The Statistical Trends of Protein Evolution: A Lesson from AlphaFold Database

Qian-Yuan Tang [1]*, Weitong Ren [2], Jun Wang [3], Kunihiko Kaneko [4]†

[1] Lab for Neural Computation and Adaptation, RIKEN Center for Brain Science, 2-1 Hirosawa, Wako, Saitama, 351-0106, Japan.

[2] Theoretical Molecular Science Laboratory, RIKEN Cluster for Pioneering Research, 2-1 Hirosawa, Wako, Saitama 351-0198, Japan.

[3] School of Physics, National Laboratory of Solid State Microstructure, and Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing, 210093, China.

[4] Center for Complex Systems Biology, Universal Biology Institute, University of Tokyo, Tokyo, Japan.

Email:

* tangqianyuan@gmail.com

† kaneko@complex.c.u-tokyo.ac.jp

Abstract

The recent development of artificial intelligence provides us with new and powerful tools for studying the mysterious relationship between organism evolution and protein evolution. In this work, based on the AlphaFold Protein Structure Database (AlphaFold DB), we perform comparative analyses of the protein structures of different species. The statistics of AlphaFold-predicted structures show that, as species evolve towards higher complexity (e.g., from prokaryotes to eukaryotes, from unicellular to multicellular organisms, etc.), the proteins within them evolve toward larger radii of gyration, higher coil fractions, higher modularity, and slower relaxations, indicating that the average flexibility of proteins gradually increases in the species evolution. With the scaling analyses, the size dependence of proteins' shape, topology, and dynamics suggest a decreasing fractal dimension of the proteins in species evolution. This evolutionary trend is accompanied by the increasing eigengaps in the vibration spectra of proteins, implying that the proteins are statistically evolving toward higher functional specificity and lower dimensionality in the native dynamics. Furthermore, we also uncover the topology and sequence bases of this evolutionary trend. The residue contact networks of proteins evolve toward higher assortativity, and the hydrophobic and hydrophilic amino acid residues evolve to have increasing segregations in the sequences. This study provides new insights into how the diversity in the functionality of proteins increases and how the dimensionality of the manifold of protein dynamics reduces during evolution. The statistical evolutionary trends also shed light on the study of protein design.

Introduction

The evolution of species takes place over time, as Darwin noted in his "*On the Origin of Species*". From prokaryotes to eukaryotes, from unicellular to multicellular organisms, from invertebrates to vertebrates, and so on, the complexity of life has increased throughout evolution (1, 2). Proteins, as the basic building blocks of lives, are also evolving throughout the evolutionary process. However, the relationship between organism evolution and protein evolution remains a mystery (3–6). As the organism evolves toward higher complexity, is there any parallel evolutionary trend related to protein evolution? Theoretically, in the microscopic view (protein level), if considering the evolution of proteins with a common ancestral sequence or structure, their sequences and structures do not necessarily follow the evolutionary path as the species evolve (7). In contrast, in the macroscopic view (organism level), if considering an ensemble of thousands of proteins within the same organism, just as thermodynamic laws and collective order can emerge from the random motions of molecules, it is possible to discover a "collective" evolution trend consistent with the direction of species evolution. Unfortunately, most studies on protein evolution today still focus on the microscopic level, that is, to elaborate on the datasets of proteins that share the common ancestral sequence or structure (8–13). It is difficult to construct a database for analyzing the evolution of all proteins in various species since it is costly and time-consuming to determine the structures of all proteins contained within the proteome of a species by experiments.

Notably, the recent development of artificial intelligence (AI) provides us with new and powerful tools to help us elucidate the trends in protein evolution on the macroscopic scale. AlphaFold, developed by DeepMind, is an artificial intelligence system based on deep learning, which performs protein structure predictions with high accuracy (14, 15) and finds many applications in medical and biological sciences (16, 17). Although some of the predictions were noted to have limitations (18, 19), AlphaFold had already won an unprecedented and overwhelming success. In CASP14, the structures predicted by AlphaFold 2 were vastly more accurate than other competing methods (20). The exceeding accuracy and the speed of AlphaFold 2 provide the possibility of generating an extensive database of structure predictions. AlphaFold Protein Structure Database (AlphaFold DB) provides open access to protein structure predictions for the proteomes of the model organisms and proteomes related to global health. The database provides free access to over 365,000 predicted protein structures across the proteomes of 21 different species ranging from bacteria, archaea, unicellular and multicellular eukaryotes to humans, and such a database is still expanding (21). The extensive structural dataset not only offers the potential to answer the critical questions about life activities and human health, but also suggests a new paradigm for studying protein evolution. Instead of focusing only on specific families of proteins (16), we can now perform comparative analyses of proteins structures in different species. The statistics of the full-proteome protein structures may indicate the trend of species evolution. With this new perspective, we can better elucidate the relationship between the evolution of proteins and species.

In this work, based on the AlphaFold DB, we perform comparative analyses of the protein structures of different species. The statistics of AlphaFold-predicted structures show that, as the species evolve toward higher complexity, the structures of the proteins evolve toward larger radii of gyration, higher coil fractions, higher modularity, and slower relaxations, indicating that the average flexibility of proteins gradually increases in the species evolution. With the scaling analyses, the size dependence of proteins' shape, topology, and dynamics suggest a decreasing fractal dimension of the proteins in species evolution. This evolutionary trend is accompanied by the increasing eigengaps in the vibration spectra of proteins, implying that the proteins are statistically evolving toward higher functional specificity and lower dimensionality in the equilibrium dynamics. Furthermore, we also uncover the topology and sequence bases of this evolutionary

trend. The residue contact networks of the proteins are evolving toward higher assortativity, and the hydrophobic and hydrophilic amino acid residues are evolving to have increasing segregations in the sequences. This study provides new insights into how the diversity in the functionality of proteins increases and how the dimensionality of the manifold of protein dynamics reduces during evolution. The statistical evolutionary trends also shed light on the study of protein design.

Results

The statistics of AlphaFold-predicted structures indicate that as species evolve, the average flexibility of proteins gradually increases.

We performed a comparative analysis of the predicted protein structures of the 21 organisms from the AlphaFold DB. The details of the dataset are listed in Supplementary Information (SI) Table S1. Let us first focus on the proteins with similar chain lengths. Namely, for the proteomes of various species, we always select the protein structures with chain lengths $N \approx 250$, calculate their structural characteristics, and conduct statistical analyses. As shown in Fig. 1A, despite having almost identical chain lengths, there are significant differences in the distribution of the radii of gyration R_g across different species. More interestingly, if we sort the different species according to the median R_g, it is observed that the trend of increasing protein radii generally correlates with the direction of species evolution. As the species evolve toward higher complexity, say, from prokaryotes to eukaryotes (e.g., *E. coli* vs. Yeast), from unicellular to multicellular organisms (e.g., from yeast or *Dictyostelium* to other multicellular organisms), from invertebrates to vertebrates (e.g., *C. elegans* vs. human), and so on, the proteins' average radii of gyration R_g gradually increase. Since the chain lengths of these proteins are similar, this result implies an evolutionary trend toward lower packing density and higher average flexibility throughout species evolution. Note that such an evolutionary trend is relatively robust. For example, if one sorts the species according to the proteins with other given chain lengths, or we select other structural descriptors such as the solvent-accessible surface areas (SASA) to quantify the flexibility of the proteins, a similar evolutionary trend can still be observed (see Fig. S1).

To demonstrate this evolutionary trend clearly, we select five species for more detailed examinations. The five selected species include archaea (*Methanocaldococcus jannaschii*), bacteria (*E. coli*), unicellular eukaryotes (budding yeast, *S. cerevisiae*), multicellular eukaryotes (*C. elegans*), and human (*Homo sapiens*). Except for the archaea, the complexity of the other four species increases sequentially. Based on the proteomes of these species, we select the proteins with chain lengths $N \approx 250$ and conduct detailed statistics for these structures. The histograms shown in Figs. 1B-E demonstrate how the tertiary structure, secondary structure, amino acid residue contact topology, and equilibrium dynamics of the proteins evolve as the species evolve.

First, as shown in Fig. 1B, as the species evolve toward higher complexity, for the proteins with similar chain lengths, the mean value of the radius of gyration R_g gradually increases. Then, we apply the DSSP algorithm (22, 23) to assign secondary structures (helices, sheets, or coils) to the proteins and investigate how protein flexibility changes during evolution. We find that the average fraction of coils in protein molecules also gradually increases as the species evolve (Fig. 1C). Next, according to the spatial distance between amino acid residues, we construct the residue contact networks for the proteins and calculate the modularity Q of these networks. A higher modularity Q indicates dense connections within the same module but sparse connections between different modules. For the proteins with similar chain lengths N from different species, the mean value of the modularity Q increases as the species evolves (Fig. 1D). The

evolutionary trend toward higher modularity is consistent with previous theoretical models (24). Furthermore, we examine the flexibility of the proteins according to their relaxations around the native structure. According to recent studies, the functional dynamics of the proteins are closely related to such relaxations and can be predicted by the elastic network model (ENM) (25–27). Usually, when a protein is more flexible, its relaxation will be slower. Based on ENM, one can conduct the normal mode analysis and obtain the eigenvalues corresponding to the relaxation modes (28–30). Among these eigenvalues, the smallest nonzero eigenvalue λ_1, which corresponds to the square of the slowest mode frequency ($\lambda_1 \sim \omega_1^2$) and is proportional to the inverse square of the amplitude of the motion ($\lambda_1 \sim A_1^{-2}$), can also be recognized as a measure of molecular flexibility. The statistics of λ_1 show an evolutionary trend consistent with the other results. As shown in Fig. 1E, statistically, as the species evolves, the proteins will exhibit slower relaxations around the equilibrium. In short, the above results clearly demonstrate an evolutionary trend toward higher flexibility. Further analyses show that if we remove the proteins with lower prediction confidence from the dataset, a similar statistical evolutionary trend can still be observed (see Discussions and SI Fig. S2). Besides, these statistical results also show an increasing variance in protein flexibility, suggesting that an expanding diversity of protein structures is accompanied by organism evolution.

The scaling analyses suggest a decreasing fractal dimension of the proteins in species evolution.

In the previous section, based on the proteins with similar chain lengths, we discuss the evolutionary trend in protein structures and dynamics during species evolution. In fact, such an evolutionary trend holds not only for proteins with specific chain lengths but also for proteins with different chain lengths in various species. Based on the Protein Data Bank (PDB) (31), recent scaling analyses (32, 33) revealed the power laws related to the size dependence of the protein shapes, modularity, and slowest vibration mode. Due to the limited number of experimentally determined protein structures, it had been unable to compare the scaling coefficients for proteomes of different species separately. AlphaFold DB, however, compensates for the lack of experimental data, enables us to perform accurate scaling analyses of proteins across different species, compare scaling coefficients, and reveal the trend in protein evolution.

In this work, based on the predicted structures from AlphaFold DB, we apply the scaling analyses to the proteins from different species. For proteins within a species, we first divide the proteins into groups according to their chain length. Then, for the proteins in the same group (i.e., with similar chain length N), we calculate the mean value, as well as its standard error, of the radius of gyration R_g, the modularity Q of the residue contact network, and the slowest mode eigenvalue λ_1. In this way, we obtain the size dependence of proteins' shape, topology, and dynamics. Figure 2 shows the results of the scaling analysis for the five selected species.

Let us first analyze the scaling relations between protein shape and chain length N for different species. Here, the protein shape is described by the length of the longest principal semi-axis L_C. There will be a scaling relation: $L_C \sim N^{1/d}$, in which d is the average fractal dimension of the proteins. For example, when the proteins are densely packed into globular shapes in the three-dimensional space, there will be $L_C \sim R_g \sim N^{1/3}$. As shown in Fig. 2A, the scaling analyses show that the average fractal dimension $d \leq 3$ for all the species. This result displays that the proteins are not always folded into densely packed globules, and it is in line with previous studies on the fractal structure of the proteins. Taking a closer look at the scaling coefficients of the different species, one can find that the fractal dimension d tends to decrease as the species evolve. These results are consistent with the evolutionary trend toward higher flexibility discussed in the previous section. Besides, if we directly calculate the fractal dimension of the proteins (average packing dimension over all the residues) with the box-counting method, a similar evolutionary trend can also be observed (see SI Fig. S3).

Then, let us examine the size dependence of the modularity Q of the proteins' residue contact networks for different species. According to the definition, a network that can be more easily divided into modules would have a higher modularity Q. Because lower-dimensional geometric graphs can be more easily divided into modules (e.g., a one-dimensional geometric graph can be divided into two modules by removing only one edge), networks embedded in a lower-dimensional space will have a higher Q. Theoretically, for a d-dimensional cubic lattice network with N nodes, the modularity Q versus N follows the scaling relation: $1 - Q \sim N^{-1/(\tilde{d}+1)}$, in which $\tilde{d}$ serves as another estimation of the average fractal dimension of the proteins (33, 34). Note that even for the same group of proteins, the estimated $\tilde{d}$ (based on the network topology related to the local contacts between residues) may differ from the average fractal dimension d (estimated from the overall shapes of the proteins). Yet, according to Fig. 2B, the size dependence of modularity Q indicates again that the residue contact networks are fractal ($\tilde{d} < 3$). Moreover, it is also observed that a decreasing fractal dimension $\tilde{d}$ correlates with the evolution of species.

Next, we investigate the size dependence of the equilibrium dynamics of the native proteins. Previous research based on the PDB had shown that as the protein chain length N increases, the relaxation of the protein would become slower, i.e., the slowest mode eigenvalue λ_1 versus chain length N obeys the scaling relation: $\lambda_1 \sim N^{-\mu}$, and $\mu \approx 1$. Based on the AlphaFold DB, our analyses can reproduce a similar scaling relation. However, as shown in Fig. 2C, detailed analyses reveal that the scaling coefficient μ may vary across different species. A roughly increasing μ is correlated with the evolution of species, suggesting that proteins are evolving toward higher flexibility and slower relaxation during species evolution.

Furthermore, in addition to the difference in scaling coefficients, the data in Fig. 2 also clearly demonstrate the evolutionary trend of proteins. For example, in Fig. 2A, the low standard errors of the means indicate a highly accurate estimation of the mean values of L_c. Besides, at any given chain length N, if we sort the species by the mean values of L_c, the ordering of the species is highly robust (e.g., the mean L_c of the human is always greater than that of the *E. coli*), showing a correlation with the direction of species evolution.

The increasing eigengaps in the vibration spectra imply an evolutionary trend toward higher functional specificity and lower dimensionality in protein dynamics.

According to the "structure-dynamics-function" paradigm (35–37), it is the dynamics which is encoded by the three-dimensional structure that determine the functionality of the protein. To better understand the impact of this evolutionary trend on protein function, we need more detailed statistics on the dynamics of proteins. As reported by previous research, the protein dynamics are suggested to be constrained to low-dimensional manifolds and could be described by a few 'slow modes' which correspond to low vibration frequency, low excitation energies, and large amplitudes motions. To obtain these slow modes, for every protein, we construct the corresponding elastic network model and conduct the normal mode analysis. In this way, the vibration spectrum can be obtained (see Method). In the vibration spectrum, the eigengaps between leading eigenvalues can work as the identifier to quantify the functional specificity of the proteins (38). The greater the ratio between the neighboring eigenvalues, the higher the functional specificity of the proteins. For example, let us consider the first and second slowest modes, if $\lambda_2/\lambda_1 = 10$, then the ratio of the square amplitudes of the two modes will be: $A_1^2/A_2^2 = \lambda_2/\lambda_1 = 10$, that is, the motion along the direction of the first mode is an order of magnitude greater than the motion along the direction of the second mode, reflecting a high functional specificity of the protein, or say, lower dimensionality in protein dynamics. In contrast, if $\lambda_2/\lambda_1 \approx 1.01$, the protein has almost the same tendency to move along the directions associated with the first and second slowest modes, which signifies a lower functional specificity and a higher dimensionality in protein dynamics.

Here, similar to previous sections, for the proteins with similar chain lengths $N \approx 250$ from the five selected species, we conduct normal mode analysis based on the elastic network model and calculate the corresponding eigenvalues. Then, the distributions of the eigengaps λ_2/λ_1, λ_3/λ_2, and λ_4/λ_3 (at logarithmic scale) are obtained. As shown in Figs. 3A-C, all the average ratios between neighboring eigenvalues increase as the species evolve, clearly demonstrating an increasing scale separation in the vibrational frequencies of the proteins. This indicates an evolutionary trend of proteins toward functional specificity and lower dimensionality. Still, it is worth noting that not all proteins in an organism will evolve in the same direction. Such an evolutionary trend is statistical. Besides, the results also indicate that the evolution toward functional specificity happens on every scale, e.g., the motions along the direction of the first modes are evolving to be much greater than the second, the motions along the second modes are evolving to be much greater than the third, etc.

To quantify such a scale-free property, it is necessary to introduce the power laws to describe the vibrations of the proteins. Previous studies have shown that the vibrational spectrum of proteins obeys a power-law distribution (39), and the rank-size distribution of the inverse of eigenvalues follows a Zipf-like distribution (40–42). In SI Fig. S4, several proteins are provided as examples. These proteins have similar chain lengths, but their vibration spectra are different. The rank-size distribution of the slow-mode eigenvalues shows a power-law-like behavior, which can be described as the Zipf's law, i.e., $\lambda_i^{-1} \sim i^{-z}$, in which z is known as the Zipf's coefficient. Generally, for the proteins with similar sizes, a larger z corresponds with larger eigengaps (Fig. 3D). In Fig. 3E, the statistics of Zipf's coefficients z show that, as the species evolves, the average z increases, indicating that the vibrational spectra of proteins are evolving toward larger eigengaps statistically. This evolutionary trend is illustrated in Fig. 3F. In the illustration, the conformational space (the space encompassing all possible structures results from thermal fluctuations) of a protein around its native structure is represented as an ellipsoid. The axes of the ellipsoids denote the normal modes, and their semi-length represents the square magnitudes of the motions in the directions of the normal modes. The functional specificity and low dimensionality correspond to the anisotropy of the conformational space. This evolutionary trend is consistent with the structure-based analysis in previous sections. As species evolve, both the structure and dynamics of proteins show a statistical trend of dimensional reduction. Yet, it is worth noting that the standard deviation of Zipf's coefficient z also increases as the species evolve, indicating an increasing diversity of protein structures and dynamics in the evolution.

Evolutionary trends of the residue contact networks and the corresponding protein sequences.

In previous studies, residue contact networks of native proteins have been shown to play a dominating role in determining protein dynamics and functions (25, 43). The residue contacts are, in turn, largely determined by the protein sequences. Therefore, we may use the residue contact network as a bridge to investigate evolutionary trends in protein sequences.

In the previous section, we show the evolutionary trend of increasing scale separation in the vibrational frequencies of the proteins. In fact, such an evolutionary trend in frequency space can correspond to the changes in the distribution of the residues' local packing density in the real space. According to the local density model (44), there will be a nearly linear relationship between the residues' square of vibration frequency and the local packing density (see Methods). As a result, the increasing scale separation in vibrational frequencies corresponds to the increasing variances of the residues' local packing density. In protein molecules, the packing density distribution is mainly determined by the hydrophobic effects. As illustrated in Fig. 4A, by avoiding exposure to water, hydrophobic residues are buried in the interior of the protein, forming the hydrophobic cores with high local packing density. In contrast, the hydrophilic (polar)

residues are likely to exposure to the solvent, so they have low local packing density. It is such hydrophobic effects that serve as the major driving force of protein folding (45, 46).

For the proteins with similar chain lengths, the variance of the local packing density is closely related to the assortativity ρ of the residue contact network. The assortativity of a network is defined as the Pearson correlation coefficient of degrees between pairs of linked nodes (47, 48). According to the definition, if the residues (nodes) with high packing density (degrees) are more likely to have contact with each other, then there will be a higher assortativity ρ and a larger variance in the packing density distribution of residues. Remarkably, as shown in Fig. 4B, for the proteins with similar chain lengths, as the species evolve, the average assortativity ρ tends to increase, indicating an evolutionary trend toward larger variance in the packing density distribution. The evolutionary trend related to the packing density is in line with the evolution of vibration spectra. As shown in Fig. 4C, for the proteins in our dataset, regardless of the species they belong to, the assortativity ρ is proportional to Zipf's coefficient z. This result indicates that we have found the topological descriptor associated with the functional specificity of the protein.

Next, let us discuss the evolutionary trend of the protein sequences. As illustrated in Fig. 4A, the spatial segregation of hydrophilic and hydrophobic residues is closely related to the sequential segregation. When the hydrophilic and hydrophobic residues are uniformly mixed and randomly distributed in the sequence, they will be less likely to have profound spatial segregation. Conversely, when there are longer hydrophilic or hydrophobic fragments in a sequence, it is more likely that hydrophobic cores will form. Here, we introduce the hydropathy variation $\mathrm{CV_{HP}}$ to quantify the sequential segregation of hydrophobicity, where $\mathrm{CV_{HP}}$ is defined as the coefficient of variation (i.e., standard deviation divided by the mean value) of the filtered hydropathy profile of a sequence (see Method). As shown in Fig. 4D, for the proteins with similar chains, the average $\mathrm{CV_{HP}}$ tends to increase as the species evolve, indicating an evolutionary trend toward more profound segregation of hydrophobic and hydrophilic residues in the sequence. As shown in Fig. 4E, for the proteins in our dataset, regardless of the species they belong to, the hydropathy variation $\mathrm{CV_{HP}}$ is proportional to the assortativity of the residue contact network, implying that the spatial segregation of hydrophilic and hydrophobic residues do correlate with the sequential segregation.

Discussion

In this work, based on the protein structures predicted by AlphaFold, we propose a new framework for studying protein evolution by comparing the protein structures in different organisms. Rather than studying specific protein families as in traditional studies of protein evolution, this study aims to elucidate the relationship between molecular evolution and species evolution. In parallel with species' evolution to higher complexity, although not all proteins in an organism will evolve in the same direction, there is a statistical evolutionary trend that the proteins evolve toward higher flexibility, higher functional specificity, and higher degrees of hydrophobicity segregation in both their structures and sequences. This result is also consistent with the fact that the intrinsic disorder is more abundant in eukaryotic proteins than prokaryotic proteins (49). For the AlphaFold-predicted structures, lower prediction confidence (which can be quantified as lower pLDDT values, see SI) usually arrives for proteins with long disordered regions (15). Interestingly, even if we remove those proteins with lower prediction confidence (i.e., with long intrinsic disorder regions) from the dataset, a similar statistical evolutionary trend can still be observed (see Fig. S2). Such insensitivity to the prediction confidence indicates that the evolution toward higher flexibility not only holds for proteins with long intrinsic disorder regions, but also holds for proteins with relatively stable native structures, strengthening our conclusions on protein evolution. Besides, since this study focuses primarily on the global

dynamics (e.g., slowest modes), which are robust to the local variations in native structures (25, 40), the evolutionary trends suggested by statistical results are nonetheless compelling.

Our results have demonstrated the differences between the proteins from archaea (*M. jannaschii*) and bacteria (*E. coli*). Although the statistics of their structures (e.g., radii of gyration) do not show significant differences (see Table S2), their sequences do show significant differences (Fig. 4D). Note that these results do not mean that bacteria have a higher complexity than archaea. It is their living environments that result in the major difference (i.e., sequence-based mechanism of thermal stabilization (50)) between *M. jannaschii* and *E. coli*. As a thermophilic methanogenic archaean, *M. jannaschii* lives in extreme environments such as hypothermal vents at the bottom of the oceans. The optimal growth temperature of *M. jannaschii* is about 85°C (51, 52), which is even higher than the melting temperature (~50°C) of most proteins in *E. coli* (53). The relatively low flexibility of the proteins in *M. jannaschii* contributes to the thermal stability, thus leading to the species' adaptation to a hot environment. In fact, as the lives on earth originated from a hot environment, so the ancient lives were inevitably adapted to the high temperatures, just like the thermophilic archaean. As the temperature of the earth started to decrease, there would be a significant drop in the activity (e.g., the catalytic speed of the enzymes) of the proteins in these ancient cells (54). In cooler environments, the thermal stability of the proteins was no longer an important constraint, but rather its activity (which related to the flexibility) became a key factor in determining the fitness of the organism.

However, the difference in the temperature of the living environment is not sufficient to explain the trend of protein evolution. For example, the temperature environments in which *E. coli* and humans live are similar, yet statistically, the structures and dynamics of the proteins in these two species are significantly different. According to our statistical results, as species evolve and organism complexity increases, the proteins will have a statistical evolutionary trend toward higher flexibility and a trend toward a lower-dimensional manifold in their dynamics. The evolutionary dimensional reduction in protein dynamics, which was also implied in the emergence of the funnel-like energy landscape (55), may contribute to the efficiency and robustness of the protein function. For example, for allosteric proteins, it is observed that evolution designs the sequence and shapes the structure of the proteins, leading toward more specific transition pathways (38, 56). This statistical trend is parallel with the phenomenon of evolutionary dimensional reduction observed in other experimental and theoretical studies (57–62).

Moreover, these results also suggest that the generalist proteins tend to evolve toward specialists statistically (63). It is in line with experimental observations showing that ancestral enzymes at the evolutionary origins are likely to have high promiscuity, as they may be able to catalyze a wide variety of chemical reactions (64–68). Our statistical results demonstrate the validity of the ancestral sequence reconstruction in designing thermally stable and promiscuous enzymes (69–71). It is important to note that the high promiscuity of ancestral enzymes is in turn compatible with the low complexity of the ancestral species. Organisms with lower complexity have relatively smaller genomes and fewer kinds of enzymes (72). Despite a small genome size, promiscuous enzymes help these species achieve a variety of life activities. Conversely, larger genomes can encode more proteins capable of performing highly specific functions and coping with more complex cellular environments. For example, from prokaryote to eukaryote, there will be more complex cellular environments, and from unicellular to multicellular organisms, the cells will be diversified due to cell differentiation. The diversification and specification of proteins enable them to function in a more complex and diverse cellular environment. Consequently, complex organisms can perform their biological functions more effectively, as they also acquire the plasticity to adapt to complex and diverse external environments. Besides, previous research had shown that the poorly packed and disordered proteins might exhibit high evolvability (73–75). Therefore, our statistical results also suggest that as the species evolves, the evolvability of proteins is also increasing.

Furthermore, we uncover the topology and sequence bases of the evolutionary trend of proteins. Such an evolutionary trend is consistent with what has been suggested by previous studies (76–80). Besides enhancing the functional sensitivity of proteins, such evolutionary trends can also prevent the dysfunctional aggregation of proteins (80). It is worth noting that the sequence analysis as conducted in this study does not depend on the AlphaFold-predicted structures. Therefore, the evolutionary trends discussed in this paper are not some kind of systematic bias due to structural prediction methods, but a robust trend that exists in nature.

In summary, based on the statistics of AlphaFold-predicted structures, we establish a new framework for studying protein evolution at the proteome level. Such a framework establishes the connections between protein evolution and species evolution, and contributes to our understanding of the correspondence between dynamics and evolution of biological systems. The statistical results show that as species evolve toward higher complexity, the proteins evolve toward higher flexibility and structural diversity. A decreasing fractal dimension and increasing functional specificity are also observed in the protein evolution. The topology and sequence bases of these evolutionary trends are also elucidated. The residue contact networks of the proteins are evolving toward higher assortativity, and the hydrophobic and hydrophilic amino acid residues are evolving to have increasing segregations in the sequences. These evolutionary trends contribute to our understanding of the origin and evolution of lives, and shed light on the sequence- and structure-based design of proteins. In the future, the proteome analysis based on AI-predicted protein structures, integrated with other kinds of bioinformation such as protein-protein interactions (81, 82), expression levels (83), evolutionary speed (84), etc., will definitely offer us new insight into the behaviors and evolution of cells and organisms.

Materials and Methods

Dataset. All the protein structures used in this study can be downloaded from Alphafold Protein Structure Database (21). A detailed description of the dataset is listed in SI Table S1.

Residue contact network. We construct the residue contact networks based on the native structure of the proteins. The residues, which are represented as their C_α atoms, are modeled as the nodes in the network. Our calculations of the network modularity and assortativity are based on the residue contact networks corresponding to the protein molecules. When the mutual distance between two residues (nodes) is smaller than the cutoff distance r_C, then the two residues are considered to be connected with an edge. In this work, we take the cutoff distance $r_C = 8$ Å.

Elastic network model (ENM). By modeling the edges in the residue contact networks as linear springs, the elastic network model can capture the equilibrium fluctuations of proteins around the native conformation. These fluctuations are closely related to the functional dynamics of the proteins. In this work, our discussions are based on the Gaussian network model (GNM) , the simplest form of the ENM, where the residue fluctuations are assumed to be Gaussian variables distributed around the equilibrium coordinates. The dynamics predicted by GNM can be well matched to experimental or simulation results (25, 26). With GNM, the potential energy of a protein with chain length N is given as: $V_{\mathrm{GNM}} = \frac{\kappa}{2}\sum_{i,j=1}^{N} \Delta r_i \Gamma_{ij} \Delta r_j$, in which κ is a uniform force constant; Δr_i and Δr_j are the displacement of residues i and j, respectively; and Γ_{ij} is the element of Kirchhoff matrix Γ. For residues i and j ($i \neq j$), if their mutual distance $r_{ij} \leq r_C$, then $\Gamma_{ij} = -1$; if $r_{ij} > r_C$, then $\Gamma_{ij} = 0$; and for the diagonal elements, $\Gamma_{ii} = -k_i = -\sum_{j \neq i} \Gamma_{ij}$, where k_i denote the degree of node i. In the calculation, we take $r_C = 8$ Å. Our results are robust to the selection of the model parameter r_C (see SI Fig. S5).

Normal mode analysis. The eigendecomposition of matrix Γ gives the eigenvalues and the corresponding eigenvectors related to the motions of normal modes (28–30). To compare the eigenvalues for the proteins with different chain lengths, in this work, the diagonal elements of matrices Γ are normalized as 1. Such a normalized matrix is also known as the symmetric normalized graph Laplacian (43): $L = D^{-1/2}\Gamma D^{-1/2}$, where D is a diagonal matrix $D = \mathrm{diag}[k_1, k_2, \cdots k_N]$ describing the local packing density of the residues. Diagonalizing matrix L, then we have $L = U\Lambda U^T$, in which matrix $\Lambda = \mathrm{diag}[\lambda_0, \lambda_1, \lambda_2, \cdots \lambda_{N-1}]$ ($0 = \lambda_0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_{N-1}$) represents the eigenvalues, and matrix $U = [u_0, u_1, u_2, \cdots, u_{N-1}]^T$ denotes to the eigenvectors. The zeroth eigenvector u_0 corresponds to the translational and rotational motions of the proteins, other nonzero modes correspond to the vibrational motions. The nonzero eigenvalue λ_i is proportional to the square of the vibrational frequency ω_i, i.e., $\lambda_i \sim \omega_i^2$. According to the equipartition of energy, the vibration amplitude A_i is inversely proportional to the frequency: $A_i \sim 1/\omega_i$. Thus, for two nonzero modes i and j, there is $A_i^2/A_j^2 = \lambda_j/\lambda_i$. In this work, we use the Python package ProDy to calculate the normal modes of the proteins (85).

Power-law fitting and the Zipf's coefficient. To fit the Zipf's law $\lambda_i^{-1} \sim i^{-z}$ and obtain the coefficient z, in the calculation, we select the top 25% of the eigenvalues to perform the power-law fitting (86). Note that the observed evolutionary trend as indicated by Zipf's coefficient z are robust to the details related to the normal mode calculation and power law fitting (see SI Fig. S5).

Local Density Model. The local density model (44) is a simplified version of the elastic network model. In the model, the proteins are modeled as a set of *unconnected* residues oscillating around the native position. The spring constant γ_i related to residue i is assumed to be proportional to its local packing density (the residue's degree k_i), that is: $\gamma_i \sim k_i$. Intuitively, if a residue only has few contacts, then its vibration is slow; if it has many contacts with other residues, its motions are highly constrained, so the vibration will fast. In this model, since the residues' motions are independent, the normal modes can directly correspond to the vibrations of the residues. Thus, the square vibration frequency of residue i is proportional to the local packing density: $\widehat{\omega_i}^2 \sim \gamma_i \sim k_i$. Note that $\widehat{\omega_i}$ (i denotes the residue index) is introduced to distinguish from ω_i (i denotes the mode index).

Modularity. Modularity is a topological descriptor that quantifies if a network can be easily divided into modules. It is defined as the fraction of the edges that fall within the given module minus the expected fraction when edges were distributed at random (87, 88). According to this definition, for a network with N nodes and M edges described by the adjacency matrix A (in which $A_{ij} = 1$ if and only if node i and j are connected), the modularity Q can be calculated as: $Q = Tr(x^T \cdot B \cdot x)/4M$, in which matrix B is given as $B_{ij} = A_{ij} - k_i k_j/2M$, and vector x is the column vector describing the network partition. For any given partition of a network, the corresponding modularity Q can be calculated, and its value lies between -1 and 1. The appropriate partition of a network would maximize the modularity Q. In this work, we employ the Louvain method (89) to partition the network and maximize the value modularity Q.

Hydropathy variation. In the calculation, we first obtain the original hydropathy data of a sequence according to the Zimmerman hydrophobicity scale of amino acid residues (90). Then, a moving average with window length $l_w = 7$ is introduced to calculate the hydropathy profile. If there are no significant hydrophobicity segregations in a sequence, the moving average will smooth out the differences between hydrophobic and hydrophilic amino acids, which will lead to little variation in the filtered hydropathy profile. In contrast, when there are significant hydrophobicity segregations in a sequence, there will be large variations in the hydropathy profile. Thus, we introduce the coefficient of variation $\mathrm{CV_{HP}}$ (defined as the standard deviation divided by the mean value of the filtered hydropathy profile) to quantify the sequential segregation of hydrophobic and hydrophilic residues.

Acknowledgments

We gratefully thank Taro Toyoizumi, Haobo Wang, Hisao Moriya, Haiguang Liu, Zhengqi He, Weiyi Qiu, Wenfei Li, Zeke Xie, Yingnan Li, Xinhong Liu, Lei-Han Tang and Xuefei Li for participating in stimulating discussions. This research was supported by Brain/MINDS from AMED under Grant no. JP21dm0207001, National Natural Science Foundation of China (No. 11774157), 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.

References

1. C. Adami, C. Ofria, T. C. Collier, Evolution of biological complexity. *Proc. Natl. Acad. Sci. USA* **97**, 4463–4468 (2000).
2. C. Furusawa, K. Kaneko, Origin of Complexity in Multicellular Organisms. *Physical Review Letters* **84**, 6130 (2000).
3. E. V. Koonin, Y. I. Wolf, G. P. Karev, The structure of the protein universe and genome evolution. *Nature* **420**, 218–223 (2002).
4. C. Pál, B. Papp, M. J. Lercher, An integrated view of protein evolution. *Nature Reviews Genetics* **7**, 337–348 (2006).
5. K. B. Zeldovich, E. I. Shakhnovich, Understanding Protein Evolution: From Protein Physics to Darwinian Selection. *Annual Review of Physical Chemistry* **59**, 105–127 (2008).
6. T. Sikosek, H. S. Chan, Biophysics of protein evolution and evolutionary protein biophysics. *Journal of the Royal Society Interface* **11** (2014).
7. I.-G. Choi, S.-H. Kim, Evolution of protein structural classes and protein sequence families. *Proc. Natl. Acad. Sci. USA* **103**, 14506–14061 (2006).
8. J. P. Pin, T. Galvez, L. Prézeau, Evolution, structure, and activation mechanism of family 3/C G-protein-coupled receptors. *Pharmacology & Therapeutics* **98**, 325–354 (2003).
9. Y. A. Labas, *et al.*, Diversity and evolution of the green fluorescent protein family. *Proc. Natl. Acad. Sci. USA* **99**, 4256–4261 (2002).
10. R. Zardoya, Phylogeny and evolution of the major intrinsic protein family. *Biology of the Cell* **97**, 397–414 (2005).
11. G. C. Finnigan, V. Hanson-Smith, T. H. Stevens, J. W. Thornton, Evolution of increased complexity in a molecular machine. *Nature* **481**, 360–364 (2012).
12. F. Morcos, *et al.*, Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proc. Natl. Acad. Sci. USA* **108**, E1293–E1301 (2011).
13. R. Espada, R. G. Parra, T. Mora, A. M. Walczak, D. U. Ferreiro, Capturing coevolutionary signals in repeat proteins. *BMC Bioinformatics* **16**, 207 (2015).
14. A. W. Senior, *et al.*, Improved protein structure prediction using potentials from deep learning. *Nature* **577**, 706–710 (2020).
15. J. Jumper, *et al.*, Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
16. C. Bayly-Jones, J. C. Whisstock, Mining folded proteomes in the era of accurate structure prediction. *PLOS Computational Biology* **18**, e1009930 (2022).
17. A. J. Robertson, J. M. Courtney, Y. Shen, J. Ying, A. Bax, Concordance of X-ray and AlphaFold2 Models of SARS-CoV-2 Main Protease with Residual Dipolar Couplings Measured in Solution. *J Am Chem Soc* **143**, 19306–19310 (2021).

18. K. M. Ruff, R. v. Pappu, AlphaFold and Implications for Intrinsically Disordered Proteins. *Journal of Molecular Biology* **433** (2021).
19. H. Bagdonas, C. A. Fogarty, E. Fadda, J. Agirre, The case for post-predictional modifications in the AlphaFold Protein Structure Database. *Nature Structural and Molecular Biology* **28**, 869–870 (2021).
20. J. Jumper, *et al.*, Applying and improving AlphaFold at CASP14. *Proteins: Structure, Function and Bioinformatics* **89**, 1711–1721 (2021).
21. M. Varadi, *et al.*, AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research* **50**, D439–D444 (2022).
22. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
23. R. P. Joosten, *et al.*, A series of PDB related databases for everyday needs. *Nucleic Acids Research* **39**, D411 (2011).
24. M. Hemery, O. Rivoire, Evolution of sparsity and modularity in a model of protein allostery. *Physical Review E* **91**, 042704 (2015).
25. I. Bahar, T. R. Lezon, L. W. Yang, E. Eyal, Global dynamics of proteins: Bridging between structure and function. *Annual Review of Biophysics* **39**, 23–42 (2010).
26. T. Haliloglu, I. Bahar, B. Erman, Gaussian Dynamics of Folded Proteins. *Physical Review Letters* **79**, 3090–3093 (1997).
27. I. Bahar, A. R. Atilgan, M. C. Demirel, B. Erman, Vibrational Dynamics of Folded Proteins: Significance of Slow and Fast Motions in Relation to Function and Stability. *Physical Review Letters* **80**, 2733 (1998).
28. D. A. Case, Normal mode analysis of protein dynamics. *Current Opinion in Structural Biology* **4**, 285–290 (1994).
29. S. Hayward, A. Kitao, N. Gō, Harmonicity and Anharmonicity in Protein Dynamics: A Normal Mode Analysis and Principal Component Analysis. *Proteins: Structure, Function, and Genetics* **23**, 177–186 (1995).
30. H. Wako, S. Endo, Normal mode analysis as a method to derive protein dynamics information from the Protein Data Bank. *Biophysical Reviews* **9**, 877–893 (2017).
31. H. M. Berman, *et al.*, The Protein Data Bank. *Nucleic Acids Research* **28**, 235–242 (2000).
32. Q.-Y. Tang, Y.-Y. Zhang, J. Wang, W. Wang, D. R. Chialvo, Critical Fluctuations in the Native State of Proteins. *Physical Review Letters* **118** (2017).
33. Q.-Y. Tang, K. Kaneko, Long-range correlation in protein dynamics: Confirmation by structural data and normal mode analysis. *PLoS Computational Biology* **16** (2020).
34. R. Guimerà, M. Sales-Pardo, L. A. N. Amaral, Modularity from fluctuations in random graphs and complex networks. *Physical Review E* **70**, 4 (2004).
35. J. M. Friedman, Structure, Dynamics, and Reactivity in Hemoglobin. *Science.* **228**, 1273–1280 (1985).
36. T. Haliloglu, I. Bahar, Adaptability of protein structures to enable functional interactions and evolutionary implications. *Current Opinion in Structural Biology* **35**, 17–23 (2015).
37. P. K. Agarwal, Enzymes: An integrated view of structure, dynamics and function. *Microbial Cell Factories* **5**, 1–12 (2006).
38. Y. Togashi, A. S. Mikhailov, G. Ertl, Nonlinear relaxation dynamics in elastic networks and design principles of molecular machines. *Proc. Natl. Acad. Sci. USA* **104**, 8697–8702 (2007).
39. S. Reuveni, R. Granek, J. Klafter, Proteins: Coexistence of stability and flexibility. *Physical Review Letters* **100** (2008).
40. Q.-Y. Tang, T. S. Hatakeyama, K. Kaneko, Functional sensitivity and mutational robustness of proteins. *Physical Review Research* **2**, 033452 (2020).
41. Q.-Y. Tang, K. Kaneko, Dynamics-Evolution Correspondence in Protein Structures. *Physical Review Letters* **127** (2021).

42. Z. Xie, Q.-Y. Tang, Y. Cai, M. Sun, P. Li, On the Power-Law Spectrum in Deep Learning: A Bridge to Protein Science. *arxiv: 2201.13011* (2022).
43. C. Atilgan, O. B. Okan, A. R. Atilgan, Network-based models as tools hinting at nonevident protein functionality. *Annual Review of Biophysics* **41**, 205–225 (2012).
44. B. Halle, Flexibility and packing in proteins. *Proc. Natl. Acad. Sci. USA* **99**, 1274–1279 (2002).
45. K. A. Dill, Dominant forces in protein folding. *Biochemistry* **29**, 7133–7155 (1990).
46. K. A. Dill, J. L. MacCallum, The protein-folding problem, 50 years on. *Science.* **338**, 1042–1046 (2012).
47. M. E. J. Newman, Assortative Mixing in Networks. *Physical Review Letters* **89**, 208701 (2002).
48. M. E. J. Newman, Mixing patterns in networks. *Physical Review E* **67**, 026126 (2003).
49. W. Basile, M. Salvatore, C. Bassot, A. Elofsson, Why do eukaryotic proteins contain more intrinsically disordered regions? *PLOS Computational Biology* **15**, e1007186 (2019).
50. I. N. Berezovsky, E. I. Shakhnovich, Physics and evolution of thermophilic adaptation. *Proc. Natl. Acad. Sci. USA* **102**, 12742–12747 (2005).
51. W. J. Jones, J. A. Leigh, F. Mayer, C. R. Woese, R. S. Wolfe, Methanococcus jannaschii sp. nov., an extremely thermophilic methanogen from a submarine hydrothermal vent. *Arch. Microbiol.* **136**, 254–261 (1983).
52. C. J. Bult, *et al.*, Complete Genome Sequence of the Methanogenic Archaeon, Methanococcus jannaschii. *Science.* **273**, 1058–1073 (1996).
53. K. Ghosh, K. Dill, Cellular proteomes have broad distributions of protein stability. *Biophysical Journal* **99**, 3996–4002 (2010).
54. V. Nguyen, *et al.*, Evolutionary drivers of thermoadaptation in enzyme catalysis. *Science.* **355**, 289–294 (2017).
55. J. N. Onuchic, Z. Luthey-Schulten, P. G. Wolynes, Theory of protein folding: the energy landscape perspective. *Annu Rev Phys Chem* **48**, 545–600 (1997).
56. W. Li, P. G. Wolynes, S. Takada, Frustration, specific sequence dependence, and nonlinearity in large-amplitude fluctuations of allosteric proteins. *Proc. Natl. Acad. Sci. USA* **108**, 3504–3509 (2011).
57. C. Furusawa, K. Kaneko, Formation of dominant mode by evolution in biological systems. *Physical Review E* **97**, 42410 (2018).
58. T. U. Sato, K. Kaneko, Evolutionary dimension reduction in phenotypic space. *Physical Review Research* **2** (2020).
59. A. Sakata, K. Kaneko, Dimensional Reduction in Evolving Spin-Glass Model: Correlation of Phenotypic Responses to Environmental and Mutational Changes. *Physical Review Letters* **124** (2020).
60. S. Dutta, J. P. Eckmann, A. Libchaber, T. Tlusty, Green function of correlated genes in a minimal mechanical model of protein evolution. *Proc. Natl. Acad. Sci. USA* **115**, E4559–E4568 (2018).
61. J. P. Eckmann, J. Rougemont, T. Tlusty, Colloquium: Proteins: the physics of amorphous evolving matter. *Reviews of Modern Physics* **91**, 031001 (2019).
62. J. P. Eckmann, T. Tlusty, Dimensional reduction in complex living systems: Where, why, and how. *BioEssays* **43**, 2100062 (2021).
63. M. Soskine, D. S. Tawfik, Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics* **11**, 572–582 (2010).
64. K. Takano, A. Aoi, Y. Koga, S. Kanaya, Evolvability of thermophilic proteins from archaea and bacteria. *Biochemistry* **52**, 4774–4780 (2013).
65. T. L. O'Loughlin, W. M. Patrick, I. Matsumura, Natural history as a predictor of protein evolvability. *Protein Engineering, Design and Selection* **19**, 439–442 (2006).
66. J. M. Gardner, M. Biler, V. A. Risso, J. M. Sanchez-Ruiz, S. C. L. Kamerlin, Manipulating Conformational Dynamics to Repurpose Ancient Proteins for Modern Catalytic Functions. *ACS Catalysis* **10**, 4863–4870 (2020).
67. D. Petrovic, V. A. Risso, S. C. L. Kamerlin, J. M. Sanchez-Ruiz, Conformational dynamics and enzyme evolution. *Journal of the Royal Society Interface* **15** (2018).

68. O. Khersonsky, D. S. Tawfik, Enzyme promiscuity: A mechanistic and evolutionary perspective. *Annual Review of Biochemistry* **79**, 471–505 (2010).
69. G. P. Pinto, M. Corbella, A. O. Demkiv, S. C. L. Kamerlin, Exploiting enzyme evolution for computational protein design. *Trends in Biochemical Sciences* **47**, 375–389 (2022).
70. L. C. Wheeler, S. A. Lim, S. Marqusee, M. J. Harms, The thermostability and specificity of ancient proteins. *Current Opinion in Structural Biology* **38**, 37–43 (2016).
71. D. L. Trudeau, D. S. Tawfik, Protein engineers turned evolutionists—the quest for the optimal starting point. *Current Opinion in Biotechnology* **60**, 46–52 (2019).
72. D. C. Gagler, *et al.*, Scaling laws in enzyme function reveal a new kind of biochemical universality. *Proc. Natl. Acad. Sci. USA* **119** (2022).
73. N. Tokuriki, F. Stricher, L. Serrano, D. S. Tawfik, How protein stability and new functions trade off. *PLoS Computational Biology* **4**, e1000002 (2008).
74. N. Tokuriki, D. S. Tawfik, Protein Dynamism and Evolvability. *Science (1979)* **324**, 203–207 (2009).
75. S. Meier, S. Özbek, A biological cosmos of parallel universes: Does protein structural plasticity facilitate evolution? *BioEssays* **29**, 1095–1104 (2007).
76. M. A. Moret, G. F. Zebende, J. C. Phillips, Hydropathic wave ordering of alpha crystallin—Membrane interactions enhances human lens transparency and resists cataracts. *Physica A: Statistical Mechanics and its Applications* **514**, 573–579 (2019).
77. J. C. Phillips, Hydropathic Self-Organized Criticality: A Magic Wand for Protein Physics. *Protein & Peptide Letters* **19**, 1089–1093 (2012).
78. J. C. Phillips, Scaling and self-organized criticality in proteins II. *Proc. Natl. Acad. Sci. USA* **106**, 3113–3118 (2009).
79. J. C. Phillips, Self-organized networks: Darwinian evolution of dynein rings, stalks, and stalk heads. *Proc. Natl. Acad. Sci. USA* **117**, 7799–7802 (2020).
80. S. G. Foy, B. A. Wilson, J. Bertram, M. H. J. Cordes, J. Masel, A Shift in Aggregation Avoidance Strategy Marks a Long-Term Direction to Protein Evolution. *Genetics* **211**, 1345 (2019).
81. S. Maslov, K. Sneppen, Specificity and stability in topology of protein networks. *Science.* **296**, 910–913 (2002).
82. J. Zhang, S. Maslov, E. I. Shakhnovich, Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. *Molecular Systems Biology* **4**, 210 (2008).
83. D. A. Drummond, J. D. Bloom, C. Adami, C. O. Wilke, F. H. Arnold, Why highly expressed proteins evolve slowly. *Proc. Natl. Acad. Sci. USA* **102**, 14338–14343 (2005).
84. L. Agozzino, K. A. Dill, Protein evolution speed depends on its stability and abundance and on chaperone concentrations. *Proc. Natl. Acad. Sci. USA* **115**, 9092–9097 (2018).
85. A. Bakan, L. M. Meireles, I. Bahar, ProDy: Protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).
86. J. Alstott, E. Bullmore, D. Plenz, Powerlaw: A python package for analysis of heavy-tailed distributions. *PLoS ONE* **9**, e95816 (2014).
87. M. E. J. Newman, Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **103**, 8577–8582 (2006).
88. M. E. J. Newman, Fast algorithm for detecting community structure in networks. *Physical Review E* **69**, 5 (2004).
89. V. D. Blondel, J. L. Guillaume, R. Lambiotte, E. Lefebvre, Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).
90. J. M. Zimmerman, N. Eliezer, R. Simha, The characterization of amino acid sequences in proteins by statistical methods. *Journal of Theoretical Biology* **21**, 170–201 (1968).

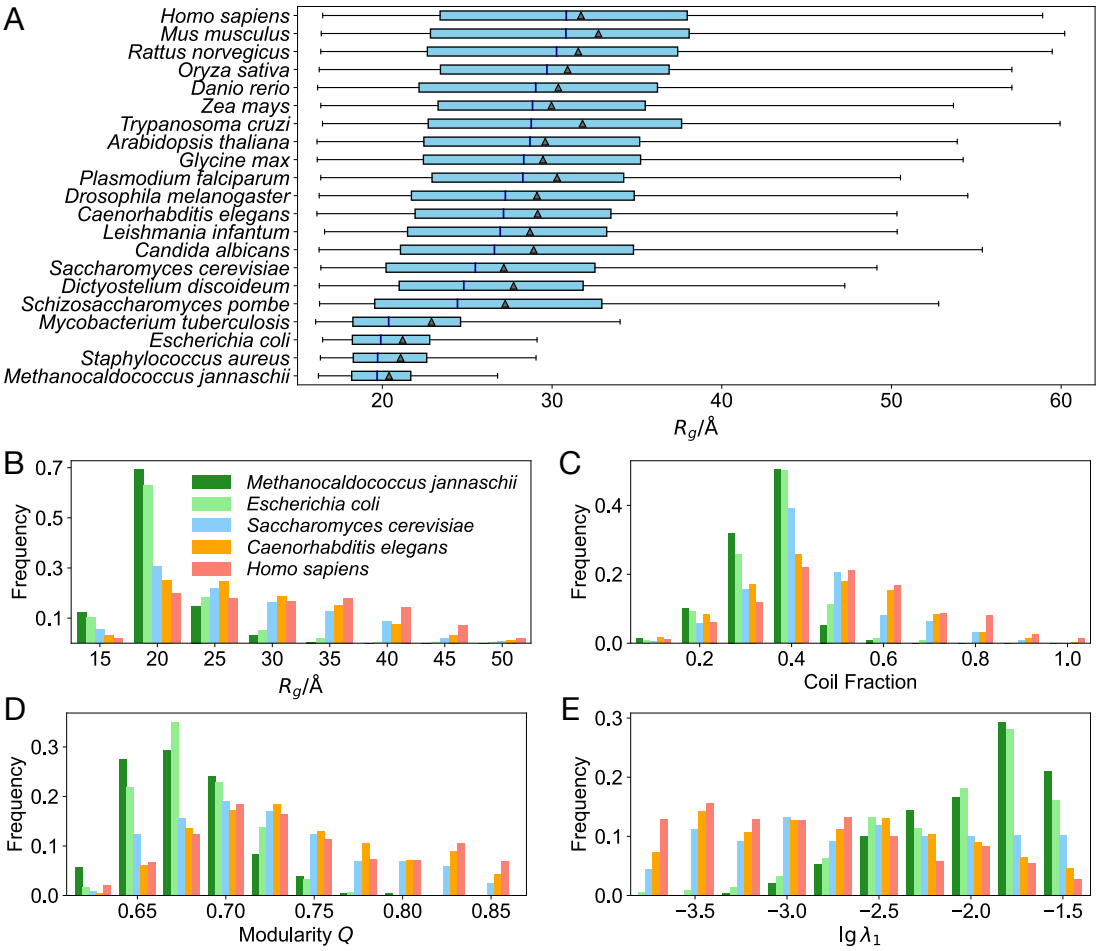

Figure 1. The statistics of AlphaFold-predicted structures indicate that as species evolve, the average flexibility of proteins gradually increases. (A) For the proteins from 21 different species and with similar chain lengths N ($225 \leq N < 275$), the distributions of the radii of gyration R_g are shown as the box-and-whiskers. Here, the triangle and the vertical bar in the box denote the mean value and the median of the R_g, respectively. For proteins from the five selected species with similar chain length N, the histogram of the (B) radii of gyration R_g, (C) coil fraction, (D) modularity Q of the residue contact network, and (E) the slowest mode eigenvalue λ_1 at the logarithmic scale.

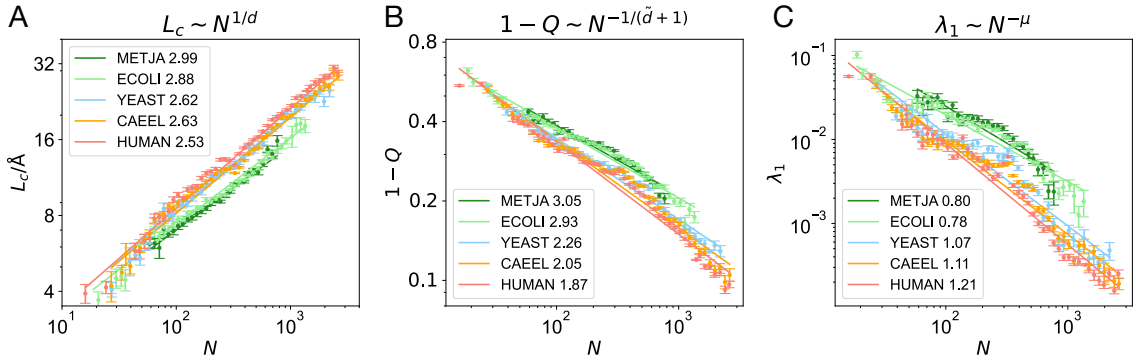

Figure 2. The scaling analyses suggest a decreasing fractal dimension of the proteins in species evolution. For the proteins from five selected species, the size (chain length N) dependence of (A) length of the longest semi-axis L_c, (B) modularity Q, and (C) the slowest mode eigenvalue λ_1. The fitted scaling coefficients d, $\tilde{d}$, and μ are listed in the legends of the subplots (A), (B), and (C), respectively. These scaling coefficients are obtained by robust fittings with Theil–Sen estimators.

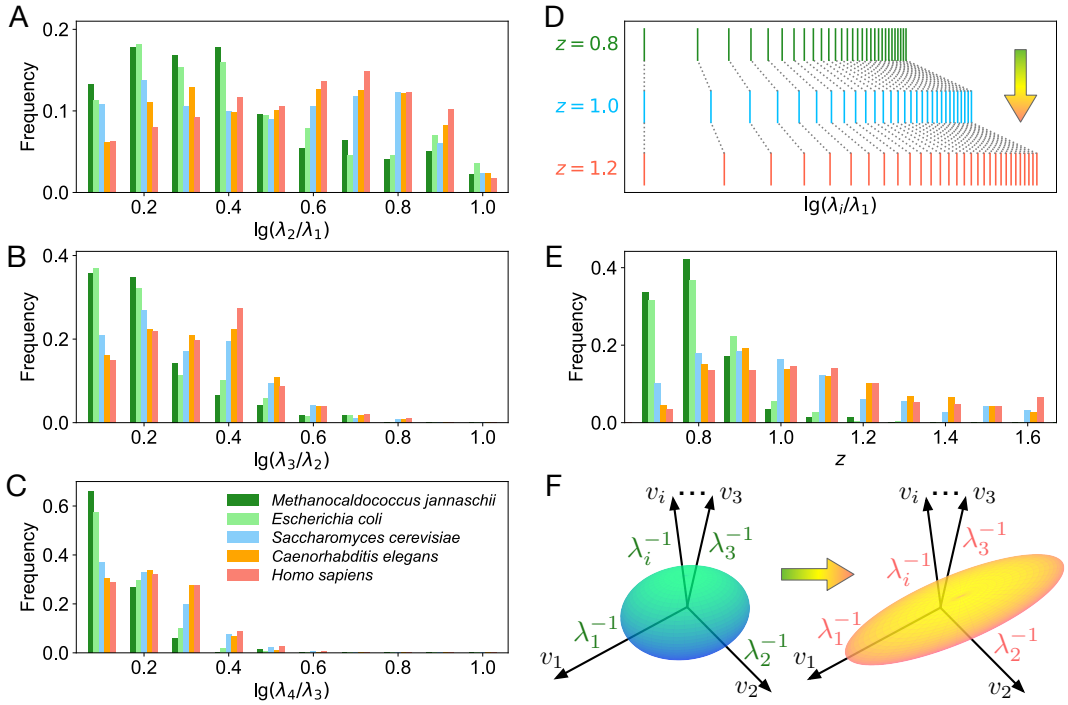

Figure 3. The increasing eigengaps in the vibration spectra imply an evolutionary trend toward higher functional specificity and lower dimensionality in protein dynamics. For proteins with similar chain lengths N ($225 \leq N < 275$) in the five selected species, the distribution of the eigengaps (A) λ_2/λ_1, (B), λ_3/λ_2 and (C) λ_4/λ_3 at the logarithmic scale. (D) Illustration of power-law spectra at the logarithmic scale with Zipf's coefficients $z = 0.8$, 1.0, and 1.2. (E) For proteins with similar chain lengths in different species, the distribution of the Zipf's coefficients z. (F) Illustration of the evolutionary trend of dimensional reduction in the conformational spaces. The arrows in subplots (D) and (F) indicate the statistical direction related to the species' evolution.

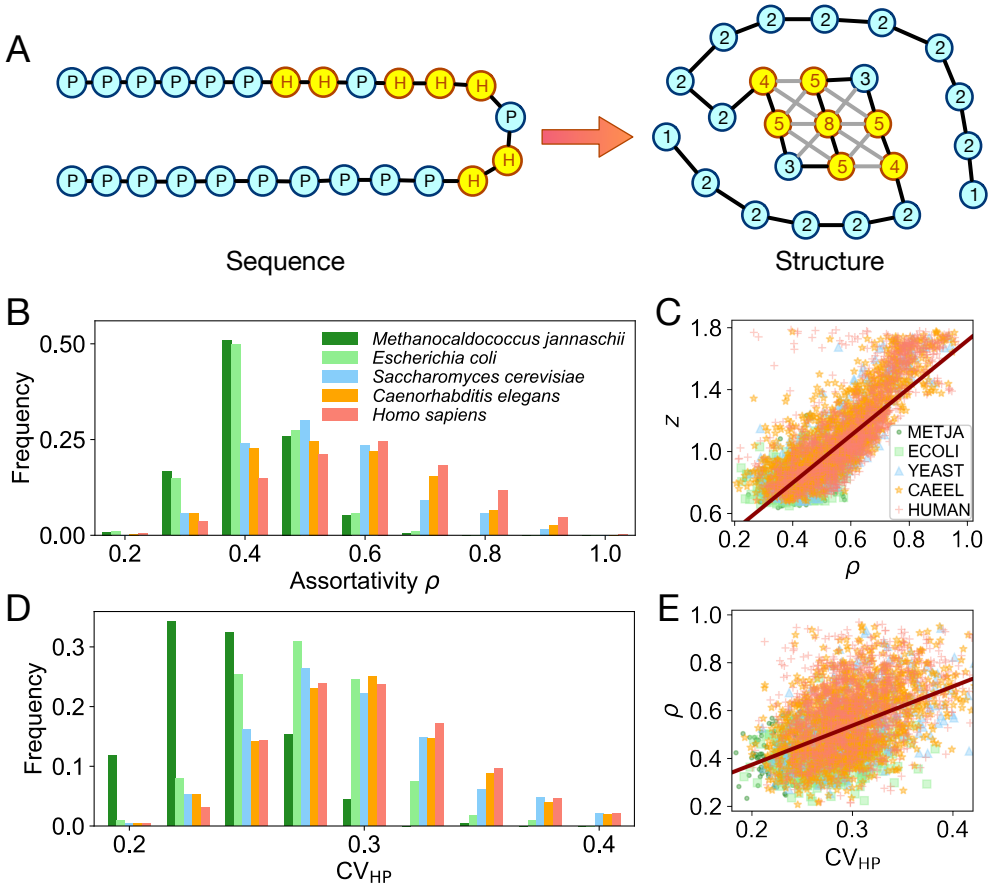

Figure 4. Evolutionary trends of the residue contact networks and the corresponding protein sequences. (A) Illustration of a protein sequence (left) folds into a native structure described as an assortative residue contact network (right). The yellow and cyan nodes denote the hydrophobic (H) and polar (P) amino acid residues, respectively. When the protein folds into the native structure, the hydrophobic residues tend to aggregate into a densely connected hydrophobic core. In the subplot (right), the numbers represent the node's degree (number of connections with other nodes). (B) For proteins with similar chain length N ($225 \leq N < 275$) in the five selected species, the histogram of residue contact network assortativity ρ. (C) The scattering plot and the fitted trend line of the assortativity ρ vs. Zipf's coefficient z. (D) For proteins with similar chain lengths N in different species, the histogram of the hydropathy variations $\mathrm{CV_{HP}}$. (E) The scattering plot and the fitted trend line of hydropathy variation $\mathrm{CV_{HP}}$ vs. residue contact network assortativity ρ.

Supplementary Information Text

Methods

Confidence metric of structure prediction – pLDDT. AlphaFold produces a per-residue confidence metric called the predicted local distance difference test (pLDDT) on a scale from 0 to 100. Usually, residues with pLDDT ≤ 50 correspond to very low confidence, residues with $50 <$ pLDDT ≤ 70 have low confidence, and residues with $70 <$ pLDDT ≤ 90 correspond to a generally correct backbone prediction, and residues with pLDDT > 90 are expected to be modeled to high accuracy. Although the residues with pLDDT < 50 are predicted with low confidence, as mentioned in Ref. (15) of the Main Text, low pLDDT values can work as reasonably strong predictors of disorder regions in the proteins, or say, those proteins with low average residue pLDDT usually have long disordered regions. In the Main Text, to show the evolutionary trend related to the increasing flexibility of the proteins, our statistics also take into account the AlphaFold-predicted structures with low average pLDDT values. It is worth noting that, even if we filter the proteins with low average pLDDT, as shown in Fig. S2, we can find that the statistical evolutionary trend is kept.

Kolmogorov-Smirnov (KS) test. The KS test is a non-parameterized test which is usually applied to compare two samples and test whether the two samples come from the same distribution. In the KS test, the KS statistic D is defined as the maximum difference (i.e., the vertical distance) between the two cumulative distribution functions (illustrated in Fig. S3D), and the p-value is a measure of the probability that such observed difference could have occurred just by random chance. In comparing two distributions, a larger KS statistic D and smaller p-values indicate more significant differences between them.

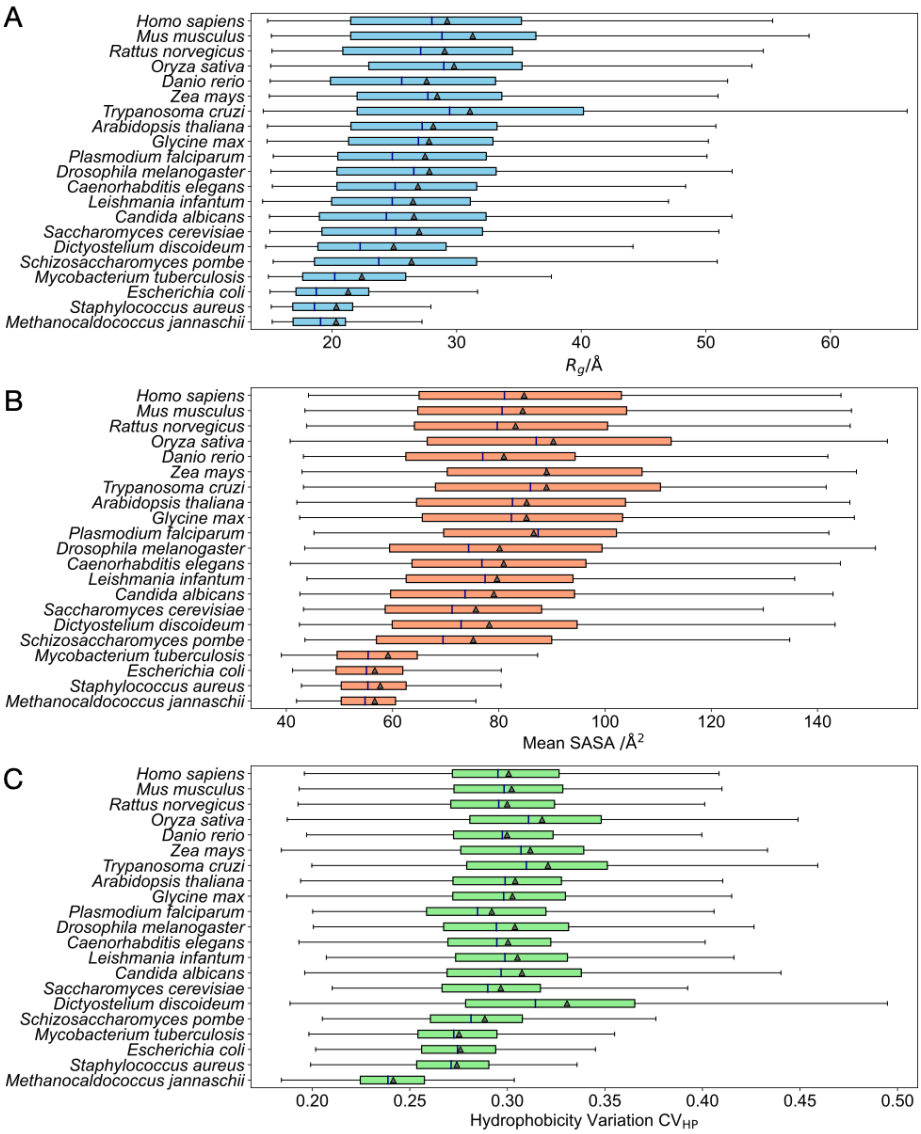

Fig. S1. The observed evolutionary trends of the proteins are robust to the selection of chain length or indicators. (A) For the 21 different species, when selecting the proteins with some other chain lengths ($175 \leq N < 225$), the distributions of the radii of gyration R_g are shown as the box-and-whiskers. To compare with the results shown in the Main Text, here in all the subplots, the orderings of the species are based on Fig. 1A. The result shows that the observed trends of the evolution are robust to the selection of the protein sizes. For the proteins with similar chain lengths ($225 \leq N < 275$) from 21 different species, (B) the distributions of the solvent-accessible surface area (SASA) per residue, and (C) the distributions of hydrophobicity variation CV_{HP} are shown as the box-and-whiskers. These results show that the observed evolutionary trends are robust to the selection of structure- or sequence-related indicators.

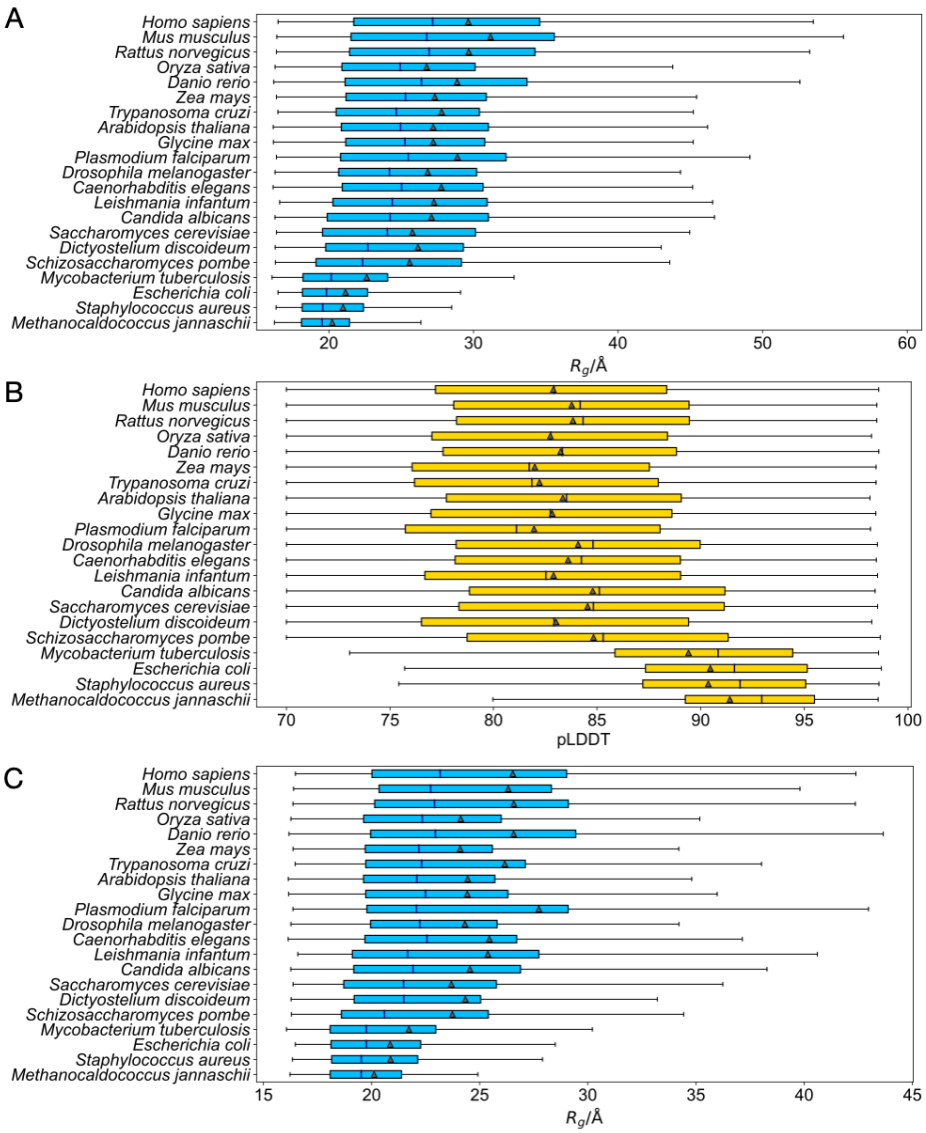

Figure S2. The observed evolutionary trends of the proteins are robust to the selection of confidence threshold of structure prediction. (A) When we filter the AlphaFold-predicted structures with average residue pLDDT ≤ 70, for the proteins with similar chain lengths ($175 \leq N < 225$), the distributions of R_g are shown as the box-and-whiskers. The result shows that the evolutionary trend toward higher flexibility can still be observed. To compare with the results shown in the Main Text, here in all the subplots, the orderings of the species are also based on Fig. 1A. (B) For the proteins in subplot (A), the distribution of their average pLDDT. It is observed that for the proteins at the same size, compared with prokaryotic proteins, the eukaryotic proteins have a lower average pLDDT. Such a result also indicates higher flexibility is accompanied by the species evolution. (C) For the proteins with similar chain lengths ($175 \leq N < 225$), when we only select those AlphaFold-predicted structures with 75% of their residues having pLDDT > 70, the distribution of R_g still indicates the evolutionary trend toward higher flexibility. These results prove that the evolution toward flexibility also holds for proteins with relatively stable native structures, strengthening our conclusions in the Main Text.

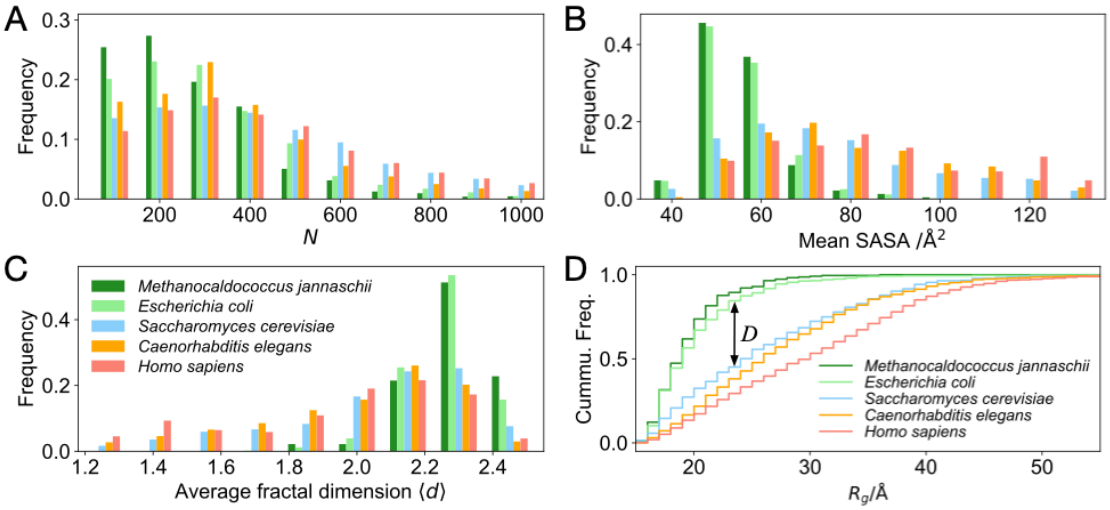

Figure S3. Some other statistics of AlphaFold-predicted structures. (A) For the five selected species, the statistics of the proteins' chain lengths. Note that the number of proteins with chain lengths N around 200 to 300 is relatively high for all the selected species. Hence, in the main text, we mainly focus on the proteins with chain lengths $225 \leq N < 275$. Besides, this subplot also shows that as species evolve, the average chain length of proteins also increases, which is another sign of the increasing complexity of the species. For proteins from the five selected species with similar chain length N $(225 \leq N < 275)$, the histogram of the (B) SASA per residue (C) average fractal dimension $\langle d \rangle$ also shows the evolutionary trends of the proteins. Here, to estimate the fractal dimension of a protein, we first calculate the fractal dimension of every residue based on the box-counting method (counting how the number of neighboring residues increases as the cutoff distance increases) and then calculate the average over all the residues. Note that there are many residues at the surface of the protein, so that this estimation will give a relatively lower value of the fractal dimension. Note that the results in subplots (B) and (C) are consistent with the results shown in Figs. 1 and 2 of the Main Text. These results also show that as the species evolves toward higher complexity, for the proteins with similar chain lengths, there is an evolutionary trend toward higher flexibility and lower fractal dimension. (D) For proteins from the five selected species with similar chain length N $(225 \leq N < 275)$, the cumulative distribution functions of the radii of gyration R_g. In this subplot, the distance D (denoted by the two-way arrow) correspond to the Kolmogorov-Smirnov (KS) statistics quantifying the distance between the distributions of the *E. coli.* and yeast (see the underlined entry in Table S2A).

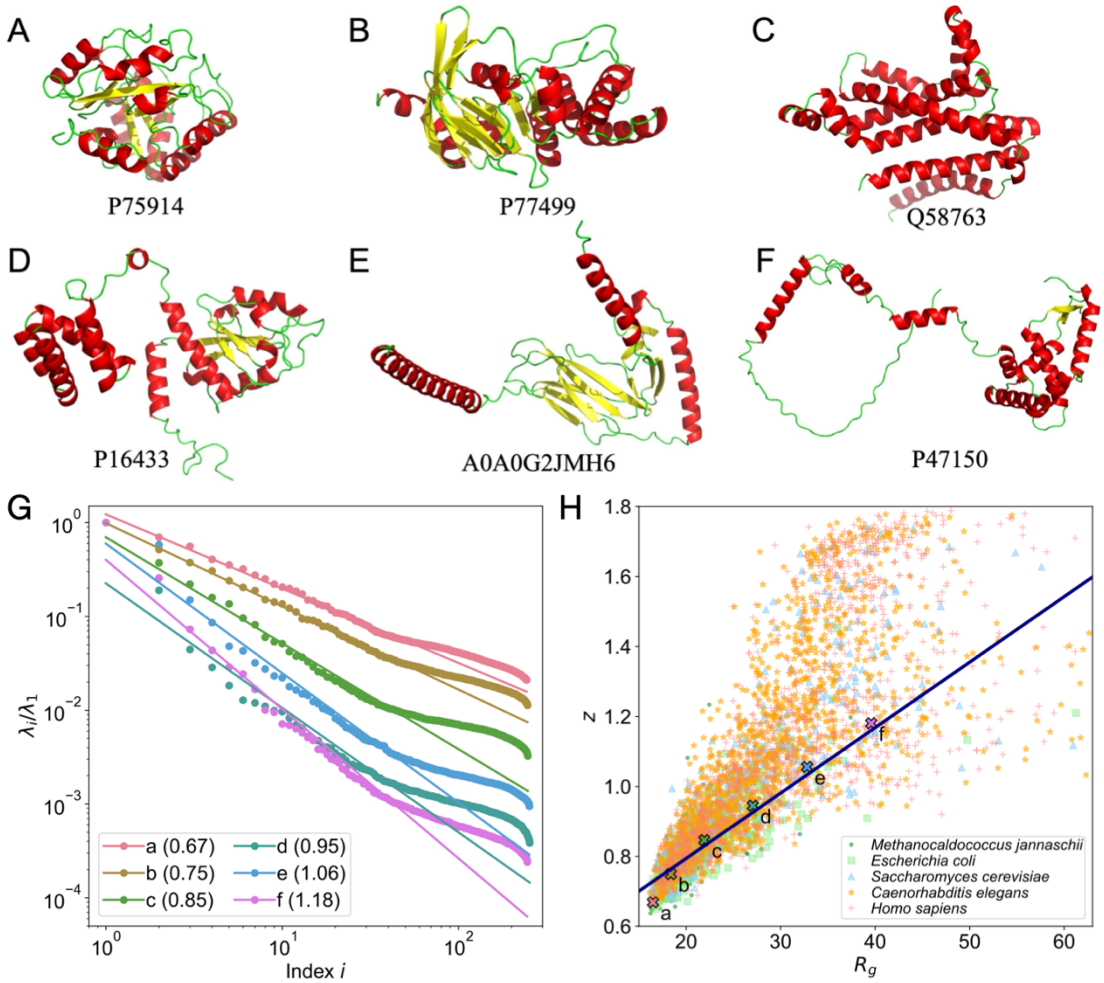

Figure S4. The proteins with similar chain lengths may have different structures and corresponding vibration spectra. (A-F) The cartoon illustrations of the six selected proteins in our dataset. Their UniProt codes are also listed in the figure. (G) For the proteins illustrated above, the rank-size distributions of the vibration spectra and the power-law fitting results are shown. The fitted Zipf's coefficient *z's* are listed in the legend. Here, the fittings are based on the top 25% of the eigenvalues. (H) For the proteins with similar chain lengths ($225 \leq N < 275$) from the five selected species, the scattering plot of their z vs. R_g. The trend line in navy shows that the Zipf's coefficient z is positively correlated with R_g. The crosses in the subplot mark the selected proteins illustrated above.

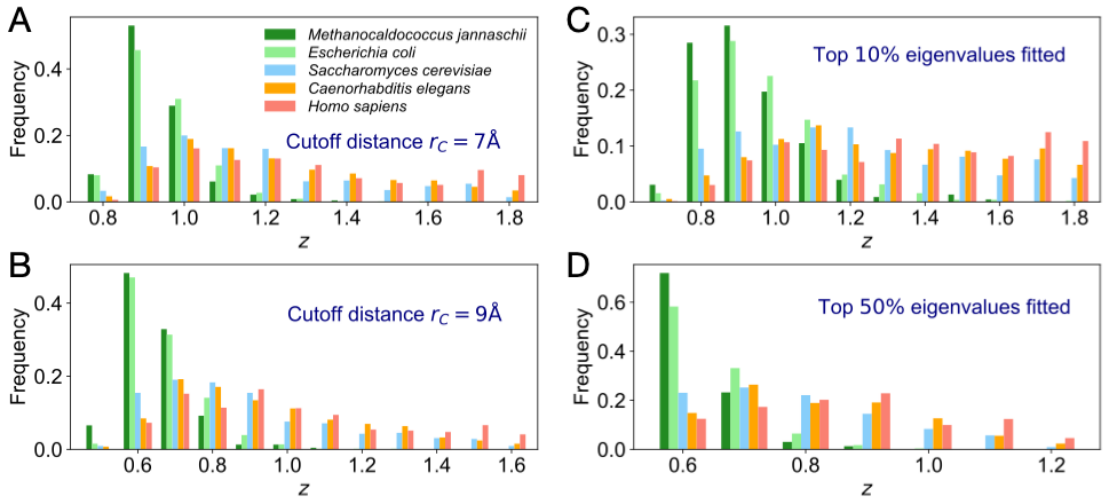

Figure S5. The observed evolutionary trend as indicated by Zipf's coefficient is robust to the fitting details. (A-D) With different calculation or fitting details, the obtained distributions of Zipf's exponent z for proteins with similar chain lengths in the five selected species. In the normal mode analysis based on the Gaussian network model, the selection of the cutoff distance (A) $r_C = 7$ Å and (B) $r_C = 9$ Å may lead to the variations in the values of the fitted z, but the observed evolutionary trends are robust to the selection of r_C. With $r_C = 8$ Å, in fitting the rank-size distributions of the vibration spectra, the selection of top (C) 10% and (D) 50% also may lead to variations in fitted z, but the observed evolutionary trend is robust to the selected number of eigenvalues.

Table S1. Detailed information of the dataset used in this study. The species names, codes, reference proteome, and the number of the protein structures are listed.

Species	**Common Name**	**Species Code**	**Reference Proteome**	**#Structures**
Arabidopsis thaliana	*Arabidopsis*	ARATH	UP000006548	27,434
Caenorhabditis elegans	Nematode worm	CAEEL	UP000001940	19,694
Candida albicans	*C. albicans*	CANAL	UP000000559	5,974
Danio rerio	Zebrafish	DANRE	UP000000437	24,664
Dictyostelium discoideum	*Dictyostelium*	DICDI	UP000002195	12,622
Drosophila melanogaster	Fruit fly	DROME	UP000000803	13,458
Escherichia coli	*E. coli*	ECOLI	UP000000625	4,363
Glycine max	Soybean	SOYBN	UP000008827	55,799
Homo sapiens	Human	HUMAN	UP000005640	23,391
Leishmania infantum	*L. infantum*	LEIIN	UP000008153	7,924
Methanocaldococcus jannaschii	*M. jannaschii*	METJA	UP000000805	1,773
Mus musculus	Mouse	MOUSE	UP000000589	21,615
Mycobacterium tuberculosis	*M. tuberculosis*	MYCTU	UP000001584	3,988
Oryza sativa	Asian rice	ORYSJ	UP000059680	43,649
Plasmodium falciparum	*P. falciparum*	PLAF7	UP000001450	5,187
Rattus norvegicus	Rat	RAT	UP000002494	21,272
Saccharomyces cerevisiae	Budding yeast	YEAST	UP000002311	6,040
Schizosaccharomyces pombe	Fission yeast	SCHPO	UP000002485	5,128
Staphylococcus aureus	*S. aureus*	STAA8	UP000008816	2,888
Trypanosoma cruzi	*T. cruzi*	TRYCC	UP000002296	19,036
Zea mays	Maize	MAIZE	UP000007305	39,299
			Total	**365,198**

Table S2. The Kolmogorov-Smirnov test for the distributions of the proteins from the five selected species. For the proteins with similar chain lengths ($225 \leq N < 275$) in different species, we compare their distributions of the (A) radii of gyration R_g, (B) Coil fraction, and (C) hydropathy variation CV_{HP}. The KS statistic D and the corresponding p-values are listed in the upper and lower triangles in the subtables, respectively. (Here, *** denotes the p-values below 0.001, indicating that there are statistically *highly significant* differences between the two distributions.) These results show that, when we compare the structure (radii of gyration or secondary structure) of proteins with similar chain lengths, there is no significant difference ($p > 0.01$) between *M. jannaschii* (archaea) and *E. coli* (bacteria). In contrast, when comparing their sequences (CV_{HP}), there is a significant difference between *M. jannaschii* and *E. coli*, while the sequences of the proteins within eukaryotes do not differ significantly ($p > 0.01$) from each other.

A

R_g	METJA	ECOLI	YEAST	CAEEL	HUMAN
METJA		0.0946	0.4657	0.5598	0.6288
ECOLI	0.1104		**0.3975**	0.4776	0.564
YEAST	***	***		0.1154	0.2076
CAEEL	***	***	***		0.166
HUMAN	***	***	***	***	

B

Coil %	METJA	ECOLI	YEAST	CAEEL	HUMAN
METJA		0.1158	0.3622	0.4128	0.5335
ECOLI	0.0265		0.2754	0.3429	0.4607
YEAST	***	***		0.1199	0.2274
CAEEL	***	***	***		0.1271
HUMAN	***	***	***	***	

C

CV_{HP}	METJA	ECOLI	YEAST	CAEEL	HUMAN
METJA		0.4962	0.5956	0.6217	0.6437
ECOLI	***		0.2056	0.2598	0.2694
YEAST	***	***		0.0683	0.0745
CAEEL	***	***	0.0895		0.0364
HUMAN	***	***	0.0525	0.291	