PLOS ONE

RESEARCH ARTICLE

Protein dynamic communities from elastic network models align closely to the communities defined by molecular dynamics

Sambit Kumar Mishra[1,2], **Robert L. Jernigan**[1,2]*

1 Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, United States of America, **2** Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology, Iowa State University, Ames, Iowa, United States of America

* jernigan@iastate.edu

OPEN ACCESS

Citation: Mishra SK, Jernigan RL (2018) Protein dynamic communities from elastic network models align closely to the communities defined by molecular dynamics. PLoS ONE 13(6): e0199225. https://doi.org/10.1371/journal.pone.0199225

Editor: Yang Zhang, University of Michigan, UNITED STATES

Received: February 6, 2018

Accepted: June 4, 2018

Published: June 20, 2018

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This research was supported by NIH grant R01-GM72014 and NSF grant DBI-1661391, as well as funds from the Carver Trust awarded to the Roy J. Carver Department of Biochemistry, Biophysics and Molecular Biology.

Competing interests: The authors have declared that no competing interests exist.

Abstract

Dynamic communities in proteins comprise the cohesive structural units that individually exhibit rigid body motions. These can correspond to structural domains, but are usually smaller parts that move with respect to one another in a protein's internal motions, key to its functional dynamics. Previous studies emphasized their importance to understand the nature of ligand-induced allosteric regulation. These studies reported that mutations to key community residues can hinder transmission of allosteric signals among the communities. Usually molecular dynamic (MD) simulations (~ 100 ns or longer) have been used to identify the communities—a demanding task for larger proteins. In the present study, we propose that dynamic communities obtained from MD simulations can also be obtained alternatively with simpler models–the elastic network models (ENMs). To verify this premise, we compare the specific communities obtained from MD and ENMs for 44 proteins. We evaluate the correspondence in communities from the two methods and compute the extent of agreement in the dynamic cross-correlation data used for community detection. Our study reveals a strong correspondence between the communities from MD and ENM and also good agreement for the residue cross-correlations. Importantly, we observe that the dynamic communities from MD can be closely reproduced with ENMs. With ENMs, we also compare the community structures of stable and unstable mutant forms of T4 Lysozyme with its wild-type. We find that communities for unstable mutants show substantially poorer agreement with the wild-type communities than do stable mutants, suggesting such ENM-based community structures can serve as a means to rapidly identify deleterious mutants.

Introduction

The dynamic nature of globular proteins allows them to sample multiple conformations around their native equilibrium conformation. Such intrinsic dynamics is conferred by their geometry and can be influenced by events such as ligand binding or even binding of a partner enzyme [1]. Such events typically shift the conformational equilibrium of proteins allowing

them to sample new conformations by lowering energy barriers, which were not accessible from the native state [2,3]. Such dynamic plasticity is characteristic for protein function [4–6]. It facilitates signal transduction through allosteric regulation as well as allowing bio-molecular machines to undergo large scale conformational changes from their native state essential for their function [7–9].

Inspecting the conformational ensemble arising due to the dynamic nature of proteins gives immediate insight into how different parts of a protein move with respect to one another. Some regions may exhibit highly correlated motions while others may be anti-correlated in their motions. A map describing the extent of inter-residue dynamical correlation between residues can then be used to create a graphical representation which portrays the dynamic nature of a protein [10]. In such a graph, the nodes represent the residues and the edges are weighted by the dynamical correlation for a residue pair. Residue blocks which are highly correlated in their motions and move as a cohesive unit can then be identified from these graphs and are commonly referred to as dynamic communities [11,12]. These communities may correspond to structural domains in proteins; however, they are often smaller modules whose motions relate to the protein's function.

Previous studies have used both normal mode analysis (NMA) and molecular dynamics (MD) approaches to detect structural domains and dynamic communities in proteins. Hinsen *et al.* [13] used normal modes to compute residue-level deformation energy and then, identified dynamically rigid segments using a threshold based on the deformation energy. Kundu and co-workers [14] used Gaussian Network Model (GNM) [15] to partition protein structures into domains using the eigenvector corresponding the lowest non-zero eigenvalue, also referred to as the Fiedler vector. In another study, Yesylevskyy *et al.* [16] used GNM to obtain a correlation matrix describing inter-residue dynamics and used it to calculate a "correlation matrix of correlation patterns" which essentially describes the overlap between the correlation patterns for different residues. Then they performed hierarchical clustering on this matrix to obtain rigid communities. A similar study used correlations in residue dynamics calculated from normal mode analysis to decompose protein kinases into residue blocks that are dynamically cohesive [17].

Other studies where MD simulations were used to identify the rigid domains have also been carried out. Potestio *et al.* [18] used MD simulations to obtain conformational ensemble describing the essential dynamics of proteins and then used dominant eigenvectors from covariance matrix describing the variation in the ensemble to identify rigid domains. McClendon and co-workers [10] performed a thorough investigation of protein kinase A using microsecond-scale MD simulations and then identified communities using inter-residue dynamical correlations from the trajectory with the Girvan-Newman clustering scheme to understand the mechanism of allostery in the enzyme. A similar study on Bruton's tyrosine kinase by Chopra *et al.* [19] revealed that inspecting the community changes for the enzyme's mutant form reveals the changes in the allosteric coupling in the enzyme. In another study, Yao and co-workers [20] performed community analysis on G proteins using 80-ns MD simulations to identify residues playing a critical role in the allosteric coupling between functional domain interfaces.

MD simulations do provide a high resolution dynamic image of a protein describing detailed motions of individual atoms at different time points. However, most proteins require energy minimization with respect to an all-atom potential prior to any simulation, a computationally demanding task for larger structures. Moreover, to observe large-scale conformational changes as often seen in the case of multi-domain proteins, simulations need to be performed on the microsecond to millisecond time-scales, which also require considerable computing power. In such cases, coarse-grained approaches like ENM have an upper hand [15,21,22].

These models adopt a coarse-grained representation for proteins by representing each residue by only its alpha carbon (C^{α}). They also implement a simplified potential that uses Hookean springs to connect residue pairs within a cutoff distance to calculate the native state dynamics for proteins. In assuming that the crystal structure of a protein corresponds to a local minimum on the energy landscape and considering it as the native state conformation, these models eliminate the necessity for energy minimization. Owing to their reduced nature, these models require minimal computational resources even for large macromolecular structures. Previous studies have shown that theoretical B-factors calculated using ENM correspond well to the experimental temperature factors [15,21,22]. A study by Leioatts *et al*. suggests that these models provide consistent outcomes irrespective of the details of their formulations and thus, do not strongly depend on their underlying parameters [23]. In addition, normal modes from ENMs show significant overlaps with principal components from both experimental sets of structures as well as with MD ensembles [24] and tuning the inter-residue Hookean springs further improves the correspondence with MD [25]. Comparing the dynamics between ENM and MD also suggests that collective motions obtained with ENM from alternate conformations of a macromolecular complex cannot be reliably obtained using multiple runs of MD simulations [26]. Besides, when supplemented with MD, ENMs have also found their applications for generating conformers along transition pathways [27].

In this study, we have performed a large set of comparisons between the dynamic communities obtained from GNM [15] (a type of ENM) and from MD for a set of 44 non-redundant proteins. After applying a systematic hierarchical clustering scheme on the dynamic cross-correlation matrices, we observe a close correspondence between the communities from GNM and MD for specific community levels, characterized by a significantly high value of Cohen's kappa coefficient [28]. Centrality measures for the weighted dynamic network from GNM and MD also reveal a strong correlation for the closeness centrality values. We also verify the extent of agreement for the inter-residue cross-correlations between GNM and MD by investigating the overlaps of the principal eigenvectors calculated from the dynamic cross-correlation matrices and observe a good overlap. A further analysis of the effect of mutations on communities derived using GNM for T4 lysozyme confirms that highly deleterious point mutations significantly alter the community structure when compared to the neutral mutations. The results from our study open up new avenues for mining dynamic communities in macromolecular structures with ENM and using their changes to screen for deleterious mutants.

Results

We perform our study on a set of 44 non-redundant proteins (see S1 Table) taken from the MOlecular Dynamics Extended Library (MODEL) database [29]. Each protein has a minimum simulation time of 100 ns for its MD trajectory. We consider only the positions of the residue alpha-carbon atoms of each protein from the trajectory file and calculate the inter-residue dynamical correlations from the respective MD trajectory (DCC_{MD}) using equation 1. In our procedure we consider only the first frame of the MD trajectory of a given protein as its representative structure to render the protein as a mass-spring system. In such a system, each residue is represented by a point mass (its C^{α} atom) and residue pairs within a given distance cutoff (r_c) are connected by hypothetical Hookean springs. Such a model is commonly referred to as an elastic network model. The Gaussian Network Model is a formulation of ENM that assumes residue fluctuations to be isotropic in nature. Details concerning the implementation of GNM are provided in the Materials and Methods section.

We construct GNM for each protein by setting the distance cutoff r_c to 7.5 Å and calculate the inter-residue dynamical correlations (DCC_{GNM}) using a subset of 5, 10, 20, 30 and 50

modes (Eq 5). The choice for r_c was based on a preliminary analysis where we identified the r_c that gave the best overlaps between simulation results from GNM and MD. This is followed by a systematic comparison between the inter-residue dynamical correlations from MD and GNM. Initially, we show how closely the dynamic cross-correlation (DCC) matrices from MD and GNM compare with each other for two randomly selected proteins. Following this, we perform more thorough comparisons using the three metrics described below.

i. **Kappa coefficient.** The DCC matrix for a protein describes the extent of correlation between the pairs of its C^{α} atoms. We identify blocks of residues that move cohesively (dynamic communities) by first clustering the DCC matrix hierarchically and then, using a cutoff on the height of the dendrogram obtained to identify the required number of communities (N_c). In the present study, we identify 2–10 communities (N_c = 2, 3, 4 . . ., 10) for a given protein. Agreement between the communities from MD and GNM is then assessed with kappa coefficient [28,30].

ii. **Network centrality.** We model each protein as a weighted network with the nodes corresponding to residues and edges between pairs of residues weighted by their distance transformed dynamical correlations (Eq 6 and Eq 7). Then, we calculate the residue-level closeness centralities and verify the correlations for the centralities obtained from MD and GNM.

iii. **Overlap between principal eigenvectors.** To assess how well the correlation matrices obtained from MD and GNM compare for a protein, we also perform singular value decompositions of the matrices and then use root-mean square inner product (RMSIP) to evaluate the extent of overlap between the principal eigenvectors from the two systems.

In the final section of this paper, we use GNM to delineate the community structure of wild-type and mutant forms of T4 Lysozyme and to show that elastic models can capture the difference in community structures for the wild-type and mutant forms.

DCC maps from MD and GNM

We perform an initial visual inspection of the dynamic maps obtained from MD and GNM to understand the overall extent of agreement for residue correlations from the two methods. Fig 1 describes the dynamic map for two randomly selected proteins from our dataset; *top*: copper transporter domain from copper transporting ATPase (PDB 1fvq), *bottom*: alpha-chymotrypsinogen (PDB 1cgi). The figure shows the distance map between C^{α} atoms (*A*, *D*), distance transformed DCC maps from MD (*B*, *E*) and GNM (*C*, *F*) for the two molecules. We calculated the DCC map for GNM by setting the distance cutoff r_c to 7.5Å and then considering only the 20 non-zero lowest frequency modes as these have often been shown to circumscribe the most energetically favorable conformation fluctuations in proteins [31]. The diagonal elements of the correlation maps describe fluctuations of individual residues while off diagonal elements describe inter-residue correlations or cross-correlations. We note from the outset that there are strong similarities among these representations, corresponding to the secondary structures present in these structures.

The distance map for a protein provides information about the spatial proximity of residues. Spatially close residues are naturally expected to have high correlations in their dynamics. For the two proteins, we observe both MD and GNM showing high inter-residue dynamical correlations for the spatially close residues. However, it is interesting to notice that correlations for residues in spatial proximity are more strongly indicated with the GNM than by MD. The distance transformed cross-correlation and hence, the corresponding cross-

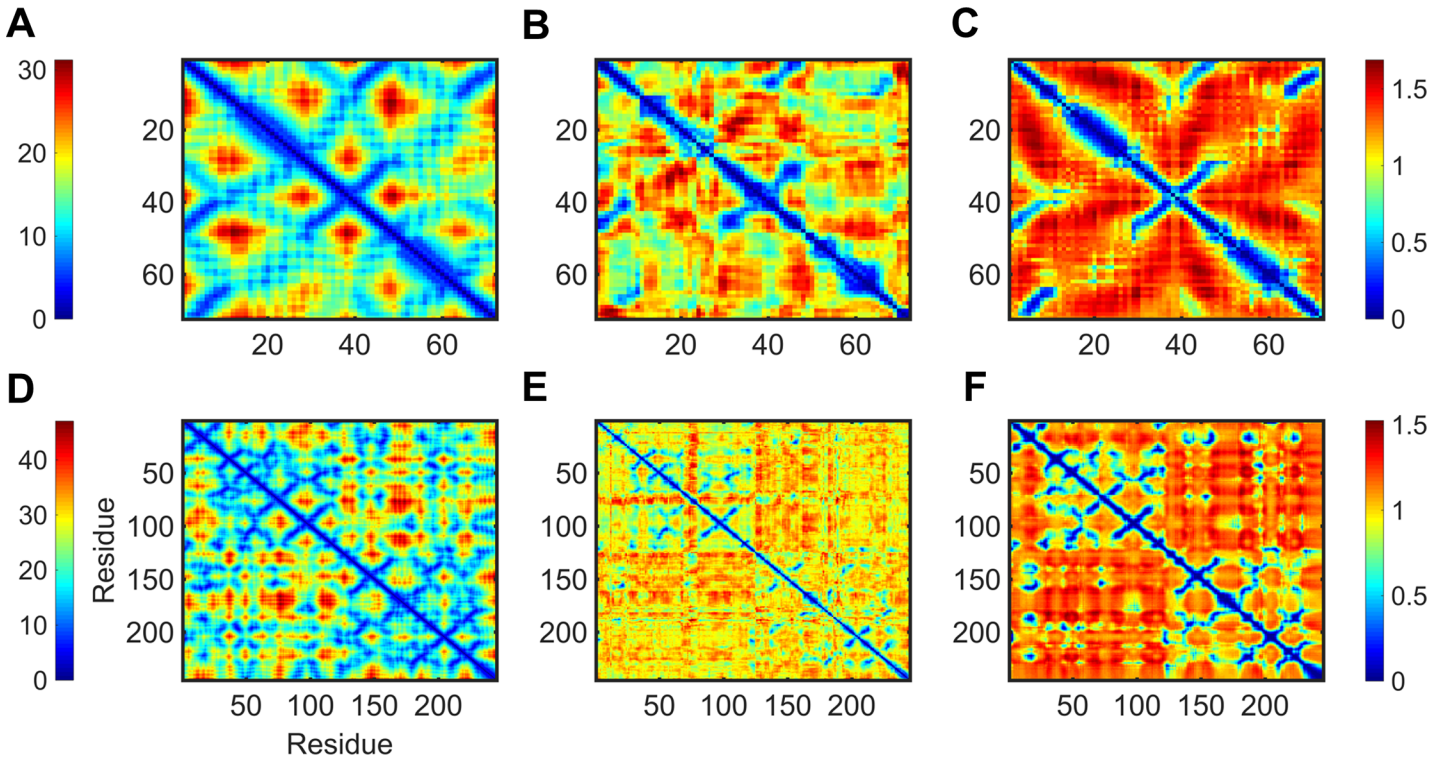

Fig 1. Examples of C^{α}-distance maps and distance transformed dynamic cross-correlations from MD and GNM for *i.* Copper transporter domain from copper transporting ATPase (*top*), and *ii.* alpha-chymotrypsinogen (*bottom*). For each protein, the figure shows the distance map for alpha-carbons (A and D), $dist\_DCC_{MD}$ (B and E) and $dist\_DCC_{GNM}$ (C and F). The color scale ranges from red (spatially distant regions and least correlated parts) to blue (regions in spatial proximity and most correlated parts). The PDB IDs of the structures used are 1fvq and 1cgi, for *i* and *ii* respectively. For ease of comparison with the C^{α}-distance maps, we use *dist_DCC* which has all values on a positive scale rather than *DCC* that has both positive and negative values.

https://doi.org/10.1371/journal.pone.0199225.g001

correlation maps from MD and GNM exhibit good overall agreement. It is also worth noting that for alpha-chymotrypsinogen, the blocks of residues with high inter-residue dynamical correlation in MD ([1–70], [80–120, 1–70] and [120–220]) are almost closely replicated by GNM. Moreover, the extent of similarity in the correlation profiles of the secondary structure elements (helical regions along the diagonal and anti-parallel beta strands perpendicular to the diagonal) for MD and GNM is quite remarkable.

Metric based comparisons

i. Kappa coefficient. Our objective is to investigate the level of similarity between the communities obtained from MD and GNM. As we identify a range of communities for a protein (N_c = 2, 3, 4 . . ., 10), we perform a one-to-one comparison between MD and GNM for a given N_c. To this end, we first calculate for each protein, the dynamic cross-correlation maps for MD (DCC_{MD}) with Eq 1. Then, we construct GNMs for all proteins using r_c = 7.5 and calculate DCC_{GNM} using a subset of the low-frequency modes: 5, 10, 20, 30 and 50 modes (Eq 4 and Eq 5). We thus have 5 correlation matrices for a given protein each calculated using a specific subset of modes described above. For a given protein, we then perform hierarchical clustering on the distance transformed DCC_{MD} (Eq 6) and DCC_{GNM} (Eq 7) and then truncate the resulting dendrogram to get 2–10 communities. Using kappa coefficient (Eq 8) [28,30], a metric which is used to test inter-rater reliability (extent of agreement between data collectors in assigning same scores to the same variables), we then determine the extent of similarity between the communities from MD and GNM.

For a given protein, we consider the N_c (2, 3, 4 . . . 10) that yields the maximum kappa coefficient ($Kappa_{max}$) for a chosen subset of modes. For example, if we choose the subset of modes used to calculate DCC_{GNM} as the first 10, we first calculate the kappa coefficient for all N_c and then choose the particular N_c that gives maximum kappa coefficient and thus, maximum agreement between MD and GNM. Fig 2 shows the median of $Kappa_{max}$ for each subset of modes used. Similar to correlation coefficients, the kappa coefficient can range from -1 to 1. A value of -1 indicates complete disagreement whereas, 0 indicates the random case. It can be seen that for all subsets of modes used, the median value for $Kappa_{max}$ is at least 0.5, indicating that the agreement is reasonably good and is not just random. Details of $Kappa_{max}$ obtained for individual proteins and the respective N_c are provided in S2 Table. We also consider all kappa coefficients for all community levels obtained using the distance cutoff 7.5Å and calculate the median kappa for each subset of modes (S3 Table and S1 Fig). As might be expected, the median kappa when considering all community levels for each subset of modes is smaller than the median of $Kappa_{max}$ (≈ 0.41). Considering the fact that the conformations sampled by MD might be limited, biased by the trajectory time scale whereas ENMs can sample a relatively broader ensemble independent of time, a kappa coefficient of 0.4 indicates fair agreement between the communities but importantly, the agreement is not random.

In Fig 3, we show the communities from MD and GNM mapped onto the structures of 5 proteins (A. Angiogenin, B. Protease, C. Guanine nucleotide dissociation inhibitor, D. Hemoglobin, and E. Ubiquitin). For each protein, the figure shows only the community level N_c that provides the best agreement with MD. The figure clearly depicts the close agreement between the communities from GNM and from MD.

ii. Network centrality. The node centrality is computed by modeling a protein as a network where nodes are the C^{α} atoms and the edges are weighted by the correlation in dynamics

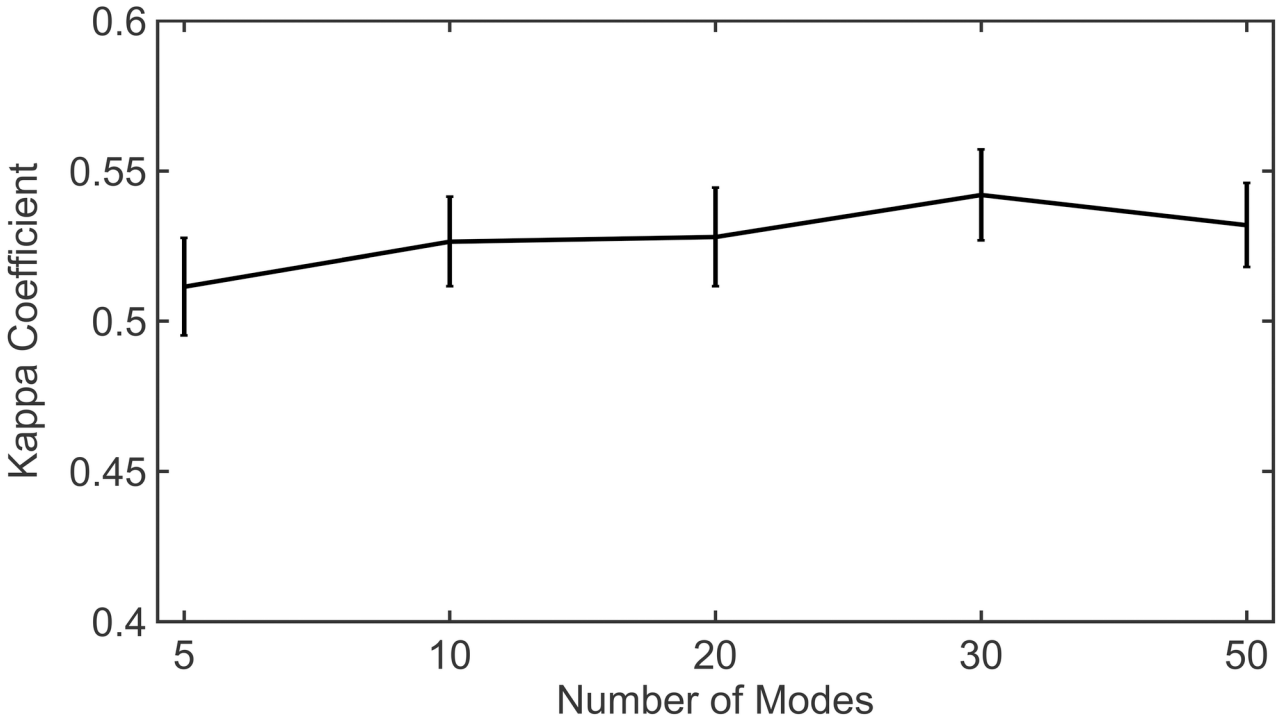

Fig 2. Variation of kappa coefficient with the number of modes. The figure shows the median $Kappa_{max}$ for all proteins in the dataset for subsets including 5, 10, 20, 30 and 50 modes. Vertical bars represent the standard error of $Kappa_{max}$.

https://doi.org/10.1371/journal.pone.0199225.g002

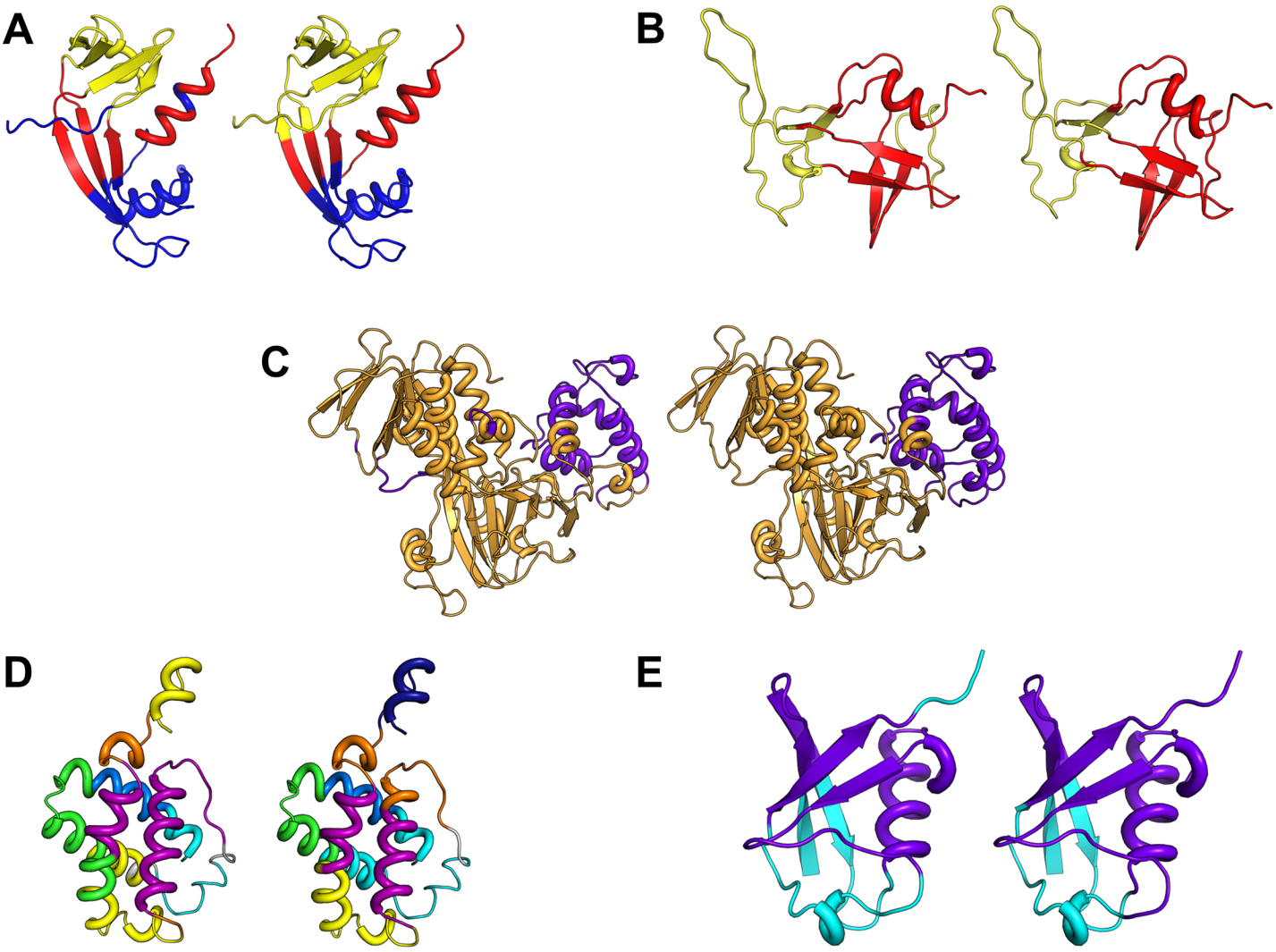

Fig 3. Comparison of communities from MD and GNM. Mapped communities for five proteins. (A) Angiogenin (PDB ID: 1agi), (B) Protease (PDB ID: 1nso), (C) Guanine nucleotide dissociation inhibitor (PDB ID: 1gnd), (D) Hemoglobin (PDB ID: 1idr), (E) Ubiquitin (PDB ID: 1ubq). The number of communities (N_c) shown for each case corresponds to the case of maximum agreement between MD and GNM given by $Kappa_{max}$. DCC_{GNM} calculated with a subset of 20 low-frequency modes was used for each protein to perform calculations for communities.

https://doi.org/10.1371/journal.pone.0199225.g003

between a residue pair. Centrality measures tell us the importance of nodes in facilitating the flow of information within the network [32–34]. The most central nodes act as hubs and can be essential to the transmission of information between nodes at the extreme ends of the network. We compare the extent of correlation for residue centralities between GNM and MD.

We consider the residue closeness centrality, which is the cumulative sum of the lengths of the shortest paths from the residue to all other residues [35,36]. It is also defined as the reciprocal of farness. The centralities calculations were performed using the distance transformed DCC_{GNM} and DCC_{MD} (Eq 6 and Eq 7). Fig 4 shows the correlations for the node closeness between MD and GNM where it can be seen that both methods show significantly high correlations in their centralities. It is worth noting that although the maximum correlation is obtained using 50 modes (≈ 0.63), a steep rise in the curve is observed only until 20 modes, after which the curve has almost converged. S4 Table describes the correlations for residue closeness centralities obtained using each subset of modes for individual proteins.

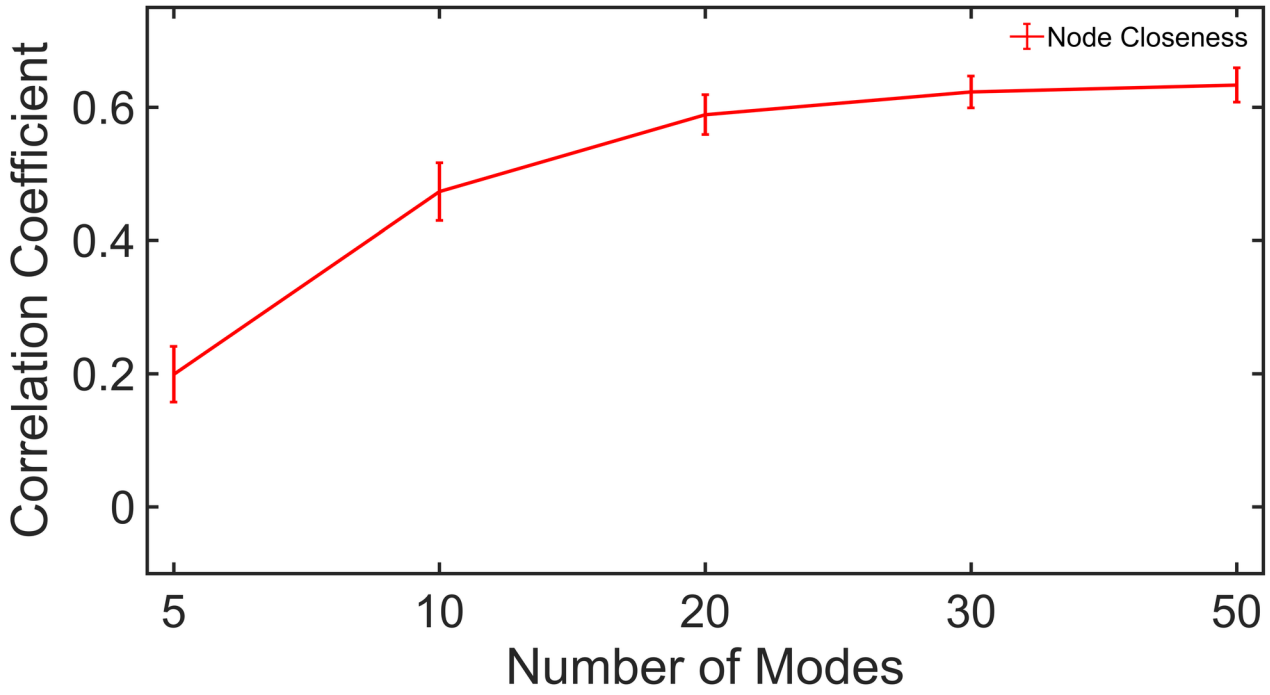

Fig 4. Node centrality correlations. The median correlation for closeness centrality from DCC_{GNM} with DCC_{MD} is shown for different subsets of modes for all proteins. Vertical bars give values of standard errors.

https://doi.org/10.1371/journal.pone.0199225.g004

iii. Overlap between principal eigen vectors. How well do the dominant motions captured from DCC_{GNM} quantitatively compare with DCC_{MD}? How many low-frequency GNM modes are required to closely reproduce the correlation pattern from MD? To answer these questions, we investigate the extent of overlap between the principal eigenvectors from DCC_{GNM} and DCC_{MD}.

Let U^N and V^N be the set of N principal eigenvectors obtained upon singular value decomposition (SVD) of DCC_{GNM} and DCC_{MD}. By principal eigenvectors we are referring to the set of eigenvectors with highest eigenvalues. Because the DCC matrix is comparable to a covariance matrix, vectors U_i and V_i are comparable to the principal components of a covariance matrix, capturing the directions of maximum variance from the residue cross-correlation matrix. We inspect the overlap between U and V using root-mean square inner product (RMSIP) (Eq 9) and quantitatively evaluate the extent of similarity between the two matrices. It is also to be noted that we consider the same number of principal eigenvectors each from U^N and V^N as the subset of modes used. Details about the calculation of RMSIP are provided in Materials and Methods. In Fig 5, we show that the overlaps between the principal eigenvectors of the DCC_{GNM} and DCC_{MD} matrices are high. The figure also depicts sharp increases in RMSIP and hence, a steep positive gradient as the subset of modes selected increases from 5 to 10 following which the curve converges. S5 Table gives the RMSIP values of individual proteins for different subsets of low-frequency modes.

Changes to dynamic communities upon mutations

Mutations can lead to changes in the structure of dynamic communities [19]. We hypothesize that highly unstable mutations tend to change the community structure in a protein more radically than mutations that are less unstable. To test this, we consider 16 mutant structures of T4 Lysozyme crystallized and reported by Mooers *et al* [37]. In their study, the authors

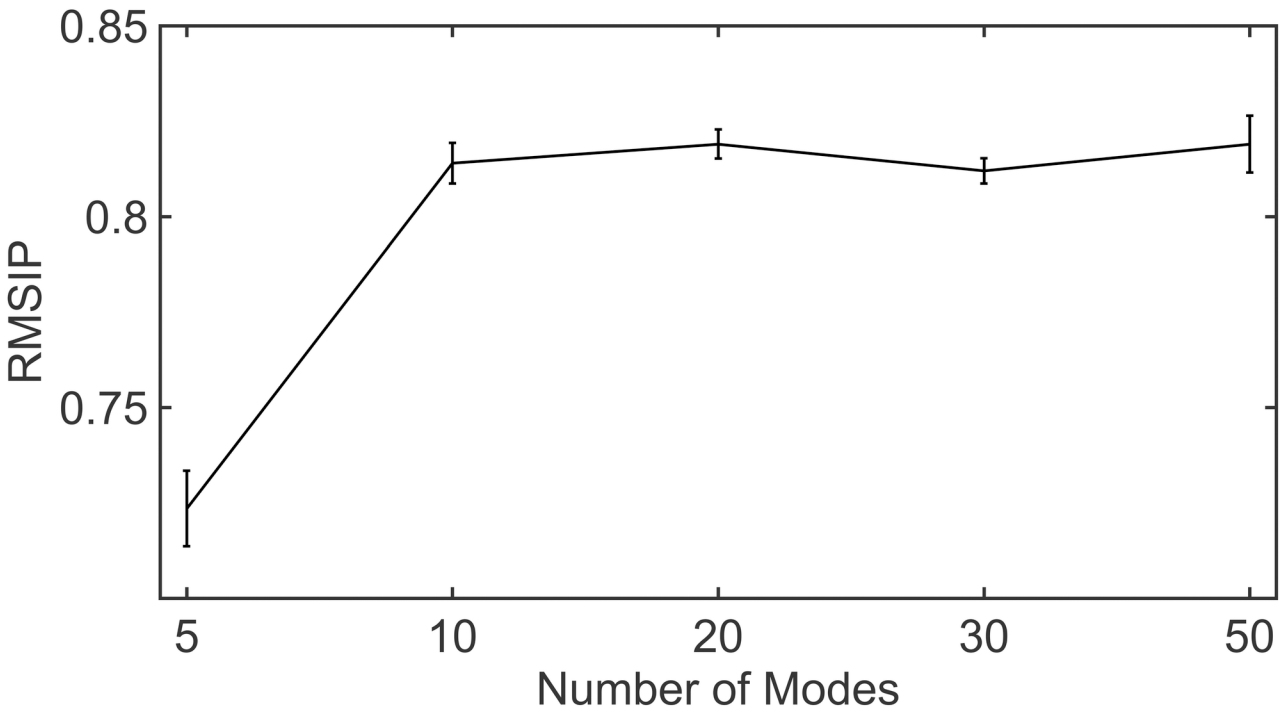

Fig 5. Overlap between principal vectors from DCC_{GNM} with DCC_{MD}. The figure shows the extent of agreement between the residue cross-correlation matrices from MD and GNM in terms of the principal eigenvectors. The principal eigenvectors are obtained from singular value decomposition of the DCC_{GNM} and DCC_{MD} matrices, respectively. The median overlap between the vectors from MD and GNM, computed with RMSIP, is shown for subsets of 5, 10, 20, 30 and 50 modes. Vertical bars represent the standard errors in RMSIP.

https://doi.org/10.1371/journal.pone.0199225.g005

investigated the effect of mutating Arg96 on the stability of the enzyme. $\Delta\Delta G$ values were reported that indicate changes in the stabilities relative to the wild-type (Table 1). The more negative numbers indicate higher instability. We arbitrarily divide the dataset into two groups: the more unstable mutants (rows 1–8) having $\Delta\Delta Gs$ between -4.7 and -2.6 and less unstable mutants (rows 9–16), $\Delta\Delta Gs$ varying between -2.6 and 0. For simplicity, we refer to the more unstable type as *unstable* and the less unstable type as *stable*. We obtain the dynamic communities with GNM using all heavy-atoms from the atomic protein structures and then, with DCC_{GNM} from 5, 10, 20, 30 and 50 modes, we verify the community agreement for each of the two mutant types with the wild-type with the kappa coefficients.

In Fig 6, we show the variation in kappa coefficient for the two mutant categories. For each category, the plot shows the median kappa for individual community levels. It is seen that the *stable* mutants (blue curve) exhibit better agreement with the wild-type than the *unstable* mutants (red curve). Also, it is interesting to note that these differences are manifested in the first 6 communities. At higher community levels, the two mutant types almost come into agreement. It is also interesting to note that this difference in community architecture is more apparent for a subset of 10 modes. To visualize these differences on the protein structures, we consider 3 pairs of *unstable* and *stable* mutants: (PDB IDs: 3c80, 3c81), (PDB IDs: 3c82, 3c81) and (PDB IDs: 3c82, 3c8s). For each pair, we identify the smallest number of communities for which the change is significant. The $\Delta\Delta G$ for each of these mutants can be seen in Table 1.

Fig 7 (3c80, 3c81), Fig 8 (3c82, 3c81) and Fig 9 (3c82, 3c8s) show communities for each mutant pair relative to the wild-type (4s0w). In each figure, the wild-type structure with the communities is shown on left, the *stable* mutant in the center and the *unstable* mutant on the right. Side chains of mutation sites are shown as sticks with the same residue side chains

Table 1. Mutants for T4 Lysozyme sorted by *ΔΔG*. The set of PDB structures used to compare the community structure of stable and unstable mutants is given below. The Mutation column gives information on the mutation and has the format "xRy", where 'x' is the residue in the wild-type, 'y' the residue in the mutant, and R is the position of mutation in the protein. More negative ***ΔΔG* values** indicate less stable mutant form.

PDB Identifier	Mutation	*ΔΔG* (pH 5.35)	Stability
3c80	R96Y	-4.7000	Unstable
3fi5	R96W	-4.5000	Unstable
3c7z	D89A, R96H	-3.8000	Unstable
3c82	K85A, R96H	-3.6000	Unstable
3c8q	R96D	-3.5000	Unstable
3cdt	R96N	-3.0000	Unstable
3cdv	R96M	-2.7000	Unstable
3c8r	R96G	-2.6000	Unstable
3cdq	R96S	-2.6000	Stable
3c8s	R96E	-2.5000	Stable
3cdo	R96V	-2.4000	Stable
3c7y	R96A	-2.0000	Stable
3c81	K85A	-0.6000	Stable
3c83	D89A	-0.5000	Stable
3cdr	R96Q	-0.3000	Stable
3c7w	R96K	0.0000	Stable
4s0w	None (wild-type)	0	Stable

https://doi.org/10.1371/journal.pone.0199225.t001

displayed in the same color. In Fig 7, the difference in community structure for 3c80 (*unstable*) and 3c81 (*stable*) is distinct showing two different communities. The *stable* and *unstable* forms differ visibly in the dynamical correlation of the N-terminal helix (residues 1–12), which is cohesive with the adjacent N-terminal beta sheets and helices in the wild-type and stable forms, while it moves in coordination with the C-terminal domain in the unstable form. The kappa coefficient for the unstable and stable mutant structures is 0.74 and 0.98, respectively. For 3c82 (*unstable*) and 3c81 (*stable*) (Fig 8), the difference is apparent at 3 communities (kappa values of 0.65 and 0.97 respectively). Again we observe a change in the N-terminal helix that moves as an independent unit in the wild-type and *stable* forms, but shows more coordinated motion with the N-terminal domain in the *unstable* form. In Fig 9, we notice the difference at 3 communities and as previously observed, the difference between the *stable* and *unstable* forms becomes visible in the N-terminal helix. The kappa coefficients for the *unstable* (3c82) and *stable* (3c8s) forms at the level of 3 communities are 0.65 and 0.94, respectively.

Discussion

In the present study, we focus on a simple approach for detecting dynamic communities in proteins with elastic networks. ENM is simpler to formulate, easier to implement and is computationally less expensive in comparison with MD. Here, we emphasize that identifying the true number of dynamic communities is largely an unsolved problem and it is not our current goal to establish ENM as a more accurate method than MD in this aspect. Rather, it is in our interest to show that this method works as well as MD does for community detection. Our results reveal that this single-parameter model can closely reproduce the results from a complex, multi-parameter model like MD, especially for community detection. Owing to its reduced nature, ENM is superior to MD in terms of execution time and thus, can contribute significantly to the investigation of the dynamic communities for larger proteins. We would also like to emphasize that simulation results from MD may not always fully capture the near-

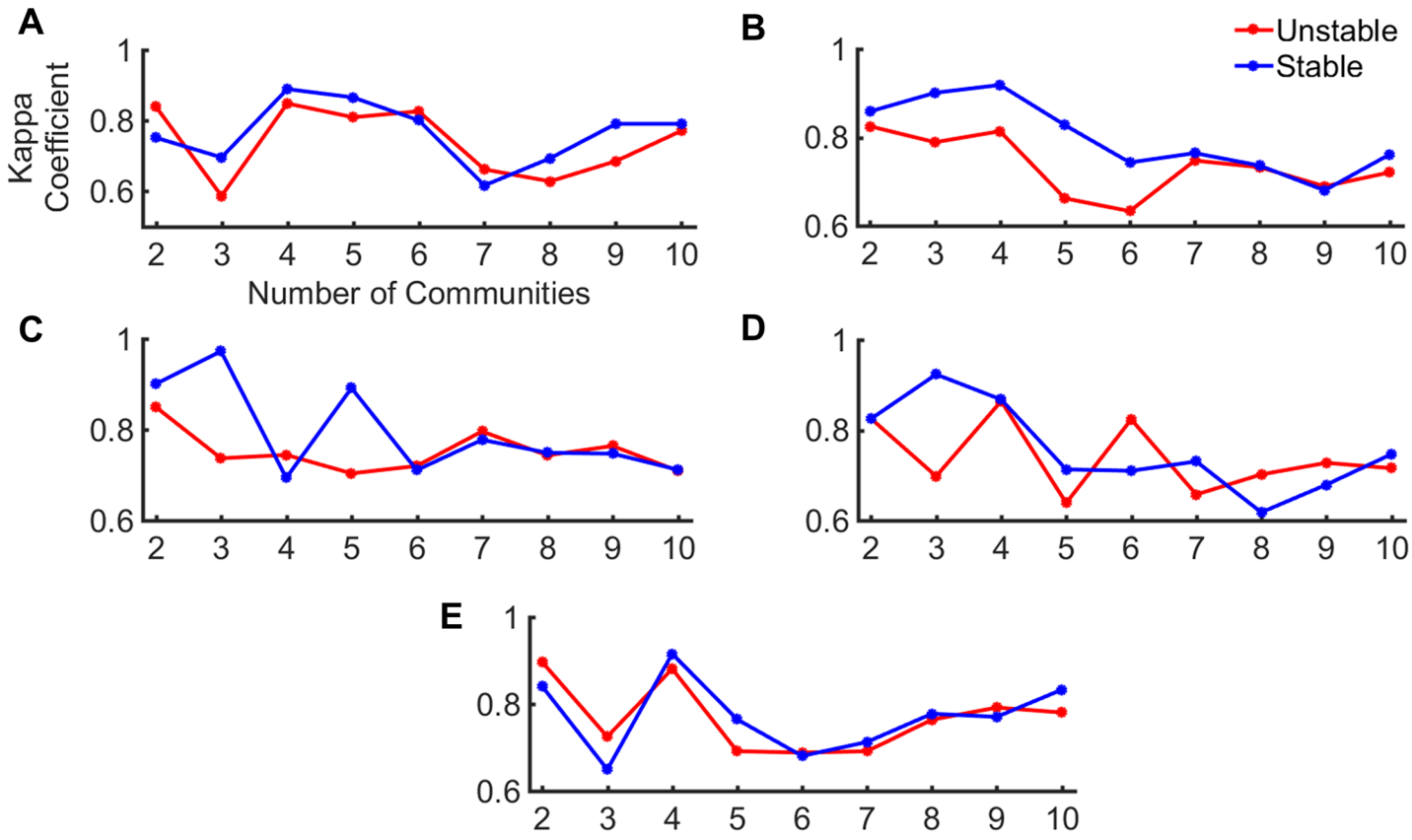

Fig 6. Community agreement for *unstable* (*red*) and *stable* (*blue*) mutants of T4 lysozyme with the wild-type. The figure shows the median kappa coefficient (agreement with wild-type) at each community level for the *unstable* and *stable* mutants. The communities were obtained with DCC_{GNM} calculated using (A) 5, (B) 10, (C) 20, (D) 30 and (E) 50 low-frequency modes. The abscissa and ordinates correspond to the number of communities and the Kappa coefficient respectively, as given in 6A.

https://doi.org/10.1371/journal.pone.0199225.g006

native conformation ensemble for a given protein and thus, one should not view results from MD as the absolute truth. The conformational sampling using MD may be highly biased by the simulation length vis a vis the size of the protein, with larger proteins requiring longer simulations to capture a fully representative ensemble of near native conformations. Thus, in our scheme of comparing communities and the underlying correlation matrices obtained from ENM with MD, a lack of agreement between MD and ENM does not necessarily imply the inability of ENM to capture the underlying conformational dynamics. Instead, in some cases, this could be related to the underlying sampling inaccuracies arising from MD.

We show that communities extracted using GNM, a simple formulation of ENM, exhibit a considerable similarity to the communities from MD. We choose GNM over its anisotropic counterpart ANM [21] because it is simpler and because previous studies have shown that GNM exhibits better correlations with experimental B-factors than ANM [38]. Moreover, in a preliminary analysis we observe that the communities obtained with GNM show better agreement with MD than does ANM. In Fig 1, the distance transformed DCC_{GNM} and DCC_{MD} matrices for two proteins selected randomly from our dataset show considerable agreement for the regions with high correlation in their dynamics. However, it is surprising to notice a better cohesive behavior, in the case of GNM, showing a close connection between inter-residue dynamical correlations and residue spatial proximity. The dispersion of close contacts suggested by the distance matrix is more closely reproduced with DCC_{GNM} than with DCC_{MD}.

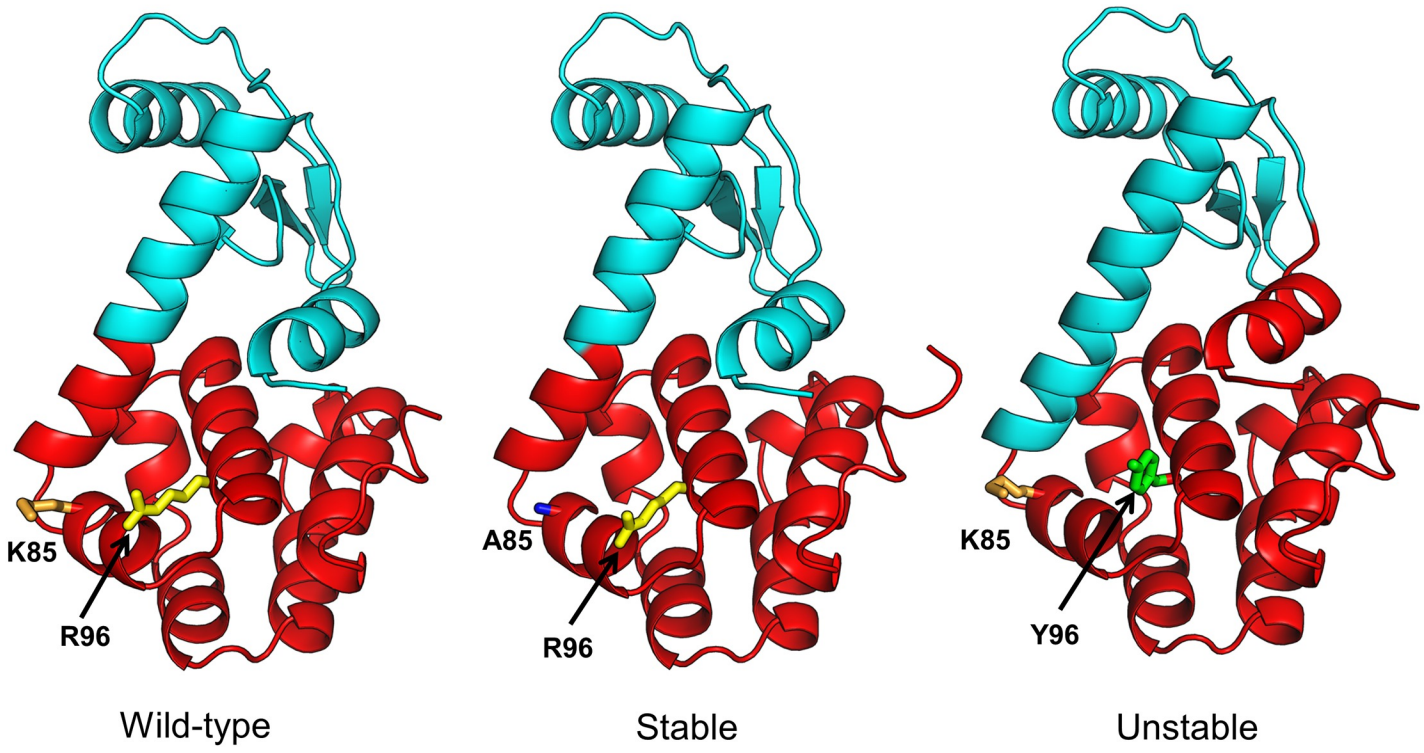

Fig 7. Comparison of community structures for wild-type (PDB: 4s0w), stable (PDB: 3c81) and unstable (PDB: 3c80) mutant forms of T4 lysozyme. Two communities (red and cyan) are shown for each structure. We choose $N_c = 2$ because the differences in community structure for the stable and unstable forms are most distinctive at this level. Similarly localized communities are colored alike. Sites of mutations are shown in sticks with the corresponding residue names labelled. Side chains of same amino acids in the sites of mutation are colored alike.

https://doi.org/10.1371/journal.pone.0199225.g007

This cohesiveness is a hallmark of the elastic network models in general, and is one reason that they can show better agreement with various protein behaviors than MD. It is however to be noted that we use only the first twenty low-frequency modes from GNM to calculate DCC_{GNM}. As we find in other analysis, the agreement between MD and GNM for different metrics mostly converges for the first 20 normal modes, with the addition of more modes not providing much significant gains.

Our approach to identify dynamic communities differs from existing methods that identify dynamic domains [39,40], which, similar to the approach taken by Kundu [14], divide the structure into rigid units (dynamic domains) based on the sign of the residue positional fluctuations given by the low frequency modes. These methods cluster residues with positive fluctuations into a single group and those with negative fluctuations into a separate group by considering each low frequency mode separately and dividing a protein structure primarily into two rigid clusters or dynamic domains. Depending on the mode that was considered, a single domain may be highly cohesive or may have individual entities that are dispersed over an entire protein structure. In contrast, our approach considers the cumulative contributions from more than one mode by calculating a cross-correlation matrix that combines multiple low frequency modes. Transforming such a matrix into a distance correlation matrix and then clustering it hierarchically, divides the structure into the desired number of dynamic communities based on the extent of inter-residue correlation. While the identification of dynamic domains chooses all residues having the same sign in their positional fluctuations and groups them into one cluster, our method could in principle divide these dynamic domains further into sub-modules, i.e., the dynamic communities.

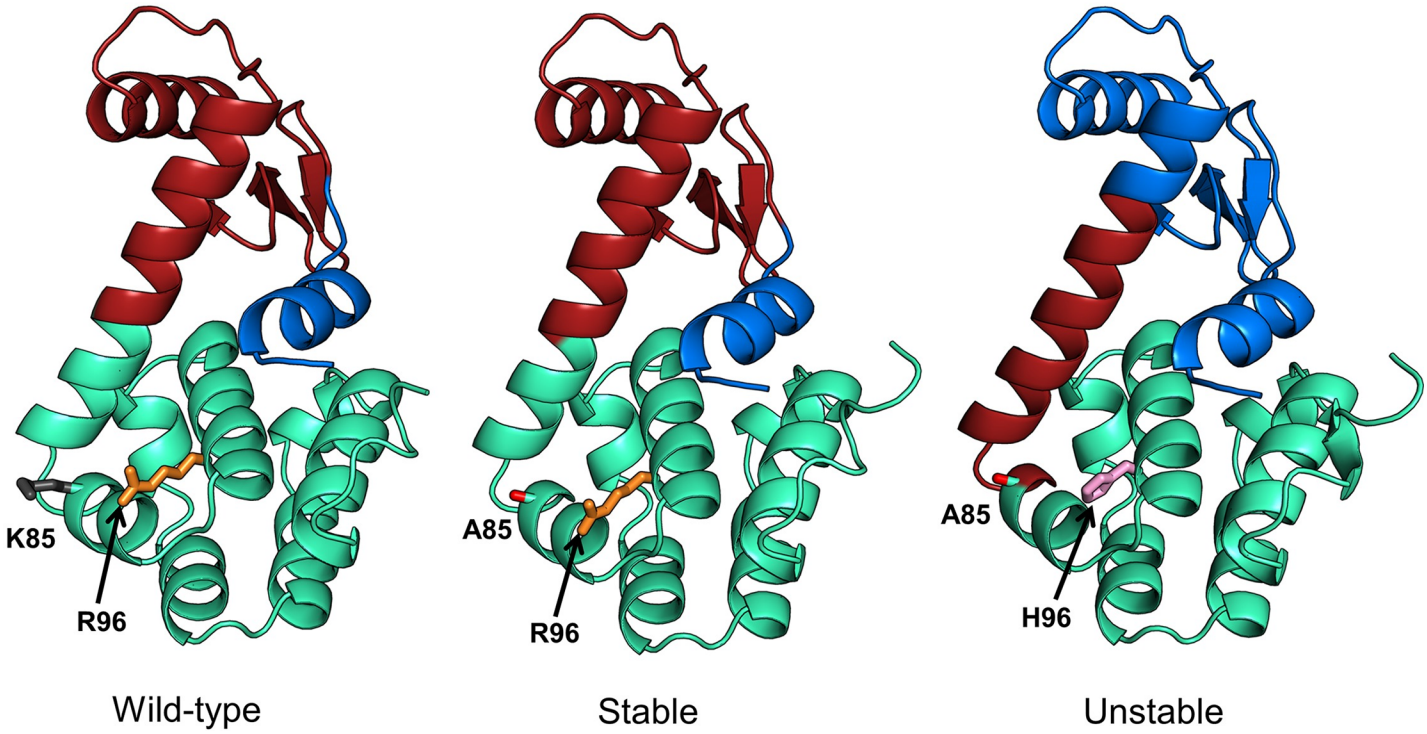

Fig 8. Comparison of community structures for wild-type (PDB: 4s0w), stable (PDB: 3c81) and unstable (PDB: 3c82) mutant forms of T4 lysozyme. Three communities (green, brown and blue) are shown for each structure. $N_c = 3$ shows maximum structural difference between the community structures of mutant and wild-type forms, hence the choice. Coloring scheme is the same as in Fig 7.

https://doi.org/10.1371/journal.pone.0199225.g008

To model the dynamics, we have considered a fixed distance cutoff $r_c = 7.5$ Å for each protein. However, it might be more realistic to use a different r_c for each protein, since using a generalized distance cutoff sometimes fails to take into account the size and variations in the packing density in different proteins and may not accurately represent the protein dynamics. Previous implementations of ENM have used a range of different r_c and then considered the r_c that best reproduces the experimental B-factors [15,21].

Our results from comparing the communities obtained upon clustering the distance transformed DCC_{GNM} and DCC_{MD} matrices hierarchically, suggest that for a certain number of communities N_c, MD and GNM show near-perfect agreement. Importantly, we observe convergence in agreement after using the first few low frequency modes. This also corroborates previous studies that showed that the first few low frequency modes are adequate to reproduce the experimentally observed conformational ensemble of proteins [31,41]. Also, in the case of GNM, though the model assumes isotropic, non-directional residue fluctuations not accounting for the directional preferences of residue mobilities, previous studies have suggested that using the first few low-frequency modes nonetheless results in good correlations with experimental B-factors [42]. When verifying the median kappa for all modes with $r_c = 7.5$ Å (S1 Fig), it is interesting to note that the median kappa for each subset of modes at all community levels is almost the same (≈ 0.41), except for the subset of 30 modes which shows highest median kappa values. While kappa coefficients of 0.41 rules out the possibility of random agreement, at the same time, one must also consider that there could be possible conformational undersampling depending on the time scale of the MD trajectory that restricts the extent of agreement between MD and GNM.

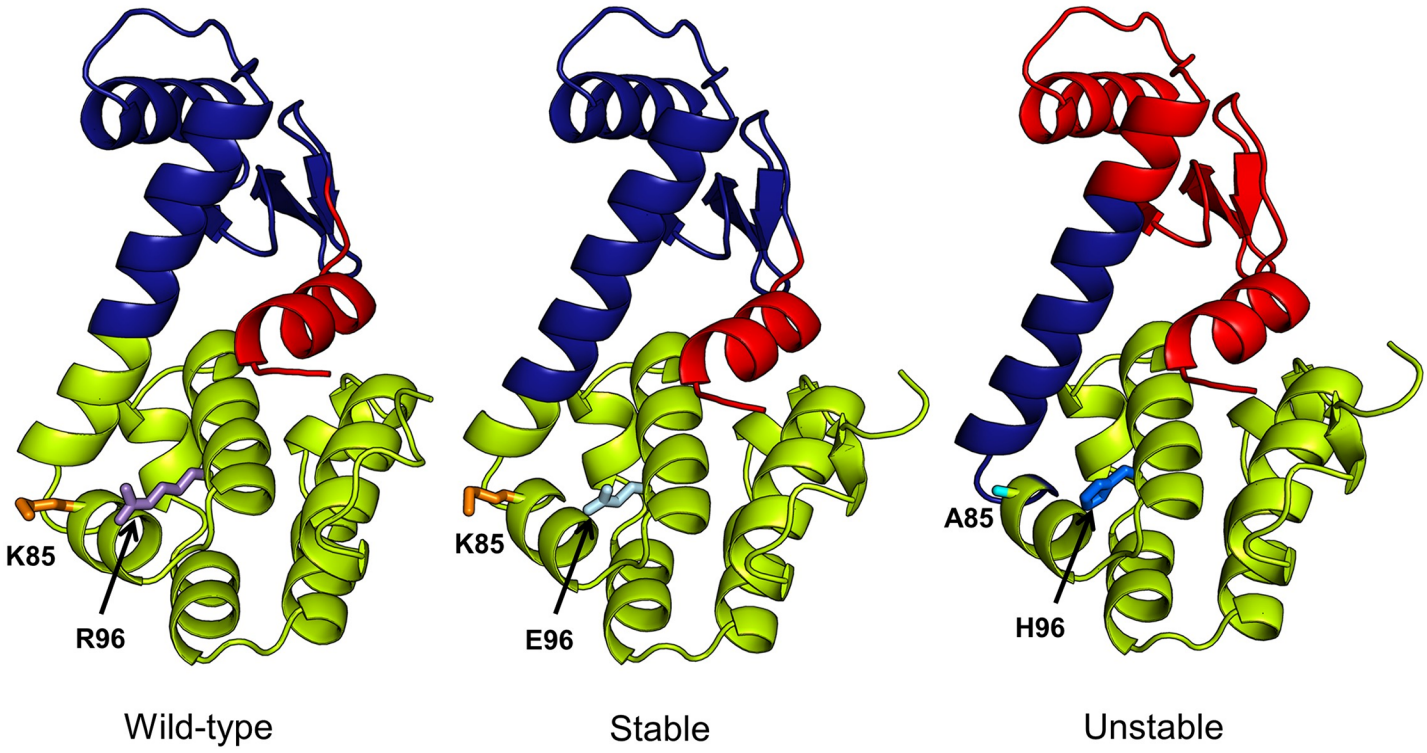

Fig 9. Comparison of community structures for wild-type (PDB: 4s0w), stable (PDB: 3c8s) and unstable (PDB: 3c82) mutant forms of T4 lysozyme. Three communities (red, blue and green) are shown for each structure. $N_c = 3$ shows the maximum structural differences for the community structures in the mutant and wild-type forms, hence its choice. The coloring scheme is same as in Figs 7 and 8.

https://doi.org/10.1371/journal.pone.0199225.g009

Near-convergence for a subset of the first-few low-frequency modes (20–30 modes) is also consistent for the correlation of node centralities and RMSIP between MD and GNM. It is interesting to observe the high correlation for node closeness (0.63), further verifying the strong correspondence between the simulation results from the two methods. However, as Fig 1 suggests, DCC_{GNM} and DCC_{MD} do not exhibit 100% agreement with each other. They agree to a large extent in the correlations of secondary structure elements and residues in spatial proximity however, they differ in their scale of inter-residue correlations which could possibly explain the lack of perfect correlation for node closeness.

Singular value decomposition of DCC_{GNM} and DCC_{MD} helps in capturing the directions of maximum variations for inter-residue correlations through its principal eigenvectors. Upon verifying the overlap of the principal eigenvectors between MD and GNM we observe an RMSIP of 0.82 (for 20 modes) followed by convergence. This confirms that the DCC_{GNM} and DCC_{MD} matrices agree to a large extent in terms of the inter-residue fluctuation correlation. It is also interesting to note that when using either a smaller number of modes (5 modes) or too many modes (50 modes) the standard error in RMSIP increases. While using very few modes possibly leads to a loss in information, including more modes in the calculations for DCC_{GNM} possibly adds to the noise, since the most reliable modes of motion for the elastic network models are those at the lower frequency end. Higher frequency modes describe local residue-level dynamics and are less reliable. Hence, including those modes in the calculation of the correlation matrix can potentially reduce the signal to noise ratio, resulting in observed lower agreement of DCC_{GNM} with DCC_{MD}.

The ability of GNM to discriminate stable mutants from unstable ones by evaluating community agreement is notable. The extent of change in community structures in unstable mutants is much greater than for stable mutants. We have used the atomic structures of T4 Lysozyme in the GNM as opposed to the coarse-grained version to account for the mutation changes. Interestingly, we observe that changes to community structures are more distinct in the higher community levels (smaller number of communities) as described by Fig 6. One should consider that we have performed this study only for a set of 16 mutant structures of T4 lysozyme, which is really a very small sample. However, we are limited in the availability of experimentally determined mutant structures for a single protein [43,44]. There is some data for the changes in free energy associated with a single point mutation in proteins [45] however, the crystal structures corresponding to these mutants are not usually available. To use this data, previous methods have considered computational approaches to mutate targeted residues in a given protein and then, used the modeled structure as a representative of the mutant form [46]. However, such computational approaches rely upon the potential function used in the modeling tool and hence, the structure of the modeled mutant (especially the sidechain positions of the mutant site and its neighbors) may be biased by the potential function. The data we have used should be more reliable because these are experimentally reported crystal structures.

Materials and methods

Dataset

We compile a set of 44 distinct proteins from the MODEL database [29] by considering only those proteins with MD trajectories of 100 ns or above. Each protein has a minimum of 50 residues. For each protein, we downloaded the all-atom trajectory from the database and parsed the all-atom trajectory into a C^{α} trajectory, having only the coordinates for residue C^{α} atoms in each frame.

Dynamic cross-correlations from MD trajectory

For each protein, we perform calculations for residue-level dynamic cross-correlations on the respective C^{α} trajectory using the *dccm* function in the Bio3D package [47] with the following equation [48,49].

$$DCC_{MD}(i,j) = \frac{< \Delta r_i(t).\Delta r_j(t) >_t}{\sqrt{< ||\Delta r_i(t)||^2 >_t}\sqrt{< ||\Delta r_j(t)||^2 >_t}} \quad (1)$$

Here, $r_i(t)$ and $r_j(t)$ refer to the coordinates of the ith and jth atoms as a function of time t, $<\cdot>$ indicates the time ensemble average and $\Delta r_i(t) = r_i(t) - (< r_i(t) >)_t$ and $\Delta r_j(t) = r_j(t) - (< r_j(t) >)_t$.

Dynamic cross-correlations from Gaussian Network Model

We use GNM [15,50], a form of ENM, to calculate the dynamic cross-correlations between residues. In GNM a protein is usually modeled as a coarse-grained system by representing individual residues by their alpha-carbons, but these points can also be atoms, which we use for the computations on the mutant proteins. Residues within a certain distance cutoff (r_c) are connected by Hookean springs. GNM assumes the protein crystal structure to be of energetic minimum conformation and doesn't require the structure to be energy minimized. It also assumes that residue fluctuations about their mean positions are isotropic and follow a

Gaussian distribution in their excursions away from the assumed minimum energy structure. The potential for GNM is given as

$$V = \frac{1}{2}\gamma \sum_{i,j}^{n} \Gamma \left[(\Delta R_i - \Delta R_j)^2 \right] \quad (2)$$

Here, ΔR_i and ΔR_j are the fluctuation vectors for residue i and j respectively, γ is the stiffness of the springs connecting residues i and j. Γ is the Kirchhoff matrix defining node connectivity and is defined as the following.

$$\Gamma = \begin{cases} -1, & if\ i \neq j\ and\ R_{ij} \leq r_c \\ 0, & if\ i \neq j\ and\ R_{ij} > r_c \\ -\sum_{j,j \neq i} \Gamma_{ij}, & if\ i = j \end{cases} \quad (3)$$

Here, R_{ij} is the distance between the alpha carbons of residues i and j while, r_c is the distance cutoff. Diagonalizing Γ yields *N-1* modes with non-zero eigenvalues. Each mode is a vector that describes the residue fluctuations about its mean position while the eigenvalues correspond to the square of the mode frequency and indicate the relative extent of motion of each point. The slow modes or the low-frequency modes describe the most energetically favorable motions of a protein.

The Kirchhoff matrix has a zero determinant and is thus, singular. The pseudo-inverse of this matrix is calculated using the *N-1* or a subset of the *N-1* modes with the following equation.

$$\Gamma^{-1} = \sum_{i=1}^{N-1} \lambda_i^{-1} V_i V_i^T \quad (4)$$

λ_i is the eigenvalue of the i*th* mode, V_i is i*th* mode and V_i^T is the transpose of V_i. The inter-residue dynamical correlation between residues i and j is then calculated as

$$DCC_{GNM}(i,j) = \frac{\Gamma^{-1}(i,j)}{\sqrt{(\Gamma^{-1}(i,i)\ \Gamma^{-1}(j,j))}} \quad (5)$$

In the present study, we first use a range of different values for the distance cutoff r_c (6, 6.5, 7, 7.5 and 8 Å) and then, select r_c = 7.5 Å, which provides high overlap for dynamics captured from GNM with MD. Using this cutoff, we calculate DCC_{GNM} using 5, 10, 20, 30 and 50 low-frequency modes.

Dynamic communities from correlation matrix

For each protein in our dataset, we convert the residue-residue dynamical correlation matrices DCC_{MD} and DCC_{GNM} into distance correlation matrices as follows

$$dist\_DCC_{MD} = 1 - DCC_{MD}, \quad (6)$$

$$dist\_DCC_{GNM} = 1 - DCC_{GNM} \quad (7)$$

We then perform hierarchical clustering on the distance correlation matrices with weighted pair-group method with arithmetic mean (WPGMA), which takes into consideration the cluster size when calculating the distance between two clusters [51]. Hierarchical clustering yields dendrograms that can be pruned at different levels to give the desired number of clusters. The clusters obtained upon pruning a dendrogram at a certain height correspond to the dynamic communities, i.e., the blocks of residues that are highly cohesive and move like a rigid body. We cut the dendrograms at different levels to obtain between 2 and 10 communities. The hierarchical clustering

was performed using the MATLAB *linkage (https://www.mathworks.com/help/stats/linkage.html)* and *cluster (https://www.mathworks.com/help/stats/cluster.html)* modules.

Comparing community assignment between MD and GNM

We use 3 metrics to assess the agreement between the communities from MD and GNM.

1. Cohen's kappa coefficient. The Cohen's kappa or simply, kappa is a statistic that is often used to evaluate the extent of agreement between data collectors or raters in their assignments to the same variables, referred to as inter-rater reliability. Kappa coefficient is considered to be more robust than percent agreement as it also takes into consideration random agreement [28]. Like correlation coefficients, the value of the kappa statistic can range from -1 to 1. A kappa of 0 indicates an agreement by chance while kappa of 1 indicates perfect agreement [28,30]. We calculate the kappa coefficient as follows

$$K = \frac{p_o - p_e}{1 - p_e} \tag{8}$$

Here, p_o is the observed probability of agreement for cluster assignment between MD and GNM while, p_e is the expected probability of agreement.

2. Network centrality. We model each protein as a weighted network in which a node represents a residue and the edge between a pair of nodes is weighted by the distance transformed correlation for the residue pair (Eq 6 and Eq 7). Then, we calculate the node closeness centralities for the networks from MD and GNM. The closeness centrality is the sum of the lengths of the shortest paths to all other nodes from the given node in the graph. We perform all calculations for network centrality using the MatLab *graph (https://www.mathworks.com/help/matlab/ref/graph.html)* and *centrality (https://www.mathworks.com/help/matlab/ref/graph.centrality.h tml)* modules.

3. Overlap between principal eigen vectors. We perform singular-value decomposition (SVD) on the DCC_{MD} and DCC_{GNM} matrices and then evaluate the overlaps between the MD and GNM eigenvector spaces for subsets of vectors having largest eigenvalues using the root-mean square inner product (RMSIP) [52] as

$$RMSIP = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n}\sum_{j=1}^{n}(V_i.U_j)^2\right.} \tag{9}$$

V and U are the principal eigenvectors obtained from SVD of the DCC_{MD} and DCC_{GNM} matrices respectively, while n is the number of vectors to be compared. We consider the same number of principal vectors for the two matrices.

Mutant dataset

We use PDB structures for the T4 lysozyme mutants crystallized by Mooers *et al.* [37]. In their study, the authors performed circular dichroism assays to estimate stability changes upon specific mutations to the enzyme and calculated the free energy change ($\Delta\Delta G$) for the mutants as $\Delta G_{mutant} - \Delta G_{wildtype}$. The authors have defined the more negative $\Delta\Delta G$ values to be the unstable mutants. The stability changes were performed at pH 5.35 and 3.05. In our study, we consider the $\Delta\Delta G$ values calculated at pH 5.35. Details of the mutant structures used and their free energy changes with respect to the wild-type are given in Table 1.

Effect of mutation on dynamic communities

We use all-atom GNM to investigate the community change in the mutant structures with respect to the wild-type. For both the mutant and wild-type forms of the enzyme, we retain all

heavy atoms in the PDB and use a distance cutoff of 3.5Å to identify interacting spring locations. Using 5, 10, 20, 30 and 50 modes, we initially calculate the inter-residue dynamical correlations and then, perform hierarchical clustering with weighted average linkage to obtain the desired number of clusters. We trim the dendrograms for each structure at specific heights to obtain 2–10 communities and then compute the agreement between the communities for the wild-type and mutant forms with the kappa coefficient.

Supporting information

S1 Table. Dataset of proteins used in the study. The MD trajectories were downloaded from the MOlecular Dynamics Extended Library (MODEL) database. We retained proteins having at least 50 residues with a minimum simulation length of 100 ns. The table is sorted by the number of residues.
(DOCX)

S2 Table. Distribution of $Kappa_{max}$ for the dataset. For each protein, we identified the community level N_c for which we obtained the maximum value for Kappa coefficient. We show the values for $Kappa_{max}$ for a subset of 5, 10, 20, 30 and 50 low frequency modes.
(DOCX)

S3 Table. Distribution of median kappa coefficient over all community levels for different subsets of modes. For each protein, the table shows the median Kappa over all community levels for each subset of modes.
(DOCX)

S4 Table. Correlation for node closeness. The table shows the correlation for node closeness between MD and GNM and the median correlations for each mode. The distance cutoff of 7.5 Å was used for GNM.
(DOCX)

S5 Table. Distribution of root-mean square inner product (RMSIP) for the dataset. The principal eigenvectors are obtained with singular-value decomposition of the cross-correlation matrices from MD and GNM. They capture the major directions of variations from the matrix. We see a considerably good overlap (median RMSIP 0.81 over all subsets of modes) between the principal eigenvectors from MD and GNM which suggests a close agreement between the two.
(DOCX)

S1 Fig. Distribution of kappa coefficient for all community levels. We verified the variation of Kappa coefficient upon choosing the generalized r_c = 7.5 Å for all proteins. The figure shows the median Kappa over all proteins for all community levels for each subset of modes. The differences in median Kappa for individual subsets of modes is not very high, however the decrease for Kappa with 50 modes following the peak at 30 modes is quite remarkable, emphasizing the importance of the first few low-frequency modes for capturing the global functional dynamics of proteins. The error bars indicate standard error for the Kappa coefficient for a given subset of modes.
(DOCX)

Author Contributions

Conceptualization: Sambit Kumar Mishra, Robert L. Jernigan.

Data curation: Sambit Kumar Mishra.

Formal analysis: Sambit Kumar Mishra.

Funding acquisition: Robert L. Jernigan.

Investigation: Sambit Kumar Mishra.

Methodology: Sambit Kumar Mishra.

Project administration: Robert L. Jernigan.

Supervision: Robert L. Jernigan.

Validation: Sambit Kumar Mishra.

Writing – original draft: Sambit Kumar Mishra.

Writing – review & editing: Sambit Kumar Mishra, Robert L. Jernigan.

References

1. Nussinov R. Introduction to Protein Ensembles and Allostery. Chem Rev. 2016; 116: 6263–6266. https://doi.org/10.1021/acs.chemrev.6b00283 PMID: 27268255
2. Alberts B, Johnson A, Lewis J, Raff M, Roberts K, Walter P. Molecular Biology of the Cell. 4th Edition, New York. 2002. 10.1091/mbc.E14-10-1437
3. Greives N, Zhou H-X. Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit. Proc Natl Acad Sci. 2014; 111: 10197–10202. https://doi.org/10.1073/pnas.1407545111 PMID: 24982141
4. Benkovic SJ, Hammes-schiffer S. R EVIEW A Perspective on Enzyme Catalysis. Science (80-). 2003; 301: 1196–1202. https://doi.org/10.1126/science.1085515 PMID: 12947189
5. Daniel RM, Dunn RV, Finney JL, Smith JC. The Role of Dynamics in Enzyme Activity. Annu Rev Biophys Biomol Struct. 2003; 32: 69–92. https://doi.org/10.1146/annurev.biophys.32.110601.142445 PMID: 12471064
6. Yon JM, Perahia D, Ghélis C. Conformational dynamics and enzyme activity. Biochimie. 1998; 80: 33–42. https://doi.org/10.1016/S0300-9084(98)80054-0 PMID: 9587660
7. Changeux J-P, Edelstein SJ. Allosteric mechanisms of signal transduction. Science. 2005; 308: 1424–8. https://doi.org/10.1126/science.1108595 PMID: 15933191
8. Brignole EJ, Smith S, Asturias FJ. Conformational flexibility of metazoan fatty acid synthase enables catalysis. Nat Struct Mol Biol. 2009; 16: 190–7. https://doi.org/10.1038/nsmb.1532 PMID: 19151726
9. Kern D, Zuiderweg ER. The role of dynamics in allosteric regulation. Curr Opin Struct Biol. 2003; 13: 748–757. https://doi.org/10.1016/j.sbi.2003.10.008 PMID: 14675554
10. McClendon CL, Kornev AP, Gilson MK, Taylor SS. Dynamic architecture of a protein kinase. Proc Natl Acad Sci U S A. 2014; 111: E4623–31. https://doi.org/10.1073/pnas.1418402111 PMID: 25319261
11. Calligari P, Gerolin M, Abergel D, Polimeno A. Decomposition of proteins into dynamic units from atomic cross-correlation functions. J Chem Theory Comput. 2017; 13: 309–319. https://doi.org/10.1021/acs.jctc.6b00702 PMID: 28068775
12. Doshi U, Holliday MJ, Eisenmesser EZ, Hamelberg D. Dynamical network of residue–residue contacts reveals coupled allosteric effects in recognition, catalysis, and mutation. Proc Natl Acad Sci. 2016; 113: 4735–4740. https://doi.org/10.1073/pnas.1523573113 PMID: 27071107
13. Hinsen K, Thomas A, Field MJ. Analysis of domain motions in large proteins. Proteins Struct Funct Genet. 1999; 34: 369–382. https://doi.org/10.1002/(SICI)1097-0134(19990215)34:3<369::AID-PROT9>3.0.CO;2-F PMID: 10024023
14. Kundu S, Sorensen DC, Phillips GN. Automatic domain decomposition of proteins by a Gaussian Network Model. Proteins Struct Funct Genet. 2004; 57: 725–733. https://doi.org/10.1002/prot.20268 PMID: 15478120
15. Bahar I, Atilgan AR, Erman B. Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential. Fold Des. 1997; 2: 173–81. https://doi.org/10.1016/S1359-0278(97)00024-2 PMID: 9218955
16. Yesylevskyy SO, Kharkyanen VN, Demchenko AP. Hierarchical clustering of the correlation patterns: New method of domain identification in proteins. Biophys Chem. 2006; 119: 84–93. https://doi.org/10.1016/j.bpc.2005.07.004 PMID: 16125297

17. Shudler M, Niv MY. Blockmaster: Partitioning protein kinase structures using normal-mode analysis. J Phys Chem A. 2009; 113: 7528–7534. https://doi.org/10.1021/jp900885w PMID: 19485335

18. Potestio R, Pontiggia F, Micheletti C. Coarse-grained description of protein internal dynamics: an optimal strategy for decomposing proteins in rigid subunits. Biophys J. Biophysical Society; 2009; 96: 4993–5002. https://doi.org/10.1016/j.bpj.2009.03.051 PMID: 19527659

19. Chopra N, Wales TE, Joseph RE, Boyken SE, Engen JR, Jernigan RL, et al. Dynamic Allostery Mediated by a Conserved Tryptophan in the Tec Family Kinases. PLoS Comput Biol. 2016; 12: 1–19. https://doi.org/10.1371/journal.pcbi.1004826 PMID: 27010561

20. Yao XQ, Malik RU, Griggs NW, Skjærven L, Traynor JR, Sivaramakrishnan S, et al. Dynamic coupling and allosteric networks in the α subunit of heterotrimeric G proteins. J Biol Chem. 2016; 291: 4742–4753. https://doi.org/10.1074/jbc.M115.702605 PMID: 26703464

21. Atilgan AR, Durell SR, Jernigan RL, Demirel MC, Keskin O, Bahar I. Anisotropy of fluctuation dynamics of proteins with an elastic network model. Biophys J. Elsevier; 2001; 80: 505–15. https://doi.org/10.1016/S0006-3495(01)76033-X PMID: 11159421

22. Tirion M. Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis. Phys Rev Lett. 1996; 77: 1905–1908. https://doi.org/10.1103/PhysRevLett.77.1905 PMID: 10063201

23. Leioatts N, Romo TD, Grossfield A. Elastic network models are robust to variations in formalism. J Chem Theory Comput. 2012; 8: 2424–2434. https://doi.org/10.1021/ct3000316 PMID: 22924033

24. Yang L, Song G, Carriquiry A, Jernigan RL. Close Correspondence between the Motions from Principal Component Analysis of Multiple HIV-1 Protease Structures and Elastic Network Modes. Structure. 2008; 16: 321–330. https://doi.org/10.1016/j.str.2007.12.011 PMID: 18275822

25. Romo TD, Grossfield A. Validating and improving elastic network models with molecular dynamics simulations. Proteins Struct Funct Bioinforma. 2011; 79: 23–34. https://doi.org/10.1002/prot.22855 PMID: 20872850

26. Doruker P, Nilsson L, Kurkcuoglu O. Collective dynamics of EcoRI-DNA complex by elastic network model and molecular dynamics simulations. J Biomol Struct Dyn. 2006; 24: 1–16. https://doi.org/10.1080/07391102.2006.10507093 PMID: 16780370

27. Gur M, Zomot E, Cheng MH, Bahar I. Energy landscape of LeuT from molecular simulations. J Chem Phys. 2015; 143. https://doi.org/10.1063/1.4936133 PMID: 26723619

28. Cohen J. A COEFFICIENT OF AGREEMENT FOR NOMINAL SCALES. Educ Psychol Meas. 1960; XX: 37–46.

29. Meyer T, D'Abramo M, Hospital A, Rueda M, Ferrer-Costa C, Pérez A, et al. MoDEL (Molecular Dynamics Extended Library): A Database of Atomistic Molecular Dynamics Trajectories. Structure. 2010; 18: 1399–1409. https://doi.org/10.1016/j.str.2010.07.013 PMID: 21070939

30. McHugh M. Interrater reliability. Biochem Medica. 2012; 22: 276–82.

31. Haliloglu T, Bahar I. Adaptability of protein structures to enable functional interactions and evolutionary implications. Curr Opin Struct Biol. Elsevier Ltd; 2015; 35: 17–23. https://doi.org/10.1016/j.sbi.2015.07.007 PMID: 26254902

32. O'Rourke KF, Gorman SD, Boehr DD. Biophysical and computational methods to analyze amino acid interaction networks in proteins. Comput Struct Biotechnol J. Natrix Separations; 2016; 14: 245–251. https://doi.org/10.1016/j.csbj.2016.06.002 PMID: 27441044

33. Bonacich P. Power and Centrality: A Family of Measures. Am J Sociol. 1987; 92: 1170–1182. https://doi.org/10.1086/228631

34. Borgatti SP. Centrality and network flow. Soc Networks. 2005; 27: 55–71. https://doi.org/10.1016/j.socnet.2004.11.008

35. Sabidussi G. The centrality index of a graph. Psychometrika. 1966; 31: 581–603. https://doi.org/10.1007/BF02289527 PMID: 5232444

36. Bavelas A. Communication Patterns in Task-Oriented Groups. J Acoust Soc Am. 1950; 22: 725–730. https://doi.org/10.1121/1.1906679

37. Mooers BHM, Baase WA, Wray JW, Matthews BW. Contributions of all 20 amino acids at site 96 to the stability and structure of T4 lysozyme. Protein Sci. 2009; 18: 871–880. https://doi.org/10.1002/pro.94 PMID: 19384988

38. Kundu S, Melton JS, Sorensen DC, Phillips GN. Dynamics of proteins in crystals: comparison of experiment with simple models. Biophys J. 2002; 83: 723–32. https://doi.org/10.1016/S0006-3495(02)75203-X PMID: 12124259

39. Li H, Chang YY, Yang LW, Bahar I. iGNM 2.0: The Gaussian network model database for biomolecular structural dynamics. Nucleic Acids Res. 2016; 44: D415–D422. https://doi.org/10.1093/nar/gkv1236 PMID: 26582920

40. Li H, Sakuraba S, Chandrasekaran A, Yang LW. Molecular binding sites are located near the interface of intrinsic dynamics domains (IDDs). J Chem Inf Model. 2014; 54: 2275–2285. https://doi.org/10.1021/ci500261z PMID: 25089914

41. Bahar I, Lezon TR, Bakan A, Shrivastava IH. Normal mode analysis of biomolecular structures: functional mechanisms of membrane proteins. Chem Rev. 2010; 110: 1463–97. https://doi.org/10.1021/cr900095e PMID: 19785456

42. Song G, Jernigan RL. vGNM: A Better Model for Understanding the Dynamics of Proteins in Crystals. J Mol Biol. 2007; 369: 880–893. https://doi.org/10.1016/j.jmb.2007.03.059 PMID: 17451743

43. Ng PC, Henikoff S. Predicting Deleterious Amino Acid Substitutions Predicting Deleterious Amino Acid Substitutions. Genome Res. 2001; 11: 863–874. https://doi.org/10.1101/gr.176601 PMID: 11337480

44. Reva B, Antipin Y, Sander C. Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Res. 2011; 39: 37–43. https://doi.org/10.1093/nar/gkr407 PMID: 21727090

45. Gromiha MM, Uedaira H, An J, Selvaraj S, Prabakaran P, Sarai A. ProTherm, Thermodynamic Database for Proteins and Mutants: developments in version 3.0. 2002; 30: 301–302.

46. Guerois R, Nielsen JE, Serrano L. Predicting changes in the stability of proteins and protein complexes: A study of more than 1000 mutations. J Mol Biol. 2002; 320: 369–387. https://doi.org/10.1016/S0022-2836(02)00442-4 PMID: 12079393

47. Grant BJ, Rodrigues APC, Elsawy KM, Mccammon JA, Caves LSD. Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics. 2006; 22: 2695–6. https://doi.org/10.1093/bioinformatics/btl461 PMID: 16940322

48. Kasahara K, Fukuda I, Nakamura H. A novel approach of dynamic cross correlation analysis on molecular dynamics simulations and its application to Ets1 dimer-DNA complex. PLoS One. 2014; 9. https://doi.org/10.1371/journal.pone.0112419 PMID: 25380315

49. McCammon JA. Protein Dynamics. Reports Prog Phys. 1984; 47: 1–46. https://doi.org/10.1007/978-1-62703-658-0

50. Rader AJ, Chennubhotla C, Yang L-W, Bahar I. The Gaussian Network Model: theory and applications. Norm Mode Anal—theory Appl to Biol Chem Syst. 2006; 10: 41–64.

51. Sokal RR, Michener CD. A statistical method for evaluating systematic relationships. Univ Kansas Sci Bull. 1958; 38: 1409–1438. citeulike-article-id:1327877

52. Amadei A, Ceruso MA, Di Nola A. On the convergence of the conformational coordinates basis set obtained by the Essential Dynamics analysis of proteins' molecular dynamics simulations. Proteins Struct Funct Genet. 1999; 36: 419–424. https://doi.org/10.1002/(SICI)1097-0134(19990901)36:4<419::AID-PROT5>3.0.CO;2-U PMID: 10450083