PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Revealing evolutionary constraints on proteins through sequence analysis

Shou-Wen Wang[1,2,3☯¤], **Anne-Florence Bitbol**[4☯*], **Ned S. Wingreen**[3,5*]

1 Department of Engineering Physics, Tsinghua University, Beijing, China, **2** Beijing Computational Science Research Center, Beijing, China, **3** Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, New Jersey, United States of America, **4** Sorbonne Université, CNRS, Laboratoire Jean Perrin (UMR 8237), F-75005 Paris, France, **5** Department of Molecular Biology, Princeton University, Princeton, New Jersey, United States of America

☯ These authors contributed equally to this work.
¤ Current address: Department of Systems Biology, Harvard Medical School, Boston, Massachusetts, United States of America
* anne-florence.bitbol@sorbonne-universite.fr (AFB); wingreen@princeton.edu (NSW)

OPEN ACCESS

Citation: Wang S-W, Bitbol A-F, Wingreen NS (2019) Revealing evolutionary constraints on proteins through sequence analysis. PLoS Comput Biol 15(4): e1007010. https://doi.org/10.1371/journal.pcbi.1007010

Editor: Faruck Morcos, University of Texas at Dallas, UNITED STATES

Received: January 5, 2019

Accepted: April 6, 2019

Published: April 24, 2019

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

Funding: S.-W. W. and N. S. W. acknowledge the Center for the Physics of Biological Function under NSF Grant PHY-1734030. A.-F. B. and N. S. W. acknowledge the Aspen Center for Physics, which is supported by NSF Grant PHY-1607611. S.-W. W. was supported by the NSFC under Grants No. U1430237 and 11635002. S.-W. W. also acknowledges Tsinghua University for supporting a half-year visit in Princeton University. N. S. W. was

Abstract

Statistical analysis of alignments of large numbers of protein sequences has revealed "sectors" of collectively coevolving amino acids in several protein families. Here, we show that selection acting on any functional property of a protein, represented by an additive trait, can give rise to such a sector. As an illustration of a selected trait, we consider the elastic energy of an important conformational change within an elastic network model, and we show that selection acting on this energy leads to correlations among residues. For this concrete example and more generally, we demonstrate that the main signature of functional sectors lies in the small-eigenvalue modes of the covariance matrix of the selected sequences. However, secondary signatures of these functional sectors also exist in the extensively-studied large-eigenvalue modes. Our simple, general model leads us to propose a principled method to identify functional sectors, along with the magnitudes of mutational effects, from sequence data. We further demonstrate the robustness of these functional sectors to various forms of selection, and the robustness of our approach to the identification of multiple selected traits.

Author summary

Proteins play crucial parts in all cellular processes, and their functions are encoded in their amino-acid sequences. Recently, statistical analyses of protein sequence alignments have demonstrated the existence of "sectors" of collectively correlated amino acids. What is the origin of these sectors? Here, we propose a simple underlying origin of protein sectors: they can arise from selection acting on any collective protein property. We find that the main signature of these functional sectors lies in the low-eigenvalue modes of the covariance matrix of the selected sequences. A better understanding of protein sectors will make it possible to discern collective protein properties directly from sequences, as well as to design new functional sequences, with far-reaching applications in synthetic biology.

supported by NSF Grant MCB-1344191 and by NIH Grant R01 GM082938. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Introduction

Proteins play crucial roles in all cellular processes, acting as enzymes, motors, receptors, regulators, and more. The function of a protein is encoded in its amino-acid sequence. In evolution, random mutations affect the sequence, while natural selection acts at the level of function, however our ability to predict a protein's function directly from its sequence has been very limited. Recently, the explosion of available sequences has inspired new data-driven approaches to uncover the principles of protein operation. At the root of these new approaches is the observation that amino-acid residues which possess related functional roles often evolve in a correlated way. In particular, analyses of large alignments of protein sequences have identified "sectors" of collectively correlated amino acids [1–6], which has enabled successful design of new functional sequences [3]. Sectors are spatially contiguous in the protein structure, and in the case of multiple sectors, each one may be associated with a distinct role [4, 7]. What is the origin of these sectors, and can we identify them from sequence data in a principled way?

To address these questions, we developed a general physical model that naturally gives rise to sectors. Specifically, motivated by the observation that many protein properties reflect additive contributions from individual amino acids [8–10], we consider any additive trait subject to natural selection. As a concrete example, we study a simple elastic-network model that quantifies the energetic cost of protein deformations [11], which we show to be an additive trait. We then demonstrate that selection acting on any such additive trait automatically yields collective correlation modes in sequence data. We show that the main signature of the selection process lies in the small-eigenvalue modes of the covariance matrix of the selected sequences, but we find that some signatures also exist in the widely-studied large-eigenvalue modes. Finally, we demonstrate a principled method to identify sectors and to quantify mutational effects from sequence data alone.

Model and methods

Selection on an additive trait

We focus on selection on an additive scalar trait

$$T(\vec{\alpha}) = \sum_{l=1}^{L} \Delta_l(\alpha_l)\,, \tag{1}$$

where $\vec{\alpha} = (\alpha_1, \ldots, \alpha_L)$ is the amino-acid sequence considered, L is its length, and $\Delta_l(\alpha_l)$ is the mutational effect on the trait T of a mutation to amino acid α_l at site l. Mutational effects can be measured with respect to a reference sequence $\vec{\alpha}^0$, satisfying $\Delta_l(\alpha_l^0) = 0$ for all l.

Eq 1 is very general as it amounts to saying that, to lowest order, mutations have an additive effect on the trait T, which can be any relevant physical property of the protein, say its binding affinity, catalytic activity, or thermal stability [12]. System-specific details are encoded by the single-site mutational effects $\Delta_l(\alpha_l)$, which can be measured experimentally. The assumption of additivity is experimentally validated in many cases. For instance, protein thermal stability, measured through folding free energy, is approximately additive [8, 13]. Importantly, we allow selection to act on a phenotype that is a nonlinear function of T. Permitting a phenotypic nonlinearity on top of our additive trait model is motivated by the fact that actual phenotype data from recent high-throughput mutagenesis experiments were accurately modeled via a nonlinear mapping of an underlying additive trait [10].

Protein sectors are usually defined operationally as collective modes of correlations in amino-acid sequences. However, the general sequence-function relation in Eq 1 suggests an

operational definition of a *functional* protein sector, namely as the set of sites with dominant mutational effects on a trait under selection. Selection can take multiple forms. To be concrete, we first consider a simple model of selection, assuming a favored value $T*$ of the trait T, and using a Gaussian selection window. We subsequently show that the conclusions obtained within this simple model are robust to different forms of selection. Our Gaussian selection model amounts to selecting sequences according to the following Boltzmann distribution:

$$P(\vec{\alpha}) = \frac{\exp(w(\vec{\alpha}))}{\sum_{\vec{\alpha}}\exp(w(\vec{\alpha}))}, \tag{2}$$

where the fitness $w(\vec{\alpha})$ of a sequence is given by

$$w(\vec{\alpha}) = -\frac{\kappa}{2}(T(\vec{\alpha}) - T^*)^2 = -\frac{\kappa}{2}\left(\sum_{l=1}^{L}\Delta_l(\alpha_l) - T^*\right)^2. \tag{3}$$

The selection strength κ sets the width of the selection window.

Such selection for intermediate values of a trait can be realistic, e.g. for protein stability [8]. However, the form of selection can vary, for example selection can be for a nonlinear transform of a trait to be above a certain threshold [10], and several relevant selection variants are investigated below. Crucially, while the trait is additive (Eq 1), the fact that fitness (Eq 3) and selection (Eq 2) are nonlinear functions of the trait leads to coupling between mutations. This phenomenon is known as global [10, 14] or nonspecific [9] epistasis, and its relevance has been shown in evolution experiments [14], over and above contributions from specific epistasis [9, 15]. The focus of this paper is on global epistasis, and we do not include specific epistasis. Studying the interplay of these two types of epistasis will be an interesting future direction.

A toy model yielding a concrete example of an additive trait

Elastic-network model. To illustrate how additive traits naturally arise, we consider the elastic energy associated with a functionally important protein deformation. We explicitly derive the additivity of this trait in the regime of small deformations and weak mutational effects. This concrete example is relevant since functional deformation modes are under selection in proteins [16–18], and dynamical domains possess a signature in sequence data [19]. Moreover, elastic-network models have elucidated a variety of protein properties [11, 20–22], including the emergence of allostery [23–29]. Thus motivated, we begin by building an elastic-network model [11, 20] for a well-studied PDZ protein domain (Fig 1(a) and 1(b)) [30, 31] and computing the relationship between its "sequence" and the energetic cost of a functionally-relevant conformational change.

To build the elastic-network model of the PDZ domain, we replace each of the $L = 76$ amino-acid residues by its corresponding alpha carbon Cα and beta carbon Cβ, as shown in Fig 1(b). Every pair of carbons within a cutoff distance d_c is then connected with a harmonic spring [11]. Following a previous analysis of the same PDZ domain [20], we set $d_c = 7.5$ Å and assign spring constants as follows: a) 2 for Cα-Cα pairs if adjacent along the backbone, 1 otherwise; b) 1 for Cα-Cβ pairs; c) 0.5 for Cβ-Cβ pairs.

Within our elastic model, the energetic cost of a small deformation from the equilibrium structure is

$$E = \frac{1}{2}\sum_{i,j}\left(\boldsymbol{r}_i - \boldsymbol{r}_i^0\right)M_{ij}\left(\boldsymbol{r}_j - \boldsymbol{r}_j^0\right) = \frac{1}{2}\,\delta\boldsymbol{r}^T M\delta\boldsymbol{r}, \tag{4}$$

where $\boldsymbol{r}_i$ is the position of the ith carbon atom, $\boldsymbol{r}_i^0$ is its equilibrium position, and the Hessian

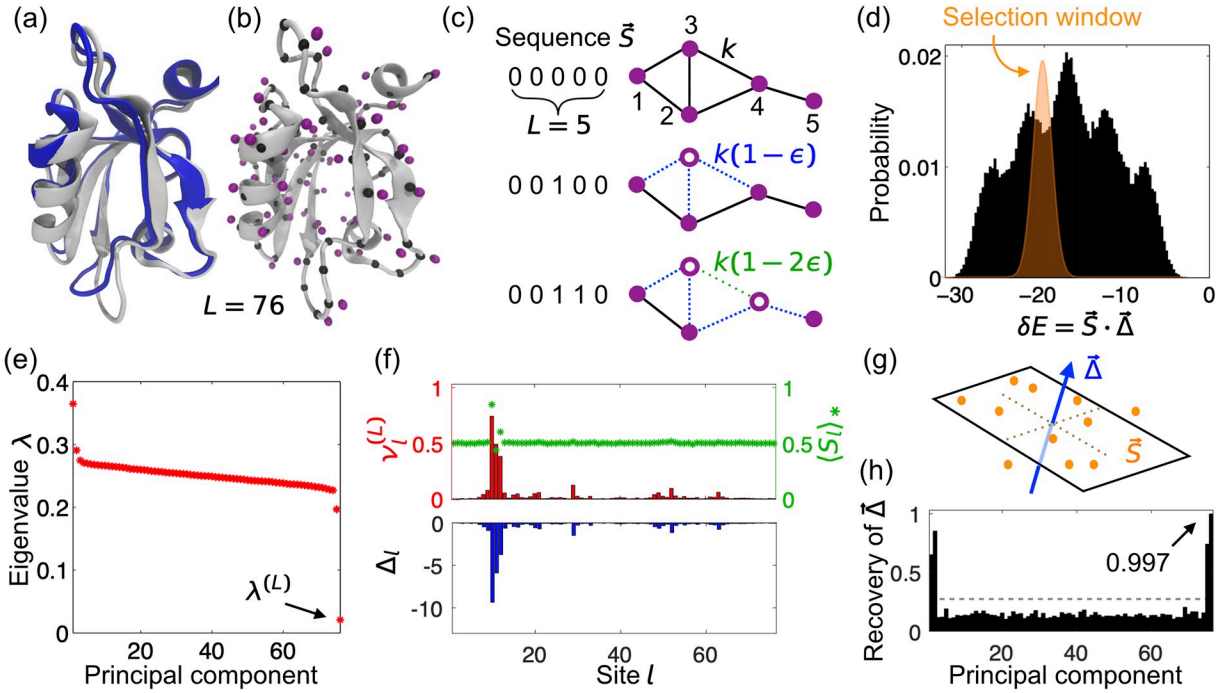

Fig 1. Selection applied to an elastic protein model leads to a statistical signature among sequences. (a) Cartoon representation of the third PDZ domain of the rat postsynaptic density protein 95 from the RSCB PDB [32] (gray: ligand free, 1BFE; blue: ligand bound, 1BE9 (ligand not shown)). (b) Elastic network model for 1BFE, where each amino-acid residue is represented by its alpha carbon (Cα, black node) and beta carbon (Cβ, purple node). Nearby nodes interact through a harmonic spring [20] (S1 Appendix). (c) Relation between protein sequence $\vec{S}$ and elastic network: 0 denotes the reference state, while 1 denotes a mutated residue, which weakens interactions of the corresponding Cβ with all its neighbors by ϵ. (d) Histogram of the energy δE required to deform the domain from its ligand-free to its ligand-bound conformation, for randomly sampled sequences where 0 and 1 are equally likely at each site. Sequences are selectively weighted using a narrow Gaussian window (orange) around $\delta E*$. (e) Eigenvalues of the covariance matrix C for the selectively weighted protein sequences. (f) Upper panel: last principal component $v_l^{(L)}$ of C (red) and average mutant fraction $\langle S_l \rangle_*$ (green) at site l after selection; lower panel: effect Δ_l of a single mutation at site l on δE. (g) Schematic representation of the selected ensemble in sequence space, where each dot is a highly-weighted sequence; thus dots are restricted to a narrow region around a plane perpendicular to $\vec{\Delta}$. (h) Recovery of $\vec{\Delta}$ for all principal components $\vec{v}^{(j)}$, with maximum Recovery = 1 (Eq 8). Gray dashed line: random expectation of Recovery (S1 Appendix).

https://doi.org/10.1371/journal.pcbi.1007010.g001

matrix M contains the second derivatives of the elastic energy with respect to atomic coordinates. Here, we take $\delta \boldsymbol{r}$ to be the conformational change from a ligand-free state (1BFE) to a ligand-bound state (1BE9) of the same PDZ domain (Fig 1(a)). This conformational change is central to PDZ function, so its energetic cost has presumably been under selection during evolution. Any other coherent conformational change would also be suitable for our analysis. Note that our aim is not to analyze conformational changes in all their richness, but to provide a minimal concrete example of a relevant additive trait, and to analyze the impact of selection acting on this trait on the associated family of sequences.

To mimic simply the effect of mutation and selection within our toy model, we introduce "mutations" of residues that weaken the spring constants involving their beta carbons by a small fraction ϵ. In practice, we take $\epsilon = 0.2$. We represent mutations using a sequence $\vec{S}$ with $S_l \in \{0, 1\}$, where l is the residue index: $S_l = 0$ denotes the reference state, while $S_l = 1$ implies a mutation (Fig 1(c)). The sequence $\vec{S}$ and the spring network fully determine the Hessian matrix M, and thus the energy cost E of a conformational change (Eq 4). Note that here $\vec{S}$ is a binary sequence, which represents a simplification compared to real protein sequences $\vec{\alpha}$ where each site can feature 21 states (20 amino acids, plus the alignment gap). We start with the binary model for simplicity, and we then extend our results to a more realistic 21-state

model. Note that binary representations of actual protein sequences, with a consensus residue state and a "mutant" state, have proved useful in sector analysis [4], although more recent approaches for diverse protein families have employed full 21-state models [7]. Binary representations are also appropriate to analyze sets of sufficiently close sequences, notably HIV proteins, allowing identification of their sectors [5] and predictions of their fitness landscapes [33].

Deformation energy as an additive trait. Focusing on mutations that weakly perturb the elastic properties of a protein, we perform first-order perturbation analysis: $M = M^{(0)} + \epsilon M^{(1)} + o(\epsilon)$. Using Eq 4 yields $E = E^{(0)} + \epsilon E^{(1)} + o(\epsilon)$, with $E^{(1)} = \boldsymbol{\delta r}^T M^{(1)} \boldsymbol{\delta r}/2$. Both $M^{(1)}$ and $E^{(1)}$ can be expressed as sums of contributions from individual mutations. We define Δ_l as the first-order energy cost $\epsilon E^{(1)}$ of a single mutation at site l of the sequence. To leading order, the effect of mutations on the energy cost of a deformation reads

$$\delta E = E - E^{(0)} = \sum_{l=1}^{L} S_l \Delta_l. \tag{5}$$

This equation corresponds to the binary-sequence case of the general additive trait defined in Eq 1. Hence, the deformation energy in our toy model of a protein as a sequence-dependent elastic network constitutes a practical example of an additive trait.

Within our functional definition, a protein sector is the set of sites with dominant mutational effects on the trait under selection. The vector $\vec{\Delta}$ of mutational effects for our elastic-network model of the PDZ domain is shown in Fig 1(f). The magnitudes of mutational effects are strongly heterogeneous (Fig. 1 in S1 Appendix). Here, the amino acids with largest effects, which constitute the sector, correspond to those that move most upon ligand binding. (Note that the ligand-binding deformation of PDZ is well-described by one low-frequency normal mode of the elastic network [20]: hence, our sector significantly overlaps with the sites that are most involved in this mode).

How is such a functionally-defined sector reflected in the statistical properties of the sequences that survive evolution? To answer this question, we next analyze sequences obtained by selecting on the trait δE. While for concreteness, we use the mutational effects obtained from our elastic model, the analysis is general and applies to any additive trait. Indeed, we later present some examples using synthetically-generated random mutational effect vectors, both binary and more realistic 21-state ones (see below and S1 Appendix).

Results

Signature of selection in sequences

For our elastic model of the PDZ domain, the distribution of the additive trait δE for random sequences is shown in Fig 1(d). We use the selection process introduced in Eqs 2 and 3 to limit sequences to a narrower distribution of δEs, corresponding, e.g., to a preferred ligand-binding affinity. The fitness of a binary sequence $\vec{S}$, a particular case of Eq 3, reads:

$$w(\vec{S}) = -\frac{\kappa}{2}\left(\sum_{l=1}^{L} \Delta_l S_l - \delta E^*\right)^2. \tag{6}$$

Here, the selection strength κ sets the width of the selection window, and $\delta E*$ is its center. For all selections, we take $\kappa = 10/(\sum_l \Delta_l^2)$, so that the width of the selection window scales with that of the unselected distribution. We have confirmed that our conclusions are robust to varying selection strength, provided $\kappa \sum_l \Delta_l^2 \gg 1$ (see Fig. 3 in S1 Appendix).

Although mutations have additive effects on the trait δE, the nonlinearities involved in fitness and selection give rise to correlations among sites. For instance, if $\delta E* = 0$ and if $\Delta_l < 0$ for all l, as in Fig 1, a mutation at a site with large $|\Delta_l|$ will decrease the likelihood of additional mutations at all other sites with large $|\Delta_l|$.

Previous approaches to identifying sectors from real protein sequences have relied on modified forms of Principal Component Analysis (PCA). So we begin by asking: can PCA identify sectors in our physical model? PCA corresponds to diagonalizing the covariance matrix C of sequences: it identifies the principal components (eigenvectors) $\vec{v}^{(j)}$ associated with progressively smaller variances (eigenvalues) $\lambda^{(j)}$. We introduce $\langle\cdot\rangle_*$ to denote ensemble averages over the selectively weighted sequences, reserving $\langle\cdot\rangle$ for averages over the unselected ensemble. The mutant fraction at site l in the selected ensemble is $\langle S_l\rangle_* = \sum_{\vec{S}} S_l P(\vec{S})$, and the covariance matrix C reads

$$C_{ll'} = \Big\langle (S_l - \langle S_l\rangle_*)\cdot(S_{l'} - \langle S_{l'}\rangle_*)\Big\rangle_* . \tag{7}$$

To test the ability of PCA to identify a functional sector, we employed the selection window shown in orange in Fig 1(d). The resulting eigenvalues are shown in Fig 1(e). One sees outliers. In particular, why is the last eigenvalue so low? Due to the narrow selection window, according to Eq 6 the highly-weighted sequences satisfy $\sum_l S_l\Delta_l = \vec{S}\cdot\vec{\Delta} \approx \delta E^*$. This means that in the L-dimensional sequence space, the data points for the highly-weighted sequences lie in a narrow region around a plane perpendicular to $\vec{\Delta}$ (Fig 1(g)). Hence, the data has exceptionally small variance in this direction, leading to a particularly small eigenvalue of C. Moreover, the corresponding last principal component $\vec{v}^{(L)}$ points in the direction with the smallest variance and is consequently parallel to $\vec{\Delta}$ (Fig 1(f)). Formally, in Eq 6, $\vec{\Delta}$ appears in a quadratic coupling term where it plays the part of a repulsive pattern in a generalized Hopfield model [34, 35]: alone, such a term would penalize sequences aligned with $\vec{\Delta}$. But here, $\vec{\Delta}$ also appears in a term linear in $\vec{S}$ and as a result Eq 6 penalizes sequences that fail to have the selected projection onto $\vec{\Delta}$.

In this example, the last principal component accurately recovers the functional sector corresponding to the largest elements of the mutational-effect vector $\vec{\Delta}$. More generally, to quantify the recovery of $\vec{\Delta}$ by a given vector $\vec{v}$, we introduce

$$\text{Recovery} = \frac{\sum_l |v_l\Delta_l|}{\sqrt{\sum_l v_l^2}\sqrt{\sum_l \Delta_l^2}}, \tag{8}$$

which is nonnegative, has a random expectation of $(\sqrt{2/\pi L})\sum_l|\Delta_l|/\sqrt{\sum_l\Delta_l^2}$ for $L \gg 1$ (S1 Appendix), and saturates at 1 (including the case of parallel vectors). For our test case, Fig 1(h) shows Recovery for all principal components. The last one features the highest Recovery, almost 1, confirming that it carries substantial information about $\vec{\Delta}$. The second-to-last principal component and the first two also provide a value of Recovery substantially above random expectation. Outlier eigenvalues arise from the sector, and accordingly, we find that the number of modes with high Recovery often corresponds to the number of sites with strong mutational effects. A more formal analysis of this effect will be an interesting topic for further study.

In our model, $\vec{\Delta}$ is fundamentally a direction of *small variance*. So why do the first principal components also carry information about $\vec{\Delta}$? Qualitatively, when variance is decreased in one direction due to a repulsive pattern $\vec{\Delta}$, variance tends to increase in orthogonal directions involving the same sites. To illustrate this effect, let $L = 3$ and $\vec{\Delta} = (-1, 1, 0)$, and consider the

sequences $\vec{S}$ satisfying $\vec{\Delta} \cdot \vec{S} = 0$ (namely (0, 0, 0); (1, 1, 0); (0, 0, 1); (1, 1, 1)). The last principal component is $\vec{\Delta}$, with zero variance, and the first principal component is (1, 1, 0): Recovery is 1 for both of them. This selection conserves the trace of the covariance matrix (i.e. the total variance), so that decreasing the variance along $\vec{\Delta} = (-1, 1, 0)$ necessarily increases it along (1, 1, 0). This simple example provides an intuitive understanding of why the large-eigenvalue modes of the covariance matrix also carry information about $\vec{\Delta}$.

It is worth remarking that Eq 6 is a particular case of a general fitness function with one- and two-body terms (known as fields and couplings in Ising or Potts models in physics). Here, the values of these one- and two-body terms are constrained by their expressions in terms of $\vec{\Delta}$. In practice, several traits might be selected simultaneously (see below), yielding more independent terms among the fields and couplings. More generally, such one- and two-body descriptions have been very successfully employed via Direct Coupling Analysis (DCA) to identify strongly coupled residues that are in contact within a folded protein [36–38], to investigate folding [39], and to predict fitness [33, 40–45] and conformational changes [46, 47], as well as protein-protein interactions [48, 49]. A complete model of protein covariation in nature should necessarily incorporate both the collective modes described here and the strongly coupled residue pairs which are the focus of DCA.

ICOD method

An important concern is whether the last principal component is robust to small and/or noisy datasets. Indeed, other directions of small variance can appear in the data. As a second example, we applied a different selection window, centered in the tail of the distribution of δEs from our elastic model of the PDZ domain (Fig 2(a), inset). This biased selection generates strong conservation, $\langle S_l \rangle_* \approx 1$, for some sites with significant mutational effects. Extreme conservation at one site now dictates the last principal component, and disrupts PCA-based recovery of $\vec{\Delta}$ (Fig 2(a) and 2(b)).

To overcome this difficulty, we developed a more robust approach that relies on inverting the covariance matrix. Previously, the inverse covariance matrix was successfully employed in Direct Coupling Analysis (DCA) to identify strongly coupled residues that are in contact within a folded protein [36–38]. The fitness in our model (Eq 6) involves one and two-body interaction terms, and constitutes a particular case of the DCA Hamiltonian (S1 Appendix). A small-coupling approximation [37, 38, 50, 51] (S1 Appendix) gives

$$C^{-1}_{ll'} \approx (1 - \delta_{ll'})\, \kappa \Delta_l \Delta_{l'} + \delta_{ll'} \left(\frac{1}{P_l} + \frac{1}{1 - P_l} \right), \tag{9}$$

where P_l denotes the probability that site l is mutated. Since we are interested in extracting $\vec{\Delta}$, we can simply set to zero the diagonal elements of C^{-1}, which are dominated by conservation effects, to obtain a new matrix

$$\tilde{C}^{-1}_{ll'} \approx (1 - \delta_{ll'}) \kappa \Delta_l \Delta_{l'}. \tag{10}$$

The first eigenvector of $\tilde{C}^{-1}$ (associated with its largest eigenvalue) should accurately report $\vec{\Delta}$ since, except for its zero diagonal, $\tilde{C}^{-1}$ is proportional to the outer product $\vec{\Delta} \otimes \vec{\Delta}$. We call this approach the *Inverse Covariance Off-Diagonal* (ICOD) method. As shown in Fig 2(d) and 2(e), ICOD overcomes the difficulty experienced by PCA for biased selection, while performing equally well as PCA for unbiased selection (Fig. 2 in S1 Appendix). Removing the diagonal elements of C^{-1} before diagonalizing is crucial: otherwise, the first eigenvector of C^{-1} is the same

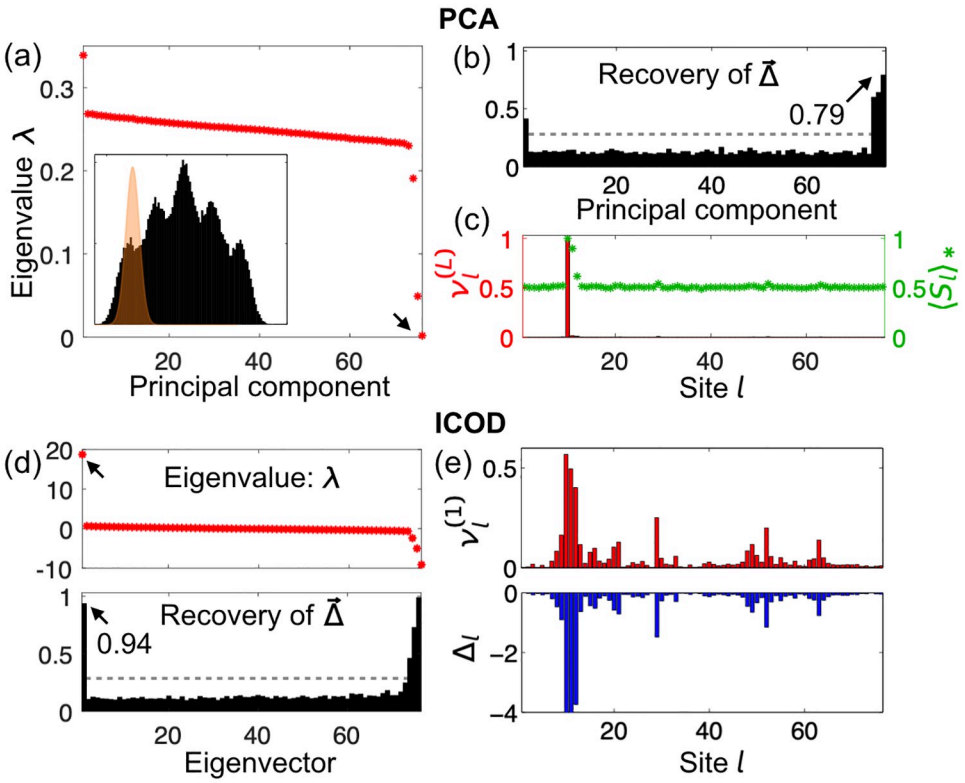

Fig 2. Recovery of mutational-effect vector $\vec{\Delta}$ from sequence analysis in the case of strongly biased selection. (a-c) Principal Component Analysis (PCA) performs poorly due to strong conservation at some sites of large mutational effect. (a) Eigenvalues of covariance matrix obtained for strongly biased selection around $\delta E^*_{\text{biased}}$ (inset, orange window) for same model proteins as in Fig 1. (b) Recovery of $\vec{\Delta}$ for all principal components. (c) Last principal component $v_l^{(L)}$ (red) and average mutant fraction $\langle S_{il}\rangle_*$ (green) at site l. (d-e) The ICOD method performs robustly. (d) Eigenvalues of $\tilde{C}^{-1}_{ll'}$ (Eq 10) (upper) and Recovery of $\vec{\Delta}$ for all eigenvectors (lower). (e) Leading eigenvector $v_l^{(1)}$ (upper) and mutational effect Δ_l at site l (lower, same as in Fig 1(f)). Gray dashed lines in (b, d): random expectation of Recovery (S1 Appendix).

https://doi.org/10.1371/journal.pcbi.1007010.g002

as the last eigenvector of C and suffers from the same shortcomings for strong conservation. Here too, eigenvectors associated to both small and large eigenvalues contain information about $\vec{\Delta}$ (Fig 2(b) and 2(d)).

Selection on multiple traits

An important challenge in sector analysis is distinguishing multiple, independently evolving sectors [4, 7, 52]. We can readily generalize our fitness function (Eqs 3 and 6) to allow for selection on multiple additive traits:

$$w(\vec{S}) = -\sum_{i=1}^{N} \frac{\kappa_i}{2} \left(\sum_{l=1}^{L} \Delta_{i,l} S_l - T_i^* \right)^2, \tag{11}$$

where N is the number of distinct additive traits $T_i(\vec{S}) = \sum_l \Delta_{i,l} S_l$ under selection, $\vec{\Delta}_i$ is the vector of mutational effects on trait T_i, κ_i is the strength of selection on this trait, and T_i^* is the associated selection bias. For example, $\vec{\Delta}_1$ might measure how mutations change a protein's binding affinity, while $\vec{\Delta}_2$ might be related to its thermal stability, etc.

In Fig. 5 in S1 Appendix, we consider selection on two distinct additive traits, using synthetically-generated random mutational-effect vectors $\vec{\Delta}_1$ and $\vec{\Delta}_2$ (S1 Appendix). Note that these mutational effects are thus unrelated to our toy model of protein elastic deformations: as stated above, our approach holds for any additive trait under selection. ICOD then yields *two* large outlier eigenvalues of the modified inverse covariance matrix $\tilde{C}^{-1}$. The associated eigenvectors accurately recover both $\vec{\Delta}_1$ and $\vec{\Delta}_2$, after a final step of Independent Component Analysis (ICA) [7, 53, 54] that successfully disentangles the contributions coming from the two constraints (see S1 Appendix).

Performance in sector recovery

We further tested the performance of ICOD by systematically varying the selection bias, both for our toy model of PDZ elastic deformations and for more general synthetically-generated random mutational-effect vectors (Fig 3). ICOD achieves high Recovery of these various mutational-effect vectors for both single and double selection over a broad range of selection biases $T*$, albeit performance falls off in the limit of extreme bias.

How does ICOD compare with other approaches to identifying sectors? We compared the performance of ICOD with Statistical Coupling Analysis (SCA), the original PCA-based method [4, 7]. In SCA, the covariance matrix C is reweighted by a site-specific conservation factor ϕ_l, the absolute value is taken, $\tilde{C}^{(\mathrm{SCA})}_{ll'} = |\phi_l C_{ll'} \phi_{l'}|$, and sectors are identified from the leading eigenvectors of $\tilde{C}^{(\mathrm{SCA})}$. We therefore tested the ability of the first eigenvector of $\tilde{C}^{(\mathrm{SCA})}$ to recover $\vec{\Delta}$ for a single selection. We found that the square root of the elements of the first SCA eigenvector can provide high Recovery of $\vec{\Delta}$ (Fig 3, and Figs. 13, 14 in S1 Appendix). However, the performance of SCA relies on conservation through ϕ_l, and it has been shown that residue conservation actually dominates sector identification by SCA in certain proteins [52]. Consequently, for unbiased selection, SCA breaks down (Fig 3(a), dashed curves) and cannot identify sector sites (Fig. 17 in S1 Appendix). ICOD does not suffer from such

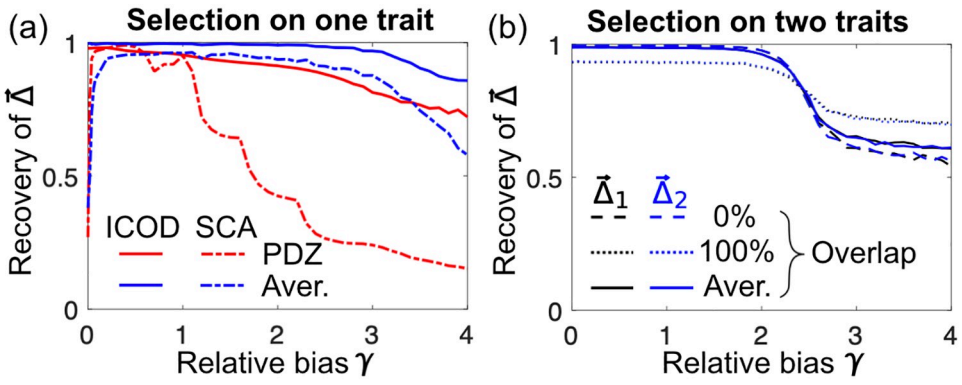

Fig 3. Average recovery of mutational-effect vectors $\vec{\Delta}$ as a function of relative selection bias $\gamma \equiv (T^* - \langle T \rangle)/\sqrt{\langle (T - \langle T \rangle)^2 \rangle}$ on the selected additive trait T. (a) Selection on a single trait. Different $\vec{\Delta}$s are used to generate sequence ensembles: the elastic-network $\vec{\Delta}$ from Fig 1 (red); synthetic $\vec{\Delta}$s (S1 Appendix) with number of sites of large mutational effect (sector sites) ranging from 1 to 100, for sequences of length $L = 100$ (blue). Recovery is shown for ICOD (solid curves) and for SCA [4, 7] (dashed curves). (b) Selection on two distinct traits. Different pairs of synthetic $\vec{\Delta}$s (S1 Appendix) are used to generate sequence ensembles (with $L = 100$): "0%" indicates two non-overlapping sectors, each with 20 sites; "100%" indicates two fully overlapping sectors, each with 100 sites; "Aver." indicates average Recovery over 100 cases of double selection, where the single-sector size increases from 1 to 100, and the overlap correspondingly increases from 0 to 100. ICA was applied to improve Recovery (S1 Appendix).

https://doi.org/10.1371/journal.pcbi.1007010.g003

shortcomings, and performs well over a large range of selection biases. Note that in SCA, only the top eigenvectors of $\tilde{C}^{(\mathrm{SCA})}$ convey information about sectors (Figs. 13, 15 in S1 Appendix).

We also compared ICOD with another PCA-based approach [34], which employs an inference method specific to the generalized Hopfield model, and should thus be well adapted to identifying sectors within our physical model (Eq 6). Overall, this specialized approach performs similarly to ICOD, being slightly better for very localized sectors, but less robust than ICOD for strong selective biases and small datasets (S1 Appendix). Exactly as for PCA and ICOD, within this method, the top Recovery is obtained for the bottom eigenvector of the (modified) covariance matrix, consistent with $\vec{\Delta}$ in our model being a repulsive pattern [34], but large Recoveries are also obtained for the top eigenvectors (Fig. 18 in S1 Appendix).

Robustness to different forms of selection

To assess the robustness of functional sectors to selections different from the simple Gaussian selection window of Eqs 2 and 3, we selected sequences with an additive trait T above a threshold T_t, and varied this threshold. For instance, a fluorescent protein might be selected to be fluorescent enough, which could be modeled by requiring that (a nonlinear transform of) an additive trait be sufficiently large [10]. As shown in Fig 4, the corresponding sectors are identified by ICOD as well as those resulting from our initial Gaussian selection window.

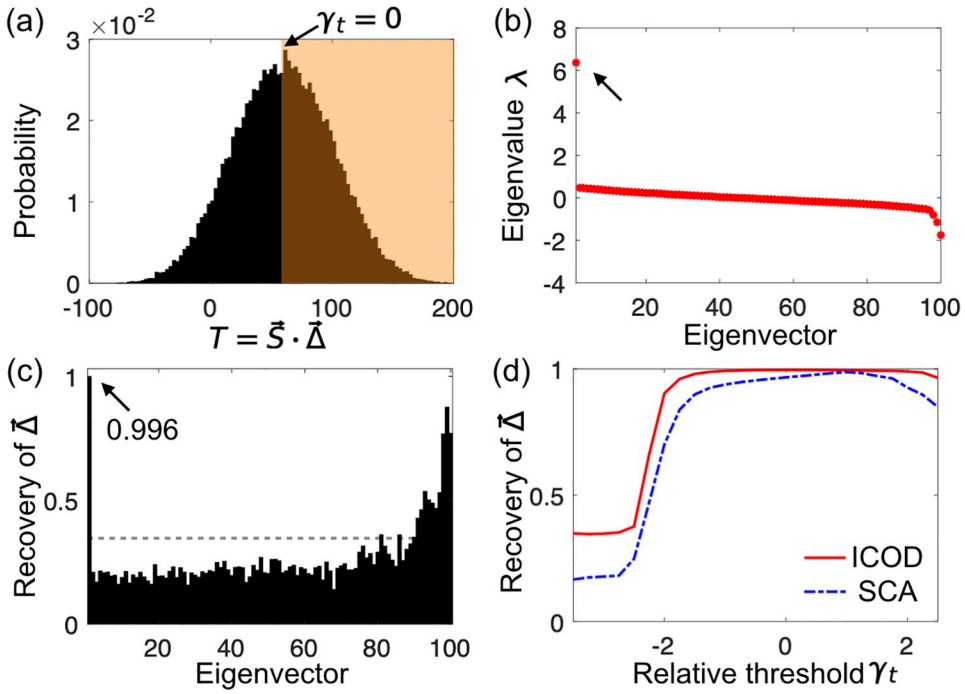

Fig 4. Identification of sectors that result from threshold-based selection. (a) Histogram of the additive trait $T(\vec{S}) = \vec{S} \cdot \vec{\Delta}$ for randomly sampled sequences where 0 and 1 are equally likely at each site. Sequence length is $L = 100$, mutational effects are synthetically generated with 20 sector sites (see S1 Appendix). Sequences are selected if they have a trait value $T(\vec{S}) > T_t$ (orange shaded region). Selection is shown for $T_t = \langle T \rangle$, or equivalently $\gamma_t = 0$, in terms of the relative threshold $\gamma_t \equiv (T_t - \langle T \rangle)/\sqrt{\langle (T - \langle T \rangle)^2 \rangle}$. (b) Eigenvalues of the ICOD-modified inverse covariance matrix $\tilde{C}^{-1}$ (Eq 10) of the selected sequences for $\gamma_t = 0$. (c) Recovery of $\vec{\Delta}$ for all eigenvectors of $\tilde{C}^{-1}$ for $\gamma_t = 0$. Gray dashed line: random expectation of Recovery. (d) Recovery of $\vec{\Delta}$ for ICOD and for SCA as functions of the relative selection threshold γ_t. The data in (d) is averaged over 100 realizations of $\vec{\Delta}$.

https://doi.org/10.1371/journal.pcbi.1007010.g004

In Fig 4(d), we show the performance of both ICOD and SCA at recovering sectors arising from selection with a threshold. Consistent with previous results (see Fig 3), we find that ICOD is more robust than SCA to extreme selections. We also successfully applied ICOD to other forms of selection: Fig. 8 in S1 Appendix shows the case of a quartic fitness function replacing the initial quadratic one (Eq 3) in the Boltzmann distribution (Eq 2) and Fig. 9 in S1 Appendix shows the case of a rectangular selection window (S1 Appendix). These results demonstrate the robustness of functional sectors, and of ICOD, to different plausible forms of selection.

Extension to 21-state sequences and to natural sequences

So far, we have considered binary sequences, with only one type of mutation with respect to the reference state. In the S1 Appendix, we demonstrate that our formalism, including the ICOD method, extends to mutations among q different states. The case $q = 21$, which includes the 20 different amino-acid types plus the alignment gap is the relevant one for real proteins. The single-site mutational effects Δ_l are then replaced by state-specific mutational effects $\Delta_l(\alpha_l)$ with $\alpha_l \in \{1, \ldots, 21\}$ (see Eq 1). Fig. 10 in S1 Appendix shows that the generalized version of ICOD performs very well on synthetic data generated for the case $q = 21$. We further demonstrate that sector identification is robust to gauge changes (reference changes) and to the use of pseudocounts (S1 Appendix).

While the main purpose of this article is to propose an operational definition of functional protein sectors and to understand how they can arise, an interesting next question will be to investigate what ICOD can teach us about real data. As a first step in this direction, we applied ICOD to a multiple sequence alignment of PDZ domains. In this analysis, we employed a complete description with $q = 21$, but we compressed the ICOD-modified inverse matrix using the Frobenius norm to focus on overall (and not residue-specific) mutational effects (see S1 Appendix for details). As shown in Fig 5(a) and 5(b), both ICOD and SCA identify one strong outlying large eigenvalue, thus confirming that PDZ has only one sector [6]. Recall that due to the inversion step, the largest eigenvalue in ICOD is related to the mode with smallest variance, whose importance was demonstrated above. Furthermore, as seen in Fig 5(c) and 5(d), both methods correctly predict the majority of residues found experimentally to have important mutational effects on ligand binding to the PDZ domain shown in Fig 1(a) [6]. For instance, over the 20 top sites identified by ICOD (resp. SCA), we find that 85% (resp. 75%) of them are also among the 20 experimentally most important sites. Note that for SCA, we recover the result from Ref. [6]. The performance of ICOD is robust to varying the cutoff for removal of sites with a large proportion of gaps (see Fig. 21 in S1 Appendix), but notably less robust than SCA to pseudocount variation (see Fig. 22 in S1 Appendix).

Importantly, both ICOD and SCA perform much better than random expectation, which is 29%. Hence, both of these methods can be useful to identify functionally important sites. The slightly greater robustness of SCA to pseudocounts on this particular dataset (see Fig. 22 in S1 Appendix) might come from the fact that many of the experimentally-identified functionally important sites in the PDZ domain are strongly conserved [52], which makes the conservation reweighting step in SCA advantageous. Since residue conservation alone is able to predict most of the experimentally important PDZ sites [52], we also compared conservation to SCA and ICOD: ranking sites by conservation (employing the conservation score of Ref. [7], see S1 Appendix) indeed identifies 70% of the top 20 experimentally-determined sites with important mutational effects. Interestingly, ICOD scores are slightly more strongly correlated with conservation than SCA scores are correlated with conservation (see Fig. 23 in S1 Appendix), despite the fact that conservation is explicitly used in SCA and not in ICOD.

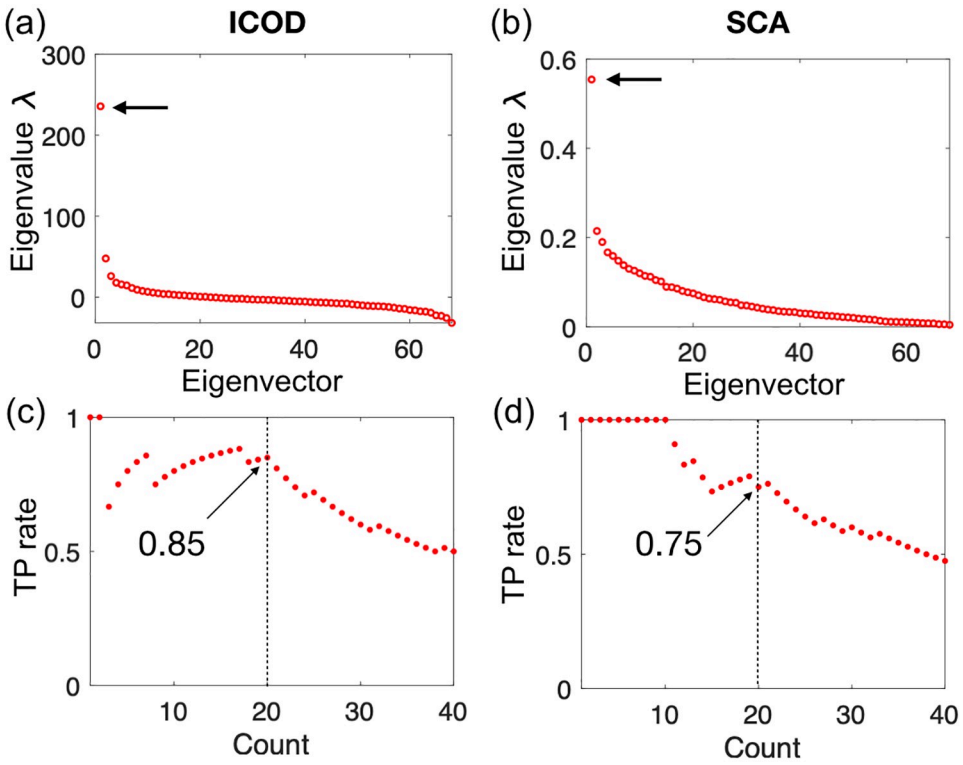

Fig 5. Performance of ICOD and SCA at predicting the 20 sites with largest experimentally-determined mutational effects in a PDZ domain. (a) Eigenvalues of the compressed ICOD-modified inverse covariance matrix $\tilde{C}^{-1}$ (S1 Appendix). (b) Eigenvalues of the SCA matrix. (c) True Positive (TP) rates obtained by taking the first eigenvector $\vec{v}^{(1)}$ from the compressed ICOD-modified inverse covariance matrix, generating a ranked list of sites of descending magnitudes of the components $||v_i^{(1)}||$ of this eigenvector at each site l (S1 Appendix), and computing the fraction of the top sites in this predicted ordering that are also among the 20 experimentally most important sites [6]. Results are shown versus the number of top predicted sites ("count"). (d) TP rates from SCA, computed as in panel (c). In panels (c) and (d), the TP rate values obtained for the top 20 predicted sites are indicated by arrows. In all panels, a pseudocount ratio $\Lambda = 0.02$ was used, and sites with more than 15% gap state were discarded (see S1 Appendix for details).

https://doi.org/10.1371/journal.pcbi.1007010.g005

Overall, this preliminary application to real data highlights the ability of ICOD to identify functionally related amino acids in a principled way that only relies on covariance. We emphasize that the main goal of this paper is to provide insight into the possible physical origins of sectors, and into the statistical signatures of these physical sectors in sequence data. A more extensive application of ICOD and related methods to real sequence data will be the subject of future work.

Discussion

We have demonstrated how sectors of collectively correlated amino acids can arise from evolutionary constraints on functional properties of proteins. Our model is very general, as it only relies on the functional property under any of various forms of selection being described by an underlying additive trait, which has proven to be valid in many relevant situations [8–10, 13].

We showed that the primary signature of functional selection acting on sequences lies in the small-eigenvalue modes of the covariance matrix. In contrast, sectors are usually identified from the large-eigenvalue modes of the SCA matrix [4, 7]. This is not in contradiction with our results because, as we showed, signatures of our functional sectors are often also found in

large-eigenvalue modes of the covariance matrix. Besides, the construction of the SCA matrix from the covariance matrix involves reweighting by conservation and taking an absolute value or a norm [4, 7], which can substantially modify its eigenvectors, eigenvalues, and their order. Conservation is certainly important in real proteins, especially in the presence of phylogeny; indeed, the SCA matrix, which includes both conservation and covariance, was recently found to capture well experimentally-measured epistasis with respect to the free energy of PDZ ligand binding [55]. However, the fundamental link we propose between functional sectors and small-eigenvalue modes of the covariance matrix is important, since large-eigenvalue modes of the covariance matrix also contain confounding information about subfamily-specific residues [56] and phylogeny [57], and consistently, some sectors identified by SCA have been found to reflect evolutionary history rather than function [4]. Interestingly, the small-eigenvalue modes are also the ones that contain most information about structural contacts in real proteins [35]. Hence, our results help explain previously observed correlations between sectors and contacts, e.g. the fact that contacts are overrepresented within a sector but not across sectors [58].

We introduced a principled method to detect functional sectors from sequence data, based on the primary signature of these sectors in the small-eigenvalue modes of the covariance matrix. We further demonstrated the robustness of our approach to the existence of multiple traits simultaneously under selection, to various forms of selection, and to data-specific questions such as reference choices and pseudocounts.

Importantly, our modeling approach allowed us to focus on functional selection alone, in the absence of historical contingency and of specific structural constraints, thus yielding insights complementary to purely data-driven methods. The collective modes investigated here are just one source of residue-residue correlations. Next, it will be interesting to study the intriguing interplay between functional sectors, phylogeny, and contacts, and to apply our methods to multiple protein families. Our results shed light on an aspect of the protein sequence-function relationship and open new directions in protein sequence analysis, with implications in synthetic biology, building toward function-driven protein design.

Supporting information

S1 Appendix. Methodological details and further results. In S1 Appendix, we present additional details about our model and methods, as well as additional results.
(PDF)

Author Contributions

Conceptualization: Shou-Wen Wang, Anne-Florence Bitbol, Ned S. Wingreen.

Investigation: Shou-Wen Wang, Anne-Florence Bitbol, Ned S. Wingreen.

Writing – original draft: Shou-Wen Wang, Anne-Florence Bitbol, Ned S. Wingreen.

Writing – review & editing: Shou-Wen Wang, Anne-Florence Bitbol, Ned S. Wingreen.

References

1. Lockless SW, Ranganathan R. Evolutionarily conserved pathways of energetic connectivity in protein families. Science. 1999; 286(5438):295–299. https://doi.org/10.1126/science.286.5438.295 PMID: 10514373
2. Süel GM, Lockless SW, Wall MA, Ranganathan R. Evolutionarily conserved networks of residues mediate allosteric communication in proteins. Nat Struct Biol. 2003; 10(1):59–69. https://doi.org/10.1038/nsb881 PMID: 12483203

3. Socolich M, Lockless SW, Russ WP, Lee H, Gardner KH, Ranganathan R. Evolutionary information for specifying a protein fold. Nature. 2005; 437(7058):512. https://doi.org/10.1038/nature03991 PMID: 16177782
4. Halabi N, Rivoire O, Leibler S, Ranganathan R. Protein sectors: evolutionary units of three-dimensional structure. Cell. 2009; 138(4):774–786. https://doi.org/10.1016/j.cell.2009.07.038 PMID: 19703402
5. Dahirel V, Shekhar K, Pereyra F, Miura T, Artyomov M, Talsania S, et al. Coordinate linkage of HIV evolution reveals regions of immunological vulnerability. Proc Natl Acad Sci USA. 2011; 108(28):11530–11535. https://doi.org/10.1073/pnas.1105315108 PMID: 21690407
6. McLaughlin RN Jr, Poelwijk FJ, Raman A, Gosal WS, Ranganathan R. The spatial architecture of protein function and adaptation. Nature. 2012; 491(7422):138. https://doi.org/10.1038/nature11500 PMID: 23041932
7. Rivoire O, Reynolds KA, Ranganathan R. Evolution-Based Functional Decomposition of Proteins. PLoS Comput Biol. 2016; 12(6):e1004817. https://doi.org/10.1371/journal.pcbi.1004817 PMID: 27254668
8. DePristo MA, Weinreich DM, Hartl DL. Missense meanderings in sequence space: a biophysical view of protein evolution. Nat Rev Genet. 2005; 6(9):678–687. https://doi.org/10.1038/nrg1672 PMID: 16074985
9. Starr TN, Thornton JW. Epistasis in protein evolution. Protein Sci. 2016; 25(7):1204–1218. https://doi.org/10.1002/pro.2897 PMID: 26833806
10. Otwinowski J, McCandlish DM, Plotkin JB. Inferring the shape of global epistasis. Proc Natl Acad Sci USA. 2018; 115(32):E7550–E7558. https://doi.org/10.1073/pnas.1804015115 PMID: 30037990
11. Bahar I, Lezon TR, Yang LW, Eyal E. Global dynamics of proteins: bridging between structure and function. Annu Rev Biophys. 2010; 39:23–42. https://doi.org/10.1146/annurev.biophys.093008.131258 PMID: 20192781
12. Cunningham AD, Colavin A, Huang KC, Mochly-Rosen D. Coupling between Protein Stability and Catalytic Activity Determines Pathogenicity of G6PD Variants. Cell Rep. 2017; 18(11):2592–2599. https://doi.org/10.1016/j.celrep.2017.02.048 PMID: 28297664
13. Wylie CS, Shakhnovich EI. A biophysical protein folding model accounts for most mutational fitness effects in viruses. Proc Natl Acad Sci USA. 2011; 108(24):9916–9921. https://doi.org/10.1073/pnas.1017572108 PMID: 21610162
14. Kryazhimskiy S, Rice DP, Jerison ER, Desai MM. Microbial evolution. Global epistasis makes adaptation predictable despite sequence-level stochasticity. Science. 2014; 344(6191):1519–1522 PMID: 24970088
15. Posfai A, Zhou J, Plotkin JB, Kinney JB, McCandlish DM. Selection for Protein Stability Enriches for Epistatic Interactions. Genes (Basel). 2018; 9(9). https://doi.org/10.3390/genes9090423
16. Zheng W, Brooks BR, Thirumalai D. Low-frequency normal modes that describe allosteric transitions in biological nanomachines are robust to sequence variations. Proc Natl Acad Sci USA. 2006; 103(20):7664–7669. https://doi.org/10.1073/pnas.0510426103 PMID: 16682636
17. Lukman S, Grant GH. A network of dynamically conserved residues deciphers the motions of maltose transporter. Proteins. 2009; 76(3):588–597. https://doi.org/10.1002/prot.22372 PMID: 19274733
18. Saldano TE, Monzon AM, Parisi G, Fernandez-Alberti S. Evolutionary Conserved Positions Define Protein Conformational Diversity. PLoS Comput Biol. 2016; 12(3):e1004775. https://doi.org/10.1371/journal.pcbi.1004775 PMID: 27008419
19. Granata D, Ponzoni L, Micheletti C, Carnevale V. Patterns of coevolving amino acids unveil structural and dynamical domains. Proc Natl Acad Sci USA. 2017; 114(50):E10612. https://doi.org/10.1073/pnas.1712021114 PMID: 29183970
20. De Los Rios P, Cecconi F, Pretre A, Dietler G, Michielin O, Piazza F, et al. Functional dynamics of PDZ binding domains: a normal-mode analysis. Biophys J. 2005; 89(1):14–21. https://doi.org/10.1529/biophysj.104.055004 PMID: 15821164
21. Delarue M, Sanejouand YH. Simplified normal mode analysis of conformational transitions in DNA-dependent polymerases: the elastic network model. J Mol Biol. 2002; 320(5):1011–1024. https://doi.org/10.1016/S0022-2836(02)00562-4 PMID: 12126621
22. Zheng W, Doniach S. A comparative study of motor-protein motions by using a simple elastic-network model. Proc Natl Acad Sci USA. 2003; 100(23):13253–13258. https://doi.org/10.1073/pnas.2235686100 PMID: 14585932
23. Yan L, Ravasio R, Brito C, Wyart M. Architecture and coevolution of allosteric materials. Proc Natl Acad Sci USA. 2017; p. 201615536. https://doi.org/10.1073/pnas.1615536114
24. Tlusty T, Libchaber A, Eckmann JP. Physical Model of the Genotype-to-Phenotype Map of Proteins. Phys Rev X. 2017; 7:021037.

25. Flechsig H. Design of Elastic Networks with Evolutionary Optimized Long-Range Communication as Mechanical Models of Allosteric Proteins. Biophys J. 2017; 113(3):558–571. https://doi.org/10.1016/j.bpj.2017.06.043 PMID: 28793211

26. Dutta S, Eckmann JP, Libchaber A, Tlusty T. Green function of correlated genes in a minimal mechanical model of protein evolution. Proc Natl Acad Sci USA. 2018; 115(20):E4559–E4568. https://doi.org/10.1073/pnas.1716215115 PMID: 29712824

27. Rocks JW, Pashine N, Bischofberger I, Goodrich CP, Liu AJ, Nagel SR. Designing allostery-inspired response in mechanical networks. Proc Natl Acad Sci USA. 2017; 114(10):2520–2525. https://doi.org/10.1073/pnas.1612139114 PMID: 28223534

28. Yan L, Ravasio R, Brito C, Wyart M. Principles for Optimal Cooperativity in Allosteric Materials. Biophys J. 2018; 114(12):2787–2798. https://doi.org/10.1016/j.bpj.2018.05.015 PMID: 29925016

29. Bravi B, Ravasio R, Brito C, Wyart M. Direct Coupling Analysis of Epistasis in Allosteric Materials. Preprint arXiv:1811.10480. 2018. https://arxiv.org/abs/1811.10480

30. Doyle DA, Lee A, Lewis J, Kim E, Sheng M, MacKinnon R. Crystal structures of a complexed and peptide-free membrane protein–binding domain: molecular basis of peptide recognition by PDZ. Cell. 1996; 85(7):1067–1076. https://doi.org/10.1016/S0092-8674(00)81307-0 PMID: 8674113

31. Hung AY, Sheng M. PDZ domains: structural modules for protein complex assembly. J Biol Chem. 2002; 277(8):5699–5702. https://doi.org/10.1074/jbc.R100065200 PMID: 11741967

32. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, Weissig H, et al. The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol. 2000; 7 Suppl:957–959. https://doi.org/10.1038/80734 PMID: 11103999

33. Mann JK, Barton JP, Ferguson AL, Omarjee S, Walker BD, Chakraborty A, et al. The fitness landscape of HIV-1 gag: advanced modeling approaches and validation of model predictions by in vitro testing. PLoS Comput Biol. 2014; 10(8):e1003776. https://doi.org/10.1371/journal.pcbi.1003776 PMID: 25102049

34. Cocco S, Monasson R, Sessak V. High-dimensional inference with the generalized Hopfield model: principal component analysis and corrections. Phys Rev E. 2011; 83(5 Pt 1):051123. https://doi.org/10.1103/PhysRevE.83.051123

35. Cocco S, Monasson R, Weigt M. From principal component to direct coupling analysis of coevolution in proteins: Low-eigenvalue modes are needed for structure prediction. PLOS Comput Biol. 2013; 9(8):e1003176. https://doi.org/10.1371/journal.pcbi.1003176 PMID: 23990764

36. Weigt M, White RA, Szurmant H, Hoch JA, Hwa T. Identification of direct residue contacts in protein-protein interaction by message passing. Proc Natl Acad Sci USA. 2009; 106(1):67–72. https://doi.org/10.1073/pnas.0805923106 PMID: 19116270

37. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA. 2011; 108(49):E1293–E1301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

38. Marks DS, Colwell LJ, Sheridan R, Hopf TA, Pagnani A, Zecchina R, et al. Protein 3D structure computed from evolutionary sequence variation. PLoS ONE. 2011; 6(12):e28766. https://doi.org/10.1371/journal.pone.0028766 PMID: 22163331

39. Morcos F, Schafer NP, Cheng RR, Onuchic JN, Wolynes PG. Coevolutionary information, protein folding landscapes, and the thermodynamics of natural selection. Proc Natl Acad Sci USA. 2014; 111(34):12408–12413. https://doi.org/10.1073/pnas.1413575111 PMID: 25114242

40. Dwyer RS, Ricci DP, Colwell LJ, Silhavy TJ, Wingreen NS. Predicting functionally informative mutations in Escherichia coli BamA using evolutionary covariance analysis. Genetics. 2013; 195(2):443–455. https://doi.org/10.1534/genetics.113.155861 PMID: 23934888

41. Cheng RR, Morcos F, Levine H, Onuchic JN. Toward rationally redesigning bacterial two-component signaling systems using coevolutionary information. Proc Natl Acad Sci USA. 2014; 111(5):E563–571. https://doi.org/10.1073/pnas.1323734111 PMID: 24449878

42. Cheng RR, Nordesjo O, Hayes RL, Levine H, Flores SC, Onuchic JN, et al. Connecting the Sequence-Space of Bacterial Signaling Proteins to Phenotypes Using Coevolutionary Landscapes. Mol Biol Evol. 2016; 33(12):3054–3064. https://doi.org/10.1093/molbev/msw188 PMID: 27604223

43. Figliuzzi M, Jacquier H, Schug A, Tenaillon O, Weigt M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. Mol Biol Evol. 2016; 33(1):268–280. https://doi.org/10.1093/molbev/msv211 PMID: 26446903

44. Barton JP, Goonetilleke N, Butler TC, Walker BD, McMichael AJ, Chakraborty AK. Relative rate and location of intra-host HIV evolution to evade cellular immunity are predictable. Nat Commun. 2016; 7:11660. https://doi.org/10.1038/ncomms11660 PMID: 27212475

45. Hopf TA, Ingraham JB, Poelwijk FJ, Scharfe CP, Springer M, Sander C, et al. Mutation effects predicted from sequence co-variation. Nat Biotechnol. 2017; 35(2):128–135. https://doi.org/10.1038/nbt.3769 PMID: 28092658

46. Morcos F, Pagnani A, Lunt B, Bertolino A, Marks DS, Sander C, et al. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. Proc Natl Acad Sci USA. 2011; 108(49):E1293–1301. https://doi.org/10.1073/pnas.1111471108 PMID: 22106262

47. Malinverni D, Marsili S, Barducci A, De Los Rios P. Large-Scale Conformational Transitions and Dimerization Are Encoded in the Amino-Acid Sequences of Hsp70 Chaperones. PLoS Comput Biol. 2015; 11 (6):e1004262. https://doi.org/10.1371/journal.pcbi.1004262 PMID: 26046683

48. Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. Proc Natl Acad Sci USA. 2016; 113(43):12180–12185. https://doi.org/10.1073/pnas.1606762113 PMID: 27663738

49. Gueudre T, Baldassi C, Zamparo M, Weigt M, Pagnani A. Simultaneous identification of specifically interacting paralogs and interprotein contacts by direct coupling analysis. Proc Natl Acad Sci USA. 2016; 113(43):12186–12191. https://doi.org/10.1073/pnas.1607570113 PMID: 27729520

50. Plefka T. Convergence condition of the TAP equation for the infinite-ranged Ising spin glass model. J Phys A: Math Gen. 1982; 15(6):1971–1978. https://doi.org/10.1088/0305-4470/15/6/035

51. Bitbol AF, Dwyer RS, Colwell LJ, Wingreen NS. Inferring interaction partners from protein sequences. Proc Natl Acad Sci USA. 2016; 113(43):12180–12185. https://doi.org/10.1073/pnas.1606762113 PMID: 27663738

52. Teşileanu T, Colwell LJ, Leibler S. Protein sectors: statistical coupling analysis versus conservation. PLoS Comput Biol. 2015; 11(2):e1004091. https://doi.org/10.1371/journal.pcbi.1004091 PMID: 25723535

53. Hyvärinen A, Karhunen J, Oja E. Independent Component Analysis. John Wiley and Sons; 2001.

54. Hansen LK, Larsen J, Kolenda T. Blind Detection of Independent Dynamic Components. In: IEEE International Conference on Acoustics, Speech, and Signal Processing 2001. vol. 5; 2001. p. 3197–3200.

55. Salinas VH, Ranganathan R. Coevolution-based inference of amino acid interactions underlying protein function. Elife. 2018; 7. https://doi.org/10.7554/eLife.34300

56. Casari G, Sander C, Valencia A. A method to predict functional residues in proteins. Nat Struct Biol. 1995; 2(2):171–178. https://doi.org/10.1038/nsb0295-171 PMID: 7749921

57. Qin C, Colwell LJ. Power law tails in phylogenetic systems. Proc Natl Acad Sci USA. 2018; 115(4):690–695. https://doi.org/10.1073/pnas.1711913115 PMID: 29311320

58. Rivoire O. Elements of coevolution in biological sequences. Phys Rev Lett. 2013; 110(17):178102. https://doi.org/10.1103/PhysRevLett.110.178102 PMID: 23679784