.

Shawe-Taylor,J. and Cristianini,N. (2004) *Kernel Methods for Pattern Analysis*. Cambridge University Press, UK.

Soltan Ghoraie,L. *et al.* (2015) Sparse networks of directly coupled, polymorphic and functional side chains in allosteric proteins. *Proteins Struct. Funct. Bioinf.*, **83**, 497–516.

Stein,M. *et al.* (2012) The interaction properties of the human Rab GTPase family: a comparative analysis reveals determinants of molecular binding selectivity. *PloS One*, **7**, e34870.

Stewart,M. *et al.* (1998) 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). *J. Mol. Biol.*, **284**, 1517–1527.

Takai,Y. *et al.* (2001) Small GTP-binding proteins. *Physiol. Rev.*, **81**,153–208.

Tsai,C.J. *et al.* (2008) Allostery: absence of a change in shape does not imply that allostery is not at play. *J. Mol. Biol.*, **378**, 1–11.

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

Van Den Bedem,H. *et al.* (2013) Automated identification of functional dynamic contact networks from X-ray crystallography. *Nat. Methods*, **10**, 896–902.

Van Den Bedem,H. *et al.* (2009) Modeling discrete heterogeneity in X-ray diffraction data by fitting multi-conformers. *Acta Crystall. Section D*, **65**, 1107–1117.

Wennerberg,K. *et al.* (2005) The Ras superfamily at a glance. *J. Cell Sci.*, **118**, 843–846.

Xu,J. (2005) Rapid protein side-chain packing via tree decomposition. In: *Research in Computational Molecular Biology*, Springer, Berlin, pp. 423–439.

Yu,Y. *et al.* (2005) Structural basis for the unique biological function of small GTPase RHEB. *J. Biol. Chem.*, **280**, 17093–17100.