Physical model of the sequence-to-function map of proteins

Tsvi Tlusty,[1, 2, 3] Albert Libchaber,[4] and Jean-Pierre Eckmann[5]

[1]*Center for Soft and Living Matter, Institute for Basic Science (IBS), Ulsan 44919, Korea*
[2]*Department of Physics, Ulsan National Institute of Science and Technology (UNIST), Ulsan 44919, Korea*
[3]*Simons Center for Systems Biology, Institute for Advanced Study, Princeton, NJ 08540, USA*
[4]*The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA*
[5]*Département de Physique Théorique and Section de Mathématiques,*
Université de Genève, CH-1211, Geneva 4, Switzerland

We treat proteins as amorphous learning matter: A ‘gene’ encodes bonds in an ‘amino acid’ network making a ‘protein’. The gene is evolved until the network forms a shear band across the protein, which allows for long-range soft modes required for protein function. The evolution projects the high-dimensional sequence space onto a low-dimensional space of mechanical modes, in accord with the observed dimensional reduction between genotype and phenotype of proteins. Spectral analysis shows correspondence between localization around the shear band of both mechanical modes and sequence ripples.

PACS numbers: 87.14.E-, 87.15.-v, 87.10.-e

DNA genes code for the three-dimensional configurations of amino acids that make functional proteins. This sequence-to-function map is hard to decrypt since it links the collective physical interactions inside the protein to the corresponding evolutionary forces acting on the gene [1–3]. Furthermore, evolution has to select the tiny fraction of functional sequences in an enormous, high-dimensional space [4], which implies that protein is a non-generic, information-rich matter, outside the scope of standard statistical methods. Therefore, although the structure and physical forces within a protein have been extensively studied, the fundamental question as to how a functional protein originates from a linear DNA sequence is still open, in particular, how the functionality constrains the accessible DNA sequences.

We examine the geometry of the sequence-to-function map, and devise a simple mechanical model of proteins as amorphous learning matter. We base our model on the growing evidence that large-scale conformational changes – where big chunks of the protein move with respect to each other – are central to function [5, 6]. In particular, allosteric proteins can be viewed as ‘mechanical transducers’ that transmit regulatory signals between distant sites [7–9]. Recent measurements showed viscoelastic flow within enzymes [10] with mechanical stress affecting catalysis [11], while analysis of structural data demonstrated localization of the strain in 2D bands across allosteric proteins [12]. All this motivates us to take as a target function to be evolved in our ‘protein’ such a large-scale dynamical mode. Other important functional constraints, such as specific chemical interactions at binding sites, are disregarded here because they are confined to a small fraction of the protein. Therefore we focus on this mechanical function, which involves many amino acids. We show that its collective nature leads to long-range correlation patterns in the gene.

In our model, the target mechanical mode is evolved by mutating the ‘gene’ that determines the connectivity in the amino acid network. During the simulated ‘evolution’, mutations eventually divide the protein into ‘rigid’ and ‘floppy’ domains, and this division enables large-scale motion in the protein [13]. The model thus provides a concrete map between the sequence, configuration, and function of the ‘protein’. The computational simplicity allows for a massive survey of the sequence universe, which reveals a strong signature of the protein’s structure and function within correlation ‘ripples’ that appear in the space of DNA sequences.

Model. Our ‘protein’ is an aggregate of amino acids (AAs), modeled as beads, with short-range interactions given as bonds (Fig. 1A). A typical protein is made of several hundred AAs, and we take $N = 540$. We layer the AAs on a cylinder, 18 high 30 wide, similar to dimensions of globular proteins. The cylindrical configuration allows for fast calculation of the low energy modes, and thereby fast evolution of the protein. Each AA may connect to the nearest five AAs in the layer below, so that we get $2^5 = 32$ AA species, which are encoded as 5-letter binary *codons*. These codons specify the bonds in the protein in a 2550-long *sequence* of the *gene* $(5 \times 30 \times (18 - 1)$, because the lowest layer is connected only upwards). To become functional, the protein has to evolve a *configuration* of AAs and *bonds* that can transduce a mechanical signal from a prescribed input at the bottom of the cylinder to a prescribed output at its top. This signal is a large-scale, low-energy deformation where one domain moves rigidly with respect to another in a shear or hinge motion, which is facilitated by the presence of a fluidized, ‘floppy’ channel separating the rigid domains [15].

The large-scale deformations are governed by the rigidity pattern of the configuration, which is determined by the connectivity of the AA network via a simple majority rule (Fig. 1A). The *input* is the rigidity state of the bottom row in which AAs can be either rigid or fluidized and potentially ‘shearable’. The rigidity state propagates upwards: Depending on the number of bonds and the state of other AAs in its immediate neighborhood, an AA will be rigidly connected, ‘shearable’, *i.e.*, loosely connected, or in a pocket of less connected AAs within a rigid neighborhood [14].

As the sequence and hence the connections mutate, the model protein adapts to a desired input-output relation spec-

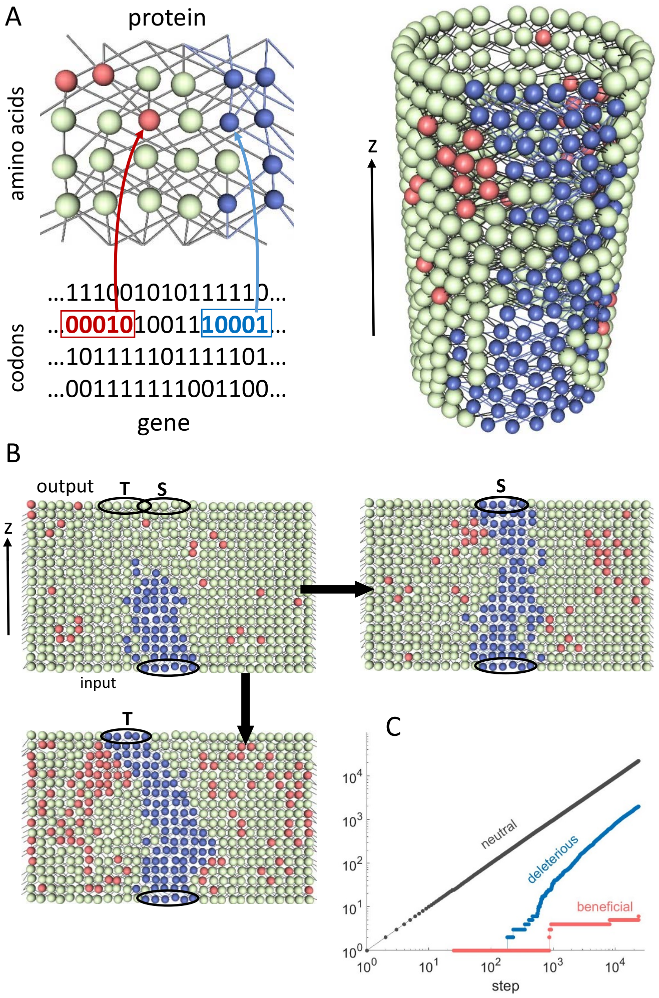

FIG. 1. (Color online) **The mechanical transducer model of proteins**. (A) The protein is made of $N = 540$ (30x18) amino acids (AAs), which are layered on a cylinder (right). Each AA may connect to the nearest five AAs in the layer below, and the bonds define $2^5 = 32$ AA species, which are encoded in the gene as 5-letter binary codons (left). To become functional, we require the connectivity of the AA network to form a 'floppy' fluid channel (or shear band) between prescribed bottom and top rows. Such configuration can transduce a mechanical signal of shear or hing motion along the fluid channel. Each AA can be in three states: rigid (gray) or fluid (*i.e.*, non-rigid), which are divided between shearable (blue) and non-shearable (red). The input is given as the rigidity state of the bottom row. The state in the rest of the protein is determined by propagation rules: to become rigid (gray), an AA should connect to a least two rigid AAs in the row below. A non-rigid AA becomes shearable (blue) if at least one of its three nearest neighbors below is shearable. Other non-rigid AAs reside in non-shearable pockets within rigid domains (red). The output is the rigidity state of the top layer, which depends on the existence of a shearable inter-domain band across the protein [14]. (B) Metropolis evolution: An initial configuration with a given input (black ellipse) and random sequence is required to evolve into a straight fluid channel (S) or a tilted one (T). (C) In each generation, a randomly drawn bit (a letter in the 5-bit codon) is flipped, and this 'point mutation' is changing exactly one bond. A typical run is a sequence of mostly neutral steps, a fraction of deleterious ones, and rare beneficial steps.

ified by the extremities of the separating fluid channel (Fig. 1B). Our simulations start from a randomized sequence and at each time step we flip a randomly drawn bit, thus adding or deleting a bond. In a zero-temperature Metropolis fashion, we keep only mutations which do not increase the distance from the target function, *i.e.*, the number of errors between the state in the top row and the prescribed outcome. Note that, following the logics of biological evolution, the 'fitness' of the 'protein' is only measured at its functional *surface* (*e.g.*, where a substrate binds to an enzyme) but not in its interior. Typically, after 10^3-10^5 mutations this input-output problem is solved (Fig. 1C). Although the functional sequences are extremely sparse among the 2^{2550} possible sequences, the small bias for getting closer to the target in configuration space directs the search rather quickly.

Dimension. The simplicity of the model allows to repeat the search numerous times (10^6). We have explored many variants of the model with strong robustness of the outcome, as we will document elsewhere. A further advantage is that the models can actually explore large parts of the universe as seen from the typical inter-sequences distance, which is comparable with the universe diameter (Fig. 2B).

The distribution of solutions in sequence space reveals a strong *dimensional reduction* when mapped to the space of functional configurations (Fig. 2A): In sequence space, the observed dimension [16, 17] is practically infinite (~ 150) [18]. This shows that the bonds are chosen basically at random, although we only consider functional sequences. On the other hand, very few among the 2^{540} configurations are solutions, owing to the physical constraints of contiguous rigid and shearable domains. As a result, when mapped to the configuration space, the solutions exhibit a dramatic dimensional reduction to a dimension of about 8-10 [19]. In our simple model, the empirical dimensional reduction between 'genotype' (sequence) and 'phenotype' (configuration, function) [20, 21] is the outcome of physical constraints on the mechanical transduction problem. In the nearly random background of sequence space, these constraints are manifested in long-range correlations among AAs on the boundary of the shearable region (Fig. 2C).

Spectral analysis (using singular value decomposition) of the solution set in both sequence and configuration spaces provides further information on the sequence-to-function map (Fig. 3). In the configuration spectrum, there are about 8-10 eigenvalues which stand out from the continuous spectrum, corresponding to the dimension 8 shown in Fig. 2A. Although the dimension of the sequence space is high (~ 150), there are again only 8-9 eigenvalues outside the continuous random spectrum [23]. These isolated eigenvectors (EVs) distill beautifully the non-random components within the mostly-random functional sequences. The EVs of both sequence and configuration are localized around the interface between the shearable and rigid domains. The similarity in number and in spatial localization of the eigenvectors reveals the tight correspondence between the configuration and sequence spaces. This duality is the outcome of the sequence-to-function map defined

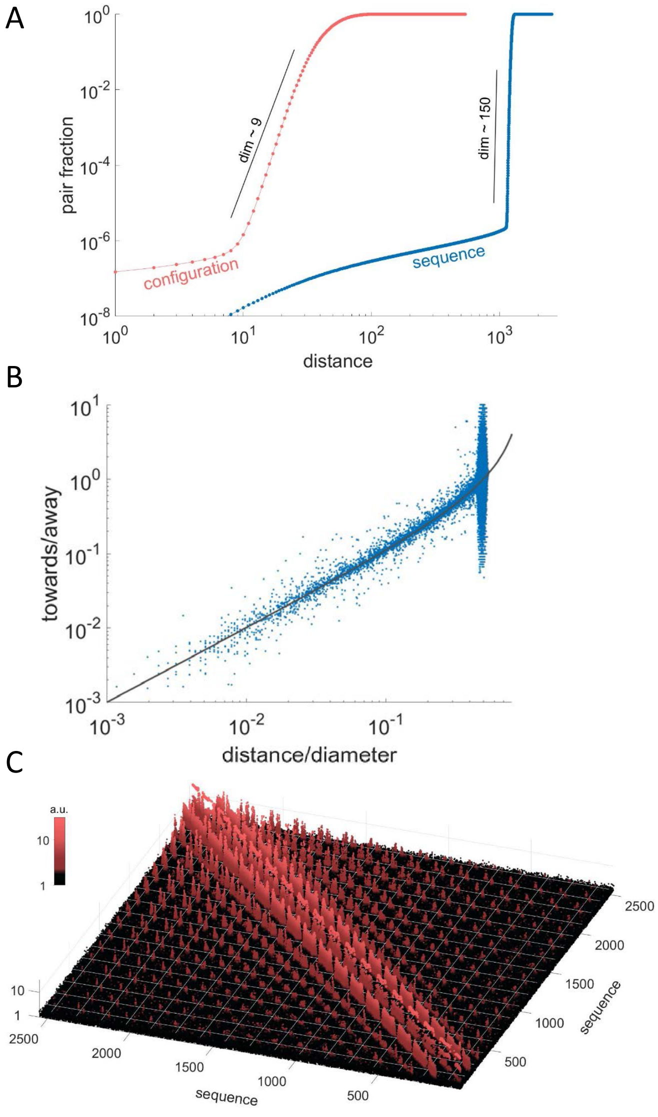

FIG. 2. (Color online) **Dimensional reduction of the sequence-to-function map**: (A) 10^6 independent functional configurations were found for the input-output problem T. An estimate for the dimension of the solutions is the correlation length, the slope of the cumulative fraction of solution pairs as a function of distance. In configuration space (red), the distance is the number of AAs (out of 540) with a different rigidity state. The estimated dimension from $10^{12}/2$ distances is about 9 (black line). In sequence space (blue), the (Hamming) distance is the number of positions which differ between two of the 2550-long sequences. The sequence space is a 2550-dimensional hypercube with 32^{510} sequences. Most distances are close to the typical distance between two random sequences (2550/2 = 1275), indicating a high-dimensional solution space. An estimate for the dimension is ~ 150 (black line). (B) A measure for the expansion in the functional sequence universe is the backward/forward ratio, the fraction of point mutations that make two sequences closer vs. the ones that increase the distance [4]. The distances d (normalized by the universe diameter = 2550) show that most sequences reach the edge of the universe, where no further expansion is possible. The black curve, $d/(1-d)$, is from purely random mutations. (C) The sequence correlation matrix across the 10^6 examples shows long-range correlations among the bits (codons) at the rigid/fluid boundary, and short-range correlations in the rigid domains (graphs for problem S in [14]).

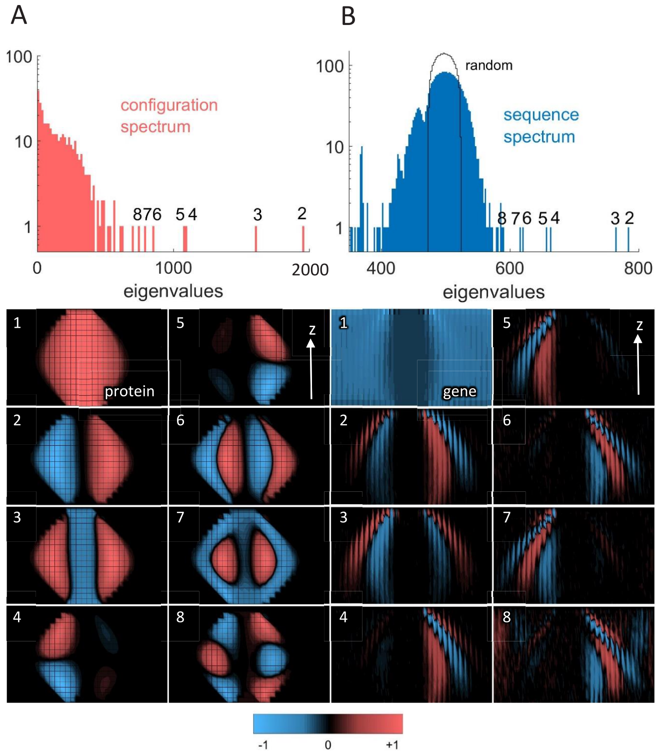

FIG. 3. (Color online) **Correspondence of modes in sequence and configuration spaces.** We produced the spectra by singular value decomposition of the 10^6 solutions of problem S (T in [14]) . (A) Top: the spectrum in configuration space exhibits about 8-10 eigenvalues outside the continuum (large 1st eigenvalue not shown). Bottom: the corresponding eigenvectors describe the basic modes of the fluid channel, such as side-to-side shift (2nd) or expansion (3rd). (B) Top: The spectrum of the solutions in sequence space is similar to that of random sequences (black line), except for about 8-9 high eigenvalues that are outside the continuous spectrum. Bottom: the first 8 eigenvectors exhibit patterns of correlation ‘ripples’ around the fluid channel region. Seeing these ripples through the random evolutionary noise required at least 10^5 independent solutions [22].

by our simple model: The geometric constraints of forming a shearable band within a rigid shell, required for inducing long-range modes, are mirrored in long-range correlations among the codons (bits) in sequence space. The corresponding sequence eigenvectors may be viewed as weak ripples of information over a sea of random sequences, as only about 8 out of 2550 modes are non-random (0.3%). These information ripples also reflect the self-reference of proteins and DNA via the feedback loops of the cell circuitry [24].

Stability under mutations. First, we determine how many mutations lead to a destruction of the solution (Fig. 4A). About 10% of all solutions are destroyed by just one random mutation. The exponentially decaying probability of surviving m mutations signals that these mutations act quite independently. Fig. 4B which shows the location of these destructive mutations around the shearable channel. We have also studied the loci where two *interacting* mutations will destroy a solution (*i.e.* none of the two is by itself destructive). In most

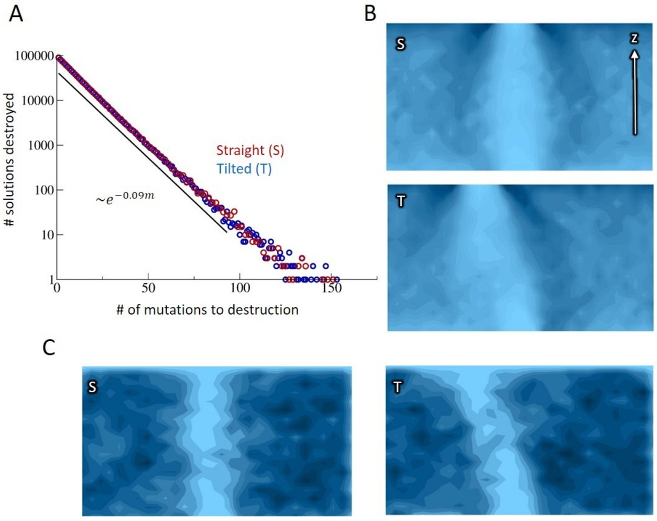

FIG. 4. (Color online) **Stability to mutations.** Mutations at sensitive positions of the sequence move the output away from the prescribed solution. (A) Fraction of runs (among 10^6) destroyed by the m-th mutation. A single mutation destroyed about 9% of solutions. The proportion decays exponentially like $\exp(-0.09m)$. (B) The density map of such mutations for problem S and T (Fig. 1B) shows accumulation around the fluid channel and at the top layer (dark regions). (C) The double mutations are evenly distributed in the rigid regions.

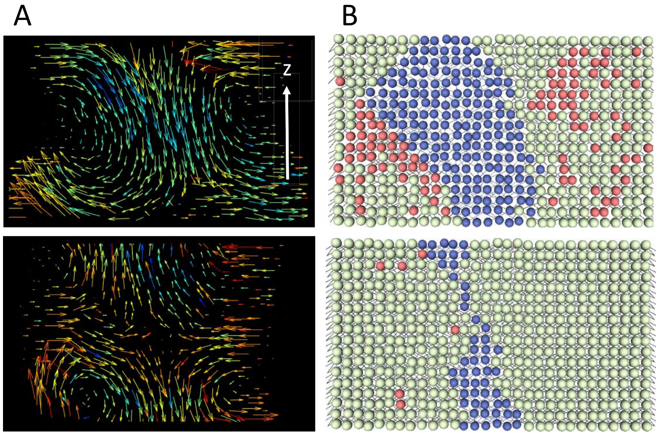

FIG. 5. (Color online) (A) **Mechanical shear modes.** Displacement and strain fields for the tilted solution T for two low eigenvalues. The vectors show the direction of the displacement and the color code denotes the stress (*i.e.*, the local change in the vector field as a function of position, maximal stress is red). (B) **Thermal Stability.** Extreme configurations, with low (50%, left) and high (95%, right) bond density, solve problem T.

cases, the two mutations are close to each other, acting on the same site. The channel is less vulnerable to such mutations, but the twin mutations are evenly distributed over the whole rigid network (Fig. 4C).

Mechanical shear modes. The evolved rigidity pattern supports low-energy modes with strain localized in the floppy, fluid channel. Since the rigidity is calculated by a simplified ad-hoc model, we tested whether the evolved AA network indeed induces such modes (Fig. 5A). A solution of the DNA/protein problem is given by a set of bonds, which defines a graph on the 540 AA nodes. This graph is embedded in 2D with AAs connected by harmonic springs (all with the same spring constant). The shear motion of such a network is characterized by the modes of its elastic tensor $\mathcal{M}$. This tensor is the $2N$x$2N$ curvature matrix in the harmonic expansion of the elastic energy $E \simeq \frac{1}{2}\delta\mathbf{r}^T\mathcal{M}\ \delta\mathbf{r}$, where $\delta\mathbf{r}$ is the $2N$-vector of the 2D displacements of the N AAs. $\mathcal{M}$ has the structure of the network Laplacian multiplied by the 2x2 tensors of directional derivatives (for details, [14] and [25, pp. 618–9]).

The first three zero EVs of $\mathcal{M}$ correspond to 2D translation and rotation symmetries of the whole protein. Another type of trivial zero EVs are associated with any patch of AAs which is totally disconnected from the rest of the network. Since the density of bonds is about $\frac{1}{2}$ and otherwise quite random, and there are 2×5 bonds at each interior AA, we expect a fraction of about $2^{-10} \sim 10^{-3}$ of isolated AAs, and even fewer patches of greater size. Further zero modes come from AAs which are connected only by one bond, and can therefore oscillate freely sideways. The probability of finding such a node is about $\binom{10}{1}/2^{10} \sim 10^{-2}$. Thus, Fig. 5A shows the EVs only for the first non-trivial eigenvalues.

Conclusion. The rigidity/shearability pattern determines the dynamical modes of the protein (Fig. 5A). The evolution of a solution with a shearable channel surrounded by rigid domains is manifested by spatially-extended low energy modes. These modes exhibit shear and hinge motions where the strain is localized in the shearable channel and where the surrounding domains translate or rotate as rigid bodies. The least random, strongly correlated sites are in the rigid shell that envelops the shearable channel. Our model predicts that these sites are also the most vulnerable to mutations (Fig. 4B), which distort the low-frequency modes and thus hamper the biological function. The large solution set allows the protein to simultaneously adapt to other tasks. For example, evolving a specific binding site, or tuning the stability to adapt to extreme temperatures by varying the bond density [26] (Fig. 5B). These effects can be examined by combining mutation surveys, biochemical assays of the function, and physical measurements of the low-frequency spectrum, especially in allosteric proteins. The model is easily extended to versions with actual springs and connections depending on pairwise interactions of neighboring sites. The concrete genotype-to-phenotype map in our simple model demonstrates that most of the gene records random evolution, while only a small non-random fraction is constrained by the biophysical function. This drastic dimensional reduction is the origin of the flexibility and evolvability in the functional solution set.

We thank S. Leibler, M.R. Mitchell and E. Moses for essential discussions and encouragement. JPE is supported by an ERC advanced grant ‘Bridges’, and TT by the Institute for Basic Science IBS-R020 and the Simons Center for Systems Biology of the Institute for Advanced Study, Princeton.

[1] E. V. Koonin, Y. I. Wolf, and G. P. Karev, Nature **420**, 218 (2002).
[2] M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, and F. A. Kondrashov, Nature **490**, 535 (2012).
[3] K. B. Zeldovich and E. I. Shakhnovich, Annu. Rev. Phys. Chem. **59**, 105 (2008).
[4] I. S. Povolotskaya and F. A. Kondrashov, Nature **465**, 922 (2010).
[5] D. Koshland, Proc. Natl. Acad. Sci. U.S.A. **44**, 98 (1958).
[6] K. A. Henzler-Wildman, V. Thai, M. Lei, M. Ott, M. Wolf-Watz, T. Fenn, E. Pozharski, M. A. Wilson, G. A. Petsko, M. Karplus, C. G. Hubner, and D. Kern, Nature **450**, 838 (2007).
[7] M. F. Perutz, Nature **228**, 726 (1970).
[8] N. M. Goodey and S. J. Benkovic, Nat. Chem. Biol. **4**, 474 (2008).
[9] S. W. Lockless and R. Ranganathan, Science **286**, 295 (1999).
[10] H. Qu and G. Zocchi, Phys. Rev. X **3** (2013).
[11] C. Joseph, C. Y. Tseng, G. Zocchi, and T. Tlusty, Plos One **9** (2014), ARTN e101442 10.1371/journal.pone.0101442.
[12] M. R. Mitchell, T. Tlusty, and S. Leibler, (2016), submitted.
[13] M. Gerstein, A. M. Lesk, and C. Chothia, Biochemistry **33**, 6739 (1994).
[14] See Supplemental Material at [URL] for further details.
[15] S. Alexander, Phys. Rep. **296**, 65 (1998).
[16] I. Procaccia, Nature **333**, 498 (1988).
[17] J.-P. Eckmann and D. Ruelle, Physica D **56**, 185 (1992).
[18] We lack sufficient data to determine such high dimensions precisely.
[19] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).
[20] Y. Savir, E. Noor, R. Milo, and T. Tlusty, Proc. Natl. Acad. Sci. U.S.A. **107**, 3475 (2010).
[21] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon, Science **336**, 1157 (2012).
[22] T. Teşileanu, L. J. Colwell, and S. Leibler, PLoS Comput. Biol. **11**, e1004091 (2015).
[23] It is tempting to also study the continuous part of this spectrum, which is not quite of the standard form. While in principle, this could be done by taking into account the known correlations, even the techniques of [27] seem difficult to implement. We thank T. Guhr for helpful discussions.
[24] T. Tlusty, Phil. Trans. Roy. Soc. A **374** (2016).
[25] F. R. K. Chung and S. Sternberg, Journal of Graph Theory **16**, 605 (1992).
[26] R. Jaenicke and G. Böhm, Curr. Opin. Struct. Biol. **8**, 738 (1998).
[27] T. Guhr, A. Müller-Groeling, and H. A. Weidenmüller, Phys. Rep. **299**, 189 (1998).

Supplemental Materials

Physical model of the sequence-to-function map of proteins

Tsvi Tlusty, Albert Libchaber, and Jean-Pierre Eckmann

CONTENTS

1. THE MODEL

We model the protein as an aggregate of amino acids (AAs) with short range interactions. In our coarse grained model, beads represent the AAs and bonds their interactions with neighboring AAs (Fig. 1A). We consider a simplified cylindrical geometry, where the AAs are layered on the surface of a cylinder at randomized positions, to represent the non-crystalline packing of this amorphous matter. Throughout this study, we examine a geometry with height $h (= 18)$, *i.e.*, the number of layers in the z direction, and width $w (= 30)$, *i.e.*, the circumference of the cylinder. When the cylinder is shown as a flat 2D surface (such as in Fig. 1B), there are still periodic boundary conditions in the horizontal w direction. The row and column coordinates of an AA are (r, c), with r for the row $(1, \ldots, h)$ and c for the column $(1, \ldots, w)$. The cylindrical periodicity is accounted for by taking the horizontal coordinate c modulo $w = 30$, $c \rightarrow \mod_w (c - 1) + 1$.

Each AA in row r can connect to any of its five nearest neighbors in the next row below, $r - 1$. This defines $2^5 = 32$ species of amino acids that differ by their ‘chemistry’, *i.e.*, by the pattern of their bonds. Therefore, in the gene, each AA at (r, c) is encoded as a 5-letter binary *codon*, ℓ_{rck}, where the *k*-th letter denotes the existence $(= 1)$ or absence $(= 0)$ of the *k*-th bond. The gene is the sequence of $N_{AA} = w \cdot h = 540$ codons which represent the AAs of the protein. It is a genetic *sequence* of $2700 = w \cdot h \cdot 5$ digits 0 or 1. Each of these numbers determines whether or not a *bond* connects two positions of the grid. Since the bonds from the bottom row do not affect the configuration of the protein and the resulting dynamical modes, the relevant length of the gene is somewhat smaller, $N_S = 2550 = w \cdot (h - 1) \cdot 5$.

Each AA position will have two binary properties, which define its state:

- The *rigidity* σ: This property can be *rigid* ($\sigma = 1$) or *fluid* ($\sigma = 0$).
- The *shearability* s: This property can be *shearable* ($s = 1$) or *non-shearable* ($s = 0$). As shown below, a non-shearable AA can be either rigid or fluid within a rigid domain of the protein. Non-shearable domains tend to move as a rigid body (*i.e.*, via translation or rotation), whereas shearable regions are easy to deform.

Only 3 of the 4 possible combinations are allowed :

1. Non-shearable and solid AA (yellow): ($\sigma = 1; s = 0$).
2. Non-shearable and fluid AA (red): ($\sigma = 0; s = 0$).
3. Shearable and fluid AA (blue): ($\sigma = 0; s = 1$).
4. Shearable solid is forbidden.

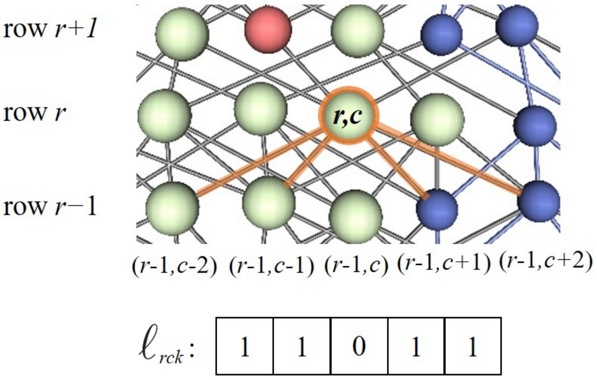

FIG. S-1. Illustration of the percolation rules for shearability and fluid/solid states. Note that site (r, c) was turned solid because it is attached to 2 solid sites below it. Also note that the red site above it is fluid, because it is attached to less than 2 solid sites below it. But it is not shearable because it does not connect to a shearable site below it. On the other hand, the top right site is shearable and fluid, since it is attached to only one solid site (namely (r, c)) and no others on the invisible part of the structure (as seen by its blue connections), and it is also connected to the blue site at $(r, c + 2)$.

Given a fixed sequence, and an *input* state in the bottom row of the cylinder, $\{\sigma_{1,c}, s_{1,c}\}$ the state of the cylinder is completely determined as follows: The three states percolate through the network, from row r to row $r + 1$ (see Fig. S-1). This propagation is directed by the presence of bonds, with a maximum of 5 bonds ending in each AA (of rows $r = 2$ to h; the state of the first row is given as input). These bonds can be *present*(=1) or *absent*(=0). according to the codon ℓ_{rck}, $k = -2, \ldots, 2$ when they point to the AA with coordinate (r, c) coming from the AA $(r-1, c+k)$.

In a first sweep through the rows, we deal with the *rigidity* property σ. In row $r = 1$ each of the w AAs is in a rigidity state rigid ($\sigma = 1$) or fluid ($\sigma = 0$). In all other rows, $r = 2$ to h, the 5 bonds determine the value of the rigidity of (r, c) through a majority rule:

$$\sigma_{r,c} = \theta\left(\sum_{k=-2}^{2} \ell_{rck}\sigma_{r-1,c+k} - \sigma_0\right), \tag{1}$$

where θ is the step function ($\theta(x \geq 0) = 1$, $\theta(x < 0) = 0$)). The parameter $\sigma_0 = 2$ is the minimum number of rigid AAs from the $r - 1$ row that are required to rigidly support AA: In 2D each AA has two coordinates which are constrained if it is connected to two or more static AAs. In this way, the rigidity property of being pinned in place propagates through the lattice, as a function of the initial row and the choice of the bonds which are present as encoded in the gene.

We next address the *shearability* property. It is determined by the rigidity of AAs as follows: We assume that all fluid AAs in row $r = 1$ are also shearable (blue: ($\sigma = 0; s = 1$)). A fluid node (r, c) in row r will become shearable exactly if at least one of its neighbors $(r-1, c)$ or $(r-1, c\pm1)$ is shearable:

$$s_{r,c} = (1 - \sigma_{r,c}) \cdot \theta\left(\sum_{k=-1}^{1} s_{r-1,c+k} - s_0\right), \tag{2}$$

where $s_0 = 1$. The first term on the lhs ensures that a solid AA can never become shearable. This completes the definition of the map from the sequence to the state.

We now define a *target*. It is a chain of w values, fluid and shearable ($\sigma = 0; s = 1$) or solid ($\sigma = 1; s = 0$), in the top row, which the protein should yield as an *output*: $\{\sigma_c^*, s_c^*\}_{c=1,\ldots,w}$. Given (i) a gene sequence, which determines the connectivity ℓ_{rck} and (ii) the *input* state, $\{\sigma_{1,c}, s_{1,c}\}_{c=1,\ldots,w}$, the algorithm described above uniquely defines the output state in the top row, $\{\sigma_{h,c}, s_{h,c}\}_{c=1,\ldots,w}$. At each step of evolution, the output state is compared to the fixed, given target, by measuring the Hamming distance, the number of positions where the output differs from the target:

$$F = \sum_{c=1}^{w} [1 - (|s_{h,c} - s_c^*| - 1) \cdot (|\sigma_{h,c} - \sigma_c^*| - 1)]. \tag{3}$$

In the biological convention $-F$ is the *fitness* that should increase towards a maximum value of $-F = 0$, when the input-output problem is *solved*.

Solutions are found by *mutations*. At each iteration, a randomly drawn digit in the gene is flipped, that is the values of 0 and 1 are exchanged. This corresponds to erasing or creating a randomly chosen link of a randomly chosen AA. After each flip, a sweep is performed, and the new output at the top row is again compared to the target. A mutation is kept *only if the Hamming distance is not increased as compared to the value before the mutation* (equivalently the fitness is not allowed to decrease). This procedure is repeated until a solution $(F = 0)$ is found. This will happen with probability 1, perhaps after very many flips, if the problem has a solution at all.

2. SIMULATIONS

All simulations are done on the $30 \times 18 = 540$ playground, as described above. We have done simulations for many variants of the model, and many targets, but we present only two specific problems, for which the most extensive study was done: In the first, the fluid regions of the input and the target are opposite and of length 6 at the bottom and length 5 at the top. In the second run, top and bottom are the same, but the top is shifted sideways by 5 units. We will call these two examples *straight* and *tilted*, denoted as S and T.

For each of these, we study 200 independent *branches*, starting from a random sequence with about 90% of the bonds present at the start. Given any fixed sequence, we sweep according to the rules of Eq.(1)-(2) through the net, and measure the Hamming distance F (Eq.(3)) between the last row and the desired target. When this Hamming distance is 0, we consider the problem as solved. If not, we flip randomly a bond (exchanging 0 with 1) and recalculate the Hamming distance. We view this flip as a *mutation* of the sequence, equivalent to mutating one nucleic base in a gene. If the Hamming distance decreases or remains unchanged, we keep the flip, otherwise we backtrack and flip another randomly chosen bond. This is repeated until a solution is found [1].

Once a solution is found, we destroy it by further mutations and then look for a new solution, as before, starting from the destroyed state. This we call a *generation*. For each of the 200 branches, we followed 5000 generations, leading to a total of 10^6 solutions. The time to recover from a destroyed state is about 1500 flips per error in that state, which is similar to time it takes to find a solution starting from a random gene. A destruction takes around 11.2 mutations on average.

We also did another 10^6 simulations starting each time from another random configuration. The statistics in both cases are very similar, but the destruction-reconstruction simulations obviously show some correlations between a generation and the next. This effect disappears after about 4 generations.

3. SHEAR

Consider now either of the two examples, straight or tilted (S and T). A solution of such an example is given by a set of bonds, and this set of bonds defines a graph on the $N_{AA} = h \cdot w = 540$ AAs. This graph is embedded in 2D where $\vec{x}_{r,c}$ are the positions of the AAs, which are connected by straight bonds. To discuss the shear modes of such a network we consider the elastic tensor, which is the tensor product of the network Laplacian with the 2 by 2 tensor of directional derivatives, as defined *e.g.*, in Chung and Sternberg [2, pp. 618–619].

To be more specific, we describe what this means component-wise. The playground $\Omega \subset \mathbf{Z}^2$ has size h in the z-direction and size w in the x direction, with periodic boundary condition in the x direction. All bonds go from some (r,c) to $(r+1,c)$, $(r+1,c\pm 1)$, $(r+1,c\pm 2)$, again with periodic boundary conditions in the c-direction. Each such bond defines a direction vector (d_z,d_x) in $\mathbf{R}^2$ which we normalize to $d_x^2+d_z^2=1$. Note that this vector depends on both the origin and the target of the bond.

If we imagine harmonic springs between the nodes connected by bonds (all with the same spring constant), then we can define the (symmetric) tensor matrix of deformation energies in the x and y direction by

$$A'_{km} = M(k,m) \text{ , with } k,m \in \Omega \text{ ,}$$

and where each element of A'_{km} is—when k and m are connected by a bond—the 2 by 2 matrix (indexed by $i,j \in \{1,2\}$)

$$M(k,m) = (d_x(k,m), d_z(k,m))^{\mathrm{T}} \otimes (d_x(k,m), d_z(k,m)) = \begin{pmatrix} d_x^2 & d_x d_z \\ d_x d_z & d_z^2 \end{pmatrix} \text{ .}$$

If k and m are not connected, then $M(k,m)$ is the 0 matrix. The elements of $M(k,m)$ are denoted $M(k,m)_{ij}$.

Finally we complete the $2N \times 2N$ matrix A' to a 'Laplacian' A by adding diagonal elements to it, so that the row (and column) sums are 0. In components, this means that we require, for each $k \in \Omega$ and each $i,j \in \{1,2\}$, the sums

$$\sum_{\ell} (A_{km})_{ij}$$

to vanish. Other properties of A are described in [2].

Since we take periodic boundary conditions in the x direction, there will always be a (simple)

0 eigenvalue of A in this direction. Other 0 eigenvalues correspond to translation in the z direction or rotation in the $x-z$ plane. Another type of (double) 0 eigenvalues are associated with any patch of nodes which is totally disconnected from the rest of the lattice. Since the density ϱ of bonds is about $1/2$ and otherwise quite random, and there are twice 5 bonds at each interior node we expect (assuming random distribution of bonds) there to be about $N \cdot 2^{-10} \sim 0.001N$ isolated nodes, *i.e.*, isolated singletons, and even fewer patches of greater size.

Further zero modes come from nodes which can oscillate sideways without first order effects. This will happen if a node is only connected by one bond. Since $\varrho \sim 1/2$, the probability of finding such a node is about

$$N\frac{\binom{10}{1}}{2^{10}} \sim 0.01N \ .$$

Thus, we show in Figure 5A (in the main text) the eigenfunctions only for the first eigenvalues after the trivial ones. Due to the tensorial nature of the problem, the eigenvectors have two components, which we show as shear-flow.

4. DIMENSION

Dimension of a space measures the number of directions in which one can move from a point. In the case of our model, since from any sequence in sequence space one can move along $N_S = 2550$ axes by flipping just one bit, we see that the sequence space has dimension 2550, and the number of different elements in this space is a hypercube with $2^{2550} \sim 10^{768}$ elements.

The set of solutions which we find, has however much smaller dimension, as we show in Fig. 2A for the straight (S) example and in Fig. S-2 for the tilted (T) one. In the case of experimental data, as ours, the dimension is most conveniently determined by the box-counting (Grassberger-Procaccia [3]) algorithm. This is obtained by just counting the number $N(\varrho)$ of pairs at distances $\leq \varrho$, and then finding the slope in a log-log plot. This is indicated by the black lines in Figure 2A and Fig. S-2 we see that, clearly, the dimension in the space of configurations is about 8-9, while, in the space of sequences, the dimension is basically ‘infinite’, namely just limited by the maximal slope one can obtain [4].

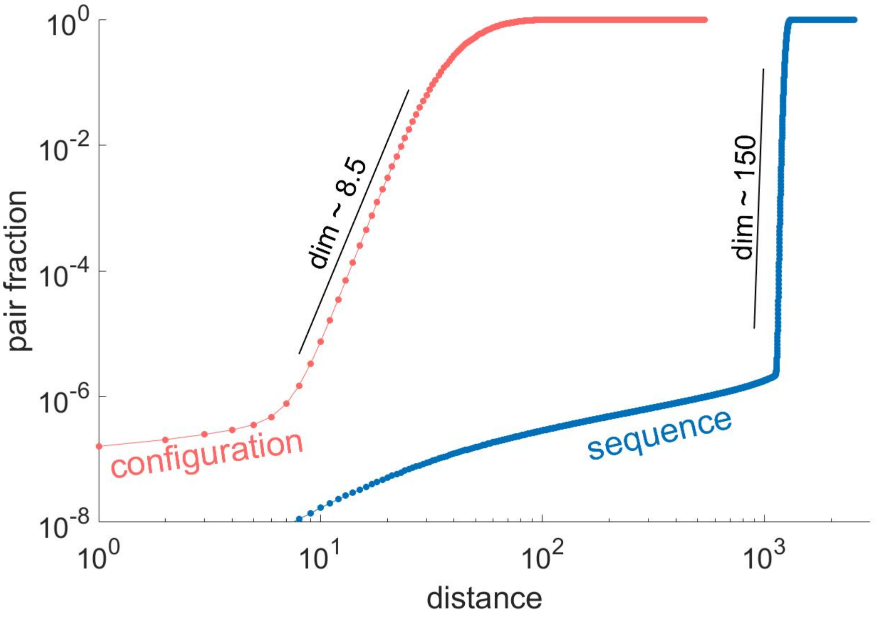

FIG. S-2. The dimensions measured for the tilted case (T).

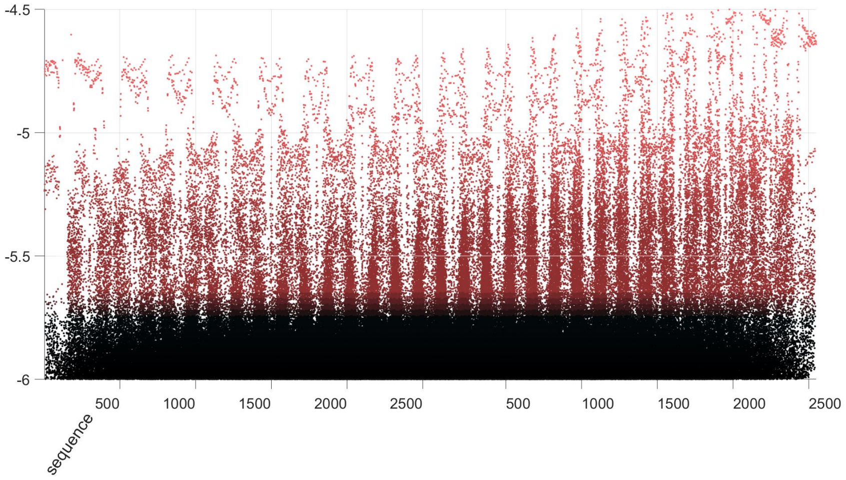

FIG. S-3. A projection of Fig. 2C perpendicular to the $j = j'$ axis.

5. CORRELATION MATRIX

In Fig. 2C, we study the correlations among the 10^6 solutions in sequence space. Given the matrix W_{ij}, of all sequences, with $i = 1, \ldots, N = 10^6$, $j = 1, \ldots, 2550$ (of binary digits), we compute the means $\langle W_{\cdot j} \rangle = \sum_{i=1}^{N} W_{ij}/N$ and the standard deviations $\mathrm{std}_j = \left(\sum_i |W_{ij} - \langle W_{\cdot j} \rangle|^2\right)^{1/2}$.

Then, in the usual way, we form $M_{ij} = W_{ij} - \langle W_{\cdot j} \rangle$ and

$$C_{j,j'} = \frac{(M^* M)_{j,j'}}{\mathrm{std}_j \mathrm{std}_{j'}} \ .$$

Figure 2C then shows $\log(|C_{j,j'}|)$, with the autocorrelation C_{jj} omitted.

Note that both, the means and the variances depend very weakly on j. Fig. 2C reveals and reinforces several observations also made in other calculations of this paper. First, looking onto the axis $j = j'$ in the figure one sees a periodicity of the patterns corresponding to the 17 gaps between the 18 rows of the configuration space. This reflects the necessity to maintain a *connected* liquid channel. Also, as seen in Fig. 2C as well as in Fig. S-3, the correlations grow somewhat towards the ends, especially toward the upper ($j = 2550$) end. This is because of the mechanical constraint which forces the channel to become more precise towards the ends, in analogy with Fig. 4B.

The periodic patterns all over the square reflect not only the natural periodicity of 150 ($= 5 \cdot w$) elements in the sequence, but also show that the boundaries of the channel form a special shell (with *two* peaks per row).

6. SPECTRUM

We compute spectra for both the sequences and the configurations, for the 10^6 solutions. Let us detail this for the case of sequences: We have 10^6 binary vectors with $N_S = 2550$ components each, and we want to know the 'typical' spectrum of such vectors. This is conveniently found with the Singular Value Decomposition (SVD), in which one forms a matrix M of size $m \times n = 10^6 \times 2550$. This matrix can be written as $U \cdot D \cdot V^*$, where U is $m \times m$, V is $n \times n$ and D is an $m \times n$ matrix which is diagonal in the sense that only the elements D_{ii} with $i = 1, \ldots, n$ are nonzero. (We assume here that we are in the case $m > n$.) The D_{ii} are in general > 0 and in this case the singular value decomposition is unique. We call the set of the D_{ii} the spectrum of the sequences, and the vectors in V the eigenvectors of the SVD. It is the first few of those which are shown in Fig. 3B.

Mutatis mutandis, we perform the same SVD for the case of the configurations, using the s-values (that is, of the shearability) of vectors of the configurations. (This is reasonable, because, in general, there are very few non-shearable and fluid AAs.)

Apart from the numerical findings, which are shown in Figure 3 for the straight (S) example and in Fig. S-4 for the tilted (T) one, some comments are in order.

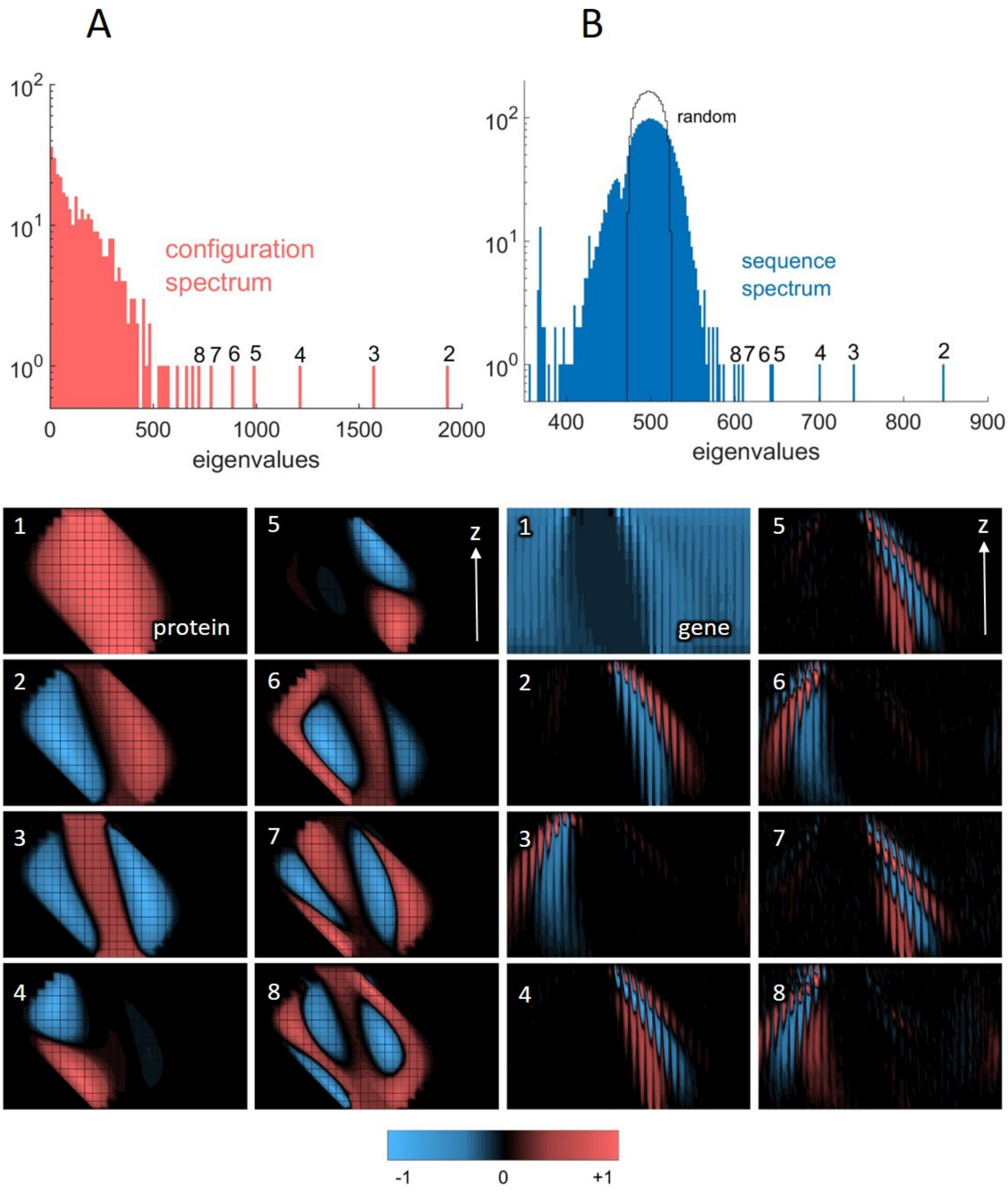

FIG. S-4. The spectra and eigenfunctions for the tilted example (T). (A) The configurations spectrum and eigenfunctions. (B) The sequence spectrum and eigenfunctions.

Configuration space (The eight figures on the bottom left): The first mode is proportional to the average configuration. The next modes reflect the basic deviations of the solution around this average. For example, the second modes is left-to-right shift, the third mode is expansion-contraction etc. Since, the shearable/non-shearable interface can move at most one AA sideways between consecutive rows, the modes are constrained to diamond-shaped areas in the center of the protein. This is the joint effect of the ‘influence zones’ of the input and output rows.

Sequence space (The eight figures on the bottom right): The first eigenvector is the average bond occupancy in the 10^6 solutions. The higher eigenvalues reflect the structure in the many-body correlations among the bonds. The typical pattern is that of 'diffraction' or 'oscillations' around the fluid channel. This pattern mirrors the biophysical constraint of constructing a rigid shell around the shearable region. Higher modes exhibit more stripes, until they become noisy, after about the tenth eigenvalue. The bond-spectrum, top right in Fig. S-4 has some outliers, which correspond to the localized modes shown in the 8 panels below. Apart from that, the majority of the eigenvalues seem to obey the Marčenko-Pastur formula, see [5]. If the matrix is $m \times n$, $m > n$, then the support of the spectrum is $\frac{1}{2}(\sqrt{m} \pm \sqrt{n})$. In our case, since we have a $10^6 \times 2550$ matrix, one expects (if they were really random) to find the spectrum at $\frac{1}{2}(\sqrt{10^6} \pm \sqrt{2550})$, which is close to the experiment, and confirms that most of the bonds are just randomly present or absent. We attribute the slight enlargement of the spectrum to memory effects between generation in the same branch. This corresponds to the well-known phylogenetic correlations among descendants in the same tree.

7. SURVIVAL UNDER MUTATIONS

Here, we ask how robust the solutions are as further mutations take place. First, we determine how many mutations lead to a destruction of the solution. The statistics of this is shown in Fig. 4 of the main text. We note that about 10% of all solutions are destroyed by just one mutation, while there is an exponential decay of survival of m mutations. This signals that the mutations act independently.

One can also ask *where* the critical mutations take place. This is illustrated in Fig. 5B, and was discussed in the main text. We have also studied the places where exactly *two* mutations will kill a solution (and none of the 2 is a single site'killer') and in these cases, one finds that the two mutations are generally close to each other, acting on the same site. Again, the channel is less vulnerable to mutations but now the mutations are evenly distributed over the rest of the network.

8. EXPANSION OF THE PROTEIN UNIVERSE

Let us explain in further detail how Figure 2B was obtained. Here, we test our model against the ideas of [6]. Our results will give some insight about the nature of the graph of solutions.

First, we describe the question as it is found in [6]. Take any two solutions and consider their gene sequences s_1 and s_2. They will have a Hamming distance $d(s_1, s_2)$, which we normalize by dividing by 2550 (the number of elements in $s_i, i = 1, 2$), which we call the protein universe diameter. The question is how much the solution following one generation after s_2 differs from s_1. If we call that solution s_3, then the observed quantity is defined as follows: Let $w_i = 1$ if $s_{1,i} = 1$ and -1 if $s_{1,i} = 0$, for $i = 1, \ldots 2550$. Then for each i let $x_i = w_i \cdot (s_{3,i} - s_{2,i})$. Note that $x_i > 0$ if the change between $s_{3,i}$ and $s_{3,i}$ is *towards* s_1 and < 0 if it is *away from* s_1. Finally, $N_{\text{away}} = \sum_{i:w_i<0} 1$ and $N_{\text{towards}} = \sum_{i:w_i>0} 1$, and we plot in Fig. S-5 $N_{\text{towards}}/N_{\text{away}}$ as a function of D.

In Fig. S-5 we show the results for data set S, (the set T is shown in Fig. 2B). The black curve is nothing but $D/(1 - D)$, where D is the normalized Hamming distance, *i.e.*, the proportion of sites which are different between s_1 and s_2. The fit to this curve tells us an important aspect about the set of possible solutions. Note that the set of all possible s forms a hypercube of dimension 2550 with 2^{2550} corners. The set of solutions is a very small subset of this hypercube, where all corners which are not solutions have been taken away, including the bonds leading to these corners. This leads to a very complicated sub-graph of the hypercube. While we do not have a good mathematical description of how it looks, the good fit shows that the comparisons between s_1, s_2, and s_3 are *as if one performed a random walk on the full cube*. (Note that such a result must be intimately connected to the high dimension of the problem, since for low dimensional hypercubes it does not hold.) Almost all solutions are at the edge of the universe, where the typical Hamming distances among the sequences are close to the typical distance between random sequences,

9. FLEXIBILITY OF SOLUTIONS: THERMAL STABILITY

The histogram of the density of links for the 10^6 solutions is shown in Fig. S-6. These distributions are obtained for simulations in which links are flipped randomly in a symmetric fashion. One can easily push these densities somewhat up or down, by favoring/restricting the flips of links towards 1. However, much more extreme solutions can be found by deterministic procedures which turn as many links to 1 resp. 0. In these cases, we have obtained densities of as high as 0.96 and as low as 0.14, that is, 2452/2550 links, resp. 372/2550 links. Two such extreme cases are illustrated in Fig. 5C.

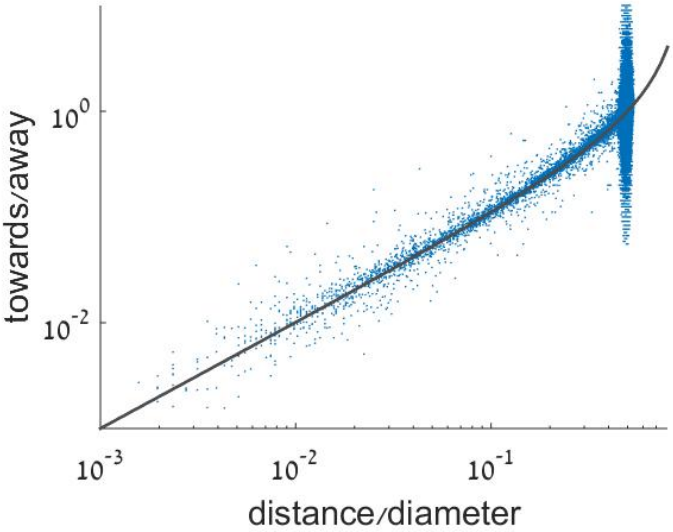

FIG. S-5. A sampling of the 10^6 solutions and the relation of changing a gene toward/away from an original one which is at a Hamming distance d. The black curve is a parameter independent fit by the function $D/(1 - D)$ with $D = d/d_{\max}$.

[1] This is really a Metropolis algorithm [7] at zero temperature.

[2] F. R. K. Chung and S. Sternberg, Journal of Graph Theory **16**, 605 (1992).

[3] P. Grassberger and I. Procaccia, Phys. Rev. Lett. **50**, 346 (1983).

[4] I. Procaccia, Nature **333**, 498 (1988).

[5] V. A. Marčenko and L. A. Pastur, Teor. Funkciĭ Funkcional. Anal. i Priložen. Vyp. **4**, 122 (1967).

[6] I. S. Povolotskaya and F. A. Kondrashov, Nature **465**, 922 (2010).

[7] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller, and E. Teller, Journal of Chemical Physics **21**, 1087 (1953).

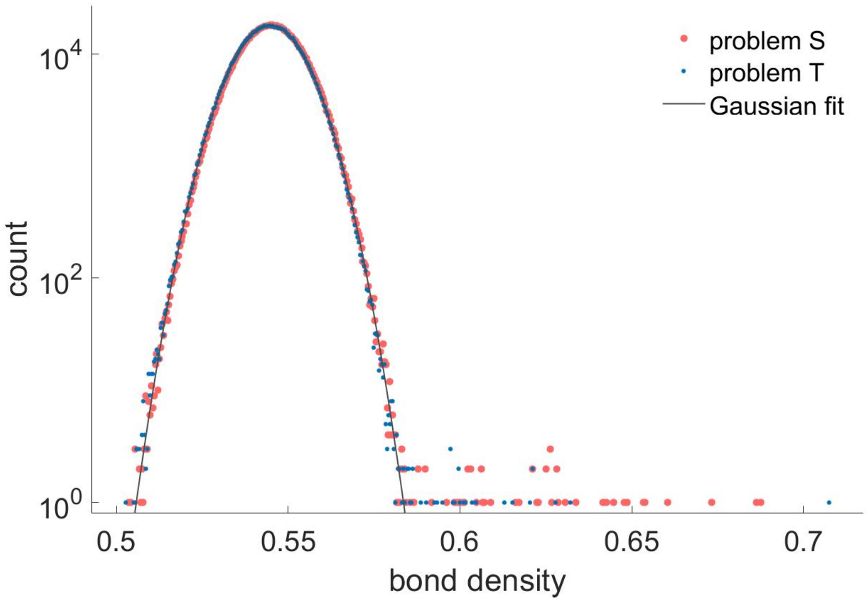

FIG. S-6. The distributions of the bond densities for the 10^6 solutions. Note that these densities are just like random Gaussian variables, except for the outliers.