Correlation, response and entropy approaches to allosteric behaviors: a critical comparison on the ubiquitin case

IOP Publishing Phys. Biol. 20 (2023) 056002 https://doi.org/10.1088/1478-3975/ace1c5

Physical Biology

OPEN ACCESS

RECEIVED
15 May 2023

REVISED
16 June 2023

ACCEPTED FOR PUBLICATION
26 June 2023

PUBLISHED
10 July 2023

PAPER

Correlation, response and entropy approaches to allosteric behaviors: a critical comparison on the ubiquitin case

Fabio Cecconi[1,2,*], **Giulio Costantini**[3], **Carlo Guardiani**[4], **Marco Baldovin**[3,5] and **Angelo Vulpiani**[6]

[1] CNR-Istituto dei Sistemi Complessi, Via dei Taurini 19, 00185 Rome, Italy
[2] INFN-Sezione di Roma1, P.le Aldo Moro, 2 00185 Rome, Italy
[3] CNR-Istituto dei Sistemi Complessi, Piazzale A. Moro 5, 00185 Rome, Italy
[4] Dipartimento di Ingegneria Meccanica e Aerospaziale, Sapienza Università di Roma, Via Eudossiana 18, 00184 Rome, Italy
[5] CNRS, LPTMS, Université Paris-Saclay, 530 Rue André Riviére, 91405 Orsay, France
[6] Dipartimento di Fisica, Università di Roma Sapienza, P.le Aldo Moro 5, 00185 Rome, Italy
[*] Author to whom any correspondence should be addressed.

E-mail: fabio.cecconi@cnr.it

Keywords: correlations, response, entropy, allosteric, allostery, ubiquitin

Abstract

Correlation analysis and its close variant principal component analysis are tools widely applied to predict the biological functions of macromolecules in terms of the relationship between fluctuation dynamics and structural properties. However, since this kind of analysis does not necessarily imply causation links among the elements of the system, its results run the risk of being biologically misinterpreted. By using as a benchmark the structure of ubiquitin, we report a critical comparison of correlation-based analysis with the analysis performed using two other indicators, response function and transfer entropy, that quantify the causal dependence. The use of ubiquitin stems from its simple structure and from recent experimental evidence of an allosteric control of its binding to target substrates. We discuss the ability of correlation, response and transfer-entropy analysis in detecting the role of the residues involved in the allosteric mechanism of ubiquitin as deduced by experiments. To maintain the comparison as much as free from the complexity of the modeling approach and the quality of time series, we describe the fluctuations of ubiquitin native state by the Gaussian network model which, being fully solvable, allows one to derive analytical expressions of the observables of interest. Our comparison suggests that a good strategy consists in combining correlation, response and transfer entropy, such that the preliminary information extracted from correlation analysis is validated by the two other indicators in order to discard those spurious correlations not associated with true causal dependencies.

1. Introduction

In proteins undergoing allosteric regulation [1], the binding of a ligand to the *regulatory site* affects the catalytic activity of the *active site*, generally placed at a distant location from the binding region [2]. The term 'allostery' was coined by Monod and Jacob [3] just to define such *long-range effects* activated across a molecule by the binding event to a specific site.

This sort of biological 'remote switching process' [4] is assumed possible as protein native states are not only structurally stable, but also susceptible enough to transfer signals among far away sites through long-range correlated fluctuations [5–8]. In practice, the release of the binding energy can trigger structural and/or dynamical changes in far regions of the biomolecule, thus allowing a fine control of the active site [9]. In this perspective, it is still unclear whether the coordinated motion of aminoacids in protein native states is a universal element to interpret such a wide phenomenology, and if allostery is a further manifestation of the structure-function relationship [10]. The theoretical methods generally applied to infer the structural origin of allostery in macromolecules are based on normal mode analysis (NMA) that can be performed either through full-atom simulations [11, 12] or within less expensive coarse-grained approaches, such as the elastic network models (ENM) [13]. Other approaches deal with allostery as a problem of information transport

across the network of interactions (contacts) defined by the topology of the native structure [14–16]. A typical scheme assigns Markov-chain like transition rules for exploring the network and then identifying allosteric paths connecting regulating and active sites, as the most probable ones [17]. This technique has been successfully applied to the study of the electro-mechanical coupling between voltage sensor domain and pore domain in voltage-gated potassium channels [18–20].

A popular tool to analyze the residue-residue coherent dynamics in a molecule with N residues is the equilibrium $3N \times 3N$-covariance matrix [10],

$$C_{ij} = \langle \Delta \mathbf{r}_i(t)\, \Delta \mathbf{r}_j^{\mathrm{T}}(t) \rangle \tag{1}$$

where $\Delta \mathbf{r}_i(t) = \mathbf{r}_i(t) - \mathbf{R}_i$ indicates the instantaneous displacement of residue i from its native position, $\mathbf{R}_i$, taken as equilibrium position. The direct study of the correlations of the free molecule (apo-structure) and their variations after the binding (holo-structure) can show how the docking of a ligand to the regulatory site produces observable changes to the dynamics of the target site [12].

Under a physical-like interpretation, allostery can be discussed in terms of the response of proteins and enzymes to a local perturbation generated by the binding of a ligand. The reaction to binding determines the release of the mechanical strain that converts into the transfer of signal at the molecular scale from the source to the target [21]. In other words, according to the *ensemble model* scenario [22, 23], allostery emerges from a modification of the native free-energy landscape of proteins under different 'effects': ligand binding, mechanic or chemical excitations, and environmental changes in pH or temperature.

More generally, allostery can also be associated with the notion of 'causation' in which regulatory and active sites are linked by a series of cause–effect pathways. There are two possible definitions of cause–effect relationship which correspond either to the *interventional view* or to the *observational view* of causation.

In the first, one directly performs local or structural modifications of a system and measures how they affect the behavior of specific system variables. This definition of causality was proposed by Pearl [24]. Conversely, the second consists in determining whether and to what extent the simple observation of certain variables is useful to predict the future of others, without manipulating the system.

The analysis of correlations, equation (1), belongs to the observational approach. However, 'correlations do not imply causation', as they measure only associations among variables without explaining their cause–effect relationship [25]. For instance, two residues i and j can move jointly not because they are in direct interaction, but because they are driven by a shared group of other residues [25–27].

As a consequence, it is reasonable to suspect that a mere investigation of allostery based on correlations only could overlook relevant biological information.

On the contrary, response theory seems not only the natural framework to understand how a static [28] or dynamic perturbation [29, 30] propagates from source to target site in proteins, but also the most reasonable approach to establish the causal influence between source and target. The detection of causation based on response theory recalls the interventional approach, according to which the coordinate x_j causally influences the coordinate x_i, if a perturbation of x_j results in a variation of the measured value of x_i. In formulae, we will say that x_i influences x_j, if

$$R_{ji}(t) = \lim_{\delta x_j(\tau)\to 0} \frac{\overline{\delta x_i(\tau + t)}}{\delta x_j(\tau)} \neq 0 \qquad \text{for some } t > \tau\,, \tag{2}$$

i.e. a small perturbation on $x_j(\tau)$ at time τ results in a non-zero future variation on the average of $x_i(t+\tau)$ over its unperturbed evolution. In equation (2), we assume steady dynamics. If δx_j is small enough, it is well known that the quantity (2) can be related to the spontaneous correlations in the unperturbed dynamics by one of the pillars of non-equilibrium statistical mechanics, the fluctuation–response theorem (FRT) [31], also known as fluctuation–dissipation theorem (FDT).

The analysis of allostery communication beyond correlations has been already suggested in the literature. For instance, some authors looked at the structural properties of molecules to predict their allosteric behaviours [32], others [33, 34] employed the transfer entropy (TE), a measure of causation borrowed from information-theory and introduced by Schreiber [35] and Paluš *et al* [36] in the context of stochastic processes and dynamical systems. The entropy transfer from the evolution of the coordinate $x_j(t)$ to the evolution of the coordinate $x_i(t)$ determines the information (uncertainty) that we gain (lose) on the future states of x_i, if we not only consider the past history of x_i, but we also include the past of x_j. It quantifies the causal influence of x_j on x_i. In formulae,

$$\begin{aligned} \mathrm{TE}_{j\to i}(t) = {} & H[x_i(\tau + t)|x_i(\tau)] \\ & - H[x_i(\tau + t)|x_i(\tau), x_j(\tau)], \end{aligned} \tag{3}$$

where $H(a|b)$ indicates the conditional Shannon entropy [37] of the state a given the state b. Equation (3) assumes stationary processes. Notice that TE is by definition asymmetric, thus naturally incorporating a direction of the entropy/information transfer from $x_j \to x_i$, that is generally different from $x_i \to x_j$.

For the sake of completeness, it should be mentioned that allosteric processes are often analyzed also in terms of pairwise *mutual information* and its high-order generalization called *interaction information* which are useful indicators of causal dependency

[38]. In particular Interaction Information is appropriate when the complexity of the allosteric process involves clusters of variables at the same time so it cannot be decomposed into elementary pairwise dependencies [39]. However, since these quantities are generally used in pure statistical formulation which does not take into account the temporal evolution, they will not be discussed here.

In this work, motivated by the importance of allosteric mechanisms, we compare the behavior of the three mentioned indicators: correlation, response, and TE when applied to the human ubiquitin (Ub), a simple and very well-studied protein, which, according to recent experiments [40], has shown regulation of allosteric nature.

Since each of the three indicators extracts specific features from the fluctuations of the residues, their combined use should provide a more robust view of those coordinated movements that play a possible role in allosteric mechanism.

Especially from the analysis of the response, we will try to identify sites of Ub that are more susceptible to perturbation, and understand if they are possibly involved in the known allosteric pathways of Ub. Following [29, 30], we employ the dynamic version of the indicators to exploit the information contained in their time dependence and in their propagation across the structure.

The exact comparison of these indicators in realistic all-atom simulations is severely limited by the system size and by the need of collecting very long time series (dimensional curse). To overcome this limitation, we describe the native state fluctuations of Ub via a coarse-grained approach, portraying the native state as a mechanical system made of nodes, representing the position of the α-carbons (C_α) of the aminoacids, connected by harmonic springs. This description is referred to as the Gaussian network model (GNM) introduced in [41]. The main advantage of using a Gaussian model relies on its full solvability, allowing the analytical expressions of correlations, responses, and TEs to be worked out.

In the literature, GNM and more generally coarse-grained elastic networks constitute simplified less-expensive alternatives to all-atom NMA, generally used for fast and easy characterization of slow and large-scale dynamics of native structures. They allow the identification of flexible and rigid regions in huge single and multi-domain proteins and are proven to be meaningful in the prediction of *functional modes* relevant for a comprehension of the structure-function relationship of biopolymers [42–45]. In addition, when the size of the system becomes inaccessible to all-atom NMA, the ENM represent the only viable approach.

It is important to recall that the GNM presents two crucial limitations. First of all, it does not apply to processes where a molecule changes its shape to visit other conformational states in order to perform its function. Therefore, the approach we follow here describes only 'allostery without conformational changes' proposed by Cooper and Dryden [22], and Ub falls into this category. Actually, there are generalizations of the ENM including also molecule transitions from one state to the other [46] by defining a GNM representation for each state and then assigning the transition rate among such states.

The second limitation of the standard GNM approach arises from neglecting the side chains, a common approximation of many coarse-graining methods. As we shall see in section 3, to partially relieve this crude approximation, we use a variant of GNM which is based on the 'heavy-atom contact map' (briefly heavy-map), that incorporates some effects of the side-chain presence.

Even if the small rearrangements in Ub allostery justify the use of the GNM-like approach improved by heavy-map representation, it should be noted that for allosteric processes occurring via transition among different metastable states, or involving not negligible nonlinear fluctuations [47], full-atomistic molecular dynamics, possibly supported by enhanced sampling techniques (e.g. conformational flooding [48]), remains the leading investigation tool.

The paper is organized as follows. In section 2 we briefly recall some structural properties of Ub and its allosteric control. section 3 reports the simplified theoretical framework for the description of Ub and the mathematical properties of the indicators (correlation, response, and entropy transfer) that we used to characterize the interplay among Ub fluctuation modes. Sections 4–6 contain the results obtained by the analysis based on these three indicators. In section 7, we show the results on the complex Ub-USP, ubiquitin and ubiquitin-specific peptidase (USP) to see how the Ub internal motion is modified by the interaction with one of its natural substrates. Final discussion and conclusions are drawn in section 8.

2. Allostery-like behavior of ubiquitin

In this section, we briefly summarize the principal biological information on Ub that has been used to orient our theoretical analysis. Post-translational modifications are covalent modifications altering the functional state of a protein; typical post-translational modifications involve the attachment of small chemical groups like acetyl, phosphate or methyl groups. Ubiquitylation, the attachment of Ub to its target substrates, can be considered an extreme case where the chemical group attached to the target protein is itself a small protein. Ubiquitylated proteins are normally targeted to degradation in the 26S proteasome, but ubiquitylation can also induce trafficking or endocytosis. Ub is bound to each target protein through

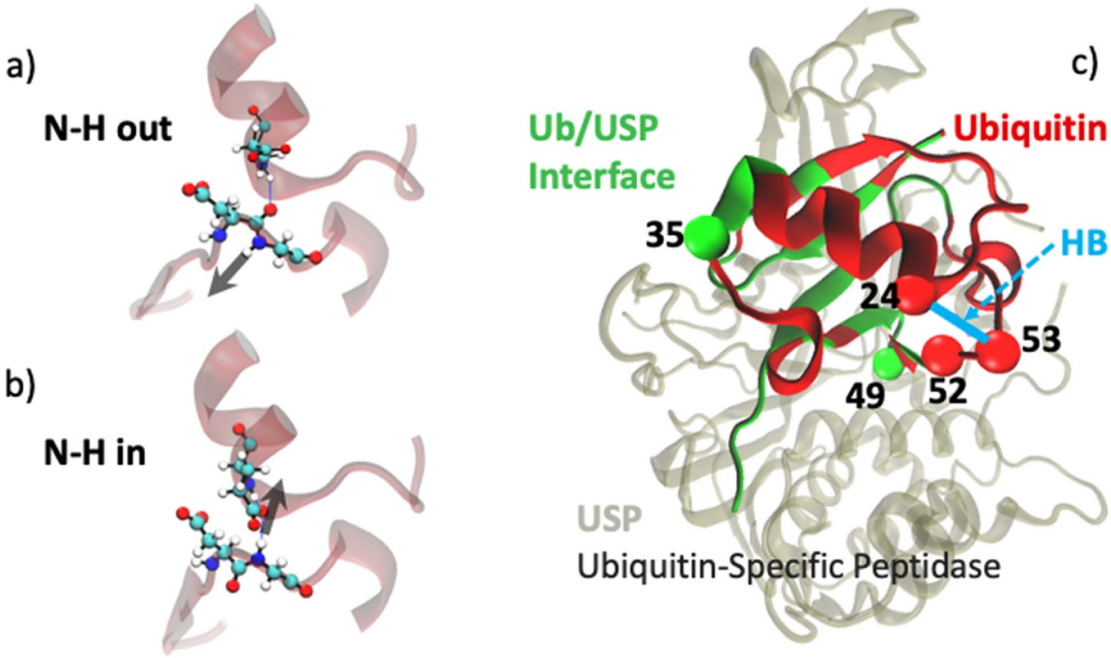

Figure 1. The conformational switch of peptide bond Asp52-Gly53. Panel (a) shows the NH-out conformation where the NH group of Gly53 points towards the solvent (direction shown by the arrow). Panel (b) shows the NH-in conformation where the NH group points towards the protein interior (as shown by the arrow). Panel (c) ubiquitin-ubiquitinase complex (PDB:2IBI). The USP is shown in the background, while Ub is displayed in red and green, the green region represents Ub binding interface with the USP (as defined through a distance cutoff of 3.0 Å). The C-terminal tail of Ub is also part of the binding interface with USP. The red spheres represent Ile23 and the Asp52-Gly53 peptide-bond, while the Glu24 and Gln49 are shown in green, and the blue rod indicates the hydrogen-bond Glu24-Gly53. Panel (c) shows that the conformational switch (Asp52-Gly53) and the Ub-USP interface are located on the opposite sides of Ub emphasizing the allosteric nature of the regulation mechanism.

the sequential action of three enzymes that ultimately connect the COO^- terminal group of Ub with the side chain of a Lysine residue of the target. Since Ub also comprises several Lysine residues, it can become the target of further ubiquitylation reactions that create a linear chain of Ub units bound to the target protein. The geometry and bonding pattern of these chains determines a different fate of the marked molecule.

A recent work by Smith *et al* [40], based on NMR relaxation dispersion experiments highlighted a conformational switch of peptide bond Asp52-Gly53 of allosteric significance. As sketched panels (a) and (b) of figure 1, showing portions of structures from PDB files 1UBI [49] and 2IBI respectively, this bond flips between two states referred to as *NH-out* (1UBI) and *NH-in* (2IBI).

In the NH-out state, the NH group of the peptide bond points towards the bulk where it forms hydrogen bonds with water molecules (black arrow). As a result, the only possible interaction with the neighboring Glu24 residue is a hydrogen bond between the CO group of the Asp52-Gly53 bond and the NH group of Glu24. The sidechain of Glu24 is therefore not involved in this interaction. By contrast, in the NH-in configuration the NH group of the peptide bond points towards the interior of the protein (black arrow), where it can H-bond the sidechain of Glu24, that is also hydrogen bonded by the NH group of Glu24 itself. To assess the functional significance of this conformational switch Smith *et al* performed a bioinformatics analysis on a database of Ub experimental structures. This analysis suggested a correlation between the NH-in conformation and the binding of the Ub to the ubiquitinase USP (ubiquitin-specific protease). Figure 1(c) shows the structure of the complex Ub-USP. This result was completely unexpected since neither Asp52 nor Gly53 is directly involved in Ub-USP binding. The finding thus led to the hypothesis that the switch of the Asp52-Gly53 bond might induce an allosteric rearrangement of the USP binding region of Ub. Indeed, further analysis showed that the NH-in and NH-out states are respectively associated to the contraction and expansion of the ubiquitinase binding interface. Moreover, it was shown that a contracted binding interface allows fewer steric clashes, energetically promoting the binding of USP. The mechanism can thus be summarized as follows: the NH-in state allosterically induces the contraction of the binding interface reducing the number of clashes and favoring the USP binding. The residues more affected by this interface deformation are Gly35 and Gln49. Interestingly, these results agree with an older work by Massi *et al* [50] that identified chemical exchange processes affecting Ile23, Asn25 (flanking Glu24), Thr55 (close to the Asp52-Gly53 bond) and Val70.

The allosteric regulation of Ub can be seen as a propagation of perturbation from the couple of aminoacids (Glu24,Gly53) that we consider as *source*

to the couple (Gly35,Gln49), that instead acts as *target*. In the following, for convenience, we shall refer to these sites as the allosteric set, ASet = {24,35,49,53}.

3. Model and methods

In a coarse-grained approach, the native state of the Ub is described as a mechanical system made of nodes, representing the position of the α-carbons (C_α) of the aminoacids, connected by harmonic springs. The potential energy of GNM is very simple and reads

$$V_{\mathrm{GNM}} = \frac{g}{2}\sum_{i,j} K_{ij}\,\Delta\mathbf{r}_i \cdot \Delta\mathbf{r}_j\,, \tag{4}$$

where $\{\Delta\mathbf{r}_1, \Delta\mathbf{r}_2, \ldots, \Delta\mathbf{r}_N\}$ are the instantaneous displacements of the N C_α atoms from their native positions taken as equilibrium states. The quantity g defines the adjustable energy scale that can be set by matching the theoretical mean square displacement of the C_α from their native positions with the experimental crystallographic B-factors [51–53]. The coefficients K_{ij} are the elements of the coupling matrix $\mathbb{K}$, often termed Kirchhoff matrix, which is defined through the contact matrix elements, A_{ij} (also connectivity matrix of the network) through the relation

$$K_{ij} = \begin{cases} -A_0 & |i-j| = 1 \\ -A_{ij} & |i-j| > 1 \\ \sum_{l=1,N} A_{il} & i = j\,, \end{cases} \tag{5}$$

where A_0 is a factor weighting the strength of the harmonic interaction along the chain (backbone) over the off-chain interaction. $A_0 \sim 10$ seems a reasonable value to distinguish the role of the backbone links from the rest of the network.

The effect of the side chains can be partially included by using a GNM approach based on a heavy-atom contact-map [54] that excludes hydrogens. In this scheme, a pair of residues $i-j$ is connected by a spring if, they have at least a couple of heavy atoms a, b in contact in the native state of Ub (PDB: 1 ubi). In formulae, this means that $A_{ij} = 1$ if the relation

$$\sum_{a,b} \Theta(r_c - |\mathbf{r}_{i,a} - \mathbf{r}_{j,b}|) \geqslant 1 \tag{6}$$

holds, with a cutoff $r_c = 5$ Å, where $\Theta(u)$ is the unitary step function.

Panel (a) of figure 2 shows the heavy contact-map representing the native interactions, panel (b) shows the topological diagram of Ub secondary structure that includes five beta-strands S1, ..., S5 and two helices H1, H2.

In the following, we will redefine $\Delta\mathbf{r}_i \to \mathbf{r}_i$ to simplify the notation.

The full NMA of Ub would require the computation of the Hessian matrix, obtained by computing the second partial derivatives of the force-field potential on the equilibrium state. The less complex Hessian of the GNM turns to be decomposed into blocks

$$\frac{\partial^2 V}{\partial r_i^\mu \partial r_j^\nu} = \delta_{\mu,\nu}\, K_{ij}$$

where $\mathbf{r}_i$ denotes the position of the ith C_α and μ and ν indicate the generic component x, y, z. In practice, the three coordinates of a C_α becomes formally equivalent $x_i \equiv y_i \equiv z_i$, hence the position vector $\mathbf{r}_i = x_i\,\mathbf{e}$ is virtually a scalar (with $\mathbf{e} = (1,1,1)$), and $\langle \mathbf{r}_i^2 \rangle = 3\langle x_i^2 \rangle$. Therefore GNM approach reduces a system of $3N$ degrees of freedom to a system in N degrees of freedom only; equivalently, it deals with protein fluctuations as a problem of *scalar elasticity*.

The equation of motion for each GNM coordinate in the overdamped regime reads

$$\gamma \dot{x}_i = -g\sum_j K_{ij} x_j + \sqrt{2\gamma k_{\mathrm{B}} T}\,\xi_i(t) \tag{7}$$

here, γ denotes the friction, k_{B} the Boltzmann constant and ξ_i is a zero-average and time delta-correlated Gaussian process. Hereafter, we set $\mu = g/\gamma$. From the solution (in vector form) of equation (7)

$$\mathbf{x}(t) = \mathrm{e}^{-\mu\mathbb{K}t}\left\{\mathbf{x}(0) + \sqrt{\frac{2k_{\mathrm{B}}T}{\gamma}}\int_0^t \mathrm{d}s\, \mathrm{e}^{-\mu\mathbb{K}s}\,\boldsymbol{\xi}(s),\right\} \tag{8}$$

correlation and response are straightforward to obtain as

$$\mathbb{C}(t) = \mathrm{e}^{-\mu\mathbb{K}t}\,\mathbb{C}(0) \tag{9}$$

$$\mathbb{R}(t) = \mathrm{e}^{-\mu\mathbb{K}t} \tag{10}$$

where $\mathbb{C}(0) = \langle \mathbf{x}(0)\,\mathbf{x}^{\mathrm{T}}(0) \rangle$ is the equal-time correlation matrix also termed equal-time covariance matrix and the average is over the thermal noise.

The advantage of using the GNM relies on its full solvability, since it defines a multivariate Onrstein–Ulehnbeck process (7) whose explicit solution only requires numerical diagonalization of the sparse matrix $\mathbb{K}$. $\mathbb{K}$ is symmetric and diagonalizable with all real not negative eigenvalues, but not invertible because it has one vanishing eigenvalue, due to the translation invariance of the system. It can be represented in the form

$$\mathbb{K} = \mathbb{U}\Lambda\mathbb{U}^\dagger \tag{11}$$

where Λ is the diagonal matrix containing all the eigenvalues $\{0, \lambda(2), \ldots, \lambda(N)\}$ of $\mathbb{K}$. $\lambda(k)$ is the kth eigenvalue of $\mathbb{K}$ associated to its kth eigenvector, $\mathbb{K}\mathbf{u}(k) = \lambda(k)\mathbf{u}(k)$. $\mathbb{U} = \{\mathbf{u}(1), \ldots, \mathbf{u}(N)\}$ is the diagonalizing matrix whose columns are the eigenvectors, $\mathbb{U}^\dagger$ is its transpose, moreover $\mathbb{U}\mathbb{U}^\dagger = \mathbb{I}$, $\mathbb{I}$ being the identity $N \times N$-matrix.

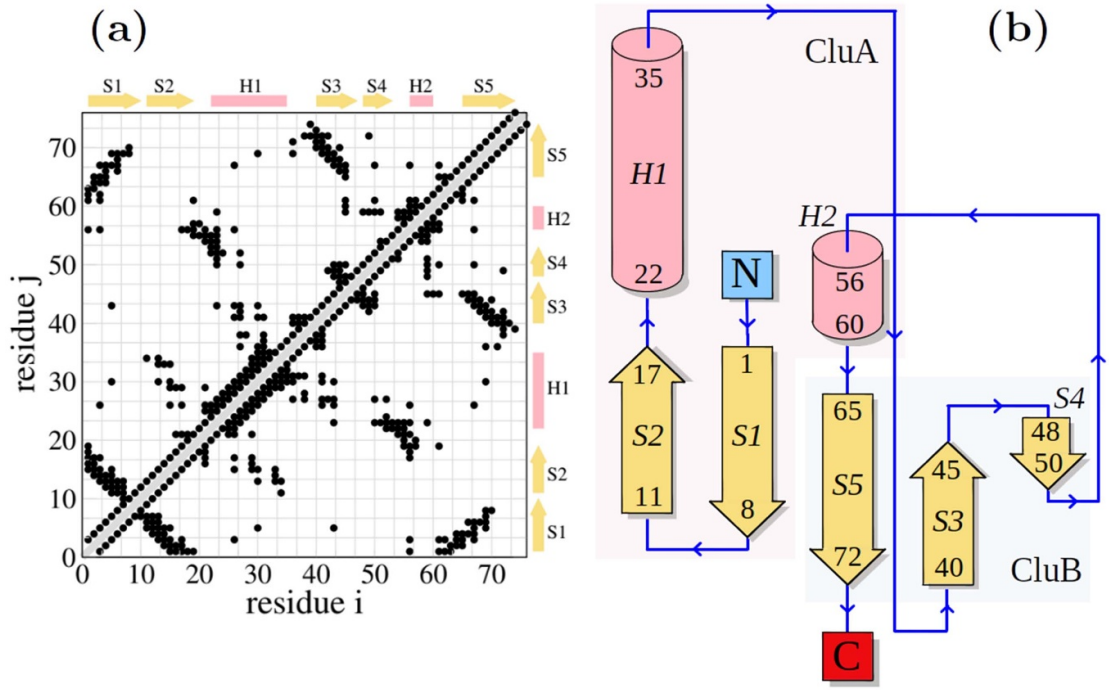

Figure 2. Topology structure of Ub. Panel (a): contact map defined by the heavy-atom contacts using a cutoff of $R_c = 5$ Å; the cutoff identifies 283 native contacts such that $|i-j| \geqslant 2$. Panel (b): topology plot of Ub showing the secondary structure content with five beta-sheets and two helices. Red and blue shadowed boxes indicate the separation of Ub in two clusters, CluA and CluB joined by the hinge residues Glu64 [55]. Reprinted figure with permission from [55] Copyright 2020 by the American Physical Society.

3.1 Correlation

The Gaussian nature of the GNM implies that the equilibrium covariance matrix $\mathbb{C}(0)$ is proportional to pseudo-inverse $\mathbb{K}^*$ of the matrix $\mathbb{K}$,

$$\mathbb{C}(0) = \frac{3k_B T}{g}\mathbb{K}^*.$$

$\mathbb{K}^*$ replaces the standard matrix inversion, as $\mathbb{K}$ is not invertible because of its vanishing determinant due to the translation invariance of the system. By definition $\mathbb{K}^* = \mathbb{U}\mathbb{D}\mathbb{U}^\dagger$, where the diagonal matrix D is $D_{ij} = \delta_{ij}/\lambda(i)$, if $\lambda(i) \neq 0$ and $D_{ii} = 0$, if $\lambda(i) = 0$.

According to equation (9), the explicit elements of the time dependent correlation matrix are

$$C_{ij}(t) = \frac{3k_B T}{g}\sum_{k=2}^{N}\frac{u_i(k)u_j(k)}{\lambda(k)}e^{-\mu\lambda(k)t}; \qquad (12)$$

the above sum excludes the $\lambda(1) = 0$ eigenvalue.

3.2 Response

The response definition equation (2) assumes a clear and general expression when the process $\mathbf{x}(t)$ is stationary with invariant probability density function (PdF) $P_s(\mathbf{x})$, [56]

$$R_{ij}(t) = -\left\langle x_j(t)\frac{\partial \ln P_s(\mathbf{x})}{\partial x_i}\Big|_{\mathbf{x}(0)}\right\rangle. \qquad (13)$$

Where the average $\langle\cdots\rangle$ is taken over the unperturbed system, whose invariant PdF, $P_s(\mathbf{x})$, is required to never vanish for any $\mathbf{x}$. $\mathbb{R}(t)$ is the matrix of the linear response functions (at time t) of the considered system. Equation (13) shows the existence of a rigorous link among responses and correlations, so that we can use such a relation to infer a degree of dependence from the observation of time series of $\mathbf{x}(t)$, without actually perturbing the system.

It should be remarked that equation (13) holds for systems with an invariant PdF and in general cases, it expresses the response in terms of complicated multivariate correlation functions. However, in systems governed by stochastic linear dynamics, such as equation (7), even with no Gaussian noise, the response turns out related only to the two-time correlation function [25, 57],

$$\mathbb{R}(t) = \mathbb{C}(t)\mathbb{C}^{-1}(0)\,. \qquad (14)$$

$\mathbb{C}^{-1}(0)$ indicates the inverse or of the correlations matrix at zero lag (i.e. the covariance matrix). Let us note that the translational invariance of the GNM does not allow to apply directly formula (13) to compute the response functions, as the invariant PdF, $P_s(\mathbf{x}) = P_s(\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$, cannot exist because is not normalizable. However, we can use equation (14), providing to employ the pseudoinverse of $\mathbb{C}(0)$.

The explicit expression of the response in equation (10) is obtained by the representation $\mathbb{R}(t) = \mathbb{U}e^{-\mu\mathbb{K}t}\mathbb{U}^T$, which in matrix elements reads

$$R_{ij}(t) = \frac{1}{N} + \sum_{k=2}^{N} u_i(k)u_j(k)e^{-\mu\lambda(k)t}. \qquad (15)$$

In equation (15), the term $1/N$ corresponds to the vanishing eigenvalue $\lambda(1) = 0$, indeed because of the translation invariance, the zero-mode is the uniform vector with components $u_i(1) = 1/\sqrt{N}$, for all i. As a consequence, $R_{ij}(\infty) = 1/N$, while, by definition, in the opposite limit $\mathbb{R}(0)$ coincides with the identity matrix, $R_{ij}(0) = \delta_{ij}$.

3.3. Transfer entropy

According to the definition (3) the explicit TE expression between residues $j \to i$, with $\tau = 0$, is the quantity

$$\mathrm{TE}_{j\to i}(t) = \left\langle \log \frac{P[x_i(t)|x_i(0), x_j(0)]}{P[x_i(t)|x_i(0)]} \right\rangle \tag{16}$$

where the angular-brackets indicate the average over the joint probability density $P[x_i(t), x_i(0), x_j(0)]$ and $P[x_i(t)|x_i(0), x_j(0)]$, $P[x_i(t)|x_i(0)]$ are the conditional probability densities of $x_i(t)$ conditioned to the previous values $x_i(0)$ and $x_j(0)$. Obviously, TE is not symmetric, $\mathrm{TE}_{j\to i} \neq \mathrm{TE}_{i\to j}$, and identically vanishes for $i = j$ as $P[x_i(t)|x_i(0), x_j(0)]$ coincides with $P[x_i(t)|x_i(0)]$. Notice that the asymmetry of TE stems from the presence of conditional probabilities, for which the conditioning and conditioned variables (or events) cannot be exchanged because they do not play equivalent roles.

For a Gaussian system like the GNM, there is a simple way to evaluate the TE among residues by using the correlations [58], equation (12),

$$T_{j\to i}(t) = -\frac{1}{2} \ln \left(1 - \frac{\alpha_{ij}(t)}{\beta_{ij}(t)} \right) \tag{17}$$

where

$$\alpha_{ij}(t) = [C_{ii}(0)C_{ij}(t) - C_{ij}(0)C_{ii}(t)]^2, \tag{18}$$

and

$$\beta_{ij}(t) = [C_{ii}(0)C_{jj}(0) - C_{ij}^2(0)][C_{ii}^2(0) - C_{ii}^2(t)], \tag{19}$$

see [58] and appendix for a sketch of the derivation. Formula (17) is a more compact notation of the expression reported in [33].

The asymmetry $\alpha_{ij}(t) \neq \alpha_{ji}(t)$ and $\beta_{ij}(t) \neq \beta_{ji}(t)$ is an immediate consequence of the TE asymmetry emerging also in the Gaussian formulation.

It is immediate to show, that $T_{j\to i}(\infty) = 0$, either by definition equation (16) invoking the independence of events far away in time, or using the correlation decay at large times in equation (17). Analogously, one expects $T_{j\to i}(0) = 0$.

In the following, we perform correlation analysis, response analysis, and transfer–entropy analysis of the GNM of the Ub, and when possible, we shall attempt a critical comparison among them.

4. Correlation analysis

Figure 3 reports the equilibrium residue–residue covariance matrix for Ub in Pearson's form, (i.e. normalized by the variances)

$$c_{ij} = \frac{C_{ij}(0)}{\sqrt{C_{ii}(0)\, C_{jj}(0)}} \tag{20}$$

using a temperature color code. The positive and negative correlations are plotted separately to facilitate the analysis. High values correspond to hotter and low values to darker colors. The lighter regions of the density plots show that positive correlations 'nucleate' around the native contacts, panel (a) of figure 3. In particular, the region around the residues Glu24 and Gly53 shows a correlation presumably driven by the native contacts in those segments of the protein. Also, the region around 40–70 results highly correlated and it involves the residues around the hydrophobic patch (HP) (Leu8-Ile44-Val70) [59]. This may suggest a sort of concerted motion among the elements of the HP associated with the binding to enzyme E1 [59]. Anti-correlated motion seems to involve mainly residues that are not in native contact and spatially distant, green and red regions of figure 3(b). Figure 4 reports the time behavior of pairwise correlations among residues of ASet that are considered interesting to the allosteric communication in Ub. Since these residues do not form any native contact with each other, for comparison, we also report the correlation between sites Ile23-Gly53, forming a native contact. The slowest and fastest autocorrelation decay (not shown), $C_{i,i}(t) \sim \exp(-t/\tau_i)$, allows us to select the interval of time, $0.065 \leqslant \tau_i \leqslant 0.726$, which can be considered the optimal time range for sampling the observables, because fluctuations are to be considered still active (see the vertical dashed lines in figure 4). In the following, we shall choose the sampling times $t = (0.20, 0.25, 0.30, 0.35)$ that are equally spaced in this interval.

In allosteric communication it is common to define *source* (allosteric site that binds the effector) and the *target* or active site. Similarly, we consider the allosteric regulation of Ub, as a propagation of perturbation from a couple of aminoacids (24, 53) (source) to the couple (35, 49) (target), which participate in the binding interface Ub-USP. Accordingly, figure 5 plots the correlation profiles, $C_{p,j}(t)$, from the sites in ASet to every site $j = 1, \ldots, N = 76$ of the Ub, at different times to monitor the evolution of the correlation spreading through the protein structure. The times at which the profiles are sampled are such that the correlations are still significantly different from zero.

The peaks of figure 5 indicate the sites that are mainly correlated, and obviously, the highest peak refers to autocorrelation $C_{p,p}(t)$ that decays as $C_{p,p}(t) \sim \exp(-t/\tau_p)$ at large time.

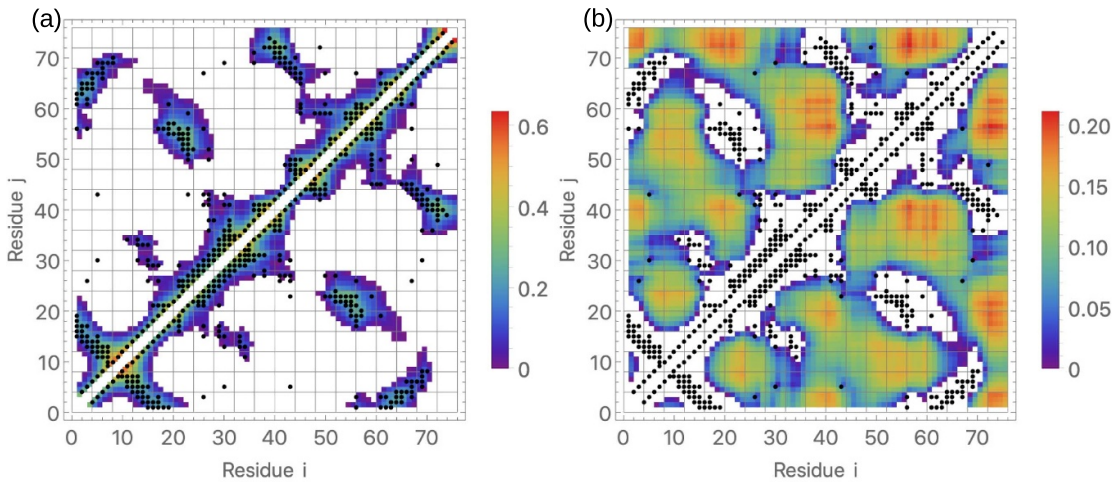

Figure 3. Density plot of equation (20). Panel (a) reports positive residue-residue correlations, black dots represent the native contacts for a direct comparison. Panel (b) shows the negative residue-residue correlations. This splitting should reveal correlated and anti-correlated residue motion. Generally, positive correlations emerge from native contacts. In particular, the region around 40–70 results in high correlations and corresponds to the residues around the HP Leu8-Ile44-Val70.

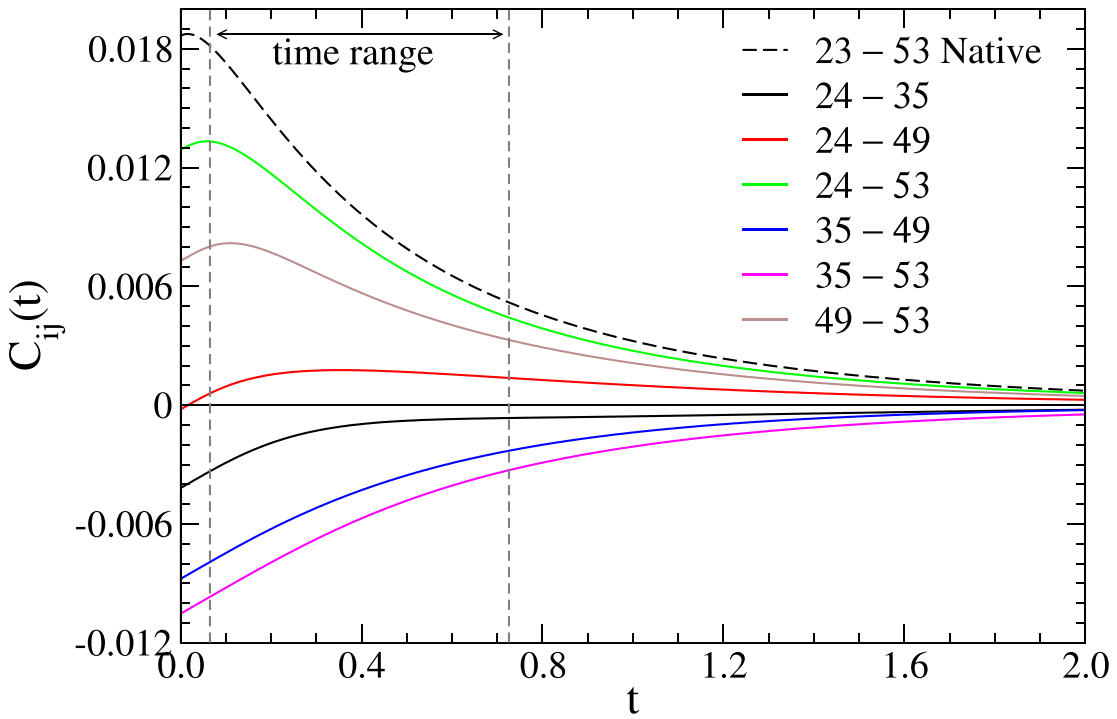

Figure 4. Instances of time behavior of correlations from equation (12) between residues that are considered relevant for the Ub allosteric properties. The correlation of the native contact 23–53 (dashed) is included for comparison. Vertical dashed lines bound the range defined by the slowest and fastest time of auto-correlation decay. In this range, we assume that fluctuations are still active.

In the chosen time interval, the global structure of the profile remains stable during the time evolution, including positive and negative correlations. For example from the top panel, a positive correlation Glu24-Gly53 is always established, whereas Glu24 is mainly negatively correlated with Thr9, Gln40, and the C-terminus tail 70–76.

5. Response analysis

The correlation analysis of the previous section identifies the presence of simultaneous coordinated displacements among residues, however, it says nothing about their effective dependencies. A better indicator for such a purpose is the response function. In

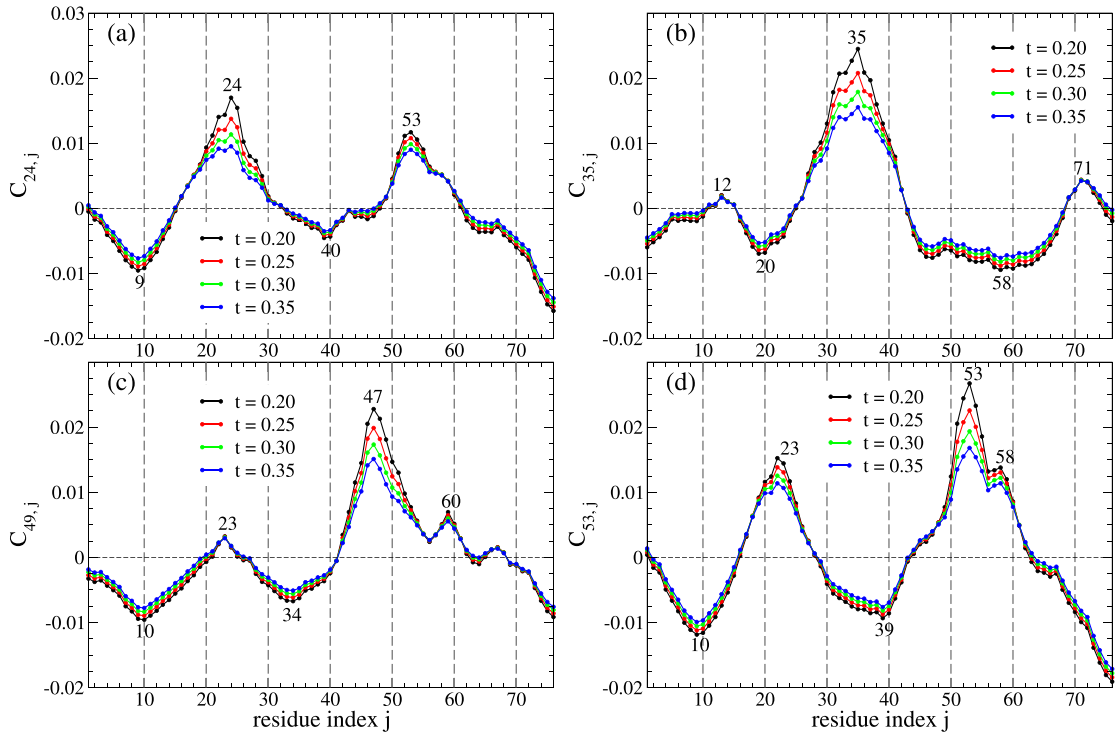

Figure 5. Correlation profiles $C_{p,j}(t)$ for sites indicated as connected allosterically by experiments [40], $p = \{24, 35, 49, 53\}$, to the other sites of the Ub. The peaks locate regions of high correlation.

analogy with correlation analysis, we plot in figure 6. the time-behavior of the responses $R_{ij}(t)$ among the usual set of residues of ASet. The threshold $1/N$ in equation (13) identifies two behaviors of response: a class of curves that grow monotonically in time to $1/N$ from below, e.g. Glu24-Gly35, Glu24-Gln49, Gly35-Gln49, and Gly35-Gly53, non-monotonic responses which may exhibit either a single or double peak. All the native responses show a common pattern with either a single pronounced peak or a less pronounced maximum at short times and a broader maximum at larger times. However, also a pair of sites not in contact can show a peaked response, as there could be a path of native contacts connecting them, but their peak is always smaller than the peak associated with a native pair. This is an obvious consequence of the 'causal nature' of the response which is more sensitive to direct interactions. We can say that the sites which can be really considered active from a response scenario are those with responses crossing the line $1/N$, reaching the maximum and decaying to $1/N$ from above.

Likewise to correlation analysis, to understand how a perturbation in a site 'p' spreads through the Ub structure, it is convenient to plot the profile $R_{p,j}(t) \equiv R_{j,p}$, for $j = 1, \ldots, N = 76$. These profiles for the residues ASet are shown in figure 7, at the same four times sampled in the correlation study (figure 5). A common feature of such profiles is the presence of a sort of 'response localization' generating peaks that survive at different times assessing the robustness of the scenario. Obviously, the most pronounced peak corresponds to the site 'p' where the perturbation is initially applied. This peak decays exponentially and spreads in time, since the perturbation naturally propagates along the backbone, covering a region of about ten sites centered on p, an immediate effect of the chain connectivity. To verify how the Ub backbone drives the response propagation, we turned off all the non-bonded interactions so that the Ub becomes a pure harmonic chain with nearest-neighbor interactions only. The gray-shadowed regions in figure 7 indicate the spreading of the perturbation along such a harmonic chain.

As already noticed, only the sites with peaks above the threshold $1/N$ (indicated by a dashed horizontal line) are considered to develop a significant response to the perturbation.

If the site p = Glu24, which is a source in Ub allosteric control, is perturbed, the maximum of the response is felt by the other source site Gly53, in agreement with NMR relaxation dispersion experiments [40].

With reference to the target sites Gly35 and Gln49, we can say that they do not have a reciprocal response action, however, the target Gln49 has a causal link with both sources Glu24 (actually site '23') and Gly53 which is included in the peak around Gln49.

Gln49, in turn, responds to perturbation of Gly53 either via the backbone or through the path Gln49-Tyr59-Gly53, in which only the link Gln49-Tyr59 is a native contact.

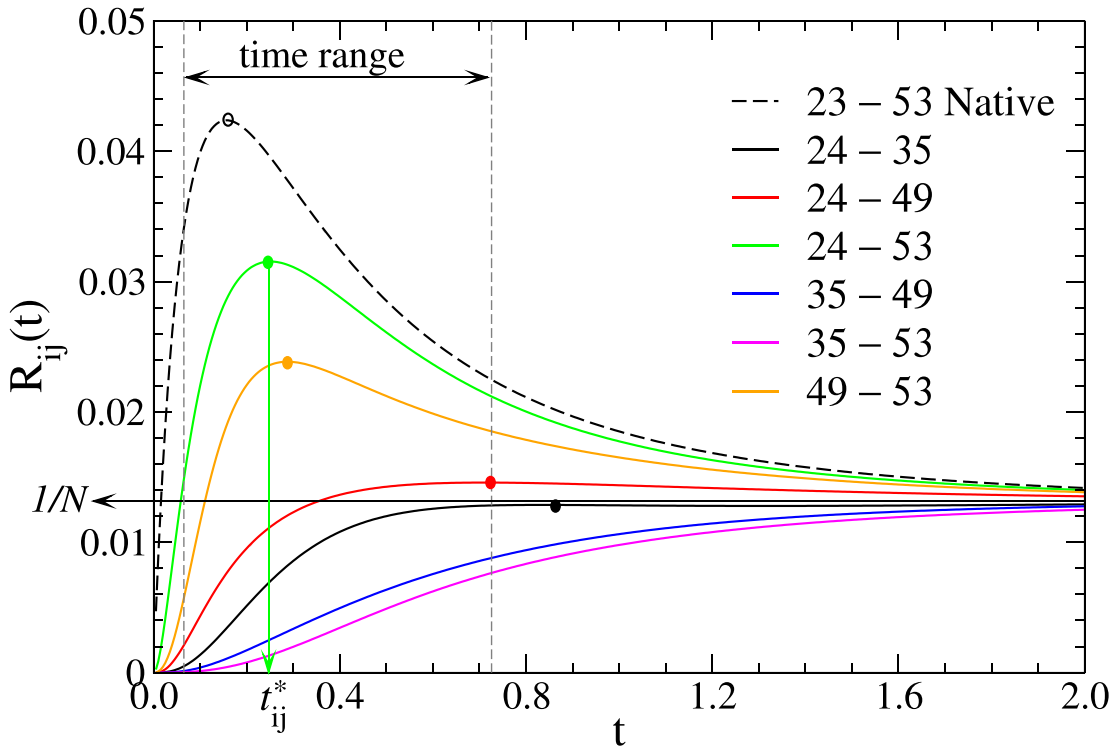

Figure 6. Time-behavior of the response functions $R_{ij}(t)$ for the sites involved in Ub allostery. The horizontal line with the arrow marks the threshold $1/N$ ($N = 76$) below which the responses are considered irrelevant. The response peaks (dots) select the times t^*_{ij} which allow the propagation of the perturbations to be traced, see figure 6. The response for the native contact Ile23-Gly53 is reported for comparison, to show that native responses are generally higher than responses between residues not in direct interaction: response is more sensitive to direct causation. Vertical dashed lines bound the range of interesting time, as describe in the caption of figure 4.

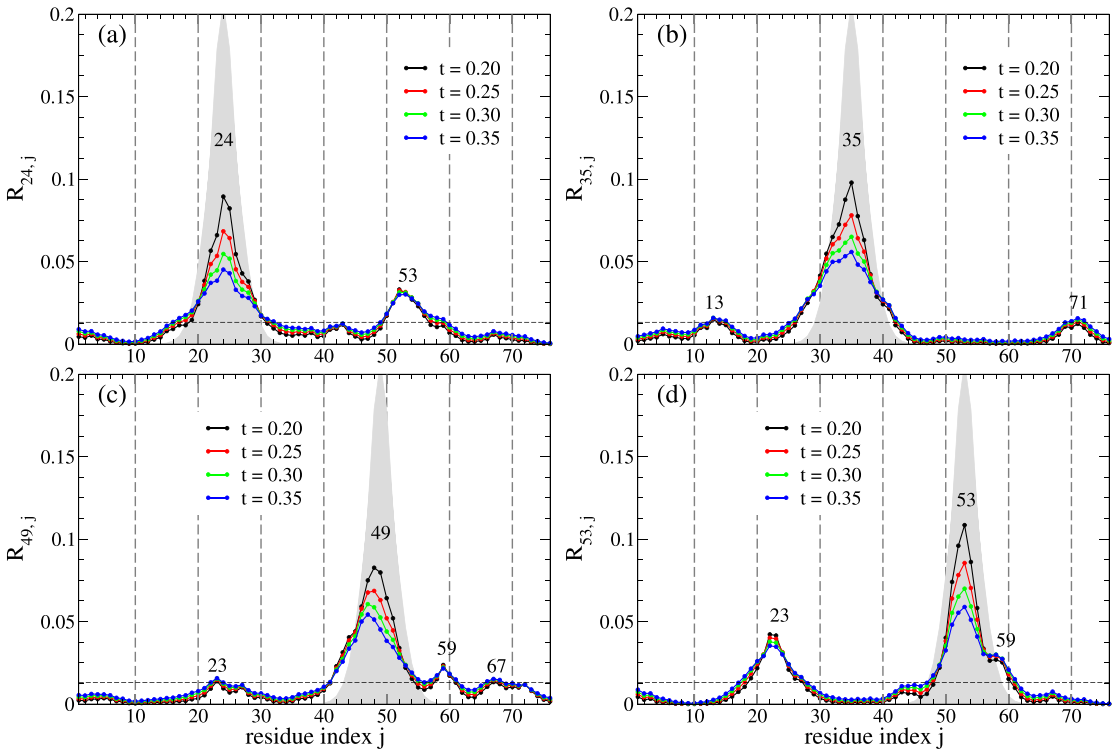

Figure 7. Profiles of response $R_{p,j}(t)$, for sites $p = \{24, 35, 49, 53\}$ at different times to assess the robustness of the scenario. In each panel, the dashed line marks the $1/N$ threshold below which the response is considered 'uneffective'. The gray-shadowed regions show the spreading of the perturbation along the chain representing the Ub harmonic backbone.

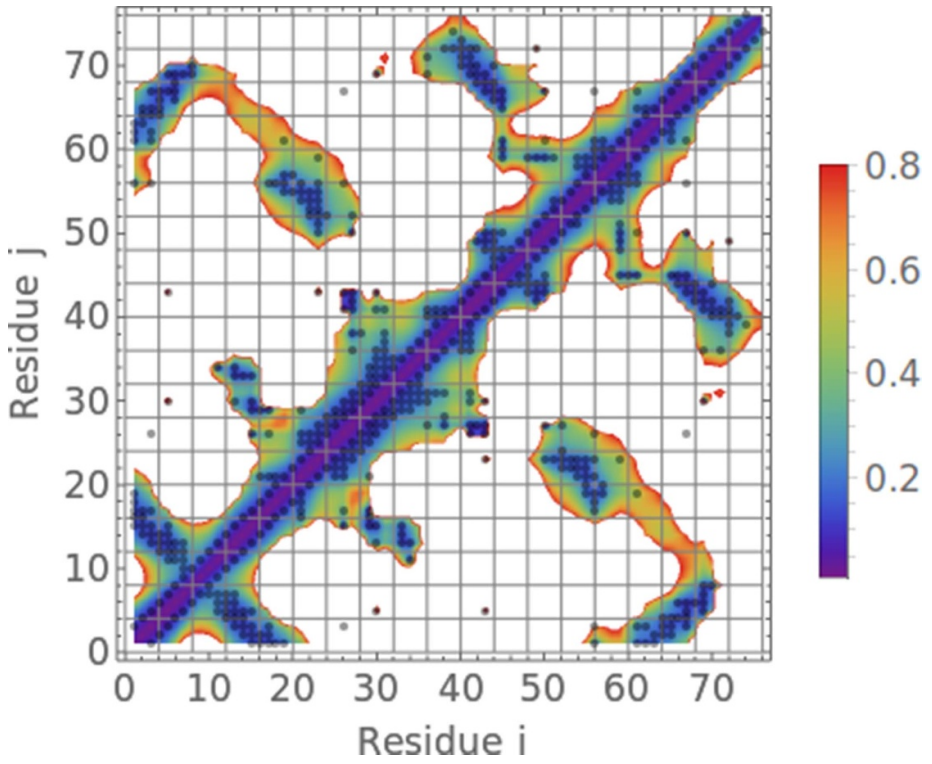

Figure 8. Density plot in temperature color-code of the time t^*_{ij}, at which each response function $R_{ij}(t)$ attains its maximum, providing a view of how the response spans the Ub structure. The response propagates by nucleation either from native contacts or from the backbone, and coalescence, see for instance the block, $1 \leqslant i \leqslant 30$ and $50 \leqslant j \leqslant 70$. The coalescence process involves the strands S1, S2 and S5 of the secondary structure reported in figure 2.

A perturbation on the site Gly35 involves mainly sites Ile13 and Leu71, although the meaning of this interplay is not biologically clear. This might induce us to suppose that the site Gln49 is somehow more relevant in the allosteric response than Phe35.

To provide a global view of how the elastic network of interactions supports the spreading of response across the Ub structure, we compute the time at which each response $R_{ij}(t)$ becomes maximal: $t^*_{ij} = \{t | R_{ij}(t) = \text{Max}\}$, see figure 6 for the definition of t^*_{ij}. The result reported in figure 8 shows that the response propagates along the proteins by nucleation from native contacts and from the backbone region as well, then it expands across the structure, giving rise to a sort of coalescence process among certain secondary structure elements. In particular in the region $1 \leqslant i \leqslant 30$ and $50 \leqslant j \leqslant 70$, such a response coalescence involves the strands S1, S2 and S5 of the secondary structure depicted in figure 2. Interestingly, the coalescence puts in connection the two clusters, CluA and CluB, in which Ub is structurally partitioned [55, 60].

6. Transfer-entropy analysis

Response functions are indicators that are easily interpreted in terms of causal dependencies, however in a GNM model at equilibrium, responses are symmetric (likewise correlations) and cannot distinguish the role of source and target of causation. For this reason, it is natural to complement response analysis with TE computation, equation (17), that is becoming an increasingly popular tool for characterizing the role of coordinated fluctuations in allostery processes [33, 34]. The behavior of some representative TEs as a function of the time-lag is reported in figure 9. Since by definition, $\text{TE}_{i\to j}$ is different from $\text{TE}_{j\to i}$, thus discriminating the concept of 'donor' and 'acceptor' of entropy, we consider both curves for each couple of residues already considered in figures 4 and 6.

This asymmetry has been used in [33]. to give each residue of the pair $i-j$ the role of 'driver' or 'driven', depending if the difference $\text{TE}_{i\to j} - \text{TE}_{j\to i}$ is positive (i driver, j driven) or negative (j driver, i driven).

The plots exhibit a typical skewed bell shape growing from zero at $t = 0$ and decaying to zero as $t \to \infty$, as discussed in section 3.3, and characterized by a single well-defined peak. Moreover, TEs between two residues that are in native contact (23-53) are larger than TEs of nonnative couples. The dependence of TEs on the time-lag can raise ambiguities in their comparison or ranking, which can be strongly affected by the chosen time interval. Since every choice of an optimal time can be questioned, it is necessary to repeat the analysis at different stages to confirm that the ranking is conserved at different times.

Instead of plotting the dominant direction of information flow $\text{TE}_{p\to j} - \text{TE}_{j\to p}$ as suggested in [33,

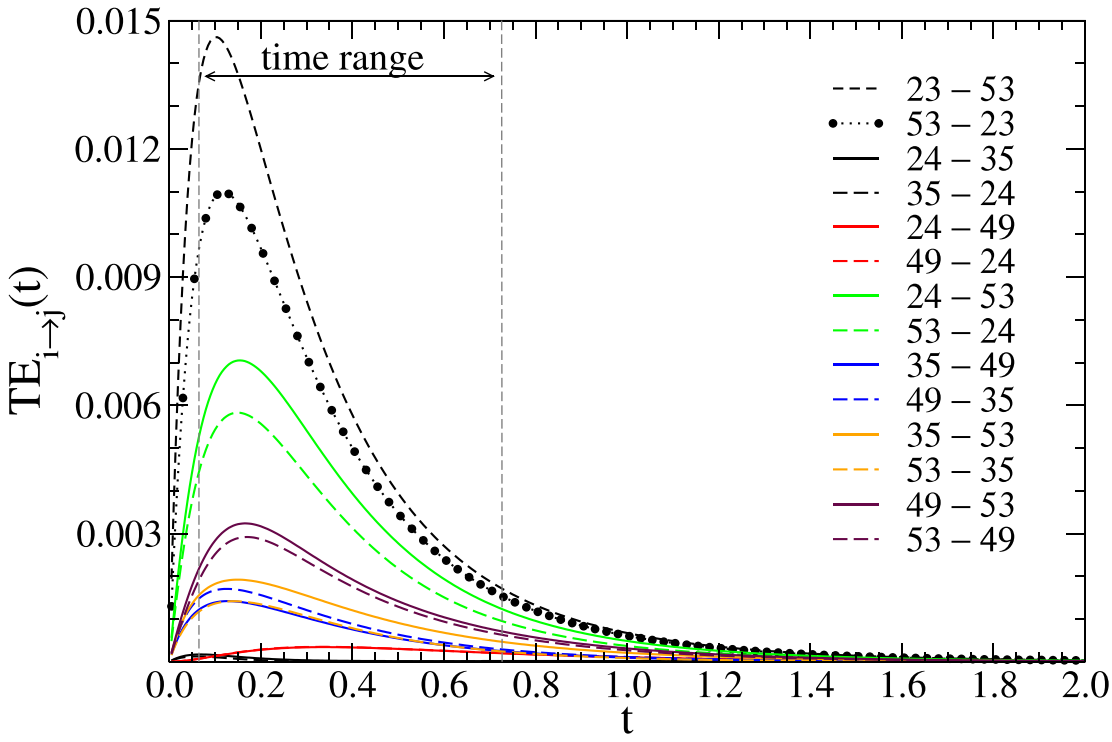

Figure 9. Time behavior of the entropy transfer between the sites underlying the allosteric communication of Ub according to experiments. Note that, unlike correlations and responses, $TE_{j\to i} \neq TE_{i\to j}$, i.e. TE by definition is not symmetric.

34], we prefer to report in figure 10 the profiles $TE_{p\to j}$ and $TE_{j\to p}$ separately, assuming that the reference site 'p' can act as a source (donor) or target (acceptor) of information transfer, respectively. In our case, this choice is due to the high sensitivity of TE differences from the sampling times, which prevents their robust interpretation.

The $TE_{p\to p}$ is trivially zero by definition, while just below and above p, the profiles show peaks corresponding to a significant flow of entropy associated with the activity of the backbone.

The TE analysis basically confirms the scenario drawn from the response functions, i.e. Gly53 and Glu24 are linked both as donors and acceptors of entropy. However, a new feature is highlighted by the TE, as the region around 1–20 (strands S1-S2 of figure 2(b)) and the C-terminal tail 70–76 displays a sensible entropy transfer, a detail not accounted for by the response analysis. This aspect is important since, as we will see in the following, such residues participate in formations of the complex Ub-USP and their capability of entropy transfer is dramatically affected by binding. The presence of humps of TE at the level of C-terminal tail might be relevant since this is the binding region of the Ub-activating enzymes, the tail fluctuations are thus expected to be tightly regulated by the allosteric switching.

Concerning Gly35 and Gln49, we can say that these two residues show a distributed entropy exchange over many segments of the Ub, including the terminal and interface region of the Ub-USP complex. In particular, panel (b) of figure 10. seems to suggest a sort of causal control between Gly35 and the wide region 45–68 of Ub, comprising the secondary motifs S4-H2-S5 of figure 2(b). Unfortunately, this connection remains unexplained in terms of mere structural properties. Conversely, Gln49 exchanges entropy with a large portion of the N-terminal region S1–S2 and with Tyr59, where there is a hump.

The profile analysis seems to indicate that TE is more sensitive than response function to the details of both molecular structure and modeling, so TE predictions need careful pondering.

7. Ubiquitin in complex with ubiquitinase: Ub-USP

In this section, we consider the complex Ub-USP to show how the behavior of the above indicators modifies upon binding and check whether this affects the role of source (24, 53) and target (35, 49) residues. While the molecular switch (peptide bond between Asp52-Gly53 and Glu24) is internal in the Ub and does not participate in the binding process, the residues Gly35 and Gln49, modulated by the switch, belong to the interface region Ub-USP, and thus are expected to undergo a higher perturbation due to the binding.

The GNM is applied to the full structure Ub-USP (pdb-id: 2IBI) that has been reconstructed via Modeller [61] because in the pdb-file 2IBI there were

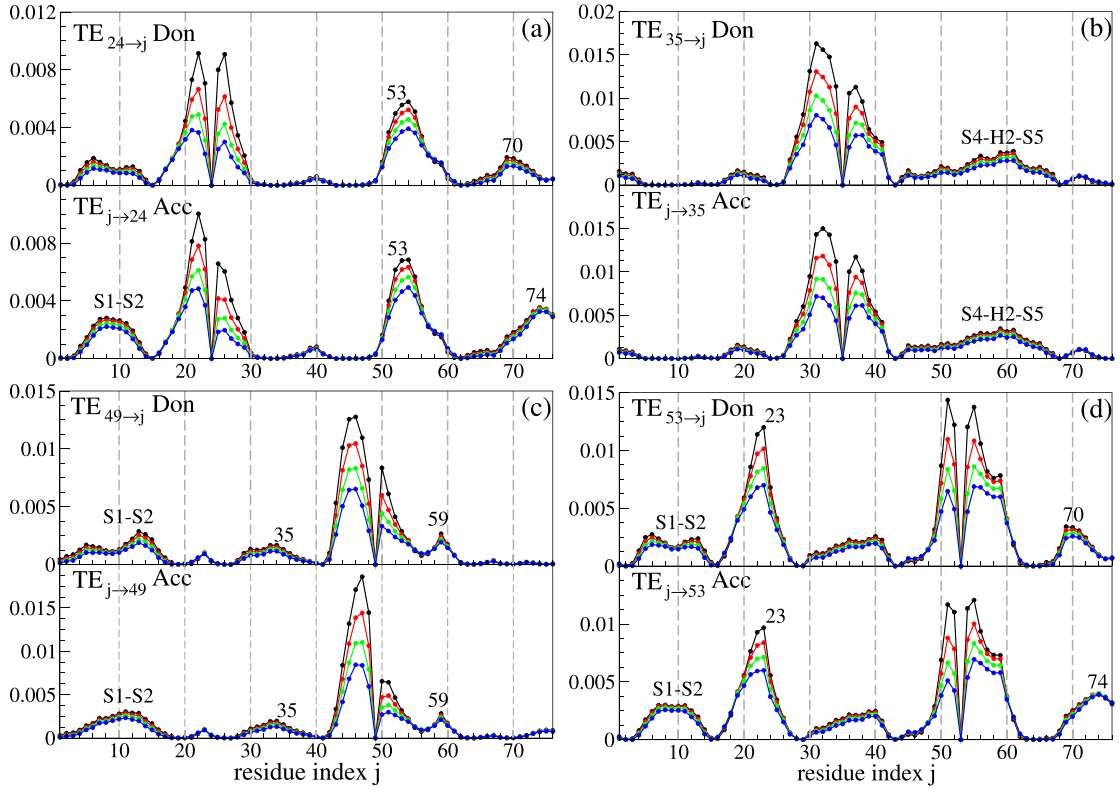

Figure 10. Profiles of transfer entropy $TE_{p\to j}$ (donor) and $TE_{j\to p}$ (acceptor) for sites $p = \{24, 35, 49, 53\}$, sampled at the same times of figures 5 and 7.

three gaps (three missing residues) Ser382, Tyr450, Asn537. The three gaps were filled using a homology modeling approach whereby the available structure USP was used as a template to model the conformation of the whole USP sequence (including the three residues whose structure was not resolved in the PDB file). Even if our GNM analysis involves the whole Ub-USP complex, we only present the results restricted to the Ub chain.

In the Ub-USP complex the correlation profile of the considered Ub residues becomes positive (figure 11), while in the isolated Ub, these residues were positively and negatively correlated with the rest of the molecule (figure 5). This evident difference in correlation behavior among internal Ub residues when coupled to USP can be explained by observing that correlations are sensitive to the collective fluctuations of the whole complex. Except for this shift toward positive values, a comparison of figure 11(a) with the profiles in figure 5(b) suggests that the landscape of correlations established by residue Glu24 with the rest of the Ub remains basically unaltered, the same happens for Gly53, not shown, since such two residues do not participate in the binding surface. On the contrary, the profiles of residue Gly35 and Gln49 are significantly modified in the bonded Ub with respect to the case of free Ub.

The inspection of the response profile of Ub within the complex Ub-USP in figure 12 does not show relevant differences with respect to the profiles in figure 7. The scenario is qualitatively reproduced, except for the intensity of the peaks and some small and irrelevant details.

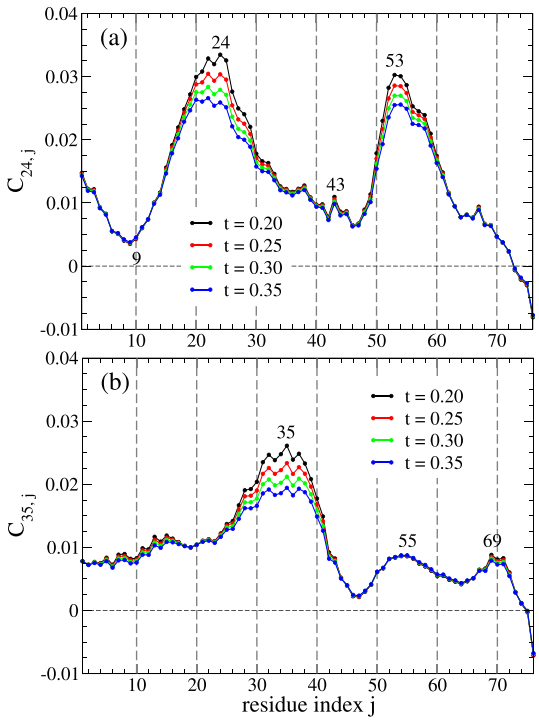

Figure 11. Correlation profiles of Ub residues extracted from the complex Ub-USP for sites $p = \{24,35\}$, sampled at representative four times.

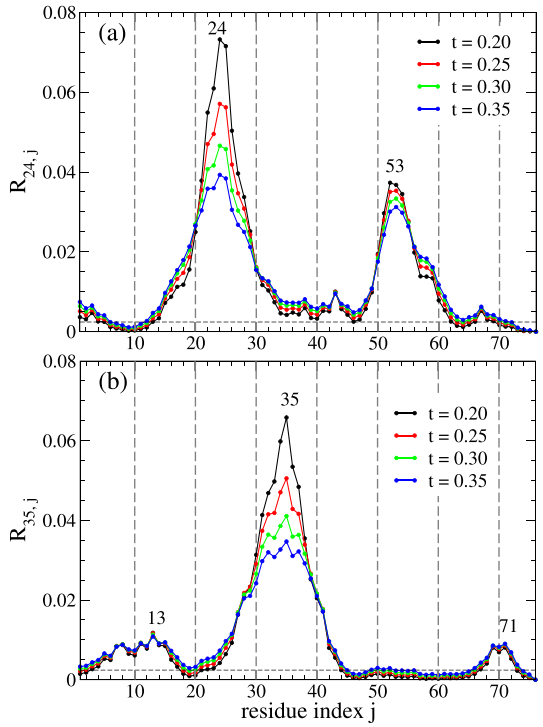

Figure 12. Response profiles of Ub residues $p = \{24,35\}$, panel (a) and (b) respectively, when Ub is bonded to USP. The comparison with the corresponding profiles for the free Ub, figures 7(a) and (b) shows a qualitative agreement, suggesting that response is less affected by the collective effects of the coupling.

We can conclude that, unlike correlations, the responses only vary because of the new interactions established by Ub with USP, most of them located at the interface Ub-USP. Thus in the Ub-USP, the response functions of the internal sites of Ub are only slightly modified by the presence of the USP, which acts as an external common driving. This is a further indication that the response function is less affected by spurious effects. In summary, one can conclude that correlations are also sensitive to long-range effects carried by collective fluctuations, whereas responses are expected to be more susceptible to relatively short paths of direct interactions.

8. Conclusions

The common use of correlation-based methods for inferring functional modes among the coordinated fluctuations of biomolecular systems, although favored by its ease of implementation, is limited by the fact that correlation not necessarily implies causation.

Therefore, especially to understand allosteric control, it could be convenient to go beyond correlation through the employment of response-function and transfer-entropy that have the advantage of inferring the causal relationships among source and target sites.

In this work, we compared correlation, response and entropy-transfer analysis taking as a benchmark the ubiquitin protein that is widely studied and whose allosteric regulation has been recently discovered. The purpose was to test the capability of these three indicators to recognize the relationships among residues identified in the experiments as underlying the Ub allosteric behavior.

To make this comparison free from the artifacts of poor statistics, we describe the Ub fluctuations around its native state through the GNM [41] that, being solvable, provides exact expressions for any observable, easily accessible to the numerical computation. GNM is considered a reasonable coarse-grained approximation of the all-atom NMA, as it is able to capture the relevant features of slowest functional modes in proteins and enzymes [43, 44].

Here, to partially include the effects of the side chains, we used the atom-wise definition of the GNM connectivity matrix, termed heavy-atom contact map, see equation (6). Of course, this represents only a caricature of the real side-chain effects which nevertheless improves the GNM reliability.

The analysis of correlation within the GNM suggests a scenario in agreement with the allosteric-switch mechanism triggered by the flipping of Asp52-Gly53 peptide bond that modulates the interaction of ubiquitin and ubiquitinase. The correlation profiles along the Ub chains show several positive and negative peaks that identify a certain set of residues including those involved in the allosteric process. However, this positive result obtained by correlations could not be considered conclusive and required validation and further insight from response theory. Response-based analysis localizes on a subset of the positive and negative peaks present in the correlation profiles. These represent the pair of residues in which the correlated motion can be safely associated with a causal relationship.

However, the responses of a GNM at equilibrium are symmetric, therefore they do not distinguish whether a residue behaves as a driver (donor) or is driven (acceptor). Therefore it has been necessary to complement response results with the TE analysis.

The TE profiles indicate that the status of donor and acceptor for a residue is not intrinsic, but may depend on the time lag between two consecutive observations. A comparison of response and TE profiles shows that TE is more sensitive than response to the finer details of the molecular structure and to the modeling approach, thus making the interpretation of the results quite delicate.

It is important to observe that when the estimation of TE has to be carried out from data, the results can be affected by various factors, such as: the dimension of the space of states, the length of time-series of data, and the procedure adopted to estimate high-dimensional conditional probabilities [62–64]. In addition, TE might be altered by spurious information arising from shared or common external inputs and drivings. Estimation of response instead is less affected by the above factors, and formula (14) allows

us to have a proxy of the response from an easy re-elaboration of the observed correlations of a system, thus reconciling the *observational* and *interventional* definitions of causation.

We remark that our comparison of causal indicators has been carried out in the case of harmonic approximations through NMA, however, there are frequent cases in which the nonlinear effects in allosteric processes are so important that NMA cannot be representative. In these conditions, the statistical indicators have to be estimated from very long molecular dynamics simulations in their atomistic [33, 65, 66] or coarse-grained [67, 68] formulation.

Finally, we can conclude that the best strategy to gain insight into allosteric sites and pathways on protein structures is a proper combination of response and transfer-entropy analysis, where correlation could represent a preliminary but unnecessary step.

Data availability statement

The data will be made publicly upon request to authors The data that support the findings of this study are available upon reasonable request from the authors.

Acknowledgments

The authors are grateful to A Giacomello for a critical reading of the manuscript and useful remarks. C G acknowledges the funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme [Grant Agreement No. 803 213].

F C and A V acknowledge the support from the MIUR PRIN 2017 Project 201798CZLJ, 'Coarse-grained description for non-equilibrium systems and transport phenomena' (CO-NEST).

Appendix. Derivation of the Gaussian TE

This appendix briefly shows how equation (17) for the TE is derived in the assumption of stationary Gaussian systems. The starting point is the entropy of a Gaussian vector $\mathbf{x} = \{x_1, \dots x_n\}$,

$$H(\mathbf{x}) = \frac{1}{2} \ln \det[2\pi e \Omega]$$

where $\Omega = \langle \mathbf{x}(t)\, \mathbf{x}^{\mathrm{T}}(t) \rangle$ is the covariance matrix of $\mathbf{x}(t)$. When computing the TE, we refer to the following coupling between the processes $x_i(t)$ and $x_j(t)$, represented schematically as follows

$$\begin{aligned} \{x_i(0), x_j(0)\} &\longrightarrow x_i(t) \\ x_j(0) &\longrightarrow x_j(t) \end{aligned}$$

where the arrows indicate that not only the ith variable, but also the jth variable is able to influence the future state $x_i(t)$. In the following, for the sake of shorthand notation, we set, $x_i^+ = x_i(t)$, $x_i = x_i(0)$ and $x_j = x_j(0)$.

The Gaussian formulation is defined once the covariance matrix is specified

$$\Omega[x_i^+, x_i, x_j] = \left[\begin{array}{c|cc} C_{ii}(0) & C_{ii}(t) & C_{ij}(t) \\ \hline C_{ii}(t) & C_{ii}(0) & C_{ij}(0) \\ C_{ij}(t) & C_{ij}(0) & C_{jj}(0) \end{array}\right] \tag{A.1}$$

where $C_{\mu\nu}(0) = \langle x_\mu(t) x_\nu(t) \rangle = \langle x_\mu(0) x_\nu(0) \rangle$ denotes the correlation at equal times (zero-lag) for a stationary process, while, $C_{\mu\nu}(t) = \langle x_\mu(t) x_\nu(0) \rangle = \langle x_\mu(0) x_\nu(t) \rangle$ is the correlation at lag t, which, at equilibrium, is symmetric under index exchange. For convenience of notation we drop the argument '0' in the zero-lag correlations by defining: $C_{\mu\nu}(0) = \sigma_{\mu\nu}$.

Using the definition of conditional distribution of a multivariate Gaussian process, we can express the TE in the form

$$\mathrm{TE}(x_j \to x_i) = \frac{1}{2} \ln \frac{\mathrm{Det}\Big(\Omega[x_i^+ | x_i]\Big)}{\mathrm{Det}\Big(\Omega[x_i^+ | x_i, x_j]\Big)} \tag{A.2}$$

where $\Omega[x_i^+ | x_i]$ and $\Omega[x_i^+ | x_i, x_j]$ are the covariance matrices of the conditioned Gaussian distribution, also termed *conditioned covariance matrices*, that can be expressed via the fundamental identity for Gaussian variables [69],

$$\Omega(a|b) = \Omega(a,a) - \Omega(a,b)\, \Omega^{-1}(b,b)\, \Omega(b,a).$$

With reference to matrix (A.1) we have, $\Omega(a,a) = \sigma_{ii}$ and $\Omega(a,b) = [C_{ii}(t), C_{ij}(t)]$, moreover

$$\Omega(b,a) = \begin{bmatrix} C_{ii}(t) \\ C_{ij}(t) \end{bmatrix} \quad \textit{and} \quad \Omega(b,b) = \begin{bmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ij} & \sigma_{jj} \end{bmatrix}$$

In explicit form we can write

$$\Omega[x_i^+ | x_i, x_j] = \sigma_{ii} - [C_{ii}(t), C_{ij}(t)] \begin{bmatrix} \sigma_{ii} & \sigma_{ij} \\ \sigma_{ij} & \sigma_{jj} \end{bmatrix}^{-1} \begin{bmatrix} C_{ii}(t) \\ C_{ij}(t) \end{bmatrix}. \tag{A.3}$$

Analogously using the conditional variance, we can write in equation (A.2)

$$\Omega[x_i^+ | x_i] = \sigma_{ii} - \frac{C_{ii}^2(t)}{\sigma_{ii}}. \tag{A.4}$$

Finally the inversion of the matrix $\Omega(b,b)$,

$$\Omega^{-1}(b,b) = \frac{1}{\sigma_{ii}\sigma_{jj} - \sigma_{ij}^2} \begin{bmatrix} \sigma_{jj} & -\sigma_{ij} \\ -\sigma_{ij} & \sigma_{ii} \end{bmatrix}$$

and the explicit computation of the matrix products provide the result

$$\Omega[x_i^+ | x_i, x_j] = \sigma_{ii} - \frac{\sigma_{jj} C_{ii}^2(t) - 2\sigma_{ij} C_{ii}(t) C_{ij}(t) + \sigma_{ii} C_{ij}^2(t)}{\sigma_{ii}\sigma_{jj} - \sigma_{ij}^2}.$$

thus

$$\mathrm{TE}_{j\to i} = \frac{1}{2}\log\frac{(\sigma_{ii}\sigma_{jj}-\sigma_{ij}^2)\,[\sigma_{ii}-C_{ii}^2(t)/\sigma_{ii}]}{\sigma_{ii}(\sigma_{ii}\sigma_{jj}-\sigma_{ij}^2)-\sigma_{jj}C_{ii}^2(t)+2\,\sigma_{ij}C_{ii}(t)C_{ij}(t)-\sigma_{ii}C_{ij}^2(t)}.$$

Now by adding and subtracting the term $\sigma_{ij}^2 C_{ij}(t)$ to the denominator of the above expression, and by defining

$$\alpha_{ij}(t) = [\sigma_{ii}C_{ij}(t)-\sigma_{ij}C_{ii}(t)]^2$$
$$\beta_{ij}(t) = (\sigma_{ii}\sigma_{jj}-\sigma_{ij}^2)\,[\sigma_{ii}^2-C_{ii}^2(t)],$$

we can recast $\mathrm{TE}_{j\to i}$ in the form (17).

ORCID iDs

Fabio Cecconi https://orcid.org/0000-0001-8351-248X
Giulio Costantini https://orcid.org/0000-0002-1536-9279
Carlo Guardiani https://orcid.org/0000-0002-8914-9260
Marco Baldovin https://orcid.org/0000-0003-3559-4032
Angelo Vulpiani https://orcid.org/0000-0002-0976-2859

References

[1] Liu J and Nussinov R 2016 Allostery: an overview of its history, concepts, methods and applications *PLoS Comput. Biol.* **12** e1004966
[2] Ribeiro A A S T and Ortiz V 2016 A chemical perspective on allostery *Chem. Rev.* **116** 6488–502
[3] Monod J and Jacob F 1961 General conclusions: teleonomic mechanisms in cellular metabolism, growth and differentiation *Cold Spring Harbor Symposia on Quantitative Biology* vol 26 (Cold Spring Harbor Laboratory Press) pp 389–401
[4] Guarnera E and Berezovsky I N 2016 Allosteric sites: remote control in regulation of protein activity *Curr. Opin. Struct. Biol.* **37** 1–8
[5] Berendsen H J 2000 Collective protein dynamics in relation to function *Curr. Opin. Struct. Biol.* **10** 165–9
[6] Fenwick R B, Esteban-Martín S, Richter B, Lee D, Walter K F, Milovanovic D, Becker S, Lakomek N A, Griesinger C and Salvatella X 2011 Weak long-range correlated motions in a surface patch of ubiquitin involved in molecular recognition *J. Am. Chem. Soc.* **133** 10336–9
[7] Tang Q-Y, Zhang Y-Y, Wang J, Wang W and Chialvo D R 2017 Critical fluctuations in the native state of proteins *Phys. Rev. Lett.* **118** 088102
[8] Tang Q-Y, Kaneko K and de Groot B L 2020 Long-range correlation in protein dynamics: confirmation by structural data and normal mode analysis *PLoS Comput. Biol.* **16** e1007670
[9] Thirumalai D, Hyeon C, Zhuravlev P I and Lorimer G H 2019 Symmetry, rigidity and allosteric signaling: from monomeric proteins to molecular machines *Chem. Rev.* **119** 6788–821
[10] Diez G, Nagel D and Stock G 2022 Correlation-based feature selection to identify functional dynamics in proteins *J. Chem. Theory Comput.* **18** 5079–88
[11] Lake P T, Davidson R B, Klem H, Hocky G M and McCullagh M 2020 Residue-level allostery propagates through the effective coarse-grained Hessian *J. Chem. Theory Comput.* **16** 3385–95
[12] De Los Rios P, Cecconi F, Pretre A, Dietler G, Michielin O, Piazza F and Juanico B 2005 Functional dynamics of PDZ binding domains: a normal-mode analysis *Biophys. J.* **89** 14–21
[13] Chennubhotla C, Yang Z and Bahar I 2008 Coupling between global dynamics and signal transduction pathways: a mechanism of allostery for chaperonin GroEL *Mol. BioSyst.* **4** 287–92
[14] Van Wart A, Durrant J, Votapka L and Amaro R 2014 Weighted implementation of suboptimal paths (WISP): an optimized algorithm and tool for dynamical network analysis *J. Chem. Theory Comput.* **10** 511–7
[15] Wang J, Jain A, McDonald L R, Gambogi C, Lee A L and Dokholyan N V 2020 Mapping allosteric communications within individual proteins *Nat. Commun.* **11** 3862
[16] Di Paola L and Giuliani A 2015 Protein contact network topology: a natural language for allostery *Curr. Opin. Struct. Biol.* **31** 43–48
[17] Chennubhotla C and Bahar I 2006 Markov propagation of allosteric effects in biomolecular systems: application to GroEL-GroES *Mol Syst. Biol.* **2** 36
[18] Bassetto C A Z, Costa F, Guardiani C, Bezanilla F and Giacomello A 2023 Noncanonical electromechanical coupling paths in cardiac hERG potassium channel *Nat. Commun.* **14** 11
[19] Costa F, Guardiani C and Giacomello A 2022 Disrupted stepwise functional brain organization in overweight individuals *Commun. Biol.* **5** 11
[20] Costa F, Guardiani C and Giacomello A 2021 Exploring Kv1.2 channel inactivation through MD simulations and network analysis *Front. Mol. Biosci.* **8** 9
[21] Rocks J W, Pashine N, Bischofberger I, Goodrich C P, Liu A J and Nagel S R 2017 Designing allostery-inspired response in mechanical networks *Proc. Natl Acad. Sci. USA* **114** 2520–5
[22] Cooper A and Dryden D 1984 Allostery without conformational change *Eur. Biophys. J.* **11** 103–9
[23] Frauenfelder H, McMahon B H, Austin R H, Chu K and Groves J T 2001 The role of structure, energy landscape, dynamics and allostery in the enzymatic function of myoglobin *Proc. Natl Acad. Sci.* **98** 2370–4
[24] Pearl J 2009 *Causality* (Cambridge University Press)
[25] Baldovin M, Cecconi F and Vulpiani A 2020 Understanding causation via correlations and linear response theory *Phys. Rev. Res.* **2** 043436
[26] Aurell E and Del Ferraro G 2016 Causal analysis, correlation-response, and dynamic cavity *J. Phys.: Conf. Ser.* **699** 012002
[27] Sarra C, Baldovin M and Vulpiani A 2021 Response and flux of information in extended nonequilibrium dynamics *Phys. Rev.* E **104** 024116
[28] Erman B 2018 A computational model for controlling conformational cooperativity and function in proteins *Proteins* **86** 1001–9
[29] Essiz S G and Coalson R D 2009 Dynamic linear response theory for conformational relaxation of proteins *J. Phys. Chem.* B **113** 10859–69
[30] Hacisuleyman A, Erkip A and Erman B 2021 Synchronous and asynchronous response in dynamically perturbed proteins *J. Phys. Chem.* B **125** 729–39
[31] Marconi Marini Bettolo U, Puglisi A, Rondoni L and Vulpiani A 2008 Fluctuation-dissipation: response theory in statistical physics *Phys. Rep.* **461** 111–95
[32] Guarnera E, Berezovsky I N and Liu J 2016 Structure-based statistical mechanical model accounts for the causality and energetics of allosteric communication *PLoS Comput. Biol.* **12** e1004678
[33] Hacisuleyman A and Erman B 2017 Entropy transfer between residue pairs and allostery in proteins: quantifying allosteric communication in ubiquitin *PLoS Comput. Biol.* **13** e1005319

[34] Hacisuleyman A and Erman B 2017 Causality, transfer entropy and allosteric communication landscapes in proteins with harmonic interactions *Proteins* **85** 1056–64
[35] Schreiber T 2000 Measuring information transfer *Phys. Rev. Lett.* **85** 461
[36] Paluš M, Komárek V, Hrnčíř Z and Štěrbová K 2001 Synchronization as adjustment of information rates: detection from bivariate time series *Phys. Rev.* E **63** 046211
[37] Stone J V 2015 *Information Theory: A Tutorial Introduction* (Sebtel Press)
[38] Matsuda H 2000 Physical nature of higher-order mutual information: intrinsic correlations and frustration *Phys. Rev.* E **62** 3096
[39] LeVine M V and Weinstein H 2014 NbIT—a new information theory-based analysis of allosteric mechanisms reveals residues that underlie function in the leucine transporter LeuT *PLoS Comput. Biol.* **10** e1003603
[40] Smith C A, Ban D, Pratihar S, Giller K, Paulat M, Becker S, Griesinger C, Lee D and de Groot B L 2016 Allosteric switch regulates protein-protein binding through collective motion *Proc. Natl Acad. Sci. USA* **113** 3269–74
[41] Haliloglu T, Bahar I and Erman B 1997 Gaussian dynamics of folded proteins *Phys. Rev. Lett.* **79** 3090
[42] Nicolaï A, Delarue P and Senet P 2014 *Low-Frequency, Functional, Modes of Proteins: All-Atom and Coarse-Grained Normal Mode Analysis* ed A Liwo (Springer) pp 483–524
[43] Dykeman E C and Sankey O F 2010 Normal mode analysis and applications in biological physics *J. Phys.: Condens. Matter* **22** 423202
[44] Bahar I, Lezon T R, Yang L-W and Eyal E 2010 Global dynamics of proteins: bridging between structure and function *Annu. Rev. Biophys.* **39** 23–42
[45] Hub J S and de Groot B L 2009 Detection of functional modes in protein dynamics *PLoS Comput. Biol.* **5** e1000480
[46] Das A, Gur M, Cheng M H, Jo S, Bahar I and Roux B 2014 Exploring the conformational transitions of biomolecular systems using a simple two-state anisotropic network model *PLoS Comput. Biol.* **10** e1003521
[47] Piazza F 2014 Nonlinear excitations match correlated motions unveiled by NMR in proteins: a new perspective on allosteric cross-talk *Phys. Biol.* **11** 036003
[48] Grubmüller H 1995 Predicting slow structural transitions in macromolecular systems: conformational flooding *Phys. Rev.* E **52** 2893
[49] Alexeev D, Bury S M, Turner M A, Ogunjobi O M, Muir T W, Ramage R and Sawyer L 1994 Synthetic, structural and biological studies of the ubiquitin system: chemically synthesized and native ubiquitin fold into identical three-dimensional structures *Biochem. J.* **299** 159–63
[50] Massi F, Grey M J and Palmer A G III 2005 Microsecond timescale backbone conformational dynamics in ubiquitin studied with NMR R1 relaxation experiments *Prot. Sci.* **14** 735–42
[51] Bahar I, Atilgan A R and Erman B 1997 Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential *Fold. Design* **2** 173–81
[52] Burioni R, Cassi D, Cecconi F and Vulpiani A 2004 Topological thermal instability and length of proteins *Proteins* **55** 529–35
[53] Kundu S, Melton J S, Sorensen D C and Phillips G N Jr 2002 Dynamics of proteins in crystals: comparison of experiment with simple models *Biophys. J.* **83** 723–32
[54] Yuan C, Chen H and Kihara D 2012 Effective inter-residue contact definitions for accurate protein fold recognition *BMC Bioinform.* **13** 1–13
[55] Cecconi F and Chinappi M 2020 Native-state fingerprint on the ubiquitin translocation across a nanopore *Phys. Rev.* E **101** 032401
[56] Falcioni M, Isola S and Vulpiani A 1990 Correlation functions and relaxation properties in chaotic dynamics and statistical mechanics *Phys. Lett.* A **144** 341–6
[57] Baldovin M, Cecconi F, Provenzale A and Vulpiani A 2022 Extracting causation from millennial-scale climate fluctuations in the last 800 kyr *Sci. Rep.* **12** 1–12
[58] Sun J, Taylor D and Bollt E M 2015 Causal network inference by optimal causation entropy *SIAM J. Appl. Dyn.* **14** 73–106
[59] Singh R K, Kazansky Y, Wathieu D and Fushman D 2017 Hydrophobic patch of ubiquitin is important for its optimal activation by ubiquitin activating enzyme E1 *Anal. Chem.* **89** 7852–60
[60] Schlierf M, Li H and Fernandez J M 2004 The unfolding kinetics of ubiquitin captured with single-molecule force-clamp techniques *Proc. Natl Acad. Sci.* **101** 7299–304
[61] Šali A and Blundell T L 1993 Comparative protein modelling by satisfaction of spatial restraints *J. Mol. Biol.* **234** 779–815
[62] Hlaváčková-Schindler K, Paluš M, Vejmelka M and Bhattacharya J 2007 Causality detection based on information-theoretic approaches in time series analysis *Phys. Rep.* **441** 1–46
[63] Vicente R, Wibral M, Lindner M and Pipa G 2011 Transfer entropy-a model-free measure of effective connectivity for the neurosciences *J. Comput. Neurosci.* **30** 45–67
[64] Ursino M, Ricci G and Magosso E 2020 Transfer entropy as a measure of brain connectivity: a critical analysis with the help of neural mass models *Front. Comput. Neurosci.* **14** 45
[65] Kamberaj H and van der Vaart A 2009 Extracting the causality of correlated motions from molecular dynamics simulations *Biophys. J.* **97** 1747–55
[66] Xu X, Guardiani C, Yan C and Ivanov I 2013 Opening pathways of the DNA clamps proliferating cell nuclear antigen and Rad9-Rad1-Hus1 *Nucl. Acids Res.* **41** 10020–31
[67] Guardiani C, Di Marino D, Tramontano A, Chinappi M and Cecconi F 2014 Exploring the unfolding pathway of maltose binding proteins: an integrated computational approach *J. Chem. Theory Comput.* **10** 3589–97
[68] Guardiani C, Cencini M and Cecconi F 2014 Coarse-grained modeling of protein unspecifically bound to DNA *Phys. Biol.* **11** 026003
[69] Prince S J 2012 *Computer Vision: Models, Learning and Inference* (Cambridge University Press)