Bioinformatics, 33(14), 2017, 2106–2113
doi: 10.1093/bioinformatics/btx124
Advance Access Publication Date: 7 March 2017
Original Paper

Structural bioinformatics

A network model predicts the intensity of residue-protein thermal coupling

Luciano Censoni[1], Heloisa dos Santos Muniz[2] and Leandro Martínez[1,]*

[1]Institute of Chemistry, University of Campinas, Campinas, SP, Brazil and [2]Institute of Physics of São Carlos, University of São Paulo, São Carlos, SP, Brazil

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on November 2, 2016; revised on January 24, 2017; editorial decision on February 27, 2017; accepted on March 1, 2017

Abstract

Motivation: The flow of vibrational energy in proteins has been shown not to obey expectations for isotropic media. The existence of preferential pathways for energy transport, with probable connections to allostery mechanisms, has been repeatedly demonstrated. Here, we investigate whether, by representing a set of protein structures as networks of interacting amino acid residues, we are able to model heat diffusion and predict residue-protein vibrational couplings, as measured by the Anisotropic Thermal Diffusion (ATD) computational protocol of modified molecular dynamics simulations.
Results: We revisit the structural rationales for the precise definition of a contact between amino acid residues. Using this definition to describe a set of proteins as contact networks where each node corresponds to a residue, we show that node centrality, particularly *closeness centrality* and *eigenvector centrality*, correlates to the strength of the vibrational coupling of each residue to the rest of the structure. We then construct an analytically solvable model of heat diffusion on a network, whose solution incorporates an explicit dependence on the connectivity of the heated node, as described by a perturbed graph Laplacian Matrix.
Availability and Implementation: An implementation of the described model is available at http://leandro.iqm.unicamp.br/atd-scripts.
Contact: leandro@iqm.unicamp.br

1 Introduction

Localized vibrational perturbations do not propagate isotropically in proteins. The existence of preferential pathways through which vibrational energy flows more rapidly has been observed in a myriad of studies, both experimental and computational (Leitner, 2008). These pathways have been implied in allostery mechanisms (Ribeiro and Ortiz, 2016), which allow perturbations caused by a local event such as the binding of an agonist to be efficiently transmitted through the protein core to effect a conformational (Laskowski *et al.*, 2009) or dynamical (Tzeng and Kalodimos, 2011) change at a distant site (Motlagh *et al.*, 2014), as well as in the maintenance of functional conformations in the face of thermal perturbations (Lampa-Pastirk and Beck, 2006).

While the efficiency of such pathways may be under specific selective pressures in proteins where they are crucial for stability or function, their existence may be a general feature not dependent on the particular characteristics of each structure. It has been noted by Liang and Dill that, although solid-like densities and compressibilities are observed, detailed investigation describes an interior that is akin to randomly packed spheres, at a density close to the percolation threshold (Liang and Dill, 2001). Transport channels arise naturally in percolation clusters (Rhodes and Blunt, 2007), which indicates that the existence of preferential transport channels in globular proteins may be a consequence of their geometry. This, in turn, would imply that the potential for allostery is present in all proteins, a hypothesis which is corroborated by the experimental observation of allosteric communication in nonallosteric proteins (Clarkson *et al.*, 2006; Gunasekaran *et al.*, 2004; Leitner, 2008).

Under the ‘preferential pathways’ picture of energy transport, one might expect that the nature and amount of connections made by each residue be one determinant factor for distinguishing those most relevant for diffusion. Here, we set out to investigate this conjecture. In order to do so, we revisit a formal definition for networks that represent protein structures and discuss a method for characterizing the connectivity of the node representing each residue. We then compare the results to a measure of the strength of each residue's vibrational coupling to the rest of the structure, as obtained by a modified molecular dynamics computational experiment.

2 Approach

Krishnan *et al.* argue that the network framework presents strong advantages as a protein modeling tool: proteins exhibit a natural dimension of discretization, the amino acid residue sequence, which models should be able to exploit by adopting a *relative* geometry where only the relationships between elements matter, such as measuring inter-residue distance as opposed to absolute residue position. If these relationships are tabulated in a square, symmetrical matrix, then such a construction affords direct interpretation as an *adjacency matrix*, which is a complete description for a network (Krishnan *et al.*, 2008). From these networks, quantitative topological descriptors may be derived which allow comparison across different proteins. Graph Theory provides a wide assortment of such measurements, and the reader is directed to, e.g. the review by Costa *et al.* for a survey (Costa *et al.*, 2007). The task remains, though, to make physical sense of those measurements, in particular identifying those which are strongly correlated with relevant physicochemical characteristics of the represented protein.

One such descriptor that has been consistently linked to relevant properties is *centrality*, formalized by Freeman in a 1978 review with respect to the structural properties of human communication networks (Freeman, 1978). Although an intuitive concept, centrality can be formalized in more than one way, with each 'flavor' possessing a distinct structural interpretation. Freeman defines three different measures of centrality, namely degree, betweenness and closeness centrality, each based on a different rationale, though their values exhibit a measure of correlation with each other. It is generally accepted that degree (i.e. number of contacts) is the least informative of the three, being focused on local interactions and blind to the network's global structure.

Of these descriptors, Vendruscolo *et al.* have shown that, in networks representing the folding transition state ensemble for a set of six proteins, high values of betweenness centrality distinguish nodes that represent residues experimentally identified as critical to the folding process. This information is, however, concealed in the native state by other high-betweenness residues that cannot be singled out as important to the folding process (Vendruscolo *et al.*, 2002). Del Sol and O'Meara have shown that, when a two-protein complex is modeled as a network, residues lying at the interface can be identified by their high betweenness, which is consistent with betweenness centrality's structural interpretation. However, most of these residues are not remarkably central when each interacting chain is analyzed by itself (del Sol and O'meara, 2005). Amitai *et al.* have also demonstrated that high closeness centrality can identify conserved positions, and predicts the active site and other functional sites of the protein, by itself or by significantly improving the performance of methods based on the conservation and surface accessibility of residues (Amitai *et al.*, 2004).

In light of these results, we propose that centrality is a suitable measure to characterize the connectivity of a residue in the context of heat propagation within the structure. To test this hypothesis, we employ a computational strategy that allows direct observation of the flow of energy in a structure, the Anisotropic Thermal Diffusion (ATD) protocol. Proposed by Ota and Agard (2005) and extended into a systematic methodology by Martínez *et al.* (2011), the ATD protocol has also been employed to observe possible directional asymmetry (rectification) in heat flow through hydrogen bonds (Miño-Galaz, 2015). An ATD computational experiment consists of cooling a protein structure to a very low temperature, then separately heating each residue and measuring the temperature of the structure after a fixed interval. In this way it is possible to quantify the strength of the vibrational coupling of each residue to the protein as a whole. Residues which are outstanding at quickly dissipating excess vibrational energy have been shown to be essential to the maintenance of protein activity, as demonstrated by mutagenesis experiments (Martínez *et al.*, 2011). We investigate whether those residues correspond to nodes of higher than average centrality in the corresponding residue networks.

3 Materials and methods

3.1 Network construction

No consensus is observed in the literature regarding the correct protocol to construct a network to represent a protein structure. Though studies mostly tend to represent the amino acid residues as the *nodes* in the network, thus favoring a scale which is coarser than the usual atomic representation, there is no agreement on what defines a *contact* between residues, to be represented by an *edge* between the corresponding pair of nodes (see Krishnan *et al.*, 2008; Böde *et al.*, 2007). Alternative definitions include, but are not limited to:

- Adding an edge between two nodes when the distance between the C_α of the corresponding residues is less than or equal to 8.5 Å (Dokholyan *et al.*, 2002).
- Adding an edge between two nodes when there exists a pair of atoms such that each belongs to one of the corresponding residues, and the distance between them is less than or equal to 5.0 Å (Greene and Higman, 2003).
- Adding an edge between two nodes when the total energy of the interaction between the corresponding residues meets a predetermined threshold that depends on the nature of the residues (Amitai et al., 2004).

No formal justification is provided for these definitions, and while it may be argued that the observed correlations between network descriptors and physicochemical properties, when they exist, may depend only weakly on the precise protocols employed to construct the networks, we assert that some protocols are to be preferred on independent theoretical grounds.

Miyazawa and Jernigan, in a well-known 1985 paper that is among the earliest investigations of internal packing in proteins (Miyazawa and Jernigan, 1985), estimate the effective inter-residue contact energy by counting contacts in crystal structures. Calculating energies from contact frequencies, they present a series of arguments to establish a definition for a contact in the first place: two interior residues are considered in contact when the centers of mass of their respective side-chains are less than 6.5 Å apart. To justify this definition, they measure the radial distribution of interior residues, each represented by the position of the center of mass of its side chain and each not counting its covalent neighbors, over a sample of 42 globular proteins (their 1996 update on the same work examines a larger sample of 1168 structures (Miyazawa and

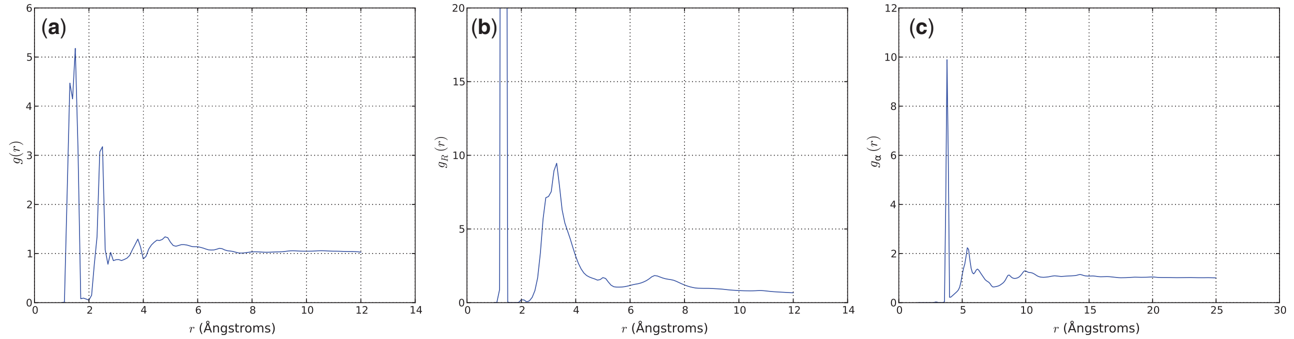

Fig. 1. (**a**) Average radial distribution function for all heavy atoms of all proteins in the ASTRAL subset of the SCOP database, release 1.75. Recognizable features include the double peak centered at 1.5 Å, compatible with a typical distance for covalent bonds, and the peak at 2.6 Å, associated to Hydrogen bonds. The peak at 3.8 Å appears to be induced by regularities in the bond network introduced by elements of secondary structure (compare panel c), and is also compatible with van der Waals contacts. (**b**) Average closest contact distribution function (*see text for description*) for the same set of structures. The vertical scale is set to one fifth of the first peak's intensity in order to improve readability. (**c**) Average radial distribution function for C_α atoms for the same set of structures. The prominent 3.8 Å peak associated to the typical C_α–C_α distance in the *trans* conformation of the peptide bond is readily recognizable

Jernigan, 1996)). They note that the first peak occurs at a distance of 5.0 Å but extends up until about 6.5 Å, which they adopt as the contact cutoff distance and report to be consistent with the calculated average residue radius of 2.91 Å, given the average residue volume and packing density. We perform similar measurements on a much larger set of 10 569 structures from the ASTRAL subset of non-redundant protein structures (release 1.75 of 2009, including only proteins with less than 40% identity to each other) (Chardonia *et al*., 2004), derived from the SCOP database (Lo Conte *et al*., 2000), and obtain distributions that are much more detailed but support similar conclusions.

We report the generic all-site radial distribution function for protein structures, the average of the all-site radial distributions calculated for each entry in the database. We follow the usual definition: for each structure, we count the number of neighbors seen at a distance between r and $r + dr$, as a function of r, by each of its N heavy atoms (ASTRAL structures are provided without Hydrogen atoms). The obtained curve is normalized by the radial distribution function of the corresponding reference state, the average of 5 curves calculated from 5 independently generated sets of N random points, each approximately contained inside the target structure's volume.

In the actual implementation, we take advantage of the relative formalism by calculating once for each structure or distribution the distance between all pairs of points and filling a *distance matrix*, then iteratively extracting from it the *total* number of contacts, defined as the number of pairs whose distance is smaller than a given cutoff, as a function of cutoff. From the total number of contacts in a structure, $C_{\mathrm{str}}(r)$, and its reference state, $C_{\mathrm{rnd}}(r)$, the radial distribution function is calculated from the definition as in Equation (1). Results are shown in Figure 1, panel a.

$$g(r) = \left\langle \frac{1}{3r^2}\frac{d}{dr}\left[r^3\,\frac{C_{\mathrm{str}}(r)}{C_{\mathrm{rnd}}(r)}\right]\right\rangle_{\text{all structures}} \qquad (1)$$

It is readily observable that ordered structure persists up to a radius of at least 5.0 Å but arguably up to 7.0 Å around each site, establishing a tentative upper value for the cutoff distance that defines a contact, though of course the range of the effective pair interaction may be shorter and still lead to observable long-range structure associated to indirect neighbors.

To provide a complementary perspective, we also calculate a variation of the radial distribution function aggregated at the residue scale, the closest contact distribution function $g_R(r)$, akin to the solvation-shell radial distribution $g_{ss}(r)$ described in (Song *et al*., 2000). We take into account only the closest contact between each pair of residues, in such a way that nearest neighbors only contribute to peaks corresponding to their actual contact distance, despite differences in size and orientation (see Figure 2a and companion text in Song *et al*., 2000 for an illustration of the sharper, more interpretable peaks obtained using this procedure).

We implement the calculation in the relative formalism; if the rows and columns of the atom distance matrix are indexed in such a way that atoms which belong to the same residue are always contiguous, then it may be partitioned in blocks, one for each residue pair, from which a residue adjacency matrix may be derived by inspection—residues i and j are in contact if, after the application of a given cutoff, there is at least one non-zero element in the ij block of the atom adjacency matrix (We note that if this procedure is used to define a *distance* measure in residue space, then this distance does not obey the triangle inequality and can't define a metric space.). The same procedure is applied to the atom adjacency matrix of each random distribution to construct the reference state (atoms in the random distribution are considered to belong to the same residue as the atom of the original structure that they are closest to when generated), and the rest of the analysis proceeds as before. Results are reported in Figure 1, panel b. Covalent neighbors must be responsible for the trivial first peak, so the second, broad peak from about 2.5–5.5 Å must be attributed to non-covalent direct contacts. Moreover, the definition of $g_R(r)$ implies that peaks corresponding to indirect (or 'second') neighbors are expected at a distance r no shorter than one full residue diameter, and are thus consistent with the third, subdued peak centered at 7.0 Å. In light of these results, we submit that a cutoff of at most 6.0 Å is the appropriate maximum distance to define a contact between two residues based on the distance between all pairs of atoms.

In the interest of completeness, we also report the radial distribution function for the C_α atoms, $g_\alpha(r)$, in order to investigate the criterion employed in (Dokholyan *et al*., 2002). The calculation is identical to the all-site radial distribution previously described, and the results are reported in Figure 1, panel c. The sharp peak at 3.8 Å is identifiable as associated to the C_α–C_α distance between consecutive residues in the chain, so the broad, double peak from 4.9 Å to 6.7 Å is likely to correspond to the C_α–C_α distance for residues in direct, non-covalent contact. The next peaks, centered at 8.6 Å and 10.0 Å, are difficult to distinguish from the baseline expectation and may correspond to indirect contacts with very low associated coupling, thermal or otherwise. Thus, we argue that the 8.5 Å value used in (Dokholyan *et al.*, 2002) is unnecessarily large and may include confounding contacts; we propose that a cutoff of 6.7 Å is adequate to determine contacts based on the distance between C_α atoms.

3.2 Correlation with heat propagation

Having established a protocol to extract representative networks from protein structures, we construct networks for a set of seven proteins (pdb: 1F5J (McCarthy *et al.*, 2000), 1M4W (Hakulinen *et al.*, 2003), 1XNB (see Wakarchuk *et al.*, 1994), 2VUJ and 2VUL (Dumon *et al.*, 2008), 1YS1 (Mezzetti *et al.*, 2005) and 2PRG (Nolte *et al.*, 1998)) for which we have access to energy diffusion data obtained via the Anisotropic Thermal Diffusion (ATD) simulation protocol (Martínez *et al.*, 2011), a modification on conventional molecular dynamics simulations designed to observe and quantify the diffusion of a localized thermal energy excess over a short time scale. For each protein, we examine the final temperature reached after a fixed time as a function of which residue is independently coupled to the hot thermal reservoir. Inspection of the obtained plots reveals that the final temperature varies (approximately) smoothly with residue index, suggesting a mechanism influenced by the local neighborhood of each residue. To investigate whether we can capture this mechanism in network descriptors, we investigate the correlation between final protein temperature and each of a set of several different centrality measures; we discuss definitions and structural rationales for the chosen measures next.

The degree centrality of a node is the number of contacts made by that node in the network, $C_\delta(k) = \delta(k)$ following usual graph theoretical notation. Where it is relevant, the measure is normalized by the maximum possible degree in a graph of N nodes, which is $N - 1$. Although somewhat rudimentary, degree centrality might capture some features of energy transfer in proteins, to the extent on which it may depend exclusively on the local neighborhood of each residue. Degree is also the simplest example of a number of centrality measures based on counting *walks* originating on the target node, as argued in (Benzi and Klymko, 2015); the degree of k, $\delta(k)$, is equivalent to the number of 1-walks which originate at k. We also investigate the longer-ranged subgraph centrality $C_S(k)$, which is the sum of the number of closed walks of *all* lengths originating at node k, weighted by a function of walk length to guarantee convergence and attribute greater importance to shorter walks, as well as eigenvector centrality, $C_E(k)$, which is the fraction of l-walks which originate at k over all l-walks, taken as l goes to infinity and therefore of a much more global character.

The closeness centrality of a node is the inverse of the sum of the distances from it to all other nodes; $C_C(k) = 1/\sum_i d(i, k)$, where $d(i,\ k)$ is the length of the shortest path between nodes i and k. Closeness centrality is normalized by the theoretical maximum $1/(N-1)$ in a graph of N nodes, corresponding to the closeness of a node which is connected to all others. The concept of closeness is introduced as being inversely proportional to the time it takes for a message to reach the entire network when originating from a given node (see Bavelas, 1948), assuming that signals travel exclusively (or highly preferentially) through shortest paths, and by taking into account global features of the network it might also correlate to heat diffusion ability better than degree.

The betweenness centrality of a node is defined as the frequency with which it lies on the shortest paths between all pairs of nodes excluding itself. Formally, the betweenness centrality of node k is given by $C_B(k) = \sum_i \sum_{j>i} \frac{g_{ij}(k)}{g_{ij}}$, where g_{ij} is the number of *geodesics*, that is, of distinct paths of length equal to distance between i and j, and $g_{ij}(k)$ is the number of such paths that contain k. Where it is relevant, betweenness centrality is normalized by the maximum possible betweenness in a graph of N nodes, which is shown to be $\frac{n^2-3n+2}{2}$ for the node in the center of a star graph. Though often included in structural analyses, it is unclear whether betweenness might be an adequate descriptor for dynamical properties such as propensity to diffuse energy. Based on a similar rationale, we also investigate the participation coefficient or P-value, as introduced in (Guimerà and Amaral, 2005). From a given, independent partition of the graph into *modules* (see e.g. Newman and Girvan, 2004), the P-value of a node is a measure of the fraction of its connections which go to nodes in different modules than itself, and is calculated as $P_k = 1 - \sum_m \delta_m(k)/\delta(k)$, where the sum is over all modules and $\delta_m(k)$ is the number of connections from node k to nodes in module m. Here, we use the software described in (Guimerà and Amaral, 2005) to obtain optimal network partitions for each structure before calculating node P-values.

4 Results

4.1 Correlation of centrality measures with heat diffusion

We constructed network models and calculated values for all centrality measures for each residue in each protein, plotting residue centrality as a function of residue index for each protein (see Fig. 2). We then calculated, for each protein, how each centrality plot correlates to a plot of final protein temperature as a function of which residue is initially heated in an ATD experiment. We report the results in Table 1. Betweenness centrality consistently reproduces the ATD data worse than degree or subgraph centrality, which in

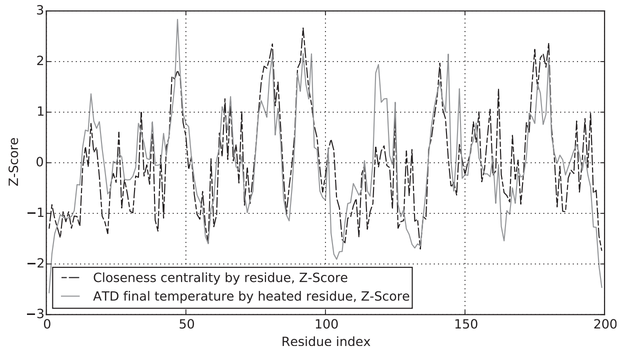

Fig. 2. Correlation between closeness centrality as a function of residue index and final protein temperature as a function of which residue is coupled to the heat bath in an ATD simulation, for the 1F5J structure. Both measures are given as Z-Score, that is, deviation from the mean expressed as number of standard deviations

Table 1. Correlation between centrality measures by residue index and final protein temperature by heated residue, for a set of seven proteins

	Pearson's correlation coefficient (r)						
Centrality Measure	1F5J	1M4W	1XNB	2VUJ	2VUL	1YS1	2PRG
Closeness (C_C)	0.754	0.741	0.723	0.771	0.754	0.746	0.634
Betweenness (C_B)	0.644	0.603	0.606	0.615	0.625	0.588	0.506
Degree (C_δ)	0.723	0.736	0.717	0.723	0.703	0.746	0.546
Subgraph (C_S)	0.708	0.686	0.698	0.724	0.712	0.578	0.542
Eigenvector (C_E)	0.757	0.734	0.724	0.766	0.738	0.689	0.589
Participation coef. (P)	0.374	0.432	0.349	0.527	0.511	0.527	0.513

turn perform worse than closeness and eigenvector centrality. Furthermore, we observe that the participation coefficient varies in this case too sharply between neighboring residues, and thus fails to adequately reproduce ATD data while correctly capturing the general direction of variation. We also produced scatter plots for each correlation (examples are shown in Fig. 3); inspection reveals betweenness centrality and subgraph centrality in particular fail to reproduce the temperature distributions associated to the subset of residues with the weakest coupling to the rest of the structure.

4.2 Model of heat diffusion in a network

Along with the aforementioned measures of centrality, we explore an analytically solvable model for heat transfer in a network, corroborating and expanding on earlier results such as (Szalay and Csermely, 2013), which reports findings similar to those presented in this section. We define a measure of 'local temperature' at each residue, calculated from the average kinetic energy of its atoms in the same fashion as the temperature of the entire system, and write the differential equation for heat diffusion on the network as a function of the local temperatures and the connectivity between residues. We then solve it for a particular case corresponding to a strong local imbalance where a single residue has a much higher initial temperature than all others.

Let $\mathbf{A}_{ij}$ be the elements of the residue adjacency matrix. The total temperature change at time t for residue i will depend on the temperature difference between residue i and all other residues it is connected to:

$$\frac{dT_i(t)}{dt} = k\sum_j \mathbf{A}_{ij}(T_j(t) - T_i(t))$$

where k is a constant analogous to the thermal diffusivity between residues. Breaking up the sum, we write:

$$\frac{dT_i(t)}{dt} = k\sum_j \mathbf{A}_{ij}T_j(t) - kT_i(t)\sum_j \mathbf{A}_{ij}$$

The sum $\sum_j \mathbf{A}_{ij}$ of row i of the adjacency matrix is the degree of node i, $\delta(i)$. We substitute and obtain:

$$\frac{dT_i(t)}{dt} = k\sum_j \mathbf{A}_{ij}T_j(t) - kT_i(t)\delta(i)$$

Solving for the set of $T_i(t)$ as a column vector, we write:

$$\frac{d\mathbf{T}(t)}{dt} = k\mathbf{A}\mathbf{T}(t) - k\mathbf{D}\mathbf{T}(t) = -k(\mathbf{D}-\mathbf{A})\mathbf{T}(t) = -k\mathbf{L}\mathbf{T}(t) \qquad (2)$$

where $\mathbf{D} = \mathrm{diag}\{\delta(1), \delta(2), \ldots, \delta(N)\}$ is the *degree matrix* and $\mathbf{L}$ the *Laplacian matrix* of the graph. This differential equation is readily solvable in terms of the eigendecomposition of $\mathbf{L}$, or in matrix exponential form:

$$\mathbf{T}(t) = e^{-\mathbf{L}kt}\mathbf{T}(0) = \sum_j \{\mathbf{v}_j^T \cdot \mathbf{T}(0)\}e^{-\lambda_j kt}\mathbf{v}_j \qquad (3)$$

where $\mathbf{v}_j$ are the eigenvectors of $\mathbf{L}$ and λ_J are the corresponding eigenvalues. If we consider that, in each run of the simulation, the protein is initially cooled to a very low temperature (10K) and subsequently a single residue is coupled to a room temperature heat reservoir for a fixed time, then we may simplify the solution further by imposing that the initial temperature vector have a single non-zero component, $\mathbf{T}(0) = [\ldots, 0, \theta, 0, \ldots]^T$, where θ is the initial temperature of the heated residue. Letting h be the index of the heated residue, follows:

$$\mathbf{T}(t) = \theta\sum_j [\mathbf{v}_j]_h e^{-\lambda_j kt}\mathbf{v}_j \qquad (4)$$

where $[\mathbf{v}_j]_h$ denotes the hth component of the jth eigenvector. Moreover, the temperature of each particular residue must obey:

$$T_i(t) = \theta\sum_j [\mathbf{v}_j]_h e^{-\lambda_j kt}[\mathbf{v}_j]_i$$

Under Equation (3) (and therefore also Equation (4)), $\mathbf{T}(t)$ reduces over time to a constant vector proportional to the $\mathbf{v}_0 = 1$ eigenvector associated to the smallest eigenvalue $\lambda_0 = 0$ (which is always a solution because every row and column of $\mathbf{L}$ has zero sum), corresponding to the situation where thermal equilibrium has been reached between all residues—evolution under Equation (4) is akin to relaxation of a concentrated pulse. In order to compare with ATD data, one might calculate the relative time to reach equilibrium as a function of which residue is initially heated, or, equivalently, the rate at which energy leaves the residue where it is initially concentrated. If residue h is initially heated, we may approximate, for $kt \ll 1$:

$$T_h(t) = \theta\sum_j e^{-\lambda_j kt}[\mathbf{v}_j]_h^2 = \theta\sum_j (1 - \lambda_j kt + \mathcal{O}((kt)^2))[\mathbf{v}_j]_h^2$$

Discarding terms on the order of $(kt)^2$ and higher:

$$T_h(t) \approx \theta\left(\sum_j [\mathbf{v}_j]_h^2 - kt\sum_j \lambda_j[\mathbf{v}_j]_h^2\right)$$

Therefore:

$$T_h(t) \approx \theta(1 - \mathbf{L}_{hh}kt) = \theta(1 - \delta(h)kt)$$

And the initial dissipation rate is dominated (to first order) by the degree of the heated residue, and more connected residues will drive the dynamics towards equilibrium faster; this observation is reminiscent of the result reported by Moreno and Pacheco (2004), where it is shown that the average time for a scale-free network of coupled oscillators to return to phase synchronization after a perturbation is a function of the degree k of the perturbed node, with exponent $\langle\tau\rangle \sim k^{-0.96}$ very close to –1, although the argument the authors present is oriented towards network topologies where loops are infrequent. Nevertheless, a negative correlation is expected and observed (data not shown) between ATD final temperature data as a function of heated residue and *time to equilibrium* (as calculated by Equation (4)) as a function of heated residue. However, evolution under Equation (3) can be shown to preserve the average temperature, $(1/N)\sum_i \mathbf{T}_i(t) = (1/N)\sum_i \mathbf{T}_i(0), \forall t$, making it in fact unsuitable to model an ATD experiment, where a residue is held coupled to a heat

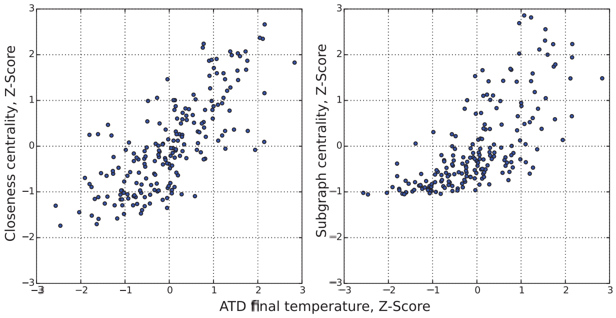

Fig. 3. Correlations between centrality measures by residue index and final protein temperature as a function of heated residue, for the 1F5J structure. Subgraph centrality fails to reproduce the distribution for the subset of residues which, when heated, result in lower than average final temperatures

reservoir for a fixed interval and the average temperature increases monotonically. We can modify the model to account for this, by including a term for the heat reservoir in the sum:

$$\frac{dT_i(t)}{dt} = k_b \mathbf{B}_{ii}(\theta - T_i) + k \sum_j \mathbf{A}_{ij}(T_j(t) - T_i(t))$$

where θ is now the temperature of the heat bath and $\mathbf{B}$ is a diagonal matrix such that $\mathbf{B}_{ii}$ is 1 if residue i is coupled to the bath and zero otherwise. Introducing the substitution $T_i^{\mathrm{rel}} = T_i - \theta$, we can rewrite:

$$\frac{dT_i^{\mathrm{rel}}(t)}{dt} = -k_b \mathbf{B}_{ii} T_i^{\mathrm{rel}} + k \sum_j \mathbf{A}_{ij}(T_j^{\mathrm{rel}}(t) - T_i^{\mathrm{rel}}(t))$$

And, in column vector form:

$$\frac{d\mathbf{T}^{\mathrm{rel}}(t)}{dt} = -k_b \mathbf{B}\mathbf{T}^{\mathrm{rel}}(t) - k\mathbf{L}\mathbf{T}^{\mathrm{rel}}(t) = (-k\mathbf{L} - k_b\mathbf{B})\mathbf{T}^{\mathrm{rel}}(t)$$

which is analogous to Equation (2) and solved by the same technique:

$$\mathbf{T}^{\mathrm{rel}}(t) = e^{(-\mathbf{L}kt - \mathbf{B}k_b t)}\mathbf{T}^{\mathrm{rel}}(0) = e^{-\mathbf{M}kt}\mathbf{T}^{\mathrm{rel}}(0)$$

And thus:

$$\mathbf{T}(t) = e^{-\mathbf{M}kt}\mathbf{T}(0) + (\mathbf{I} - e^{-\mathbf{M}kt})\boldsymbol{\theta} \tag{5}$$

where $\boldsymbol{\theta}$ denotes the vector $[\ldots, \theta, \theta, \theta, \ldots]^T$. Under Equation (5), the average temperature increases monotonically and $\mathbf{T}(t)$ reduces to $\boldsymbol{\theta}$ for large t, adequately modeling ATD behavior. The $\mathbf{M}$ matrix is a perturbation of the Laplacian matrix which includes the influence of the heat reservoir and depends on two free parameters; it is given by $\mathbf{M}kt = \mathbf{L}kt + \mathbf{B}k_b t = (\mathbf{L} + k_r\mathbf{B})\tau$, where $\tau = kt$ is the characteristic time for the coupling between residues and $k_r = k_b/k$ is the relative intensity of the coupling to the heat bath compared to the coupling between residues. Once again, we expect that heating residues which occupy privileged positions in the network will increase the average temperature faster than heating poorly connected ones; to investigate, we approximate, to second order and for $\tau \ll 1$:

Table 2. Correlation between final ATD temperature as a function of heated residue and the same data as calculated by equation (5) and averaged

Model parameters	Protein						
	1F5J	1M4W	1XNB	2VUJ	2VUL	1YS1	2PRG
τ	15.0	18.0	9.5	14.5	11.0	17.0	32.0
k_r	10^3	10^3	10^3	10^3	10^3	10^3	10^3
Pearson's r	0.772	0.774	0.751	0.775	0.749	0.793	0.627

Note: Also presented are the parameters that maximize the observed correlation in each case; increasing k_r further does not alter the correlation coefficients, suggesting that 10^3 is large enough to represent an infinitely strong coupling to the heat reservoir relative to the coupling between residues.

$$\mathbf{T}(\tau) - \mathbf{T}(0) \approx (-\mathbf{M}\tau + \mathbf{M}^2\frac{\tau^2}{2})\mathbf{T}(0) + (\mathbf{M}\tau - \mathbf{M}^2\frac{\tau^2}{2})\boldsymbol{\theta}$$

which averages to:

$$\frac{1}{N}\sum \Delta\mathbf{T}(\tau) \approx \frac{\tau}{N}\left[\sum \mathbf{M}\boldsymbol{\theta} - \sum \mathbf{M}\mathbf{T}(0)\right] + \frac{\tau^2}{2N}\left[\sum \mathbf{M}^2\mathbf{T}(0) - \sum \mathbf{M}^2\boldsymbol{\theta}\right]$$

For simplicity we calculate each term separately. If residue h is initially heated, the action of $\mathbf{M}$ on $\boldsymbol{\theta}$ amounts to the row sum of $k_r\mathbf{B}$, since every row and column of $\mathbf{L}$ has mathvariant="bold"zero sum, and sums to $k_r\theta$. Similarly, the action of $\mathbf{M}$ on $\mathbf{T}(0)$ is the h-th column of $k_r\mathbf{B}$, and sums to $k_r\theta$ as well. In order to calculate the terms involving $\mathbf{M}^2$, we expand it into $\mathbf{M}^2 = (\mathbf{L}^2 + k_r\mathbf{L}\mathbf{B} + k_r\mathbf{B}\mathbf{L} + k_r^2\mathbf{B}^2)$. The only terms that do not sum to zero are $\sum \mathbf{B}^2 \boldsymbol{\theta} = \theta$, $\sum \mathbf{B}^2\mathbf{T}(0) = \theta$ and $\sum \mathbf{B}\mathbf{L}\mathbf{T}(0) = \delta(h)\theta$. We introduce the results in the average above to obtain:

$$\Delta T_{\mathrm{avg}}(\tau) \approx \frac{\boldsymbol{\tau}}{N}\left[k_r\theta - k_r\theta\right] + \frac{\tau^2}{2N}\left[k_r\theta\delta(h) + k_r^2\theta - k_r^2\theta\right]$$

which simplifies to:

$$\Delta T_{\mathrm{avg}}(\tau) \approx \frac{\tau^2}{2N}k_r\theta\delta(h)$$

Therefore, the average temperature rises at an initial rate proportional (to second order) to the bath temperature, the intensity of the coupling to the bath and the degree of the heated residue. Expanding to higher orders reveals the influence of progressively wider neighborhoods around the heated residue; at infinity, $\sum [e^{-\mathbf{M}\tau}]_{hj}$ is recovered as a complete characterization of the connectivity of residue h. We tested the model's predictions for the final average temperature as a function of heated residue (as given by Equation (5)) against ATD data; the parameter space for τ and k_r was scanned independently for each protein in order to maximize the observed correlations, but the obtained values all fell within a very short range of each other. Results are presented in Table 2, and representative plots are shown in Figure 4. The observed correlation

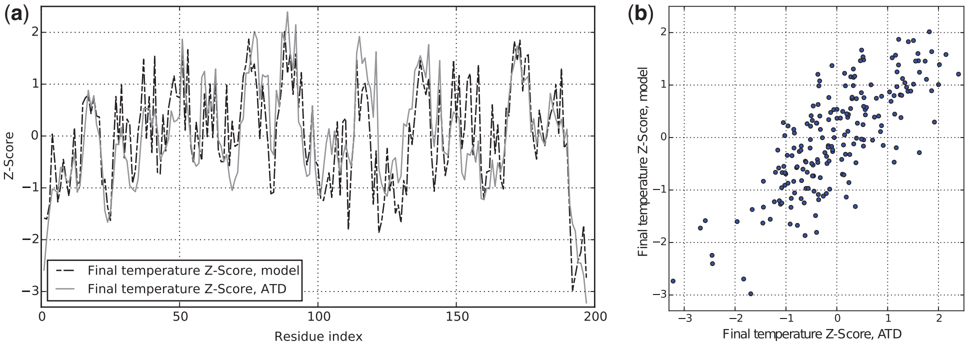

Fig. 4. (**a**) Correlation between ATD final temperature as a function of heated residue and the same data as calculated by equation (5) and averaged, for the 1M4W structure. Values are given as Z-Score, i.e. deviation from the mean expressed as number of standard deviations. (**b**) The same data in scatter plot form

coefficients lie in the same range as those reported in Table 1, but compare favorably to both closeness and eigenvector centrality.

5 Conclusion

Here we have shown that an amino acid residue's ability to efficiently transfer vibrational energy to its neighbors and to the rest of the structure is predicted by its connectivity, which was made evident by the application of graph-theoretic tools on networks representing the contacts between residues in each protein. To construct these networks, we revisited multiple 'contact' definitions encountered in the literature, and developed structural arguments to support a distance criterion of at most 6 Å between any pair of atoms for a thermal coupling to exist. Comparison of the network models with heat transfer data obtained via ATD molecular dynamics experiments identified closeness centrality and the column sums of the $e^{-M\tau}$ matrix as the best descriptors of connectivity with regards to thermal diffusion, which is well supported by their structural interpretations.

It remains open to investigation whether a simple discrimination between covalent and non-covalent contacts, associating a different coupling constant to each type of interaction, can significantly improve the correlation between connectivity and thermal diffusion; many similarly simple refinements may be proposed. It may also be of particular interest to identify those residues for which connectivity *fails* to predict strong (or weak) thermal coupling to the structure; measures such as betweenness centrality or subgraph centrality tend to underperform when predicting weak couplings (see e.g. Fig. 3). Preliminary analyses performed on a set of xylanases considered in this work (data not shown) indicate that there exists a patch of significantly well connected residues along the catalytic cleft which nevertheless exhibit weak thermal coupling to the rest of the protein. This observation, if confirmed, may be related to the maintenance of adequate shape or mobility of the catalytic cleft in the face of perturbations, and would be reminiscent of the results reported in (Bleicher *et al.*, 2011). The connection with heat diffusion established by this work may, then, help shed light on the observations that associate residues of outstanding centrality with folding nuclei and functional sites.

Funding

This work was supported by FAPESP (grants 2010/16947-9, 2013/05475-7 and 2013/08293-7) CNPq (grant 470374/2013-6) and graduate fellowships from CAPES.

Conflict of Interest: none declared.

References

Amitai,G. *et al.* (2004) Network analysis of protein structures identifies functional residues. *J. Mol. Biol.*, **344**, 1135–1146.

Bavelas,A. (1948) A mathematical model for group structures. *Hum. Org.*, **7**, 16–30.

Benzi,M. and Klymko,C. (2015) On the limiting behavior of parameter-dependent network centrality measures. *SIAM J. Anal. Appl.*, **36**, 686–706.

Bleicher,L. *et al.* (2011) Molecular basis of the thermostability and thermophilicity of laminarinases: X-ray structure of the hyperthermostable laminarinase from Rhodothermus marinus and molecular dynamics simulations. *J. Phys. Chem. B*, **115**, 7940–7949.

Böde. *et al.* (2007) Network analysis of protein dynamics. *FEBS Lett.*, **581**, 2776–2782.

Chardonia,J. *et al.* (2004) The ASTRAL compendium in 2004. *Nucleic Acids Res.*, **32**, D189–D192.

Clarkson,M.W. *et al.* (2006) Dynamic coupling and allosteric behavior in a nonallosteric protein. *Biochemistry*, **45**, 7693–7699.

Costa,L.F. *et al.* (2007) Characterization of complex networks: a survey of measurements. *Adv. Phys.*, **56**, 167–242.

del Sol,A. and O'meara,P. (2005) Small-world network approach to identify key residues in protein-protein interaction. *Proteins Struct. Funct. Bioinf.*, **58**, 672–682.

Dokholyan,N.V. *et al.* (2002) Topological determinants of protein folding. *Proceed. Natl. Acad. Sci. U. S. A.*, **99**, 8637–8641.

Dumon. *et al.* (2008) Engineering hyperthermostability Into a GH11 xylanase is mediated by subtle changes to protein structure. *J. Biol. Chem.*, **283**, 22557–22564.

Freeman,L.C. (1978) Centrality in social networks: conceptual clarification. *Soc. Netw.*, **1**, 215–239.

Greene,L.H. and Higman,V.A. (2003) Uncovering network systems within protein structures. *J. Mol. Biol.*, **334**, 781–791.

Guimerà,R. and Amaral,L.A.N. (2005) Cartography of complex networks: modules and universal roles. *J. Stat. Mech.*, P02001.

Gunasekaran,K. *et al.* (2004) Is allostery an intrinsic property of all dynamic proteins? *Proteins Struct. Funct. Bioinf.*, **57**, 433–443.

Hakulinen. *et al.* (2003) Three-dimensional structures of thermophilic beta-1,4-xylanases from *Chaetomium thermophilum* and *Nonomuraea flexuosa*. Comparison of twelve xylanases in relation to their thermal stability. *Eur. J. Biochem.*, **270**, 1399–1412.

Krishnan,A. *et al.* (2008) Proteins as networks: usefulness of graph theory in protein science. *Curr. Protein Pept. Sci.*, **9**, 28–38.

Lampa-Pastirk,S. and Beck,W.F. (2006) Intramolecular vibrational preparation of the unfolding transition state of Zn^{II}-substituted cytochrome *c*. *J. Phys. Chem. B*, **110**, 22971–22974.

Laskowski,R.A. *et al.* (2009) The structural basis of allosteric regulation in proteins. *FEBS Lett.*, **583**, 1692–1698.

Leitner,D. (2008) Energy flow in proteins. *Ann. Rev. Phys. Chem.*, **59**, 233–259.

Liang,J. and Dill,K.A. (2001) Are proteins well-packed?, *Biophys. J.*, **81**, 751–766.

Lo Conte,L. *et al.* (2000) SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res.*, **28**, 257–259.

Martínez,L. *et al.* (2011) Mapping the intramolecular vibrational energy flow in proteins reveals functionally important residues. *J. Phys. Chem. Lett.*, **2**, 2073–2078.

McCarthy,A.A. *et al.* (2000) Structure of XynB, a highly thermostable beta-1,4-xylanase from *Dictyoglomus thermophilum* Rt46B.1, at 1.8 A resolution. *Acta Crystallogr. Sect. D*, **56**, 1367–1375.

Mezzetti,A. *et al.* (2005) Mirror-image packing in enantiomer discrimination molecular basis for the enantioselectivity of *B. cepacia* lipase toward 2-methyl-3-phenyl-1-propanol. *Chem. Biol.*, **12**, 427–437.

Miño-Galaz,G.A. (2015) Allosteric communication pathways and thermal rectification in PDZ-2 protein: a computational study. *J. Phys. Chem. B*, **119**, 6179–6189.

Miyazawa,S. and Jernigan,R.L. (1985) Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules*, **18**, 534–552.

Miyazawa,S. and Jernigan,R.L. (1996) Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J. Mol. Biol.*, **256**, 623–644.

Moreno,Y. and Pacheco,A.F. (2004) Synchronization of Kuramoto oscillators in scale-free networks. *Europhys. Lett.*, **68**, 603–609.

Motlagh,H.N. *et al.* (2014) The ensemble nature of allostery. *Nature*, **508**, 331–339.

Newman,M.E.J. and Girvan,M. (2004) Finding and evaluating community structure in networks. *Phys. Rev. E*, **69**, 2, 026113.

Nolte,R.T. *et al.* (1998) Ligand binding and co-activator assembly of the peroxisome proliferator-activated receptor-gamma. *Nature*, **395**, 137–143.

Ota,N. and Agard,D.A. (2005) Intramolecular signaling pathways revealed by modeling anisotropic thermal diffusion. *J. Mol. Biol.*, **351**, 345–354.

Rhodes,M.E. and Blunt,M.J. (2007) Advective transport in percolation clusters. *Phys. Rev. E*, **75**, 1, 011124.

Ribeiro,A.A.S.T. and Ortiz,V. (2016) A chemical perspective on allostery. *Chem. Rev.*, **116**, 6488–6502.

Song,W. *et al.* (2000) Intermolecular interactions and local density augmentation in supercritical solvation: a survey of simulation and experimental results. *J. Phys. Chem. A*, **104**, 6924–6939.

Szalay,K.Z. and Csermely,P. (2013) Perturbation centrality and turbine: a novel centrality measure obtained using a versatile network dynamics tool. *PLoS ONE*, **8**, e78059.

Tzeng,S. and Kalodimos,C.G. (2011) Protein dynamics and allostery: an NMR view. *Curr. Opin. Struct. Biol.*, **21**, 62–67.

Vendruscolo,M. *et al.* (2002) Small-world view of the amino acids that play a key role in protein folding. *Phys. Rev. E*, **65**, 6, 061910.

Wakarchuk. *et al.* (1994) Mutational and crystallographic analyses of the active site residues of the *Bacillus circulans* xylanase. *Protein Sci.*, **3**, 467–475.