Supporting information for: Detection of long-range concerted motions in protein by a distance covariance

Amitava Roy* and Carol Beth Post

Medicinal Chemistry and Molecular Pharmacology, Purdue University, West Lafayette, USA

E-mail: amitroy@purdue.edu

1 Generalized correlation coefficient and canonical correlation

Let $\{\mathbf{A}\}$ and $\{\mathbf{B}\}$ be two d-dimensional Gaussian random vector series with zero mean. The total covariance matrix

$$\mathbf{C} = \begin{bmatrix} \mathbf{C_{AA}} & \mathbf{C_{AB}} \\ \mathbf{C_{BA}} & \mathbf{C_{BB}} \end{bmatrix} \tag{1}$$

is a block matrix where $\mathbf{C_{AA}}$ and $\mathbf{C_{BB}}$ are within-vector covariance matrices of $\{\mathbf{A}\}$ and $\{\mathbf{B}\}$ respectively and $\mathbf{C_{AB}} = \mathbf{C_{BA}^{\dagger}}$ is between-vector covariance matrix.

The mutual information (MI) between $\mathbf{A}$ and $\mathbf{B}$ is[1]

$$MI(\mathbf{A},\mathbf{B}) = H(\mathbf{A}) + H(\mathbf{B}) - H(\mathbf{A},\mathbf{B}) \tag{2}$$

*To whom correspondence should be addressed

where H is entropy and for a Gaussian distribution the entropy can be written as[1]

$$\begin{aligned} H(\mathbf{A}) &= \frac{1}{2}\ln(2\pi e)^d \parallel \mathbf{C_{AA}} \parallel \\ H(\mathbf{B}) &= \frac{1}{2}\ln(2\pi e)^d \parallel \mathbf{C_{BB}} \parallel \\ H(\mathbf{A},\mathbf{B}) &= \frac{1}{2}\ln(2\pi e)^{2d} \parallel \mathbf{C} \parallel \end{aligned} \tag{3}$$

with $\parallel . \parallel$ denoting the determinant. From Eq. (3) the MI between **A** and **B** is

$$MI(\mathbf{A},\mathbf{B}) = -\frac{1}{2}\ln\left(\frac{\parallel \mathbf{C} \parallel}{\parallel \mathbf{C_{AA}} \parallel \parallel \mathbf{C_{BB}} \parallel}\right) \tag{4}$$

The generalized correlation coefficient, GCC,[2] of **A** and **B** is then

$$\begin{aligned} GCC(\mathbf{A},\mathbf{B}) &= \sqrt{1-e^{\frac{-2MI}{d}}} \\ &= \sqrt{1-\left(\frac{\parallel \mathbf{C} \parallel}{\parallel \mathbf{C_{AA}} \parallel \parallel \mathbf{C_{BB}} \parallel}\right)^{\frac{1}{d}}} \end{aligned} \tag{5}$$

One can write

$$\begin{aligned} &\frac{\parallel \mathbf{C} \parallel}{\parallel \mathbf{C_{AA}} \parallel \parallel \mathbf{C_{BB}} \parallel} \\ &= \frac{\parallel \mathbf{C_{AA}} \parallel \parallel \mathbf{C_{BB}} - \mathbf{C_{BA}}\mathbf{C_{AA}}^{-1}\mathbf{C_{AB}} \parallel}{\parallel \mathbf{C_{AA}} \parallel \parallel \mathbf{C_{BB}} \parallel} \\ &= \parallel \mathbf{I} - \mathbf{C_{BB}}^{-1}\mathbf{C_{BA}}\mathbf{C_{AA}}^{-1}\mathbf{C_{AB}} \parallel \\ &= \prod_i (1-\lambda_i^2). \end{aligned}$$

$$\text{Hence} \quad GCC(\mathbf{A},\mathbf{B}) = \sqrt{1-\left(\prod_i (1-\lambda_i^2)\right)^{\frac{1}{d}}}, \tag{6}$$

where the λ_i^2 are eigenvalues of $\mathbf{C_{BB}}^{-1}\mathbf{C_{BA}}\mathbf{C_{AA}}^{-1}\mathbf{C_{AB}}$ and λ_i are called canonical correlations[3] between **A** and **B**. λ_i^2 have values between zero and one. See Ref.[4] for a short review on canonical correlations. The values λ_i^2's are also the eigenvalues of $\mathbf{C_{AA}}^{-1}\mathbf{C_{AB}}\mathbf{C_{BB}}^{-1}\mathbf{C_{BA}}$. The eigenvectors

of the former matrix are the basis vectors for **B** and of the later matrix the basis vectors for **A**. For reference if $\mathbf{C_{BB}}$, $\mathbf{C_{AA}}$ and $\mathbf{C_{AB}}$ are all diagonal then $\mathbf{C_{BB}}^{-1}\mathbf{C_{BA}}\mathbf{C_{AA}}^{-1}\mathbf{C_{AB}}$ becomes $\mathbf{R^2}$ matrix introduced in the main article. Let us assume λ_i^2 values are ordered, $\mathbf{V}_i^A$, $\mathbf{V}_i^B$ are the corresponding eigenvectors and λ_1^2 is the largest eigenvalue. Then $\mathbf{V}_1^A$ and $\mathbf{V}_1^B$ are linear combinations of $A_1,...,A_d$ and $B_1,...,B_d$ such that the correlation between them is maximum among all possible combinations of $A_1,...,A_d$ and $B_1,...,B_d$ and λ_1 is the Pearson's correlation coefficient (PCC) between them. Similarly $\mathbf{V}_2^A$ and $\mathbf{V}_2^B$ are linear combinations of $A_1,...,A_d$ and $B_1,...,B_d$, which have a PCC of zero with $\mathbf{V}_1^A$ and $\mathbf{V}_1^B$ respectively, with the second largest PCC λ_2 between them. While canonical correlations have a definite physical meaning, the value $\sqrt{1-\left(\prod_i(1-\lambda_i^2)\right)^{\frac{1}{d}}}$ is always dominated by the largest λ^2 and cannot be considered a proper correlation coefficient (CC) between position vectors. Consider the case where λ_1^2 is close to 1 while other λ^2's are close to zero. Then the product term will be close to zero and GCC will be close to 1. In such a case, although **A** and **B** are highly correlated in one direction, they are not correlated at all in the other $d-1$ directions, yet the GCC value is still 1.

For example, if we take the series $\{\mathbf{A}\}$ and $\{\mathbf{B}\}$ generated from the model in Figure 1 with $\Delta\theta = \pi/6$ we have

$$\mathbf{C_{AA}} = \begin{bmatrix} 18.4 & 17.3 \\ 17.3 & 18.4 \end{bmatrix}, \mathbf{C_{AB}} = \begin{bmatrix} 7.4 & 24.2 \\ 5.7 & 24.7 \end{bmatrix},$$
$$\mathbf{C_{BB}} = \begin{bmatrix} 5.0 & 12.6 \\ 12.6 & 48.5 \end{bmatrix}. \tag{7}$$

And we have $\lambda_1^2 = 0.90$ and $\lambda_2^2 = 0.68$, so that

$$\mathbf{V}_1^A = \begin{bmatrix} -0.64 \\ 0.77 \end{bmatrix}, \mathbf{V}_2^A = \begin{bmatrix} -0.77 \\ -0.64 \end{bmatrix},$$
$$\mathbf{V}_1^B = \begin{bmatrix} -0.91 \\ 0.41 \end{bmatrix}, \mathbf{V}_2^B = \begin{bmatrix} 0.68 \\ -0.41 \end{bmatrix}. \quad (8)$$

$\mathbf{V}_1^A$ and $\mathbf{V}_1^B$ are almost perpendicular to **A** and **B** respectively. So λ_1^2 is reflecting $R^2(B_\theta, A_\theta)$ which is 1 in the model in Figure 1. Similarly $\mathbf{V}_2^A$ and $\mathbf{V}_2^B$ are almost parallel to **A** and **B** respectively and λ_2^2 reflects $R^2(B_r, A_r)$ which is 0.69. And GCC of $(\mathbf{B}, \mathbf{A})$ becomes 0.91. GCC calculated by methods developed by Kraskov et al.[5] gave a value of 0.96 as plotted in Figure 1, which is not an accurate reflection of the correlation in $\{\mathbf{A}\}$ and $\{\mathbf{B}\}$ or B_r and A_r.

2 Correlation between components of vector

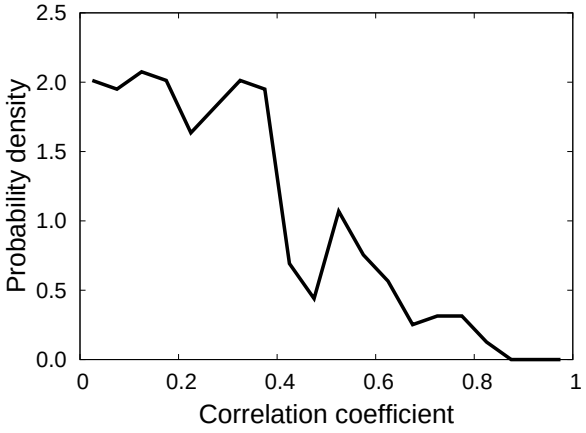

Figure S1: Pearson's correlation coefficient between $\hat{x}$, $\hat{y}$ and $\hat{z}$ of the position vectors of C_α atoms calculated from the molecular dynamics simulation of Src SH2 domain. About 40% of the calculated PCC values are more than 0.3 and about 10% are more than 0.6.

We calculated Pearson's correlation coefficient (PCC) between $\hat{x}$, $\hat{y}$ and $\hat{z}$ components of the position vectors of C_α atoms in the 80,000 conformations saved during the molecular dynamics simulation of Src SH2 domain described in the main article. Figure S1 shows the probability

density of these PCC values. About 40% of the calculated PCC values are more than 0.3 and about 10% are more than 0.6.

3 Comparison among correlation coefficients

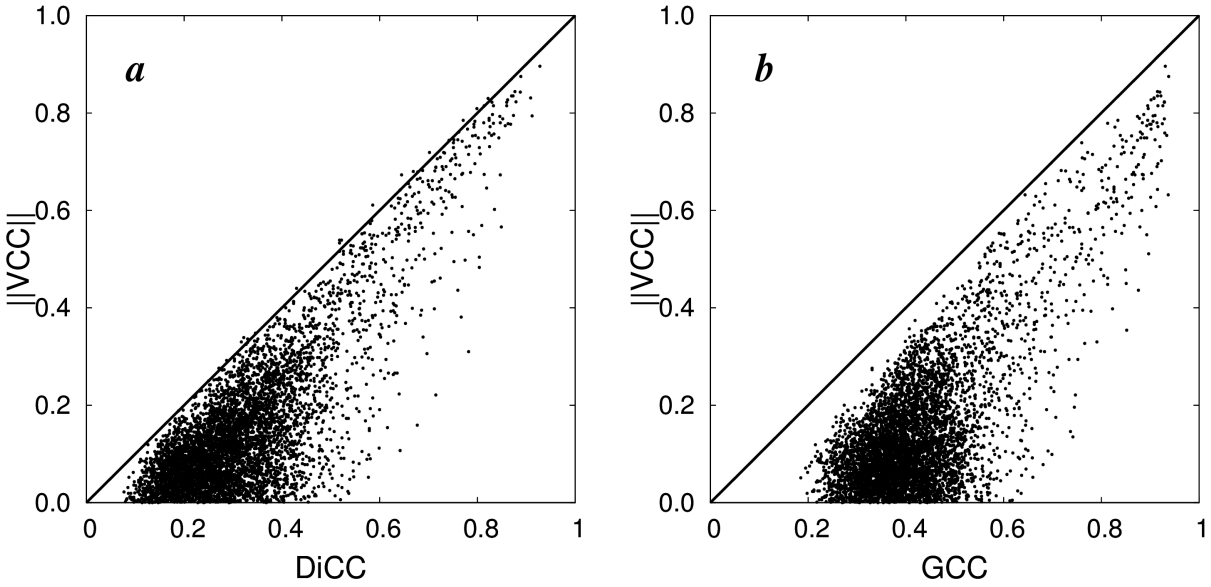

Figure S2: (**a**) DiCC and ||VCC|| of pairs of C_α atoms calculated from 40x2 ns long trajectory of Src SH2 domain in complex with the ligand pYEEI. (**b**) GCC and ||VCC|| of pairs of C_α atoms.

References

(1) Cover T.; Thomas J. Elements of Information Theory; Wiley & Sons: New York, 1991; p 230.

(2) Lange F.O.; Grubmüller H. Generalized Correlation for Biomolecular Dynamics. *Proteins* **2006,** *62,* 1053–1061.

(3) Hotelling H. The most predictable criterion. *J. Edu. Psych.* **1935,** *26,* 139–142.

(4) Kettenring J.R. Canonical analysis of several sets of variables. *Biometrika* **1971,** *58,* 433–451.

(5) Kraskov A.; Stogbauer H.; Grassberger P. Estimating mutual information. *Phys. Rev. E* **2004,** *69*, 066138.