Dynamics-Evolution Correspondence in Protein Structures

Tang, Qian-Yuan; Kaneko, Kunihiko

Published in:
Physical Review Letters

DOI:
10.1103/PhysRevLett.127.098103

Publication date:
2021

Document version
Publisher’s PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Tang, Q.-Y., & Kaneko, K. (2021). Dynamics-Evolution Correspondence in Protein Structures. Physical Review Letters, 127, [098103]. https://doi.org/10.1103/PhysRevLett.127.098103
Dynamics-Evolution Correspondence in Protein Structures

Qian-Yuan Tang$^{1,2,*}$ and Kunihiro Kaneko$^{1,†}$

$^1$Center for Complex Systems Biology, Universal Biology Institute, University of Tokyo, Komaba 3-8-1, Meguro-ku, Tokyo 153-8902, Japan
$^2$Lab for Neural Computation and Adaptation, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan

(Received 30 March 2021; accepted 28 July 2021; published 26 August 2021)

The genotype-phenotype mapping of proteins is a fundamental question in structural biology. In this Letter, with the analysis of a large dataset of proteins from hundreds of protein families, we quantitatively demonstrate the correlations between the noise-induced protein dynamics and mutation-induced variations of native structures, indicating the dynamics-evolution correspondence of proteins. Based on the investigations of the linear responses of native proteins, the origin of such a correspondence is elucidated. It is essential that the noise- and mutation-induced deformations of the proteins are restricted on a common low-dimensional subspace, as confirmed from the data. These results suggest an evolutionary mechanism of the proteins gaining both dynamical flexibility and evolutionary structural variability.

DOI: 10.1103/PhysRevLett.127.098103

Phenotypic states of all living systems are shaped through dynamics, which are changed through evolution. The possible connection between the changes by dynamics in a short term and those by genetic evolution in a long term has been a crucial topic in evolutionary and developmental biology [1–3]. Recent developments in computational, systems, and theoretical biophysics have enabled us to investigate the issue quantitatively in terms of phenotypic variations. The state variables are subjected to both external noises and internal genetic changes. The possible relationships between the variations induced by noise and genetic mutations, i.e., correlations between the changes by dynamics and evolution, have been investigated at the molecular, cellular, and organism levels both theoretically [4,5] and experimentally [6–8].

Native structures of globular proteins are also shaped as a result of folding dynamics. Herein, the amino acid sequences are determined by genes. Understanding the genotype-phenotype mapping of proteins is one of the fundamental challenges in structural biology. The conformational dynamics are subjected to thermal noises, which can trigger functional motions [9–13], whereas the mutations in the sequences may lead to deformations in their native structure, resulting in changes in the equilibrium dynamics of the proteins.

Interestingly, the dynamics and the evolution of proteins share many similarities [14,15]. The functional dynamics of proteins were reported to be constrained to low-dimensional motions, as described by the slow dynamical modes with low excitation energies [16–18], whereas other studies have shown that the sequence space [19,20] and the mutation-induced deformations of a protein family are also low-dimensional [21,22]. These suggest possible connections between the dynamics (phenotype) and evolution (genotype) of the proteins [23–28] in manners similar to other biological systems [29–33].

In this Letter, with the database of hundreds of thousands of proteins from hundreds of protein families, we analyze the correlated deformations of proteins subjected to both thermal noises and mutations. The data analysis quantitatively demonstrates the correlations between the noise-induced protein dynamics and mutation-induced variations of native structures, indicating the dynamics-evolution correspondences of proteins. Subsequently, from the study of the linear responses of native proteins, the origin of such correspondence is elucidated, to which dimensional reduction of the noise- or mutation-induced deformations of the proteins is shown to be related.

To characterize the fluctuations of a protein subjected to external noises, we introduce the vector $\vec{r} = [\vec{r}_1, \vec{r}_2, \ldots, \vec{r}_N]^\dagger$ as the variable to describe the phenotypic state of the proteins, that is, the three-dimensional Cartesian coordinates of the $N$ amino acid residues (represented by their $C_{\alpha}$ atoms). To describe the variations in the native structure subjected to internal mutations, we take the native structure $\vec{r}^0 = [\vec{r}^0_1, \vec{r}^0_2, \ldots, \vec{r}^0_N]^\dagger$, which is determined by the gene, as the model parameter. Here, the variable $\vec{r}$ and parameter $\vec{r}^0$ are a pair of conjugated variables. We denote the noise-induced deformation vector as $\Delta\vec{r} = \vec{r} - \vec{r}^0$. The covariance...
We first compare the diagonal elements of $C^{D}$ and $C^{E}$ to unity (see Ref. [37], Sec. 4.1) and obtain the cross-correlation matrices $\hat{C}^{D}$ (dynamics) and $\hat{C}^{E}$ (evolution), respectively. As shown in Fig. 1(f), matrices $\hat{C}^{D}$ and $\hat{C}^{E}$ share similar patterns. These results indicate a correspondence between the dynamics and evolution of protein structures.

To further validate this correspondence, we employ the DALI server [35] and the ProDY package [78,79] to fetch and align the structures [80] from more protein families. Our database contains protein structures from 415 protein families. In each protein family, there are more than 20 structural homologs (see Ref. [37], Sec. 1). For demonstration, we take four protein families with different chain lengths as examples to analyze the correlation lengths in protein dynamics and evolution. Similar to previous studies [81–83], we assume that the entries $\hat{C}_{ij}^{D}$ and $\hat{C}_{ij}^{E}$ are the functions of the mutual distance $r_{ij}$ between the residue $i$ and $j$. Then, we introduce the distance-dependent correlation functions $\phi^{D}(r)$ and $\phi^{E}(r)$ by averaging $\hat{C}_{ij}^{D}$ (or $\hat{C}_{ij}^{E}$) for residue pairs at mutual distance $r_{ij} \approx r$. As shown in Figs. 2(a) and 2(b), the correlation functions $\phi^{D}(r)$ and $\phi^{E}(r)$ behave similarly. We define their first zero points as the correlation lengths $\xi^{D}$ and $\xi^{E}$, respectively. As shown in the insets of Figs. 2(a) and 2(b), $\phi^{D}(r)$ and $\phi^{E}(r)$ can be scaled by their correlation lengths, respectively. As shown in Fig. 2(c), for protein families at different sizes, the correlation length $\xi^{D}$ (dynamics) is always proportional to $\xi^{E}$ (evolution). Such a relationship suggests the correspondence between the long-range correlations in protein dynamics [81,84] and the epistasis (mutations at one site may alter mutations at a distant site) in protein evolution [26–28,85].

Previously, power laws were observed in the vibration spectra of the protein dynamics [86,87]. Herein, we evaluate the eigenvalues ($\sigma_{1}^{E} \geq \sigma_{2}^{E} \geq \cdots \geq \sigma_{N-1}^{E}$) of the evolutionary cross-correlation matrix $\hat{C}^{E}$. As shown in Fig. 2(d), the rank-size distribution of the largest 20 eigenvalues of the four different protein families is plotted at a double logarithmic scale, and the slopes of the fitted curves are close to $-1$, demonstrating a power-law-like behavior. Such a distribution, which is known as Zipf’s law,
is observed for $\sigma_i^D$ and $\phi_i^D$ as $i^*$, where $z = -1$, indicating the sloppiness in parameter selection [88,89].

The aforementioned data analysis has clearly demonstrated the correspondence between dynamics and evolution. However, the origin of this correspondence is still unclear. To elucidate that, let us introduce the potential energy $E(\vec{x}; \vec{\theta})$, which is the function of the gene-related state $\vec{\theta}$ and phenotypic state $\vec{x}$ of the proteins. Here, parameter $\vec{\theta}$ represents a group of genetic control parameters that determine the native structure, and variable $\vec{x}$ consists of a set of generalized coordinates describing the conformational dynamics of a protein [90]. We take structure $\vec{r}^D(\vec{\theta})$ as the native state where $E(\vec{x}; \vec{\theta})$ reaches the minima. With the linear approximations, we neglect the nonlinear interactions related to protein dynamics and evolution. In this way, the energy function $E(\vec{x}; \vec{\theta})$ can be approximated as

$$E(\vec{x}; \vec{\theta}) = \frac{1}{2} [\vec{r}(\vec{x}) - \vec{r}^D(\vec{\theta})]^T \cdot H \cdot [\vec{r}(\vec{x}) - \vec{r}^D(\vec{\theta})],$$

where the Hessian matrix $H$ with entries $H_{ij} = \langle \partial^2 E / \partial r_i \partial r_j \rangle$ serves as the elasticity matrix. In this way, the protein dynamics and evolution are simplified as ENMs. Referring to the properties of Gaussian integrals, we have $C_D \sim H^{-1}$, where $H^{-1}$ denotes the pseudoinverse of $H$ [66,67]. Similarly, matrix $F$ (with entries

$$F_{ij} = \langle \partial^2 E / \partial r_i \partial r_j \rangle$$

is introduced to characterize the energy change by mutation-induced deformations of the native structure, where $C_E \sim F^{-1}$. The matrix $F$ is known as the observed Fisher information matrix in statistical inference [91] and information geometry [92] (see Ref. [37], Sec. 3.2). According to Eq. [66], around the native state, we have $\langle \partial^2 E / \partial r_i \partial r_j \rangle = \langle \partial^2 E / \partial r_i \partial r_j \rangle$, that is, $H = F$. Thus, a direct correspondence between dynamics and evolution around the native structure is shown.

With the correspondence $H = F$, and referring to the reciprocal relations $C_D \sim H^{-1}$ and $C_E \sim F^{-1}$, one would expect the relation $C_D \sim C_E$. In fact, due to the nonlinearity in the real protein data, such a relation may not be strictly valid. Still, we can prove that the determinant of the dynamical covariance matrix $C_D$ is proportional to the determinant of the evolutionary covariance matrix $C_E$ [93], that is, $\det C_D \sim \det C_E$ or $\prod_{i \geq 1} 1/\lambda_i^D \sim \prod_{i \geq 1} 1/\lambda_i^E$, where the eigenvalues $\lambda_i^D$ and $\lambda_i^E$ can be obtained by the eigendecomposition $H = \sum_i \lambda_i^D \vec{v}_i^D \cdot (\vec{v}_i^D)^T$, and $F = \sum_i \lambda_i^E \vec{v}_i^E \cdot (\vec{v}_i^E)^T$. Here, $0 < \lambda_1^D \leq \lambda_2^D \leq \cdots \leq \lambda_{N-6}^D$, and similarly for $\lambda_i^E$. As illustrated in Fig. 3(a), $1/\lambda_i^D$ (or $1/\lambda_i^E$) provides an estimation of the magnitude of fluctuation along the direction of the eigenvector $\vec{v}_i^D$ (or $\vec{v}_i^E$). Thus, the determinant (product of the nonzero eigenvalues) gives an estimation of the total "volume" of the conformational space accessible by dynamical or evolutionary changes in the vicinity of a given native structure (see Ref. [37], Sec. 5.1). The correspondence between the determinants indicates that the total volume of the conformational space accessed in the dynamics is proportional to that accessed in the evolution. The former is related to the sensitivity to variable perturbations [87], and the latter is associated with the robustness to parameter mutations [88,89]. Hence, high sensitivity in variable perturbations...
corresponds to high robustness in parameter mutations. A conceptual sketch is provided in Fig. 3(b). For the flat basin, thermal noises can trigger large fluctuations, while mutations can hardly change the energy around the minima. For the sharp basin, thermal noises can only trigger small fluctuations, while mutations can lead to large energy differences. Similar relations were also observed in the evolution-development correspondence [5, 94, 95] and the generalization of deep learning [96, 97].

More generally, one may expect to analyze the dynamics-evolution correspondence with variables \( \vec{x} \) and parameters \( \vec{\theta} \). For instance, with parameters \( \vec{\theta} \), one shall introduce the corresponding Fisher information matrix \( F_{\vec{\theta}} \) with entries \( F_{\vec{\theta}}^{ij} = (\partial^2 E/\partial \theta_i \partial \theta_j) \). By introducing the Jacobian matrix \( J \) with entries \( J_{ij} = \partial r_i / \partial \theta_j \), according to the chain rule, we have

\[
F_{\vec{\theta}} = J^T F J = J^T H J. \tag{2}
\]

Similarly, one may introduce another Jacobian matrix \( J' \) to reparameterize \( \vec{r} \) as \( \vec{x} \). The fluctuations of \( \vec{x} \) and \( \vec{\theta} \) can be described as the covariance matrices \( C_{\vec{x}} \sim (H_{\vec{x}})^{-1} \) and \( C_{\vec{\theta}} \sim (F_{\vec{\theta}})^{-1} \), respectively. If vectors \( \vec{x} \), \( \vec{\theta} \), \( \vec{r} \), and \( \vec{r}' \) are all in the same dimension, then \( \det C_{\vec{x}} \sim \det C_{\vec{\theta}} \) is still valid. For more general cases, when the vectors are not in the same dimension, if covariance matrices \( C_{\vec{x}} \) and \( C_{\vec{\theta}} \) are dominated by their first \( d \) modes, and if \( d \leq \dim(\vec{x}) \) and \( d \leq \dim(\vec{\theta}) \), one can prove that \( \det C_{\vec{x}} \sim \det C_{\vec{\theta}} \) (see Ref. [37], Sec. 5.2). These results suggest that the correspondence between the determinants is universal and does not rely on the selection of the coordinates [98].

The correspondence between the determinants reflects the connection between the dynamical and evolutionary conformational spaces in the measure of volume. However, it is not sufficient to explain the high coincidence in the principal components of the dynamics and evolution of proteins. To demonstrate such a coincidence, we first take 3CLpro as an example. By comparing the dynamical modes given by ENM with the evolutionary modes

\[ \text{FIG. 4.} \] For 3CLpro (and its protein family), (a) the mode-by-mode overlap between the first six dynamical modes (predicted by ENM) and evolutionary modes (given by the principal component analysis), and (b) the scattering plot of the first two dynamical modes vs evolutionary modes. For protein families in our database, (c) the dynamics-evolution subspace (spanned by the first six modes) overlap vs \( \langle \lambda^D \rangle_E \), (d) modularity \( Q \) vs \( \langle \lambda^D \rangle_E \), (e) the log volume of the conformational space \( -\ln(\lambda_{1}^D \lambda_{2}^E) \) vs \( \langle \lambda^D \rangle_E \), and (f) the chain length \( N \) dependence of the first nonzero eigenvalue \( \lambda_1^D \) and the Rayleigh quotient \( \lambda^D \).
obtained by the principal component analysis, a mode-by-mode overlap matrix can be obtained. The overlap matrix of the first six modes is shown in Fig. 4(a). One can find that the first (or the second) dynamical mode $v_1^D$ (or $v_2^D$) has high overlap ($>0.7$) with the second (or the first) evolutionary mode $v_2^E$ (or $v_1^E$). In Fig. 4(b), the scattering plots of the first two evolutionary modes ($v_2^E$ and $v_1^E$) vs dynamical modes ($v_2^D$ and $v_1^D$) show a high correlation, indicating the correspondence in the subspaces spanned by the first two modes.

To further quantify the overlap between the subspaces spanned by the dynamical and evolutionary modes, let us consider the projection of the mutation-induced deformations $\Delta \theta$ on the dynamical modes $\vec{v}^D$. We first normalize the Hessian matrix $H$ so that the average eigenvalue $\langle \lambda^D \rangle = 1$ \cite{100}. Thus, we introduce the average Rayleigh quotient

$$\langle \lambda^D \rangle_E = \frac{1}{m} \sum_{qp} \frac{\Delta \hat{\theta}_{qp}^H \cdot H \cdot \Delta \hat{\theta}_{qp}}{||\Delta \hat{\theta}_{qp}||^2},$$

where the deformation from structure $p$ to $q$ is denoted as $\Delta \hat{\theta}_{qp}$, and $m$ denotes the total number of $q$-$p$ pairs. Notably, when $\Delta \hat{\theta}_{qp}$ is parallel to the slow dynamical modes, say, $\vec{v}^D$, then, $\Delta \hat{\theta}_{qp}^H \cdot H \cdot \Delta \hat{\theta}_{qp} = \sum (\lambda_i^D \Delta \hat{\theta}_{qp} \cdot v^D_i)^2 = \lambda_i^D ||\Delta \hat{\theta}_{qp}||^2$, so $\langle \lambda^D \rangle_E = \lambda_i^D < 1 = \langle \lambda^D \rangle$. Thus, if the mutation-induced deformation can be well described by a few dynamical modes, $\langle \lambda^D \rangle_E$ yields a small value. The magnitude of $\langle \lambda^D \rangle_E$ is consistent with the measurement of the overlap between the $d$-dimensional (herein we set $d = 6$) subspaces \cite{100} spanned by dynamical and evolutionary modes (see Ref. \cite{37}, Sec. 6). As shown in Fig. 4(c), for protein families with smaller $\langle \lambda^D \rangle_E$ there is an increasing overlap between the subspaces spanned by dynamical and evolutionary modes. For our dataset, the average $\langle \lambda^D \rangle_E \approx 0.26$, and the mean subspace overlap is approximately 0.41. Compared with random systems or permuted systems with similar degrees of freedom, the dynamical and evolutionary modes of proteins exhibit a significantly high overlap in the low-dimensional subspaces \cite{101}. These results are in line with previous observations on the genotype and phenotype of proteins \cite{25,26,28}.

To understand the structural (topological) origin of low dimensionality, we analyze the modularity $Q$ of the elastic network of the proteins (see Ref. \cite{37}, Sec. 7). Intuitively, a network that can be divided more easily into modules would have higher modularity $Q$. As shown in Fig. 4(d), $\langle \lambda^D \rangle_E$ has a negative correlation with $Q$, indicating that the modularized structures contribute to the low dimensionality in protein dynamics and evolution. Owing to such low dimensionality, the total conformational space can be approximated as the subspace spanned by a few leading dynamical or evolutionary modes. For example, $-\ln(\lambda_1^D \lambda_2^D)$ or $-\ln(\lambda_1^E \lambda_2^E)$ can be assumed to approximate the log volume of the dynamical or evolutionary conformational space, respectively. As shown in Fig. 4(e), $-\ln(\lambda_1^D \lambda_1^E)$ has a negative correlation with $\langle \lambda^D \rangle_E$, suggesting that proteins maximize their conformational changes by reducing the dimensionality of the dynamics and evolution. In this way, proteins gain both flexibility in dynamics and structural variability in evolution.

Finally, we concentrate on the size (chain length $N$) dependence of the $\langle \lambda^D \rangle_E$ of different protein families. As shown in Fig. 4(f), $\langle \lambda^D \rangle_E$ decays gradually as the chain length $N$ increases, and a power-law relation $\langle \lambda^D \rangle_E \sim N^{-\xi}$ ($\xi \approx 0.23$) holds. As a comparison, the size dependence of the slowest mode ($\lambda_1^D \sim N^{-1}$) is also plotted. These scaling relations reveal an evolutionary constraint for proteins of different sizes. Larger proteins exhibit slower relaxations, and their mutation-induced deformations show higher overlap with slow dynamical modes.

Overall, our data analysis quantitatively demonstrates the dynamics-evolution correspondence in protein structures. Our theoretical study elucidates the origin of this correspondence and establishes the plasticity-robustness relation between dynamics and evolution. These results suggest a universal link between the functionality and evolvability of proteins, which provides a quantitative framework leading toward the rational function-driven protein design \cite{102,103}. In such a framework, the designed function will determine the mobility of the residues, and the information of residue motions will determine the features in the sequences together with evolutionary constraints.

We thank Jun Wang, Wei Wang, Zeke Xie, Tao Wang, Yunqiang Bian, Tetsuhiro S Hatakeyama, Sosuke Ito, Ayaka Sakata, Takuya U Sato, Weitong Ren, Liang Tian, Changsong Zhou, and Taro Toyoizumi for participating in stimulating discussions. We also thank the anonymous reviewers for their great suggestions and comments. This research was partially supported by a Grant-in-Aid for Scientific Research on Innovative Areas (17H06386) from the Ministry of Education, Culture, Sports, Science, and Technology of Japan, and a Grant-in-Aid for Scientific Research (A)20H00123 from the Japanese Society for the Promotion of Science.

\*Corresponding author.
qianyuan.tang@riken.jp
\*Corresponding author.
kaneko@complex.c.u-tokyo.ac.jp

[1] C. H. Waddington, *The Strategy of the Genes* (George Allen & Unwin, Ltd., London, 1957).

[2] J. A. G. de Visser, J. Hermisson, G. P. Wagner, L. A. Meyers, H. Bagheri-Chaichian, J. L. Blanchard, L. Chao, J. M. Cheverud, S. F. Elena, W. Fontana et al., *Perspective:
[40] M. Gurusaran, M. Shankar, R. Nagarajan, J. R. Helliwell, and K. Sekar, Do we see what we should see? describing non-covalent interactions in protein structures including precision, IUCrJ 1, 74 (2014).

[41] I. Kufareva and R. Abagyan, Methods of protein structure comparison, in Homology Modeling (Springer, New York, 2011), ch. 10, pp. 231–257.

[42] M. Iyer, Z. Li, L. Jaroszewski, M. Sedova, and A. Godzik, Difference contact maps: From what to why in the analysis of the conformational flexibility of proteins, PLoS One 15, e0226702 (2020).

[43] E. F. Garman and T. R. Schneider, Macromolecular cryocystallography, J. Appl. Crystallogr. 30, 211 (1997).

[44] A. L. Tournier, J. Xu, and J. C. Smith, Translational mobility of proteins with an elastic network model, Biophys. J. 80, 505 (2001).

[45] A. R. Atilgan, S. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Anisotropy of fluctuation dynamics of proteins on the proteome scale, Evol. Appl. 6, 423 (2013).

[46] P. Sneath, Relations between chemical structure and biological activity in peptides, J. Theor. Biol. 12, 157 (1966).

[47] C. J. Epstein, Non-randomness of amino-acid changes in the evolution of homologous proteins, Nature (London) 215, 355 (1967).

[48] R. Grantham, Amino acid difference formula to help explain protein evolution, Science 185, 862 (1974).

[49] T. Haliloglu, I. Bahar, and B. Erman, Gaussian Dynamics of Folded Proteins, Phys. Rev. Lett. 79, 3090 (1997).

[50] T. Haliloglu and B. Erman, Analysis of Correlations between Energy and Residue Fluctuations in Native Proteins and Determination of Specific Sites for Binding, Phys. Rev. Lett. 102, 088103 (2009).

[51] A. R. Atilgan, S. Durell, R. L. Jernigan, M. C. Demirel, O. Keskin, and I. Bahar, Anisotropy of fluctuation dynamics of proteins with an elastic network model, Biophys. J. 80, 505 (2001).

[52] E. Eyal, L.-W. Yang, and I. Bahar, Anisotropic network model: Systematic evaluation and a new web interface, Bioinformatics 22, 2619 (2006).

[53] I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability, Phys. Rev. Lett. 80, 2733 (1998).

[54] I. Bahar, T. R. Lezon, L.-W. Yang, and E. Eyal, Global dynamics of proteins: Bridging between structure and function, Annu. Rev. Biophys. 39, 23 (2010).

[55] Y. Togashi and H. Flechsig, Coarse-grained protein dynamics studies using elastic network models, Int. J. Mol. Sci. 19, 3899 (2018).

[56] V. Alexandrov, U. Lehnhert, N. Echols, D. Milburn, D. Engelman, and M. Gerstein, Normal modes for predicting protein motions: A comprehensive database assessment and associated web tool, Protein Sci. 14, 633 (2005).

[57] H. Wako and S. Endo, Normal mode analysis as a method to derive protein dynamics information from the protein data bank, Biophys. Rev. Lett. 9, 877 (2017).

[58] M. Ikeguchi, J. Ueno, M. Sato, and A. Kidera, Protein Structural Change Upon Ligand Binding: Linear Response Theory, Phys. Rev. Lett. 94, 078102 (2005).

[59] L. Yang, G. Song, and R. L. Jernigan, How well can we understand large-scale protein motions using normal modes of elastic network models?, Biophys. J. 93, 920 (2007).

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

[61] M. E. Newman, Modularity and community structure in networks, Proc. Natl. Acad. Sci. U.S.A. 103, 8577 (2006).

[62] M. E. J. Newman, Equivalence between modularity optimization and maximum likelihood methods for community detection, Phys. Rev. E 94, 052315 (2016).

[63] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, Fast unfolding of communities in large networks, J. Stat. Mech. 10 (2008) P10008.

[64] A. Lancichinetti and S. Fortunato, Community detection algorithms: A comparative analysis, Phys. Rev. E 80, 056117 (2009).

[65] R. Guimera, M. Sales-Pardo, and L. A. N. Amaral, Modularity from fluctuations in random graphs and complex networks, Phys. Rev. E 70, 025101(R) (2004).

[66] M. M. Tirion, Large Amplitude Elastic Motions in Proteins from a Single-Parameter, Atomic Analysis, Phys. Rev. Lett. 77, 1905 (1996).

[67] C. Chenubhnotla, A. Rader, L.-W. Yang, and I. Bahar, Elastic network models for understanding biomolecular machinery: From enzymes to supramolecular assemblies, Phys. Biol. 2, S173 (2005).

[68] Y. Togashi and A. S. Mikhailov, Nonlinear relaxation dynamics in elastic networks and design principles of molecular machines, Proc. Natl. Acad. Sci. U.S.A. 104, 8697 (2007).

[69] Note that here the noise- and mutation-induced structural variations are much greater than the expected error in structure determination and the thermal fluctuations in crystallized structures (for details, see Ref. [37], Sec. 1.2). The z score represents the distance between the raw data and the mean value in units of the standard deviation. For raw data \{x_i\}, the z score \{z_i\} is calculated by subtracting the mean value \((\bar{x})\) from the \(x_i\) and then dividing by the standard deviation \(s_x\), i.e., \(z_i = (x_i - \bar{x}) / s_x\).

[70] The Spearman correlation coefficient \(\rho_s\) is defined as the Pearson correlation coefficient between the rank variables. It describes how well the relationship between two variables can be characterized with a monotonic function.

[71] V. S. Pande, A. Y. Grosberg, and T. Tanaka, How accurate must potentials be for successful modeling of protein folding?, J. Chem. Phys. 103, 9482 (1995).

[72] V. S. Pande, A. Y. Grosberg, and T. Tanaka, Statistical mechanics of simple models of protein folding and design, Biophys. J. 73, 3192 (1997).

[73] D. J. Earl and M. W. Deem, Evolvability is a selectable trait, Proc. Natl. Acad. Sci. U.S.A. 101, 11531 (2004).

[74] N. Tokuriki and D. S. Tawfik, Protein dynamism and evolvability, Science 324, 203 (2009).

[75] Y. Liu and I. Bahar, Sequence evolution correlates with structural dynamics, Mol. Biol. Evol. 29, 2253 (2012).

[76] Z. Nevin Gerek, S. Kumar, and S. Banu Ozkan, Structural dynamics flexibility informs function and evolution at a proteome scale, Evol. Appl. 6, 423 (2013).
[78] A. Bakan, L. M. Meireles, and I. Bahar, Prodyp: Protein dynamics inferred from theory and experiments, Bioinformatics 27, 1575 (2011).
[79] S. Zhang, H. Li, J. M. Krieger, and I. Bahar, Shared signature dynamics tempered by local fluctuations enables fold adaptability and specificity, Mol. Biol. Evol. 36, 2053 (2019).
[80] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The protein data bank, Nucleic Acids Res. 28, 235 (2000).
[81] Q.-Y. Tang, Y.-Y. Zhang, J. Wang, W. Wang, and D. R. Chialvo, Critical Fluctuations in the Native State of Proteins, Phys. Rev. Lett. 118, 088102 (2017).
[82] A. Cavagna, A. Cimarelli, I. Giardina, G. Parisi, R. Santagati, F. Stefanini, and M. Viale, Scale-free correlations in starling flocks, Proc. Natl. Acad. Sci. U.S.A. 107, 11865 (2010).
[83] A. Attanasi, A. Cavagna, L. Del Castello, I. Giardina, S. Melillo, L. Parisi, O. Pohl, B. Rossaro, E. Shen, E. Silvestri et al., Finite-Size Scaling as a Way to Probe Near-Criticality in Natural Swarms, Phys. Rev. Lett. 113, 238102 (2014).
[84] Q.-Y. Tang and K. Kaneko, Long-range correlation in protein dynamics: Confirmation by structural data and normal mode analysis, PLoS Comput. Biol. 16, e1007670 (2020).
[85] X.-L. Xu, J.-X. Shi, J. Wang, and W. Li, Long-range correlation and critical fluctuations in coevolution networks of protein sequences, Physica (Amsterdam) 562A, 125339 (2021).
[86] S. Reuveni, R. Granek, and J. Klafter, Proteins: Coexistence of Stability and Flexibility, Phys. Rev. Lett. 100, 208101 (2008).
[87] Q.-Y. Tang, T. S. Hatakemura, and K. Kaneko, Functional sensitivity and mutational robustness of proteins, Phys. Rev. Research 2, 033452 (2020).
[88] R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, Universally sloppy parameter sensitivities in systems biology models, PLoS Comput. Biol. 3, e189 (2007).
[89] B. C. Daniels, Y.-J. Chen, J. P. Sethna, R. N. Gutenkunst, and C. R. Myers, Sloppiness, robustness, and evolvability in systems biology, Curr. Opin. Biotechnol. 19, 389 (2008).
[90] For example, the sequence, hydrophobicity profile, or the “distance” from the optimal sequence to the given one can be taken as the parameter \( \theta \) (for details, see Ref. [37], Sec. 3.1), and the variable \( \bar{x} \) can be a combination of Cartesian coordinates, internal coordinates (bond lengths, bond angles, and torsional angles, etc.), or other reaction coordinates. The dimension of \( \bar{\theta} \) (or \( \bar{x} \)) need not be the same as \( \vec{p} \) (or \( \vec{r} \)).
[91] M. Abt and W. J. Welch, Fisher information and maximum-likelihood estimation of covariance parameters in Gaussian stochastic processes, Can. J. Stat. 26, 127 (1998).
[92] S.-i. Amari, Differential-Geometrical Methods in Statistics (Springer Science & Business Media, New York, 2012), Vol. 28.
[93] More precisely speaking, the determinants we mention here should be pseudodeterminants (defined as the product of the nonzero eigenvalues) in real applications.
[94] K. Kaneko, Phenotypic plasticity and robustness: Evolutionary stability theory, gene expression dynamics model, and laboratory experiments, in Evolutionary Systems Biology (Springer, New York, 2012), pp. 249–278.
[95] K. Kaneko, Evolution of robustness and plasticity under environmental fluctuation: Formulation in terms of phenotypic variances, J. Stat. Phys. 148, 687 (2012).
[96] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, and P. T. P. Tang, On large-batch training for deep learning: Generalization gap and sharp minima, International Conference on Learning Representations (2017), https://openreview.net/forum?id=H1oyRlYgg.
[97] Z. Xie, I. Sato, and M. Sugiyama, A diffusion theory for deep learning dynamics: Stochastic gradient descent escapes from sharp minima exponentially fast, International Conference on Learning Representations (2021), https://openreview.net/forum?id=wXgk_iCiYGo.
[98] H. Jeffreys, An invariant form for the prior probability in estimation problems, Proc. R. Soc. Ser A. 186, 453 (1946).
[99] In the calculation of \( \langle \vec{p} \rangle_F \), we take the matrix \( H \) as the normalized graph Laplacian of the elastic network.
[100] A. Amadei, M.A. Ceruso, and A. Di Nola, On the convergence of the conformational coordinates basis set obtained by the essential dynamics analysis of proteins’ molecular dynamics simulations, Proteins 36, 419 (1999).
[101] For high-dimensional random systems, the randomly drawn vectors are almost orthogonal, so the subspace overlap is close to zero. For permuted systems with similar degrees of freedom (\( N = 300 \)), the subspace overlap (\( d = 6 \)) can be estimated as \( d/N = 0.02 \), which is much lower than that of the proteins.
[102] J. Zhou, A. E. Panaitiu, and G. Grigoryan, A general-purpose protein design framework based on mining sequence–structure relationships in known protein structures, Proc. Natl. Acad. Sci. U.S.A. 117, 1059 (2020).
[103] C. Yang, F. Sesterhenn, J. Bonet, E. A. van Aalen, L. Scheller, L. A. Abriata, J. T. Cramer, X. Wen, S. Rosset, S. Georgeon et al., Bottom-up de novo design of functional proteins with complex structural features, Nat. Chem. Biol. 17, 492 (2021).