



Protein as evolvable functionally constrained amorphous matter

MADHUSMITA TRIPATHY^{1,2}, ANAND SRIVASTAVA², SRIKANTH SASTRY³ and
MADAN RAO^{4*} 

¹Present Address: Eduard-Zintl-Institut für Anorganische und Physikalische Chemie, Technische Universität Darmstadt, 64287 Darmstadt, Germany

²Molecular Biophysics Unit, Indian Institute of Science, Bengaluru 560012, India

³Theoretical Sciences Unit, Jawaharlal Nehru Centre for Advanced Scientific Research, Bengaluru 560064, India

⁴Simons Centre for the Study of Living Machines, National Centre for Biological Sciences (TIFR), Bengaluru 560065, India

*Corresponding author (Email, madan@ncbs.res.in)

MS received 17 July 2022; accepted 14 October 2022

We explore current ideas around the representation of a protein as an amorphous material, in turn represented by an abstract graph \mathcal{G} with edges weighted by elastic stiffnesses. By embedding this graph in physical space, we can map every graph to a spectrum of conformational fluctuations and responses (as a result of, say, ligand-binding). This sets up a ‘genotype–phenotype’ map, which we use to evolve the amorphous material to select for fitness. Using this, we study the emergence of allosteric interaction, hinge joint, crack formation and a slide bolt in functional proteins such as adenylate kinase, HSP90, calmodulin and GPCR proteins. We find that these emergent features are associated with specific geometries and mode spectra of floppy or liquid-like regions. Our analysis provides insight into understanding the architectural demands on a protein that enable a prescribed function and its stability to mutations.

Keywords. Conformation fluctuations; fitness function; graph network; localized soft channels; protein allostery

1. Introduction

Only a small fraction of the allowable protein ‘universe’ constitutes real biological proteins (Anfinsen 1973; Koonin *et al.* 2002). For example, of the 20^{300} number of possible sequences of a polypeptide chain with ~ 300 residues that can potentially be generated from the naturally available 20 amino acids, living systems such as *Saccharomyces cerevisiae* exhibit only $\sim 10^4$ (Koonin *et al.* 2002; Milo and Phillips 2015; Sartori and Leibler 2020). This *dimensional reduction* comes about because, out of the numerous possible proteins, only a small subset are functionally relevant,

robust and explored by evolution. We have, for many years, been interested in understanding the architectural demands on a protein that enable a specific function, and its stability to mutations, fluctuations and cycles of performance. Some aspects of this program are not new and, recently, rather elegant theoretical formalisms have emerged (Yan *et al.* 2017; Tlusty *et al.* 2017; Dutta *et al.* 2018; Yan *et al.* 2018; Eckmann *et al.* 2019). Here we offer our perspective on this problem.

We focus on proteins that undergo significant conformational changes between their native and functional states. We first consider ‘allosteric proteins’, where the intriguing mechanism of ‘action-at-a-distance’ drives function. Motivated by the tantalising similarities between functional proteins and amorphous materials, in terms of molecular packing (Liang and Dill 2001), free energy landscape (Frauenfelder *et al.*

This article is part of the Topical Collection: Emergent dynamics of biological networks.

1991) and relaxation mechanisms (Iben et al. 1989), we explore if allosteric regulation proceeds via emergence of ‘allosteric chains’, reminiscent of ‘force chains’ in granular media (Cates et al. 1998). There are two proposed mechanisms of allostery – the *induced-fit* mechanism, where the conformational switch depends on a ligand-induced change in protein conformation that leads to specificity of enzyme action (Koshland et al. 1966), and the *conformational selection* mechanism, where the enzyme explores a multiplicity of conformation states, independent of ligand structure and occupancy, which are then differentially stabilized by the ligand (Monod et al. 1965; Changeux 2012). Since allosteric propagation and binding scenarios in proteins span a repertoire of selection and adjustment processes, it is likely that both these mechanisms could be operative in the same protein in physiological settings (Tsai et al. 1999; Ramanoudjame et al. 2006; Csermely et al. 2010; Rajasekaran and Naganathan 2017). Here we focus on induced fit proteins, such as adenylate and guanylate kinase (Müller and Schulz 1992; Stehle and Schulz 1992; Müller et al. 1996; Maragakis and Karplus 2005; Chu and Voth 2007), HSP90 (Shiau et al. 2006), calmodulin (Babu et al. 1988; Osawa et al. 1999; Stefan et al. 2008) and GPCR proteins (Cherezov et al. 2007; Hilger et al. 2018; Weis and Kobilka 2018), and ask what are the necessary physical (architectural) features that the protein must have in order to perform a specific function with high fidelity.

To do this, we need a coarse-grained representation of a protein that is appropriate for this task. A protein represented as a heteropolymer (Garel et al. 1997) is indeed a convenient starting point if the question pertains to the dynamics of folding into a native state, or to the dynamics of assembly driven by multivalent interactions of intrinsically disordered proteins (Socci and Onuchic 1994). However, a coarse-grained description of changes in protein conformation in the native state, either as a result of spontaneous fluctuations or induced by ligand binding, or during the process of chemical reaction, requires a different starting point. We need a representation that enables a classification of the low-energy excitations and modes of deformation about the native state of a protein (Maragakis and Karplus 2005). This would involve accounting for inter-monomer (or inter-sector) (Halabi et al. 2009; Smock et al. 2010) interactions of varying strengths, both along the heteropolymer backbone and across it, giving it a three-dimensional character. This suggests that the appropriate coarse-grained description for deformations of a functional protein is to treat the protein as a three-

dimensional amorphous solid with heterogeneous interactions that have been designed to facilitate a prescribed function with high fidelity. The strategy that we will use to *design* the heterogeneous interactions is akin in spirit to a ‘*gain of function*’ approach (Kuhlman et al. 2003; Ahmed et al. 2022). The ability to render a specific function with high fidelity puts constraints on the free energy landscape explored by the amorphous solid.

A key result is that in order for the protein (represented as an amorphous solid) to render a prescribed function (such as allostery) with high fidelity, it must possess ‘liquid-like’ channels of a specific geometry and orientation. The low-energy excitations of such a channel can be described by the spectrum of the graph Laplacian or equivalently of a pinned liquid–gas interface (Jasnow 1984). Alternately, one may think of the design process as a ‘pruning’ of an amorphous solid described by non-affine elasticity (DiDonna and Lubensky 2005).

2. Representation of a protein as an amorphous material

Here we make precise the representation of a protein as an amorphous solid. For simplicity, we will consider proteins that have a large molecular weight and are globular, with a well-defined ‘bulk’ and ‘surface’. A globular protein is a linear heteropolymer with side groups, which in its native conformation is folded up in a ball. This enables each monomer to interact with the rest of the monomers across three-dimensional space, via interactions of varying bond strengths. It is in this setting that we define the genotype–phenotype space and the representation as an amorphous solid.

2.1 ‘Genotype’ space

Let the set of *amino acids* (monomer types) be $\{A_i\} : i = 1, \dots, K$, each characterised by a hydrodynamic radius $\{a_i\}$ and the set of *bond types* be $\{B_\alpha\} : \alpha = 1, \dots, M$, with $M \ll K$, each characterised by a bond strength $\{b_\alpha\}$ (figure 1a). A realisation of a ‘protein’ is a weighted graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, with the vertices \mathcal{V} taken from $\{A_i\}$ and edges \mathcal{E} taken from $\{B_\alpha\}$. Note that a given vertex can have any number of edges emanating from it; the number of edges can be greater than 1 (if surface vertex) or 2 (if bulk vertex) and less than a maximum E_{max} . Together these constitute the *genotype space* \mathcal{G} .

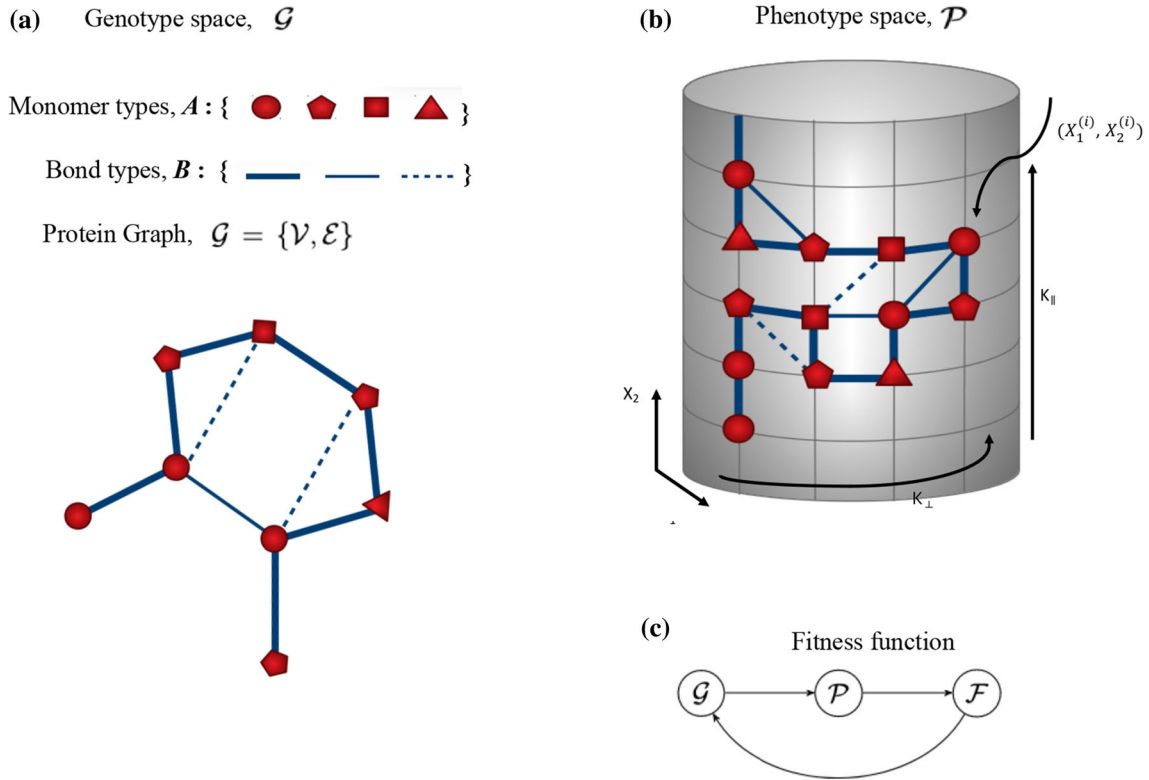


Figure 1. (a) *Genotype space* \mathcal{G} constructed from the set of monomers and bonds with different stiffnesses to generate a weighted protein graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ representing the abstract protein network. (b) *Phenotype space* \mathcal{P} obtained by embedding the graph \mathcal{G} in physical space. This embedding assigns coordinates to the vertices. We draw it on a cylinder to depict that we impose periodic boundary conditions in the x_1 -direction and free boundaries in the x_2 -direction. This choice of boundary conditions is dictated by the nature of the fitness function \mathcal{F} . (c) Every network in \mathcal{G} gets embedded in \mathcal{P} , from which we compute the fitness function \mathcal{F} . We then change the network in \mathcal{G} and iterate until we reach the optimum fitness.

2.2 ‘Phenotype’ space – Embedding in physical space

As shown in figure 1b, we embed this graph \mathcal{G} in *physical space*, that is to say, the N vertices are embedded in Euclidean space of d dimensions \mathbb{R}^d (with coordinates $\{\mathbf{x}_0^i\} : i = 1, \dots, N$). With this embedding, each vertex is subjected to forces arising from steric repulsion upon contact and short-range harmonic extensional springs from the connecting bonds. In addition, one could include contributions to the force, such as bending and torsion. This sets the stage for viewing the protein as an amorphous solid with heterogeneous spring constants.

Because the protein is a polymer with a defined backbone characterised by stronger peptide bonds, the energy scales associated with the extensional springs in the above representation will show a clear separation in bond strengths. We will refer to the peptide bonds of the backbone as *strong* bonds, and the interactions such as electrostatic, hydrophobic, hydrogen bonding,

disulphide and salt bridges, and van der Waals collectively as *weak* bonds. Neighbouring monomers that do not interact will be connected by a *non-bonding* edge. Given that the protein is a linear polymer, every bulk vertex will have two strong bonds emanating from it. Together these define the *phenotype space* \mathcal{P} .

2.3 Fidelity of function as fitness

Having established the genotype–phenotype map $\mathcal{G} \rightarrow \mathcal{P}$, we would like to drive changes in the genotype space to arrive at a desired phenotype. We do this by defining a *fitness function*.

Since we will be concerned with native proteins that undergo specific conformational change in response to a local external stimulus, such as ligand binding, the fitness function must describe the fidelity and specificity of the conformational change. Thus, in general, we define fitness as a scalar function of the displacements of the vertices of the physical graph, i.e.,

$\mathcal{F} : \mathcal{P} \rightarrow \mathbb{R}$. This function has as *input* \mathcal{I} , the prescribed displacement vectors of a subset of vertices $i \in \mathcal{I} \subset \mathcal{P}$, and as *output* \mathcal{O} , a scalar function of the displacement vectors of a different subset of vertices $j \in \mathcal{O} \subset \mathcal{P}$. The goal is to sample the genotype space \mathcal{G} and optimise the fitness function \mathcal{F} over the space of phenotypes \mathcal{P} . In section 3 we will consider several examples of this fitness function \mathcal{F} .

2.4 Optimisation algorithm

While our proposed optimisation algorithm should hold in any dimension, we will, for convenience, describe the procedure in two spatial dimensions. We start with a phenotype graph \mathcal{P} with vertices on a triangular lattice of dimension $K_{\parallel} \times K_{\perp}$ (figure 1b), and edges connecting nearest neighbour vertices, with periodic boundary conditions. Let the initial coordinates of the vertices be $\{\mathbf{x}_0^i\}$.

For the problem at hand, we can, without loss of generality, take all the monomers to be the same and assign all the genotypic diversity to the bonds. Thus, we randomly assign the weight of an edge to be $\{b_x\} : \alpha = 1, \dots, M$ with probability p_{α} , where $b_1 \equiv 0$ corresponds to the non-bonded edges. A useful parameter in the model is the number fraction of bonded edges ϕ_0 . This assignment should be subjected to constraints, such as ensuring a polymer backbone, i.e., that there exists one and only one path in \mathcal{P} comprising strong covalent bonds alone that spans all vertices, but for now we will ignore this constraint.

Given a realisation of bond strength b_s on the phenotype graph \mathcal{P} , we can compute real space displacements \mathbf{u}^i of every vertex by minimising the total elastic energy $E = \frac{1}{2} \sum_{i,j} b_{ij} (\mathbf{x}^i - \mathbf{x}^j - \mathbf{a})^2$, with respect to \mathbf{u}^i , where $\mathbf{x}^i = \mathbf{x}_0^i + \mathbf{u}^i$ and $\mathbf{a} \equiv \mathbf{x}_0^i - \mathbf{x}_0^j$. If our physical embedding was associated with a bath of temperature T , we could in principle even compute the displacement fluctuations at every vertex. These measurable physical quantities will depend on the spring constants that reside in the bonds and in the hydrodynamic radii that reside in the vertex.

Now for every realisation of bond strengths $\{b_{ij}\}$ on the phenotype graph \mathcal{P} , we can compute the fitness function \mathcal{F} for the prescribed input. We then change the realisation of $\{b_{ij}\}$ and repeat the calculation. By sampling over all the realisations of b , we arrive at one that optimises \mathcal{F} for the same fixed input. In practice this is hard because the dimensionality of the search

space goes as M^N , a very large number. We will therefore restrict the bond strengths to $\{b = 0, 1\}$, in units of a typical energy scale, and sample the genotype space \mathcal{G} using a Metropolis Monte Carlo sampling scheme.

We implement the algorithm as follows:

We first prepare the system by distributing the bond strengths $\{0, 1\}$ randomly, such that with probability p , the bond strength is 1; this specifies the number fraction of bonded edges ϕ_0 . We construct a well-defined, physically motivated, fitness function \mathcal{F} (with nice convergence properties), and choose a large N , a large enough p to ensure percolation and boundary conditions that are either open or periodic. Then,

1. We provide fixed displacement vectors for the input vertices. In response to this localized strain, all bonds with nonzero stiffness will elastically deform. We then compute the displacements $\{\mathbf{u}^i\}$ of all the vertices that minimise the total elastic energy,

$$E = \frac{1}{2} \sum_{i,j} b_{ij} (\mathbf{u}^i - \mathbf{u}^j)^2. \quad (1)$$

2. Using this energy minimized displacement vectors of the output vertices, we compute the fitness function \mathcal{F} . This will be large in general.
3. We now make moves in genotype space \mathcal{G} (mutations), which corresponds to moving in bond space $\{B_i\}$. For simplicity, we restrict the space of moves to those that interchange the 0s and 1s (bond exchange moves). This fixes the number fraction of bonded edges at its initial value ϕ_0 . This is not necessary; one could easily study moves which sample number fractions spread about ϕ_0 (as an aside, altering the value of ϕ_0 can lead us to study issues surrounding isostaticity or overconstrained configurations).
4. We then repeat the calculation and determine the new fitness. We follow this procedure until the fitness \mathcal{F} is maximized.

In order to efficiently sample \mathcal{G} to maximise \mathcal{F} , especially when N is large, one might choose a simulated annealing scheme, with a fictitious temperature T_f . For any nonzero T_f , there will be a distribution of optimal configurations; the true optimal network will be obtained by slowly taking $T_f \rightarrow 0$.

In practice, we have implemented the above algorithm on a triangular lattice with the number of vertices $N = 156$ arranged in a 12×13 grid. We have used a

slightly distorted lattice to avoid straight lines of vertices and that result in the appearance of floppy modes (Yan *et al.* 2017). The number of strength-1 bonds $N_S = 360$, which we fix throughout the simulation. This in turn fixes the average coordination number, $z = 2N_S/N = 5$. In addition, the vertices are also connected to their next neighbours via weak springs with stiffness 10^{-4} . Periodic boundary condition is imposed based on the specific case being modelled, as specified in section 3.

The binding of the ligand is modelled by imposing a displacement field, $\{\mathbf{u}^{\mathcal{I}}\}$, at the input vertices $i \in \mathcal{I}$ (we take it to be 4 adjacent vertices located at the centre of the lower boundary of the grid). Such an imposed displacement results in a deformation of the entire network, leading to a displacement, $\{\mathbf{u}^{\mathcal{I}'}\}$, at every other vertex of the network. We numerically evaluate $\{\mathbf{u}^{\mathcal{I}'}\}$ by solving the corresponding global stiffness matrix.

All vertices obey local force balance. Thus, for the vertices $i \in \mathcal{I}$, the external forces required to impose the displacements should balance the internal elastic forces, while for the vertices $j \in \mathcal{I}'$ (the complement of \mathcal{I}), the internal elastic forces should add up to zero. In block matrix form,

$$\begin{bmatrix} \mathbf{F}^{\mathcal{I}} \\ 0 \end{bmatrix} = \begin{bmatrix} \mathbf{B}^{\mathcal{I}\mathcal{I}} & \mathbf{B}^{\mathcal{I}\mathcal{I}'} \\ \mathbf{B}^{\mathcal{I}'\mathcal{I}} & \mathbf{B}^{\mathcal{I}'\mathcal{I}'} \end{bmatrix} \begin{bmatrix} \mathbf{u}^{\mathcal{I}} \\ \mathbf{u}^{\mathcal{I}'} \end{bmatrix}, \quad (2)$$

where \mathbf{B} is the block stiffness matrix. The unknown displacements can be obtained by simple matrix inversion.

Every time we move through the genotype space, we change the topology of the network, and construct a new \mathbf{B} , which is then used to calculate the unknown displacements $\mathbf{u}^{\mathcal{I}'}$. Under this evolution, we search for networks that generate a response, which matches a target displacement, $\mathbf{u}_{\mathcal{T}}^j$, at sites $j \in \mathcal{O}$ located far from the input stimulus \mathcal{I} . The fitness of the network is evaluated in terms of the deviation of the displacement field at the output sites from its target value,

$$\mathcal{F} = - \left(\sum_{j \in \mathcal{O}} (\mathbf{u}_{\mathcal{T}}^j - \mathbf{u}^j)^2 \right)^{1/2}. \quad (3)$$

To evolve towards the optimum in this non-convex optimization problem, we perform a Monte Carlo simulation using Metropolis sampling at a fictitious temperature $T_f = 0.01$. The simulation is performed for 5×10^5 steps, where the fitness value usually converges within 100 Monte Carlo steps. We present a

movie of the evolution of the network towards optimality in *Network Evolution* (<https://github.com/codesrivastavalab/allostery-theory/blob/main/convergence.gif>).

In the following section, we employ this algorithm to study four different functional proteins. We then characterize the optimised network in terms of the spatial profiles of the mean coordination number and displacement.

3. Emergence of functional proteins

Among the quantities we measure are the distributions means and fluctuations of scalars such as (i) averaged local coordination number (number of bonds per site with weight 1) and (ii) mean square displacement (SD) at every vertex ($\langle \frac{|\mathbf{u}^i|^2}{\sum_{i \in \mathcal{I}'} |\mathbf{u}^i|^2} \rangle$). This allows us to classify the variety of protein types according to the relative fraction of liquid to solid regions and geometry of these liquid regions. Using the above genotype–phenotype map, we study the emergence of allosteric interaction, hinge joint, crack formation and a slide bolt in functional proteins, such as adenylate kinase, HSP90, calmodulin and so on.

3.1 Allosteric proteins with slide bolt behaviour

In this case, the active site consists of 4 consecutive vertices on the top boundary. Such a representation models the case of globular allosteric proteins, where the active and allosteric sites are located at specific distant sites, each comprising a small part of the protein surface. In the abstract network, the stimulus site can thus be considered as an ‘allosteric’ site, while the site for targeted response is the ‘active’ site of an allosteric protein. A periodic boundary condition is imposed on the side boundaries along the x_1 -direction.

In figure 2, we show the typical structure of a fit network, and the mean coordination and squared displacement maps. In the fit network, the displacements at the response site are found to be close to the expected values. The mean coordination map indicates the presence of a less coordinated region connecting the stimuli and response sites, which is surrounded by two comparatively better connected regions. The shape of this ‘floppy’ region is similar to a ‘trumpet’, with the narrow end connecting the stimuli site and the wide end connecting the response site, as observed earlier in Yan *et al.* (2017). This observation indicates the possible presence of allosteric chains – highly deformable

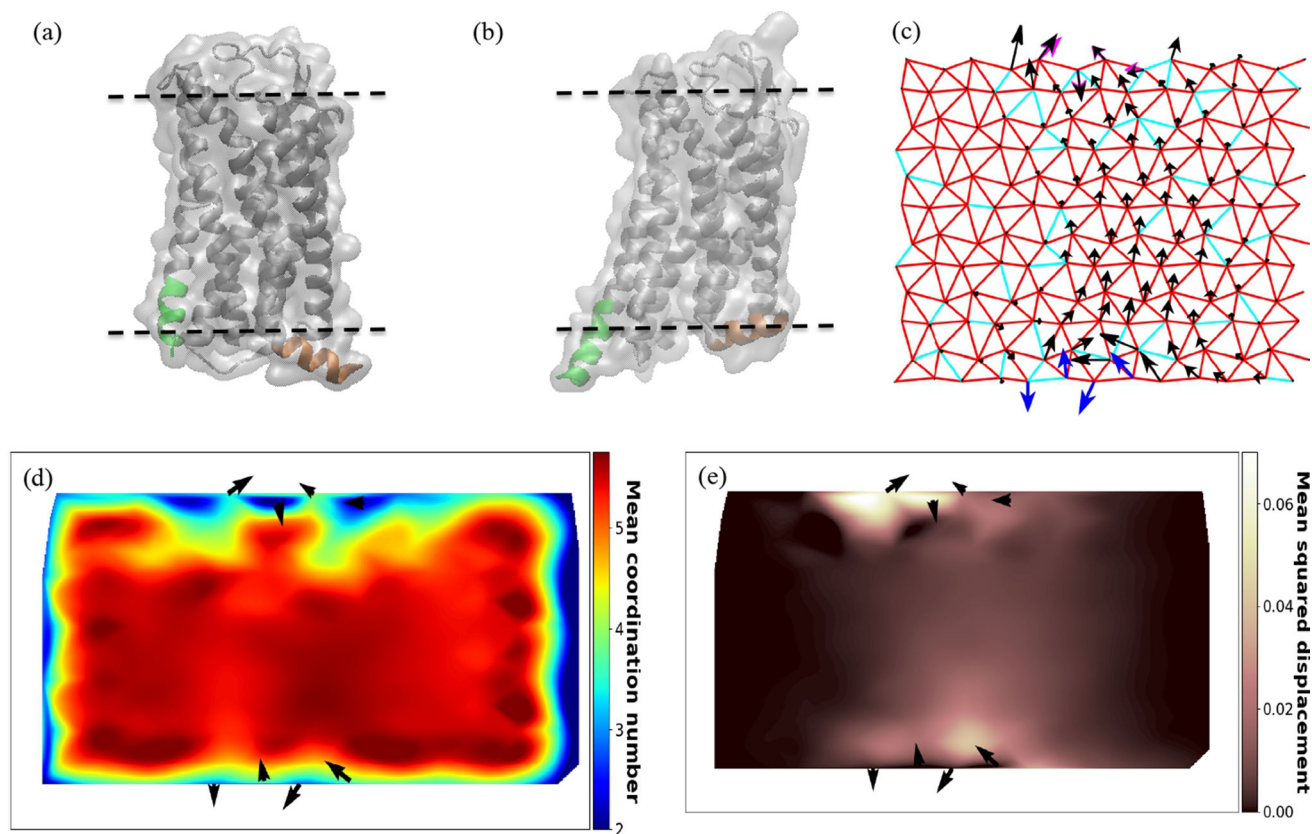


Figure 2. (a) The closed antagonist-bound inactive state conformation (PDB ID: 4YAY) and the (b) open agonist-bound fully active state conformation (PDB ID: 6DO1) of a GPCR protein (Lu *et al.* 2021) (c) An evolved optimised network. Red and cyan in the network indicate strong and zero bonds, respectively. The blue arrows indicate the imposed stimuli at the allosteric site (4 nodes at the bottom boundary), the magenta arrows on the top boundary indicate the expected response at the active sites, and the black arrows are the the response field of the optimised network. (d) Average coordination number map and (e) mean squared displacement map of the optimised network.

or ‘liquid-like’ regions in allosteric proteins whose orientation, geometry and fluctuations are tuned to the desired functionality of the protein.

In a strained elastic network, away from the site of the applied strain, the deformations die down fast. However, in this case, the deformations, measured in terms of the mean squared displacements at all the vertices of the network, decrease far away from the stimuli sites and peak again near the response sites. This feature is also noticed for the fit abstract networks in all the other cases considered. Such an observation again indicates the presence of highly deformable regions in the protein, which can allow the strain to propagate.

Implications for structure of potentially allosteric proteins are oligomers resulting from the assembly of proteomers associated in such a way that the molecule possesses at least one axis of symmetry. The oligomeric structure creates a potentially cooperative assembly of subunits (as noted by the Monod–Wyman–Changeux

(MWC) model). It remains to be seen from a detailed finite size analysis whether this continuous pathway of soft interaction from \mathcal{I} to \mathcal{O} will be retained when we increase the size of the protein.

3.2 Hinge behaviour commonly found in kinases

Proteins such as adenylate kinase (ADK) and guanylate kinase undergo open-to-closed state structural transition in order to perform their catalytic action. We model such conformational change in our abstract model by fixing the response sites at the top boundary of the network, where half of the vertices have expected displacements that are rotated relative to the other half. Through this, we intend to model the open-close motion of multi-domain proteins, such as ADK. The other two boundaries along the x_1 -direction are kept open with no periodic boundary condition.

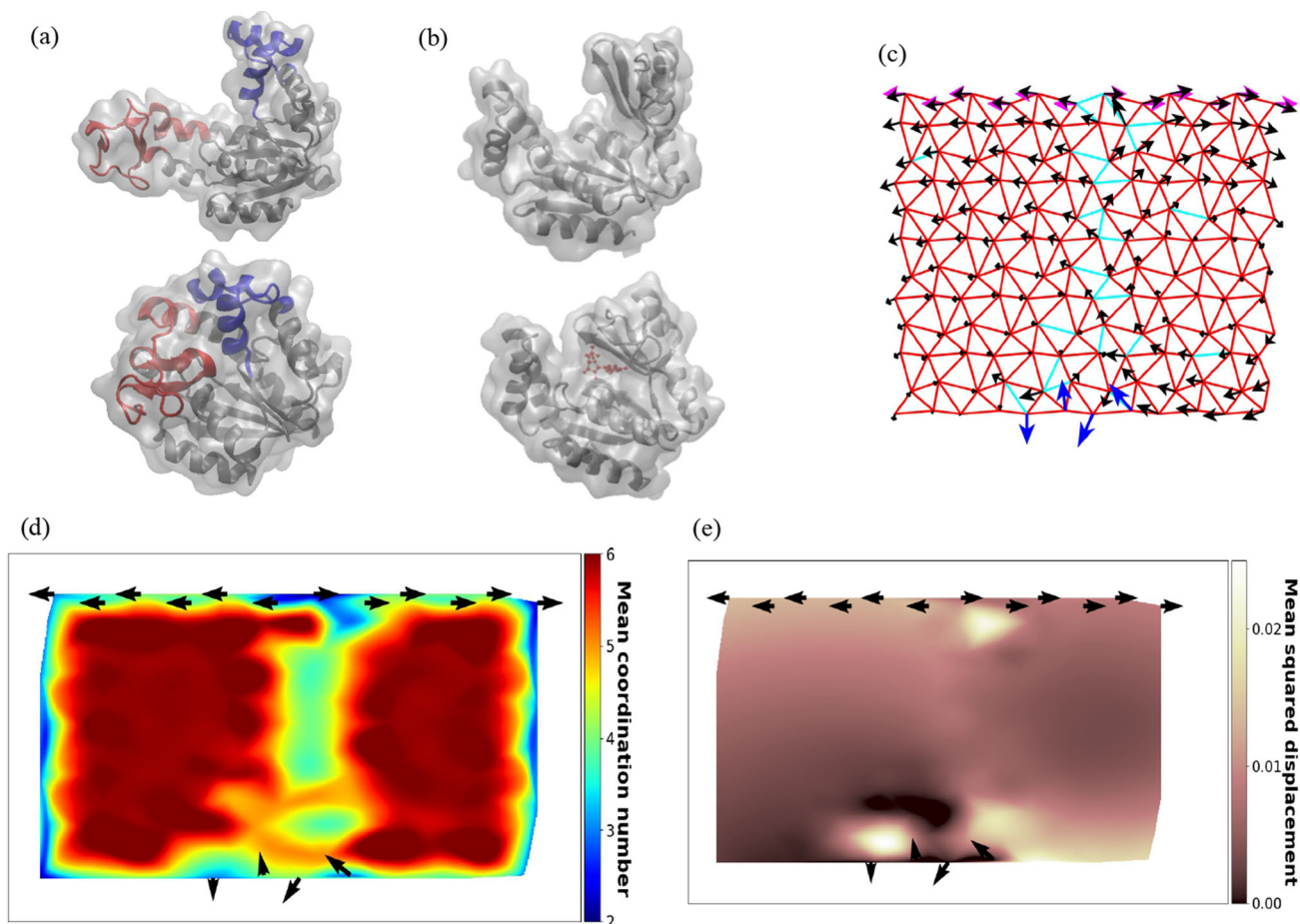


Figure 3. (a) The open-state conformation [PDB ID: 4AKE](top) and the closed-state conformation [PDB ID: 1EX6] of adenylate kinase (Müller and Schulz 1992). (b) The open-state conformation [PDB ID: 1EX6] (top) and the ligand-bound closed-state conformation [PDB ID: 1EX7] of guanylate kinase (Stehle and Schulz 1992). (c) An evolved optimised network with all the top boundary nodes as active sites. Red and cyan in the network indicate strong and zero bonds, respectively. The blue arrows indicate the imposed stimuli at the allosteric site (4 nodes at the bottom boundary), the magenta arrows on the top boundary indicate the expected response at the active sites, and the black arrows are the response field of the optimised network. (d) Average coordination number map and (e) mean squared displacement map of the optimised network.

In figure 3, we show the structure of a fit network and the mean coordination and squared displacement maps. The fit network is observed to be divided into two very rigid domains by a weakly connected liquid-like region that connects the stimuli and response sites. The two rigid domains are weakly connected near the allosteric (stimuli) sites, which mimics the hinge region of the kinases around which the rigid domains opens and closes (figure 3a and b).

3.3 Conformation changes due to ‘buried’ active sites becoming solvent-exposed

In this case, we intend to model the subsequent exposure of buried residues upon ligand binding at the target sites, such as in case of GTPase, maltose

binding protein (MBP) and calmodulin. We do this by fixing the response site at 4 consecutive vertices in the bulk of the network with target displacements perpendicular to the bottom boundary. A periodic boundary condition is imposed along the x_1 -direction as in case A (section 3.1), for globular allosteric proteins.

Figure 4 shows a fit network and the mean coordination and squared displacement maps. The mean coordination map in this case is seen to be very different from the earlier two cases. The response site is located within a strongly connected region, with a weakly coordinated region around it. This liquid-like region surrounds the response region on both sides and is connected at the site of stimuli. One can think of the rigid response region as the calcium binding sites of calmodulin that stay on the rigid surface of the protein,

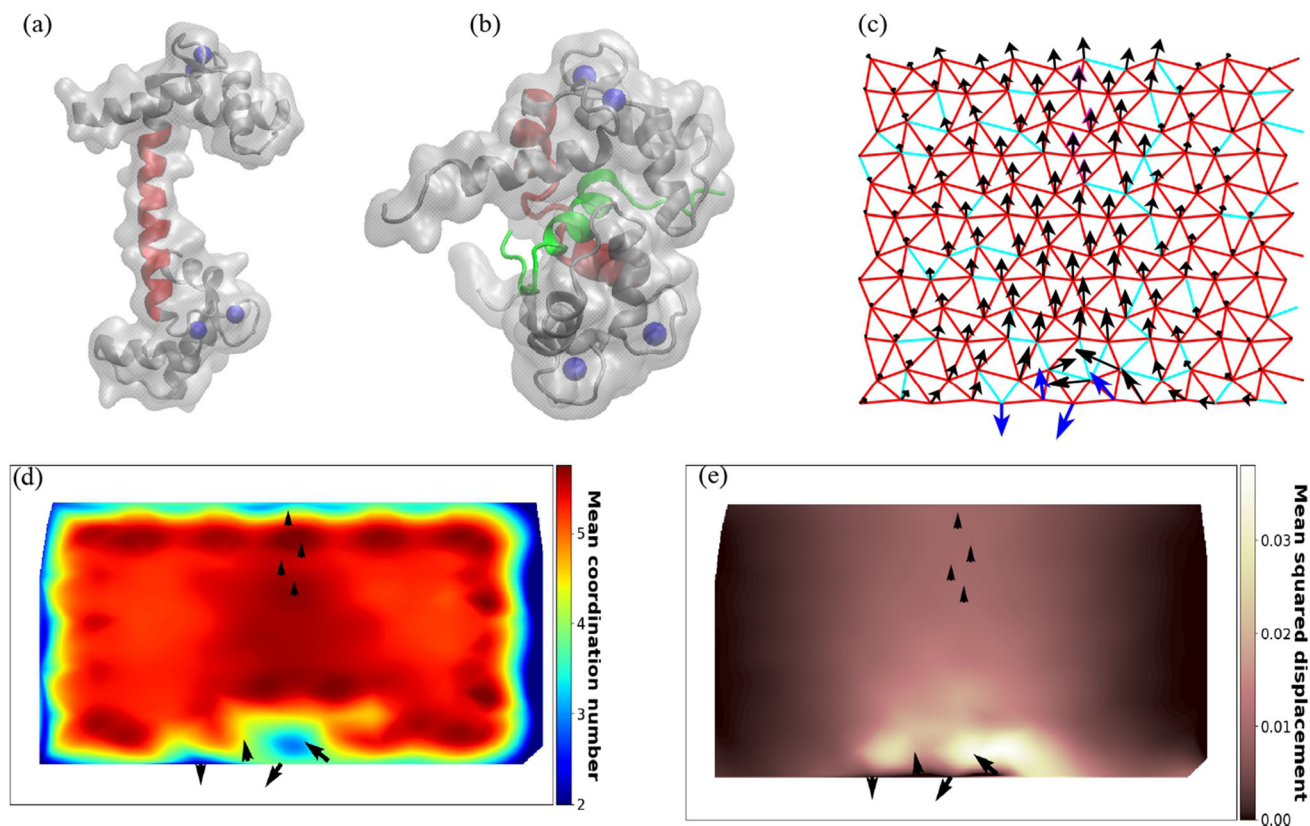


Figure 4. (a) The open-state conformation of Calmodulin (PDB ID: 3CLN) and (b) the peptide-bound state conformation (PDB ID: 1CKK) (Babu *et al.* 1988; Osawa *et al.* 1999). (c) An evolved optimised network with 4 nodes in the bulk as active sites. Red and cyan in the network indicate strong and zero bonds, respectively. The blue arrows indicate the imposed stimuli at the allosteric site (4 nodes at the bottom boundary), the magenta arrows in the bulk indicate the expected response at the active sites, and the black arrows are the the response field of the optimised network. (d) Average coordination number map and (e) mean squared displacement map of the optimised network.

while the low connected regions are the two target sites that open up when calcium is bound.

3.4 Hinge and twist motion as in chaperone proteins

Molecular chaperones like HSP90 undergo open-to-closed state structural transition that involve large domain movements. Here we model such functional proteins in terms of the abstract network, where the response site consists of the two side boundaries with target displacements that are rotated with respect to each other. Through this representation, we try to model the hinge motions of proteins consisting of two distinct domains. As the boundaries along the x_1 -direction serve as the response sites, no periodic boundary condition is applied in this case.

Figure 5 shows the structure of a fit network and the mean coordination and squared displacement maps. The displacements at the two boundaries of the fit

network are found to be very close to the expected response. The mean coordination map indicates a very weakly connected region in the middle of the network, similar to that observed in case B (section 3.2). However, unlike the former, the liquid-like region does not connect the stimuli and response sites. Rather, the network is divided into two very rigid domains which move in opposite directions. As in case B, the liquid-like region is connected at the site of applied stimuli, which acts like the hinge region. In terms of the HSP90 example, the two rigidly connected regions can be thought of as the two flexing arms, which render the open and close form of the protein (figure 5a and b).

4. Localized soft channels and non-affine elasticity

The measured quantities evaluated on the configuration or graph that optimizes the fitness have distinct features in each of the examples studied. Each of them have a

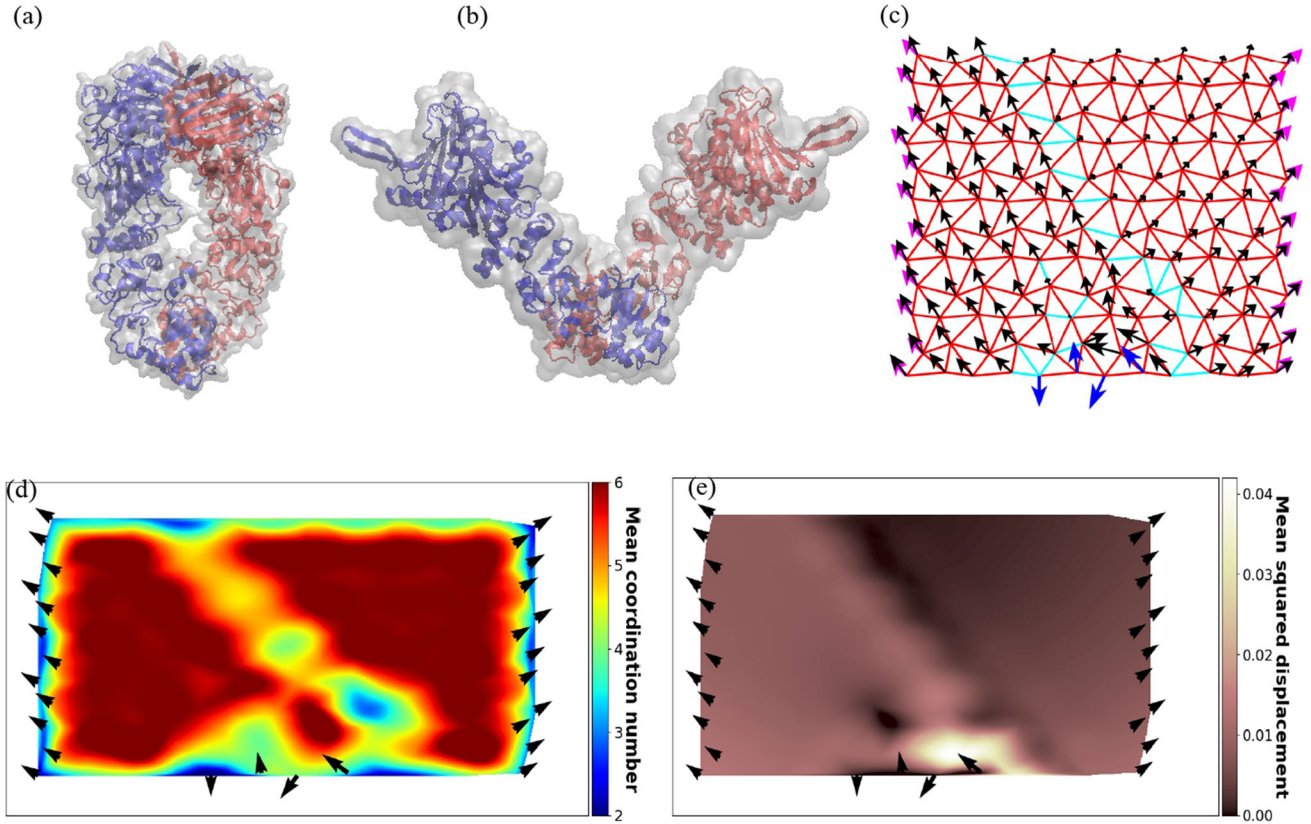


Figure 5. (a) The closed-state conformation of HSP70 (PDB ID: 2IOP) and (b) the open active state conformation (PDB ID: 2IOQ) (Shiau *et al.* 2006). (c) An evolved optimised network with all nodes on the side boundaries as the active sites. Red and cyan in the network indicate strong and zero bonds, respectively. The blue arrows indicate the imposed stimuli at the allosteric site (4 nodes at the bottom boundary), the magenta arrows on the side boundaries indicate the expected response at the active sites, and the black arrows are the response field of the optimised network. (d) Average coordination number map and (e) mean squared displacement map of the optimised network.

contiguous channel comprising vertices with low coordination number (relatively low constrained vertices) and large displacements, sharply separated from regions with high coordination number (highly constrained vertices) and low displacements. When embedded in a bath of temperature T , these low coordination number channels will be associated with large volume fluctuations; such volume fluctuations have been observed to accompany structural changes along allosteric paths (Law *et al.* 2017). The channels resemble a liquid channel embedded in an amorphous solid, and exhibit a distinct geometry and orientation. These liquid-like regions represent soft or flexible parts of the ‘evolved’ protein that drive the input–output response as encoded by the fitness function.

To proceed with this intuition, we first note from equation 1 that the optimal configurations are minimisers of the ‘energy’ $E = \frac{1}{2} \sum_{i,j} b_{ij} (\mathbf{x}^i - \mathbf{x}^j)^2$, subject to constraints implied by the fixed input and desired output. These constraints can be either taken

to be hard constraints, in which case these vertices are pinned, or soft constraints, represented as a term in the energy that represents the fitness function. This harmonic energy E can be formally represented through the spectral properties of the graph Laplacian L (Banerjee and Jost 2008). The graph Laplacian L acts on functions defined on the graph \mathcal{G} . Let u be a real-valued function on \mathcal{G} , i.e., $u : \mathcal{V} \rightarrow \mathbb{R}$, with inner product

$$(u, v) = \sum_i n_i u(i)v(i) \quad (4)$$

where n_i is the degree of i . Consider an operator Δ on this space of functions whose action on function u is

$$\Delta u(i) = u(i) - \frac{1}{n_i} \sum_{j \sim i} u(j) \quad (5)$$

If g is an arbitrary function on \mathcal{G} (and therefore, one can view g as a column vector), then

$$\frac{(g, Lg)}{(g, g)} = \frac{\sum_{i \sim j} (g(i) - g(j))^2}{\sum_i n(i)g(i)^2} \quad (6)$$

which will clearly highlight the interface of the liquid–solid regions. The spectrum of the graph Laplacian describes the interface fluctuations. One can study the evolution of the eigenvalues and eigenvectors of L as one moves through the genotype space towards the optimal configuration.

To this graph Laplacian we add the constraints implied by the fixed input and desired output. The corresponding ‘Hamiltonian’ graph operator that acts on functions on the graph is described by an elliptical operator of the form $L_G + V$, where L_G is the graph Laplacian on the network G and V is the potential that imposes this constraint in \mathcal{P} . A simple choice for V in section 3.1 is

$$V(\mathcal{P}) = \sum_{i \in \mathcal{I}} K_i (\phi_i - \phi_i^l)^2 + \sum_{j \in \mathcal{O}} J_j (\phi_j - \phi_j^a)^2 \quad (7)$$

where ϕ is the scalar function defined on G (e.g., local coordination number (density) or root square displacement) and the coefficients K_i, J_k are large so as to impose the constraint strongly. This acts like a pinning potential in the target space of \mathcal{I} and \mathcal{O} .

The Hamiltonian we have constructed bears a close resemblance to the Cahn–Hilliard theory describing the fluctuation spectrum of a pinned liquid–gas interface,

$$H[\phi(x)] = \int d^2x \left[\frac{\sigma}{2} (\nabla \phi)^2 + f(\phi) + V_{pin}(\phi) \right] \quad (8)$$

The last term is a pinning potential that breaks the Euclidean invariance of the interface (Jasnow 1984). The lowest eigenvalues of this model (Jasnow 1984) includes a capillary and peristaltic mode, which resembles the liquid-like excitations of the channel shown in figure 2.

Another perspective is from the theory of amorphous solids. One may think of the elastic network as a realisation of an amorphous solid, and ask how one may systematically tune the properties of the amorphous solid so as to get the desired phenotype (Rocks et al. 2017; Hexner et al. 2018). The ‘energy’, $E = \frac{1}{2} \sum_{i,j} b_{ij} (\mathbf{x}^i - \mathbf{x}^j)^2$, is equivalent to an elastic energy functional $\int_x \mathcal{B}(x) (\nabla u)^2$, where u is the local displacement field and \mathcal{B} are the local elastic moduli. With \mathcal{B} taken to be randomly distributed about a mean, this is equivalent to the non-affine elastic theory of amorphous solids (DiDonna and Lubensky 2005).

Now starting with a network where all the bonds are stiff, one imposes the local stress and response displacements at \mathcal{I} and \mathcal{O} . All the bonds in the network will then undergo deformation, resulting in a high elastic energy. We then make the stiffnesses of the most deformed bonds weaker ensuring that the constraints at \mathcal{I} and \mathcal{O} are maintained – this results in a lowering of the energy. The network obtained as a result of this ‘pruning’ (Hexner et al. 2018) will be the optimal network described above. This procedure corresponds to a random annealing of the elastic moduli to arrive at the optimal protein. The optimal solution arrived at in the example of the allosteric protein is akin to shear-banding in amorphous solids (Barbot et al. 2020).

5. Discussion

In this study, we explored ideas around a functional protein as an amorphous solid, designed to perform a specific function with high fidelity. The examples we studied include proteins that exhibit allosteric changes such as hinge joint (e.g., adenylate kinase and HSP90), crack formation (e.g., calmodulin) and slide bolt (e.g., GPCR). Here, we explored the mechanical rather than the chemical facets of such a mechano-chemical machine.

This mechanical approach highlights some general points of principle. For instance, it is generally believed that in the native state, the packing density is high, making it too restricted to exhibit the variety of ways in which allostery manifests. Our analysis suggests that the native state should be allowed to be locally compressible (looser packing), thus exploring a higher dimensional low energy landscape.

Our results should remind us of the concept of *sectors* (Reynolds et al. 2011), envisaged as evolutionarily conserved, spatially organized molecular motifs that can enable perturbations at specific surface positions to rapidly initiate conformational control over protein function.

The optimization of fitness \mathcal{F} over the space of phenotypes is not convex, implying that there will be many solutions to the optimisation problem. In future work, we will study the geometry of the fitness landscape, the number of minima and maxima and their proximity to one another. If there are a small number of optimal solutions, then one might expect these optimal features have been arrived at multiple times in the evolutionary history of proteins, thereby explaining the frequent reemergence of protein architectural motifs.

Many extensions of this work can be envisaged, such as extension to three dimensions, separating the backbone covalent interactions from the rest of the interactions, and including nematic correlations representing the effect of secondary structures (Chakraborty *et al.* 2021). We hope to take up these questions in the future.

Acknowledgements

It is a pleasure to present this article as part of the thematic issue titled “Emergent dynamics of biological networks” in honor of Prof. Somdatta Sinha. At a time when Theoretical Biology was not very popular in India, it was scientists like Somdatta who bravely kept the intellectual flame burning. Somdatta continues to be a gracious mentor to the younger generation of biophysicists.

MR acknowledges support from the Department of Atomic Energy (India), under project no. RTI4006, and the Simons Foundation (Grant No. 287975).

MR and SS acknowledge the award of JC Bose Fellowships, JCB/2018/000030 and JBR/2020/000015, respectively, from SERB-DST, India. AS thanks the Department of Science and Technology, India, for the early career reward (ECR) grant. MT acknowledges the research fellowship from the Department of Biotechnology, India. This research was also supported by the Department of Biotechnology, Government of India, in the form of IISc-DBT partnership programme.

References

- Ahmed S, Manjunath K, Chattopadhyay G and Varadarajan R 2022 Identification of stabilizing point mutations through mutagenesis of destabilized protein libraries. *J. Biol. Chem.* **298** 101785
- Anfinsen CB 1973 Principles that govern the folding of protein chains. *Science* **181** 223–230
- Babu Y, Bugg CE and Cook WJ 1988 Structure of calmodulin refined at 2.2 Å resolution. *J. Mol. Biol.* **204** 191–204
- Banerjee A and Jost J 2008 On the spectrum of the normalized graph laplacian. *Lin. Algebra Appl.* **428** 3015–3022
- Barbot A, Lerbinger M, Lemâitre A, Vandembroucq D and Patinet S 2020 Rejuvenation and shear banding in model amorphous solids. *Phys. Rev. E* **101** 033001
- Cates ME, Wittmer JP, Bouchaud J-P and Claudin P 1998 Jamming, force chains, and fragile matter. *Phys. Rev. Lett.* **81** 1841–1844
- Chakraborty D, Mugnai ML and Thirumalai D 2021 On the emergence of orientational order in folded proteins with implications for allostery. *Symmetry* **13** 770
- Changeux JP 2012 Allostery and the Monod-Wyman-Changeux model after 50 years. *Annu. Rev. Biophys.* **41** 103–133
- Cherezov V, Rosenbaum DM, Hanson MA, *et al.* 2007 High Resolution crystal structure of an engineered human 2-adrenergic G protein-coupled receptor. *Science* **318** 1258–1265
- Chu J-W and Voth GA 2007 Coarse-grained free energy functions for studying protein conformational changes: A double-well network model. *Biophys. J.* **93** 3860–3871
- Csermely P, Palotai R and Nussinov R 2010 Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Trends Biochem. Sci.* **35** 539–546
- DiDonna BA and Lubensky TC 2005 Nonaffine correlations in random elastic media. *Phys. Rev. E* **72** 066619
- Dutta S, Eckmann J-P, Libchaber A and Tlusty T 2018 Green function of correlated genes in a minimal mechanical model of protein evolution. *Proc. Natl. Acad. Sci. USA* **115** E4559–E4568
- Eckmann J-P, Rougemont J and Tlusty T 2019 Proteins: The physics of amorphous evolving matter. *Rev. Mod. Phys.* **91** 031001
- Frauenfelder H, Sligar S and Wolynes P 1991 The energy landscapes and motions of proteins. *Science* **254** 1598–1603
- Garel T, Orland H and Pitard E 1997 Protein Folding and Heteropolymers. in A P Young (ed) *Spin Glasses and Random Fields, Series on Directions in Condensed Matter Physics* vol. 12 (World Scientific) pp. 387–443
- Halabi N, Rivoire O, Leibler S and Ranganathan R 2009 Protein sectors: Evolutionary units of three-dimensional structure. *Cell* **138** 774–786
- Hexner D, Liu AJ and Nagel SR 2018 Role of local response in manipulating the elastic properties of disordered solids by bond removal. *Soft Matter* **14** 312–318
- Hilger D, Masureel M and Kobilka BK 2018 Structure and dynamics of GPCR signaling complexes. *Nat. Struct. Mol. Biol.* **25** 4–12
- Iben IET, Braunstein D, Doster W, *et al.* 1989 Glassy behavior of a protein. *Phys. Rev. Lett.* **62** 1916–1919
- Jasnow D 1984 Critical phenomena at interfaces. *Rep. Prog. Phys.* **47** 1059–1132
- Koonin EV, Wolf YI and Karev GP 2002 The structure of the protein universe and genome evolution. *Nature* **420** 218–223
- Koshland DE, Nemethy G and Filmer D 1966 Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* **5** 365–385
- Kuhlman B, Dantas G, Ireton GC, *et al.* 2003 Design of a novel globular protein fold with atomic-level accuracy. *Science* **302** 1364–1368

- Law AB, Sapienza PJ, Zhang J, Zuo X and Petit CM 2017 Native state volume fluctuations in proteins as a mechanism for dynamic allostery. *J. Am. Chem. Soc.* **139** 3599–3602
- Liang J and Dill KA 2001 Are proteins well-packed? *Biophys. J.* **81** 751–766
- Lu S, He X, Yang Z, et al. 2021 Activation pathway of a g protein-coupled receptor uncovers conformational intermediates as targets for allosteric drug design. *Nat. Commun.* **12** 4721
- Maragakis P and Karplus M 2005 Large amplitude conformational change in proteins explored with a plastic network model: Adenylate kinase. *J. Mol. Biol.* **352** 807–822
- Milo R and Phillips R 2015 *Cell biology by the numbers* (Garland Science)
- Monod J, Wyman J and Changeux J-P 1965 On the nature of allosteric transitions: A plausible model. *J. Mol. Biol.* **12** 88–118
- Müller C, Schlauderer G, Reinstein J and Schulz G, 1996 Adenylate kinase motions during catalysis: an energetic counterweight balancing substrate binding. *Structure* **4** 147–156
- Müller CW and Schulz GE, 1992 Structure of the complex between adenylate kinase from *Escherichia coli* and the inhibitor ap5a refined at 1.9 Å in solution: A model for acatalytic transition state. *J. Mol. Biol.* **224** 159–177
- Osawa M, Tokumitsu H, Swindells M, et al. 1999 A novel target recognition revealed by calmodulin in complex with Ca²⁺-calmodulin-dependent kinase kinase. *Nat. Struct. Biol.* **6** 819–824
- Rajasekaran N and Naganathan AN 2017 A self-consistent structural perturbation approach for determining the magnitude and extent of allosteric coupling in proteins. *Biochem. J.* **474** 22
- Ramanoudjame G, Du M, Mankiewicz KA and Jayaraman V 2006 Allosteric mechanism in ampa receptors: A fret-based investigation of conformational changes. *Proc. Natl. Acad. Sci. USA* **103** 10473–10478
- Reynolds K, Mclaughlin R and Ranganathan R 2011 Hot spots for allosteric regulation on protein surfaces. *Cell* **147** 1564–1575
- Rocks JW, Pashine N, Bischofberger I, et al. 2017 Designing allostery-inspired response in mechanical networks. *Proc. Natl. Acad. Sci. USA* **114** 2520–2525
- Sartori P and Leibler S 2020 Lessons from equilibrium statistical physics regarding the assembly of protein complexes. *Proc. Natl. Acad. Sci. USA* **117** 114–120
- Shiau AK, Harris SF, Southworth DR and Agard DA 2006 Structural analysis of *E. coli* hsp90 reveals dramatic nucleotide-dependent conformational rearrangements. *Cell* **127** 329–340
- Smock RG, Rivoire O, Russ WP, et al. 2010 An interdomain sector mediating allostery in hsp70 molecular chaperones. *Mol. Syst. Biol.* **6** 414
- Socci ND and Onuchic JN 1994 Folding kinetics of protein like heteropolymers. *J. Chem. Phys.* **101** 1519–1528
- Stefan MI, Edelstein SJ and Nov'ere NL 2008 An allosteric model of calmodulin explains differential activation of pp2b and camkii. *Proc. Natl. Acad. Sci. USA* **105** 10768–10773
- Stehle T and Schulz GE 1992 Refined structure of the complex between guanylate kinase and its substrate gmp at 2.0 Å in solution. *J. Mol. Biol.* **224** 1127–1141
- Thlusty T, Libchaber A and Eckmann J-P 2017 Physical model of the genotype-to-phenotype map of proteins. *Phys. Rev. X* **7** 021037
- Tsai C-J, Ma B and Nussinov R 1999 Folding and binding cascades: Shifts in energy landscapes. *Proc. Natl. Acad. Sci. USA* **96** 9970–9972
- Weis WI and Kobilka BK 2018 The molecular basis of G protein-coupled receptor activation. *Annu. Rev. Biochem.* **87** 897–919
- Yan L, Ravasio R, Brito C and Wyart M 2017 Architecture and coevolution of allosteric materials. *Proc. Natl. Acad. Sci. USA* **114** 2526–2531
- Yan L, Ravasio R, Brito C and Wyart M 2018 Principles for optimal cooperativity in allosteric materials. *Biophys. J.* **114** 2787–2798

Corresponding editor: SUSMITA ROY