Evolutionary tinkering enriches the hierarchical and nested structures in amino acid sequences

Zecheng Zhang,[1,2] Chunxiuzi Liu,[1,2,3] Yingjun Zhu,[4] Lu Peng,[1,2,3] Weiyi Qiu,[5] Qianyuan Tang,[6]
He Liu,[1,2] Ke Zhang,[1,2] Zengru Di,[1,2] and Yu Liu[1,2,*]

[1]*Department of Systems Science, Faculty of Arts and Sciences, Beijing Normal University, Zhuhai 519087, China*
[2]*International Academic Center of Complex Systems, Beijing Normal University, Zhuhai 519087, China*
[3]*School of Systems Science, Beijing Normal University, Beijing 100875, China*
[4]*School of Mathematical Sciences, Xiamen University, Xiamen 361005, China*
[5]*Swarma Research, Beijing 102300, China*
[6]*Department of Physics, Hong Kong Baptist University, Hong Kong SAR, China*

(Received 14 February 2024; accepted 9 May 2024; published 28 May 2024)

Genetic information often exhibits hierarchical and nested relationships, achieved through the reuse of repetitive subsequences such as duplicons and transposable elements, a concept termed "evolutionary tinkering" by François Jacob. Current bioinformatics tools often struggle to capture these, particularly the nested, relationships. To address this, we utilized ladderpath, an approach within the broader category of algorithmic information theory, introducing two key measures: order rate η for characterizing sequence pattern repetitions and regularities, and ladderpath-complexity κ for assessing hierarchical and nested richness. Our analysis of amino acid sequences revealed that humans have more sequences with higher κ values, and proteins with many intrinsically disordered regions exhibit increased η values. Additionally, it was found that extremely long sequences with low η are rare. We hypothesize that this arises from varied duplication and mutation frequencies across different evolutionary stages, which in turn suggests a zigzag pattern for the evolution of protein complexity. This is supported by simulations and studies of protein families such as ubiquitin and NBPF, implying species-specific or environment-influenced protein elongation strategies. The ladderpath approach offers a quantitative lens to understand evolutionary tinkering and reuse, shedding light on the generative aspects of biological structures.

DOI: 10.1103/PhysRevResearch.6.023215

I. INTRODUCTION

Bioinformatics approaches based on sequencing data have effectively demonstrated that DNA and amino acid sequences are encodable. This encodability has been illuminated by employing a range of potent mathematical and statistical techniques, revealing their biological significance. Various studies have suggested strong correlations between the structural features in sequences (such as regularity and nestedness) and the functional properties of proteins [1–3]. One commonly used approach to characterize the sequential features is the Shannon entropy (defined as $H = -\sum p_i \log_2 p_i$, where p_i is the probability of observing letter i) and its variants [4,5]. It was originally proposed to describe the uncertainty of a random variable, but later adopted to characterize the sequential randomness, behind the idea that a sequence can be thought of as a realization of a sequential array of this random variable. It represents a statistical notion of information and is insensitive to the internal structure and pattern of an individual sequence, but it can also be pushed forward to analyze the frequency distribution of short subsequences—namely, the k-mer method—instead of individual letters, and to investigate simple and nonoverlapping repetitions [5,6]. Shannon entropy has found extensive applications in biology and biochemistry, such as identifying genetic motifs [7], analyzing the evolution of genes [8], and describing the complexity of chemical molecules [9]. Nevertheless, this type of approach overlooks the internal hierarchical and nested relationships in a sequence that are found to be very important at the protein domain level [2,10,11] or even in language [12].

On the other hand, algorithmic information theory (AIT), established by Kolmogorov and Chaitin [13,14], serves as another powerful tool for characterizing structural features and complexity. It aims to provide the shortest description (on a universal computer) for a specific sequence or object, known as algorithmic complexity. Several effective methods have been developed to approach algorithmic complexity [15], and these have been applied to a wide range of questions, including characterizing the topological properties of real-world networks [16], investigating whether biological mutations are uniformly randomly distributed [17], and reprogramming the system by steering its algorithmic information content through controlled interventions [18]. Built upon AIT, compression algorithms that utilize repetitive subsequences have been widely used not only for practical applications like compressing sequences and images [19], but also for defining

[*] yu.ernest.liu@bnu.edu.cn

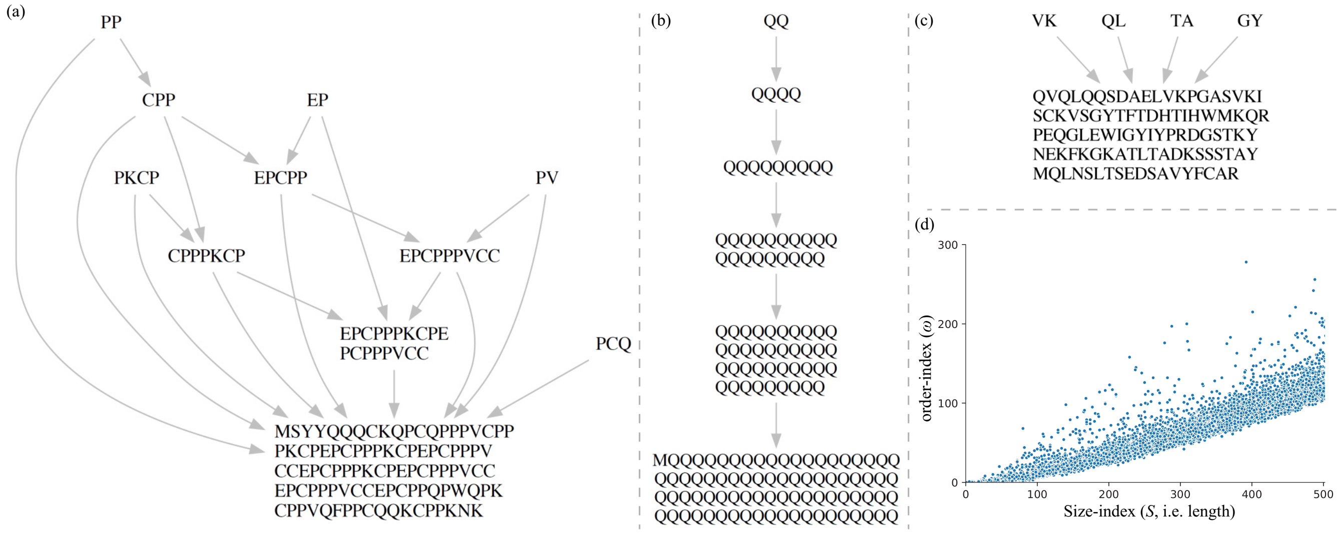

FIG. 1. Laddergraphs and a distribution for human proteins. (a) The laddergraph for the protein SPR2B_MOUSE, where the string at the bottom represents this target protein and shorter strings above are ladderons. The most basic building blocks, namely, individual amino acid, are omitted for better visualization. $S = 98$, $\lambda = 50$, and $\omega = 48$. (b) The laddergraph for the protein ATX8_HUMAN, with $S = 80$, $\lambda = 12$, and $\omega = 68$. (c) The laddergraph for the protein A0A075B674_MOUSE, with $S = 98$, $\lambda = 94$, and $\omega = 4$. (d) The distribution of order-index ω vs size-index S, for human proteins with lengths below 500 AA.

the normalized compression distance, used for clustering in various scenarios, including phylogenetics, languages, and even music [20]. A conceptually analogous measure has also been applied to characterize the structural complexity of two-dimensional (2D) and three-dimensional (3D) structures, defined within a specific set of building blocks and rules, by identifying symmetry and modularity [21]; and more recently, a study based on this measure analyzed the evolution of protein complexes, leading to the conclusion that symmetric structures are likely to appear preferentially because they require less information to encode, making them easier to emerge [22]. All these works highlight the significant value of AIT in evolutionary research.

The amino acid sequence of a protein not only embodies information about its thermodynamics, folding, and other properties (Anfinsen's principle) but also encapsulates details related to its evolutionary trajectory and history, which could be extracted. In 1977, François Jacob posited the abstract idea that evolution is akin to "tinkering" [23], or more specifically, innovations arise from the opportunistic reuse or recombination of existing elements [24]. This process will result in the information accumulated exhibiting hierarchical and nested relationships. Much of this tinkering occurs during replication errors, for example, through point mutation and DNA duplications. The latter is associated with various replication events, such as duplicons and transposable element expansion [25–27], leading to increased complexity in both protein families and genomes [28,29]. Various examples that reflect this tinkering process exist: the length of bacteriophage tails determined by tape measure protein (TMP) [30], the needle length of bacterial injectisome by Yop secretion protein P (YscP) [31], antifreeze glycoprotein in codfish [32], the widespread presence of zinc finger proteins [33], and extensive core duplications in primates [34]. Many of these proteins have undergone significant expansion and mutation, either actively or passively. Yet, the challenge of quantifying such a tinkering process and nested relationships remains.

A recently proposed approach named "ladderpath," within the broader category of AIT, can be used to quantitatively describe the structural information of objects such as sequences, molecules, proteins, and images [35]. It considers the shortest path to generate the target object as the way to characterize it, with the key assumption that the building blocks, once generated, can be reused in any amount in subsequent steps. These reused building blocks are called *ladderons*, which can also be viewed as modules, as defined in Ref. [35]. This aligns with the "tinkering" process proposed by François Jacob [23]. The number of steps required for *ab initio* generation of a target object indicates its generation difficulty, defined as the ladderpath-index λ. Hence, when considering a set of amino acid sequences with the same length, one can discern which sequence is more straightforward or easier to generate. Additionally, to characterize the degree of order in sequences of varying lengths, another useful index, called the order-index, is defined as $\omega := S - \lambda$, where S is the size-index (namely, the length) of the amino acid sequence. By deconstructing the target object into a *partially ordered multiset*—or equivalently, the *laddergraph* [as shown in Figs. 1(a)–1(c)]—the ladderpath approach characterizes the structural intricacies rooted in the hierarchical and nested relationships formed by the target object's repetitive substructures.

The concept of ladderpath also aligns with several other theories, such as addition chain, assembly theory, and the "adjacent possible" [24,36–39]. While these theories have their own measures of complexity, the ladderpath approach posits that "complexity" should be assessed using both the ladderpath-index and the order-index [35]. A sequence is not necessarily complex if it only has a high ladderpath-index with a low order-index [Fig. 1(c)], or vice versa [Fig. 1(b)]. A sequence can be deemed complex if both indices are simultaneously high. Of the three real proteins examined, the one with both a high ladderpath-index and order-index [Fig. 1(a)] exhibits the most intricate and complex hierarchies. For a more encompassing view, Fig. 1(d) shows the distribution of human proteins with lengths below 500 amino acids (AA). The ladderpath approach underscores nature's propensity to innovate through tinkering and reusing existing structures, a trend exemplified in processes like the evolutionary creation of new proteins.

This paper is organized as follows. In Sec. II A, we provide a rigorous definition of the order rate and ladderpath-complexity, and present a systematic comparison with a commonly used k-mer related method. Sections II B and II C present two statistical observations. The former reveals that human protein sequences exhibit higher ladderpath-complexity. The latter notes that proteins containing a significant portion of intrinsically disordered regions, on average, possess a higher order rate. Both observations are statistically significant. In Sec. II D, we begin by detailing a statistical observation that there are almost no superlong sequences with low order-rate values. We speculate that this might be due to the different frequencies of duplication and mutation across different evolutionary stages. This, in turn, suggests that the evolution of protein complexity follows a zigzag pattern. We offer several examples of protein families to support this speculation. The paper concludes with a discussion and a methods section that describes the algorithm for computing ladderpath-associated information. Open-source code is also available.

II. RESULTS

A. Two indicators that characterize amino acid sequences

Firstly, we have developed an efficient algorithm to compute the ladderpath-associated information of sequences, details of which can be found in Sec. IV, with codes available on GitHub for immediate use. This algorithm can effectively handle sequences of around or below 10 000 AA. For sequences extending beyond this but below 40 000 AA, longer running times are required but remain tolerable, and all sequences discussed in this paper fall within this range. In contrast, the previous algorithm (see Ref. [35]) was limited to handling sequences of approximately 20 AA. The statistics displayed in Fig. 1(d) were derived using this new algorithm.

Moving on to Fig. 1(d), we noted a distinct lower boundary for the order-index ω as the sequence length S increases. This lower boundary stems from the finite number of basic building block types (which in this context are the 20 amino acid types), because as the length of amino acid sequences increases, repetitive subsequences become inevitable, resulting in a nonzero value for ω. This is purely a mathematical property, for which we need to compensate. Hence, we introduce two new indicators—the order rate η and ladderpath-complexity κ—to better characterize the system with a finite number of basic building block types.

1. Order rate η

We define the *order rate* η of a sequence x as

$$\eta(x) := \frac{\omega(x) - \omega_0(S)}{\omega_{\max}(S) - \omega_0(S)}, \tag{1}$$

where $\omega(x)$ is the order-index of sequence x, S is the size-index of x (namely, the length of x), $\omega_{\max}(S)$ is the maximum order-index among all the sequences with length S, and $\omega_0(S)$ is the average order-index of all possible sequences with length S, roughly corresponding to the average level of the least ordered sequences; refer to Supplemental Material (SM) [40] Sec. 1 for the calculations of ω_0 and $\omega_{\max}$.

The order rate η characterizes the hierarchical and nested relationships among the subsequences of a sequence, describing the pattern regularities and repetition in the target sequence. Values of η close to zero mean that the degree of order of the sequence is close to the average level of random sequences, indicating that the sequence does not exhibit any significant pattern. As η gets larger and larger, the repetitive parts become more dominant and the sequence exhibits more hierarchical structures [see Fig. 1(a)]. η reaches 1 only when the sequence exhibits exponential elongation of a single letter, e.g., T → TT → TTTT → TTTTTTTT.

2. ladderpath-complexity κ

Another indicator we put forward to characterize the internal structure of sequences is the *ladderpath-complexity* κ, defined as

$$\kappa(x) := [\lambda(x)][\eta(x)], \tag{2}$$

where $\lambda(x)$ is the ladderpath-index of sequence x, and $\eta(x)$ is the order rate of x. As mentioned, the order rate η is a relative indicator of the regularities (compared with the average level of totally random sequences and the most ordered sequence), so its relevance might diminish across sequences of disparate lengths. This indicator ladderpath-complexity κ instead takes into account the minimum number of steps required for the generation of the sequence that is characterized by λ, thereby including the length effect. As demonstrated in the ladderpath approach, the "complexity" of a sequence should incorporate two aspects, that is, one is the difficulty in generating the target, and the other aspect focuses on the hierarchical and nested relationships within the internal sequential structure [35]; the definition of κ integrates these two aspects, hence its name: ladderpath-complexity.

For a given length (namely, size-index S), the maximum value of the ladderpath-complexity κ can be anticipated (see SM [40] Sec. 2 for the mathematical properties of κ). That is, when $\omega = (S + \omega_0)/2$ and $\lambda = (S - \omega_0)/2$, the ladderpath-complexity $\kappa(S)$ reaches its maximum value $(S - \omega_0)^2/[4(\omega_{\max} - \omega_0)]$. In the special case where $\omega_0 = 0$, κ reaches its maximum when $\omega = \lambda = S/2$ (note that ω_0 appears in the general case because of the baseline effect mentioned above). It indicates that when both ω and λ are large, the ladderpath-complexity κ could be large (if only one of ω or λ is large, κ cannot reach its maximum). This is consistent with the notion that complexity incorporate two aspects.

3. Examples and comparative analysis

Next, we take a few protein sequences as examples (with diverse η and κ values) to more clearly and intuitively illustrate what η and κ characterize (Table I and Fig. 2). We can observe that (1) PO5F1_MOUSE has an order rate η

TABLE I. Indicators characterizing protein sequences.

	Examples of proteins sequences (entry name)			
Indicator	PO5F1_MOUSE	SRY_MOUSE	UBC_HUMAN	SDK2_MOUSE
size-index (S)	352	392	685	2176
ladderpath-index (λ)	279	210	73	1379
order-index (ω)	73	182	612	797
order rate (η)	0.0442	0.3581	0.8870	0.0545
ladderpath-complexity (κ)	12.3181	75.1944	64.7484	75.1940

close to 0, meaning that the characteristic features of its internal structure are indistinguishable from those of random sequences [from Fig. 2(a) we can see its few hierarchical structures]; (2) as the order rate η increases, the sequence starts to exhibit richer hierarchical and nested structures, with diverse and overlapping ladderons [Fig. 2(b)], while, as η approaches 1, the hierarchy becomes more like a simple layer-by-layer structure [Fig. 2(c)]; (3) although PO5F1_MOUSE and SDK2_MOUSE have similar small order rate η, the latter has a much higher ladderpath-complexity κ, just because the latter is much longer. Meanwhile, although SRY_MOUSE is much shorter than SDK2_MOUSE, its ladderpath-complexity κ is even slightly higher because of its greater order rate η [from Fig. 2(b) we can see its much richer hierarchical and nested structures]. This indicates that length affects complexity but is not the sole determinant.

Now, we will compare the indicators proposed in this study with another commonly used method. As mentioned, a commonly used tool to describe the sequential feature is the Shannon entropy, which is, however, based on the statistical notion of the frequency and the uncertainty of single letters, rather than the internal structure of a sequence. Nevertheless, the k-mer method has been employed to extend the notion for single letters to substrings of a certain length. Chen *et al.* introduced a normalized indicator named *informational complexity* (C) to characterize the relative uncertainty of substrings [5]. C_k is calculated based on a sliding window of a fixed length k, and thus, the internal sequential structure has been taken into account, at least within the range of k. In fact, C_1 is the Shannon entropy of the sequence (because 1-mer is just the single letter), normalized to the maximum Shannon entropy of the same length. To draw a linguistic analogy, the Shannon entropy functions at the alphabet level, while the k-mer version C_k constructs a dictionary comprising words of a certain length k, quantifying the Shannon information conveyed by these fixed-length words. Consequently, the quantity $(1 - C_k)$, denoted as R_k, represents the degree of regularity, partially aligning with what the order rate η describes.

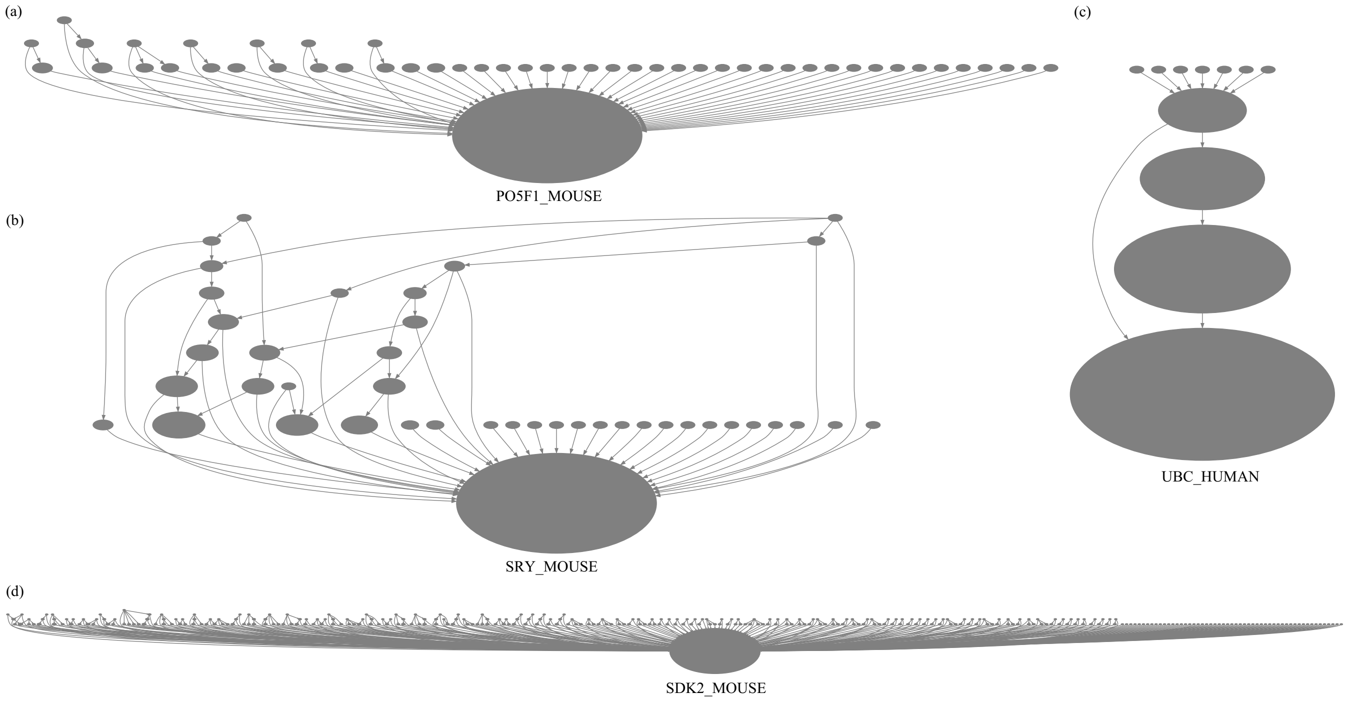

FIG. 2. Laddergraphs of the four example protein sequences presented in Table I. Unlike in Fig. 1, space constraints prevent the explicit display of ladderons in this figure. Instead, ellipses are used to symbolize ladderons, with the size of each ellipse corresponding to the length of the ladderon. (a)–(c) are scaled identically, as evidenced by the corresponding size of the largest ellipse that represents the target sequence in each. In (d), due to the excessive length of the protein SDK2_MOUSE, only a zoomed-out version of its laddergraph is displayed. A detailed version can be found in SM [40] Sec. 3.

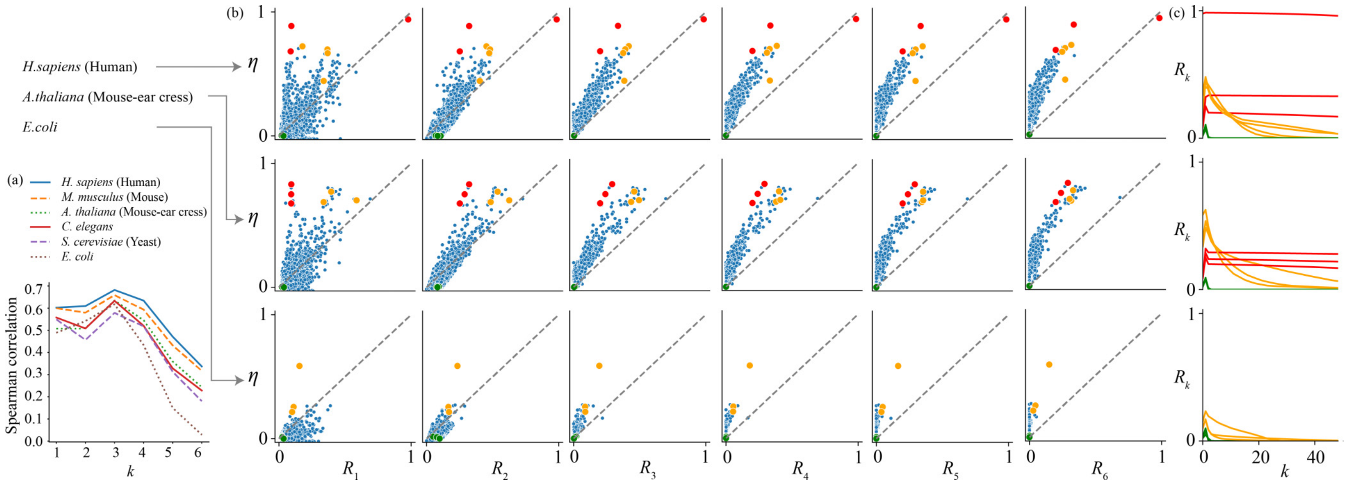

FIG. 3. Systematic comparison between R_k and the order rate η. (a) Spearman correlation between η and R_k, as k increases, for six distinct species. (b) Scatter plots of η vs R_k for $k = 1, 2, 3, 4, 5$, and 6. Each row corresponds to a different species. Individual dots within the plots represent individual proteins. (c) Several representative proteins are chosen (denoted in red, green, and yellow colors) to show how R_k changes as k increases up to 50. Note that the red curves in this subfigure correspond to the red dots in (b), and similar associations are made for the green and yellow curves; each row represents a species, corresponding to (b).

Then, we systematically compare the order rate η with R_k (Fig. 3). We observe a correlation between η and R_1, and the correlation increases as k increases to 2 and 3; After $k > 3$, the correlation begins to drop sharply [Fig. 3(a)]. The correlation exists when $k = 1, 2, 3$ because both indicators, η and R_k, correctly describe certain aspects of the sequence's regularity. Note that the order rate η quantitatively describes the hierarchical and nested relationships among the substructures of a sequence. Therefore, it has a higher correlation with R_3 and R_2, while the correlation with R_1 is lower. This is because 3-mer and 2-mer take substructures into account, while R_1 merely focuses on single letters, neglecting the internal structure. Further, the correlation decreases after $k > 3$ because the whole set of all possible k-mers expands exponentially with k, and thus the Shannon information contained in k-mers becomes submerged in the whole set, resulting in R_k becoming less and less informative.

Another observation is that while a general correlation exists, different proteins exhibit varying tendencies as k increases. For instance, the proteins represented by the red points in Fig. 3(b), which have large η values, tend to retain their position along the x axis as k increases from 2 to 6; in contrast, proteins represented by the blue points descend rapidly along the x axis. This suggests that these different protein sequences have distinct internal structures. To further probe the influence of these internal structures, we chose several representative proteins to analyze how R_k changes as k increases up to 50. Figure 3(c) illustrates this, where red curves correspond to the proteins represented by the red points in Fig. 3(b), and similar associations are made for the green and yellow curves (refer to SM Sec. 4 for the ladderpath-associated indicators of these representative proteins). We observe the following: (1) The red proteins are actually those that have large repetitive segments, but lack rich hierarchical and nested relationships (e.g., ubiquitins), and thus have a relatively high η but low κ. For them, we observe that R_k remains virtually unchanged as k increases. (2) The green proteins have very "chaotic" sequences (i.e., almost no repetitive subsequences), resulting in a low η. For these proteins, R_k approaches zero after $k > 3$. (3) The yellow points fall between these two categories, exhibiting a distinct feature: they decrease slowly with k, hinting at intriguing internal structures.

To summarize, for proteins with distinct internal structures (such as the three exemplified categories), the characterizing capability of different R_k varies. As a species likely contains at least these three categories of proteins, it remains largely arbitrary to determine which k should be used to characterize the sequential features of the species as a whole. Our approach, instead, effectively characterizes the internal structure and provides a global indicator without predefining a characterizing range. Intuitively, the ladderpath-associated indicators liberate "confined-length words" (k-mers) to "variable-length words" (the so-called ladderons, as defined in Ref. [35]), adeptly capturing the hierarchical and nested structures within sequences.

B. Statistical observation: Human protein sequences have higher ladderpath-complexity κ

Here, we present the density distribution of ladderpath-complexity κ for sequences with lengths below 2500 AA across six typical species [Fig. 4(a)]. The statistical differences between distributions reflect species-specific features. We observe that the distribution for human is the flattest, i.e., having the highest proportion of proteins with large κ. In contrast, the distribution for *E. coli* appears to be more concentrated, i.e., having the highest proportion of proteins with small κ. To put it another way (referring to Table II), in terms of the density of proteins with large κ (e.g., $\kappa = 80, 60, 40$), human and mouse rank at the top, forming the first group, followed by the second group (yeast, mouse-ear cress, and *C. elegans*), and finally, *E. coli*. However, for proteins with small κ (e.g., $\kappa = 5, 10$), the first group consists of *E. coli*,

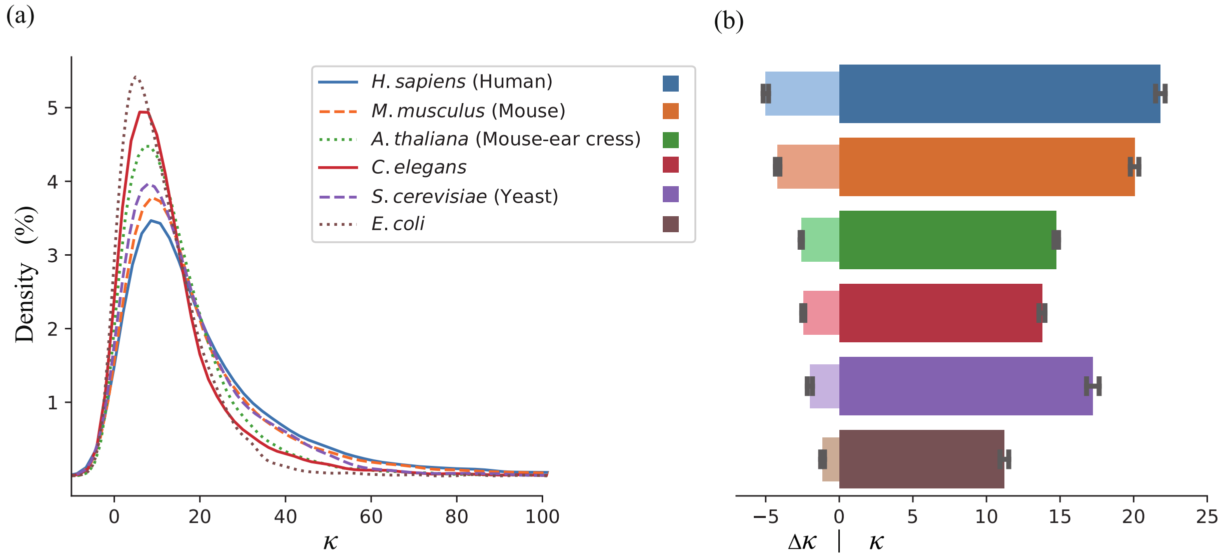

FIG. 4. Overview of the ladderpath-complexity of protein sequences across six typical species. (a) Density distribution of protein sequences with lengths below 2500 AA, with respect to ladderpath-complexity κ. (b) The average κ and the change in κ after shuffling.

C. elegans, and mouse-ear cress, followed by yeast, and finally the third group of mouse and human. This implies that the more complex a species is, the more likely it is to have a higher proportion of proteins with large κ.

Considering that ω_0 comes from the average ω of numerous random sequences with homogenous amino acid content, could the large difference primarily result from the species-specific and inhomogeneous content rather than the internal sequence structure? To test this speculation, we randomly shuffled all sequences—aiming to preserve the amino acid composition but disrupt the internal structures—then recalculated their ladderpath-complexity, and compared the changes before and after, denoted as $\Delta\kappa$ [Fig. 4(b)]. The results indicate that human sequences, followed by those of the mouse, exhibit the most significant reduction, suggesting that these sequences possess the richest hierarchical and nested structures overall. Next are multicellular organisms *A. thaliana* and *C. elegans*, forming the second group, with unicellular organisms yeast and *E. coli* at the base. This correlation is consistent with many literature sources on species complexity, particularly those based on the "total number of interacting proteins" [41]. So, we confirmed that the statistical differences in ladderpath-complexity κ stem from the internal sequence structure rather than the inhomogeneous content; the findings suggest that more complex species, such as human and mice, tend to have a higher proportion of proteins with large κ.

Nevertheless, it is important to mention that there is currently no consensus about the definition of species complexity in the literature. Various metrics, such as proteome size, the number of cell types [42,43], and the total number of interacting proteins [41], have been proposed to gauge species complexity (for more methods mentioned in the literature, see [40] SM Sec. 5). Therefore, our indicator is meant to propose an alternative method from the perspective of sequence analysis, emphasizing the internal structure of sequences, which correlates with species complexity in certain aspects.

1. Showcase: Top list of large-κ proteins

Let us now examine the list of proteins with the highest κ values, considering only those sequences with lengths below 2500 AA (Table III). Interestingly, despite this length limitation, our κ-selection results show a similarity to the findings of the repeat finder: human proteins dominate [44]. Adjusting this length limit to 2000, 1500, or even 10 000 AA does not change this observation (see SM [40] Sec. 6 for more data).

Another notable observation is the length range that spans from 1457 to 2496, indicating that length is not the

TABLE II. Data from the density distribution in Fig. 4, for particular κ values.

	Density (%), for specific κ value					
Organism	$\kappa = 5$	$\kappa = 10$	$\kappa = 20$	$\kappa = 40$	$\kappa = 60$	$\kappa = 80$
H. sapiens (human)	3.04	3.46	2.15	0.66	0.21	0.10
M. musculus (mouse)	3.32	3.77	2.11	0.57	0.19	0.08
A. thaliana (mouse-ear cress)	4.24	4.34	2.19	0.33	0.08	0.02
C. elegans	4.82	4.62	1.65	0.30	0.08	0.03
S. cerevisiae (yeast)	3.67	3.88	2.16	0.60	0.10	0.04
E. coli	5.42	4.43	1.82	0.13	0.02	0.00

TABLE III. Top 25 large-κ protein sequences with length limit below 2500 AA.

Protein (entry name)	Organism	S	λ	ω	η	κ
DMBT1_HUMAN[a]	*H. sapiens* (human)	2413	779	1634	0.518	403.40
FILA2_HUMAN	*H. sapiens* (human)	2391	931	1460	0.417	388.04
Q20007_CAEEL	*C. elegans*	2311	889	1422	0.426	378.69
FILA2_MOUSE	*M. musculus* (mouse)	2362	949	1413	0.399	378.23
DMBT1_MOUSE[a]	*M. musculus* (mouse)	2085	752	1333	0.470	353.09
APOA_HUMAN	*H. sapiens* (human)	2040	632	1408	0.548	346.64
HORN_MOUSE	*M. musculus* (mouse)	2496	481	2015	0.714	343.44
CR1_HUMAN	*H. sapiens* (human)	2039	824	1215	0.408	335.89
TRHY_HUMAN	*H. sapiens* (human)	1943	811	1132	0.391	317.06
Q6DIC6_MOUSE	*M. musculus* (mouse)	2087	969	1118	0.314	304.41
F186A_MOUSE	*M. musculus* (mouse)	1790	716	1074	0.424	303.47
MUC22_HUMAN[b]	*H. sapiens* (human)	1773	700	1073	0.432	302.34
Q9LH98_ARATH	*A. thaliana* (mouse-ear cress)	2081	1032	1049	0.267	275.09
A0A0B4J1F9_MOUSE	*M. musculus* (mouse)	1599	665	934	0.409	272.18
PWWP4_HUMAN	*H. sapiens* (human)	2061	1031	1030	0.261	269.30
FLO1_YEAST[c]	*S. cerevisiae* (yeast)	1537	596	941	0.451	268.92
NACAM_HUMAN	*H. sapiens* (human)	2078	1049	1029	0.254	266.16
CO4A2_CAEEL	*C. elegans*	1758	828	930	0.320	265.13
F7C950_MOUSE	*M. musculus* (mouse)	1606	410	1196	0.642	263.12
Q9LIE8_ARATH	*A. thaliana* (mouse-ear cress)	1480	477	1003	0.549	261.97
NBPFC_HUMAN	*H. sapiens* (human)	1457	532	925	0.489	260.13
TARA_HUMAN	*H. sapiens* (human)	2365	1249	1116	0.206	257.28
SON_HUMAN	*H. sapiens* (human)	2426	1289	1137	0.199	255.93
Q63ZW6_MOUSE	*M. musculus* (mouse)	1691	805	886	0.316	254.43
CO4A5_HUMAN	*H. sapiens* (human)	1685	801	884	0.317	254.01

[a] Belongs to the DMBT1 family.

[b] Mucin protein.

[c] Belongs to the flocculin family.

determining factor for κ; instead, repetition in the sequence plays a significant role. For example, DMBT1, flocculin, and mucin in Table III are protein classes that are famous for tandem repeats [45–47].

C. Statistical observation: Proteins containing intrinsically disordered regions have higher order rate η

Now, let us consider the relationship between the amino acid sequence and its corresponding 3D structure. Intuitively, duplicated sequences could be expected to adopt identical structures. Therefore, a long sequence with many duplicated subsequences (thereby tending to have higher order rate η) may be considered to have a consistent structure comprising explicit identical substructures [48]. For instance, the protein depicted in Fig. 5(a) exhibits a consistent and regular structure [49]; another notable example is the much larger protein DMBT1_HUMAN, shown in Fig. 5(b), which also has a high η value, as shown in Table III. Nevertheless, there are proteins with high η values but are structurally disordered, as the example depicted in Fig. 5(c), which exhibits regions predicted by AlphaFold2 with low confidence, implying structural disorder.

To uncover statistical patterns, we utilized data from the DisProt database [50] to calculate the average order rate η for proteins with a significant proportion of intrinsically disordered regions (IDRs), and compared it with other proteins without a significant proportion of IDRs. The results are shown in Fig. 5(d) (the right part with darker colors). It is evident that, generally, proteins with IDRs have higher η values than those without such regions, which is statistically significant for four out of six species analyzed. Yet, due to the limited data available in DisProt, we also employed the METAPREDICT software [51] to predict the presence of IDRs in all proteins of their proteomes for these species, and then matched them with their respective η values. The outcomes of this analysis are presented in Fig. 5(e). The pattern remains consistent, with clear statistical significance observed for five out of these six species. As we know, the order rate η is associated with the presence of repetitive substructures in the sequence; therefore, a higher η implies the presence of more repetitive substructures. Indeed, quite a few studies have found that intrinsically disordered proteins (IDPs) or IDRs contain more repeats, such as tandem repeats [52–54] and segmental duplications [55]. For example, Simon *et al.* reported that tandem and cryptic amino acid repeats often accumulate in disordered regions of proteins [53]. Similarly, Jorda *et al.* found that perfect or nearly perfect tandem repeats exhibit a strong tendency to be unstructured in various species [52].

Nevertheless, previous studies show that proteins containing IDRs have a greater amino acid abundance bias [56,57]. It is thus possible that the high η value arises from this bias rather than from the orderliness of the internal sequential structure. To investigate this, we compared the η value before and after shuffling, denoted as $\Delta\eta$, as shown in Fig. 5(d) (the left part with lighter colors). We observed that, statistically

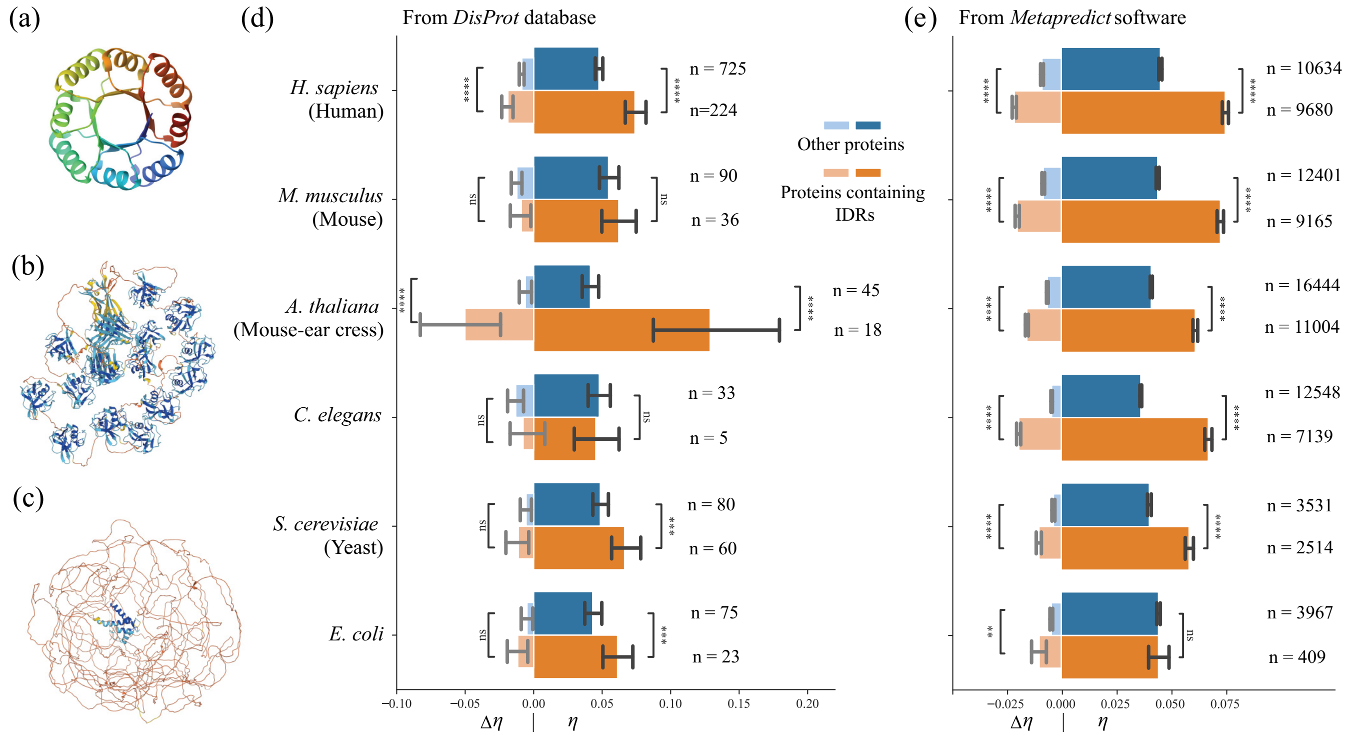

FIG. 5. Statistics related to the order rate η of proteins containing a significant proportion of IDRs. (a) Structure of an artificial protein DeNovoTIM15, from the Protein Data Bank (PDB:6wvs). (b) Predicted structure of the protein DMBT1_HUMAN by AlphaFold2. (c) Predicted structure of HORN_MOUSE by AlphaFold2. (d) The right part with darker colors shows the average η for proteins containing a significant proportion of IDRs compared to proteins without a significant proportion of IDRs. The left part shows the changes in η after shuffling the sequences. The corresponding data size n from the DisProt database is indicated. (e) This is similar to (d), but the data are from calculations using the disorder predictor software METAPREDICT for the six proteomes. Note that **** means $p < 0.0001$, *** means $p < 0.001$, ** means $p < 0.01$, and ns means "no significance."

speaking, $\Delta\eta$ is larger for proteins containing IDRs. From this observation, we can suggest that the internal sequential structure plays a role in the high η values. Therefore, the degree of orderliness may serve as a different feature of disordered regions at the sequence level.

D. Evolution of the complexity of protein sequences follows a zigzag pattern

We now present statistics that encompass all sequences (Fig. 6), not just those shorter than 2500 AA. Generally, most of these sequences have η values confined below 0.1 [Fig. 6(a)]. However, a closer examination of the η distribution [Fig. 6(c)] reveals that the proteins of human, mouse, and *C. elegans* exhibit a significant tendency: as the length of the sequence increases, there tend to be a higher number of sequences with larger η, indicating more ordered sequences. This trend is also observable in Fig. 6(a), where, for extremely long lengths, sequences with low η values are even absent.

An immediate question is why there are almost no superlong but low-η sequences. Later, we shall see that this question strongly relates to how protein sequences elongate. Now, imagine there is initially a short sequence or segment, and consider how this sequence elongates and how the order rate η evolves, via specific biological processes:

(i) Duplication: This refers to the process where a segment of a sequence, either short or long, is copied onto itself. This creates a repetitive subsequence, corresponding to a ladderon as defined in ladderpath. As a result, η of this sequence increases. The longer the segment, the greater the increase in η.

(ii) Substitution: This refers to the replacement of a base. This does not alter the sequence's length, but it may disrupt a ladderon, thereby slightly decreasing the value of η.

(iii) Insertion: This could be thought of as either the addition of a foreign segment or a single amino acid, or as a duplication of a segment immediately followed by substitutions occurring at every base.

Note that for simplicity, we only consider the processes that do not shorten the sequence, thus neglecting deletion.

We now simulate the process of elongation in three cases: (1) completely driven by duplication, (2) completely driven by insertion, or (3) driven by a combination of duplication and substitution. The simulation results are displayed in Fig. 6(d) (refer to SM [40] Sec. 7 for details on how the simulation was conducted). Although the simulation focuses solely on the elongation of protein sequences, it provides insight into the question of why there are virtually no extremely long sequences with low η values. The red trajectories in Fig. 6(d) represent case (1), where η increases the most rapidly during elongation. The green trajectory represents case (2), where the order rate η remains consistently low. The yellow trajectories, representing case (3), lie in between and closely resemble real-world scenarios where infrequent duplications

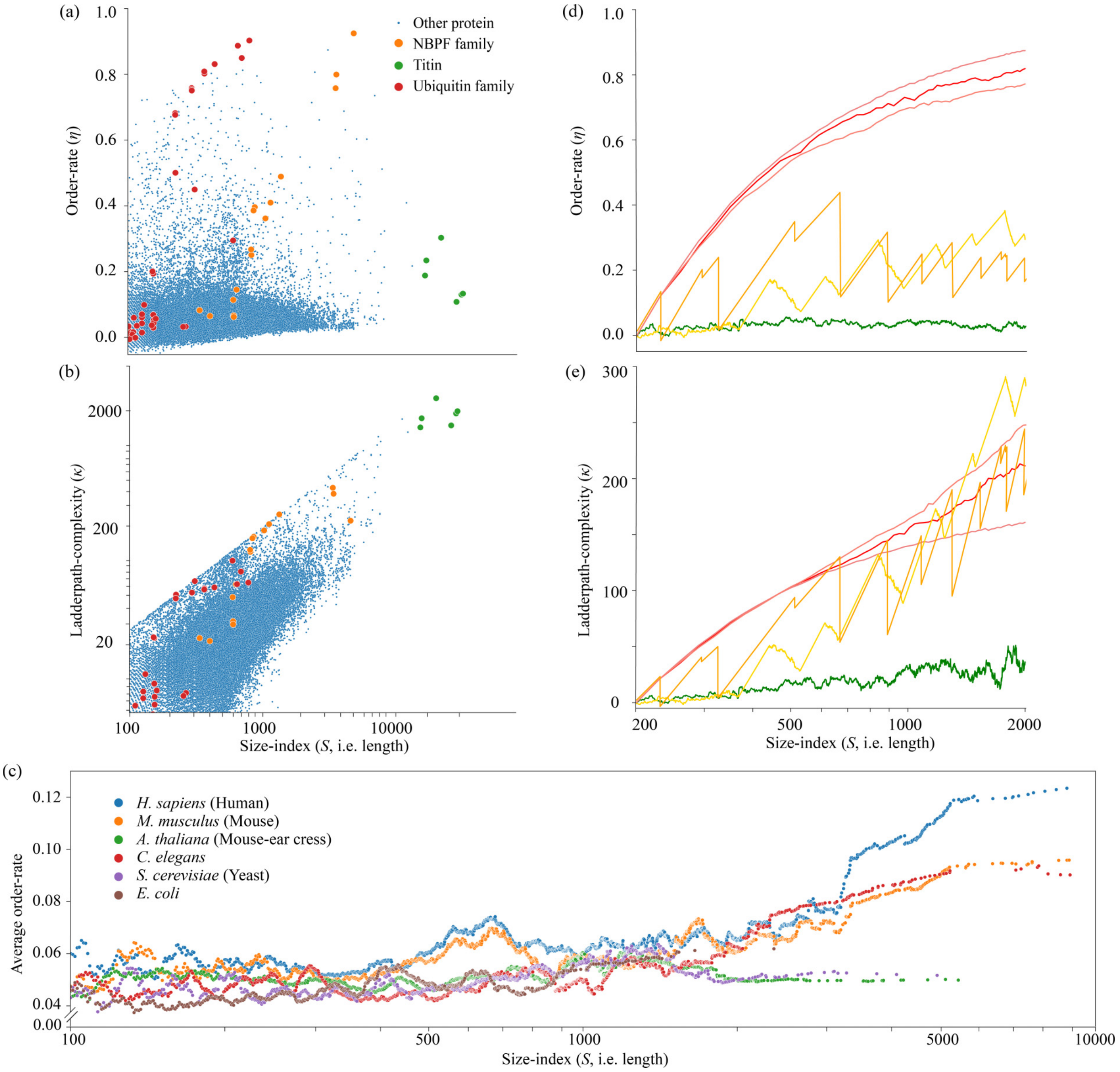

FIG. 6. Observations and simulation experiments related to protein elongation. (a) Scatter plot of protein lengths S vs order rate η, for all proteins across the six species. See the legend of (c) for the six species. (b) Similarly, a scatter plot of S vs ladderpath-complexity κ. (c) The average η values of proteins vs protein length S for the six species. Each dot (S, average η) is calculated using proteins within a sliding window centered at a specific length S. (d) Results of simulation experiments showing how η evolves as the protein sequence elongates, for the three different cases elaborated in the main text. (e) Similarly, simulation experiments showing how κ evolves.

of relatively large segments heavily increase η, while frequent substitutions consistently reduce η, forming a zigzag pattern. We also see from Fig. 6(e) that κ increases the most in case (3), namely, the yellow trajectories.

In summary, the evolution of protein sequences follows a zigzag pattern. Specifically, the duplication of segments increases the order rate of the sequence as it elongates, while this increment in order rate is gradually counteracted by various mutations, either partially or completely, depending on the relative frequencies of duplications and mutations. Now, we could consider the emergence of a new gene or pseudogene: (1) Occasionally, a replication error leads to the duplication of a segment at a different location within the sequence, resulting in higher η and κ values and contributing richer raw materials for further evolution. (2) Subsequently, this elongated sequence undergoes various "tinkering" processes across generations, reducing η and κ. Over time, this sequence gradually diverges from its ancestor and may eventually become a new gene or a pseudogene.

1. Examples: ubiquitin, titin, and NBPF family

Now we can return to the observation mentioned at the beginning of Sec. II D and ask why there are almost no extremely long but low-η proteins. Here, we provide three representative examples to address this question.

The first example is ubiquitin, which is used to emphasize the effect of duplication. Ubiquitin is a highly conserved, small regulatory protein widely found in eukaryotes, which functions as a post-translational modifier, mainly in protein

TABLE IV. Ladderpath-associated indicators of titin proteins.

Protein (entry name)	Organism	S	λ	ω	η	κ
TITIN_DROME	*D. melanogaster* (fruit fly)	18 141	7843	10 298	0.188	1472.60
TTN1_CAEEL	*C. elegans*	18 562	7545	11 017	0.234	1768.37
A0A7M7N314_STRPU	*S. purpuratus* (sea urchin)	24 046	8692	15 354	0.303	2635.61
A0A8M9QKG2_DANRE	*D. rerio* (zebrafish)	31 468	14 202	17 266	0.108	1532.84
TITIN_HUMAN	*H. sapiens* (human)	34 350	15 001	19 349	0.130	1946.69
TITIN_MOUSE	*M. musculus* (mouse)	35 213	15 295	19 918	0.133	2034.22

degradation. Polyubiquitin (UBB and UBC) has an extremely high η value because it contains almost no mutations and has several tandem head-to-tail repeats of ubiquitin, each being 76 AA long [58] [refer to Fig. 2(c) for the laddergraph of UBC_HUMAN]. The distribution of η and κ values for this protein family is shown in Figs. 6(a) and 6(b). We can see that while some members of this family have an extremely high η value approaching 1, their corresponding ladderpath-complexity κ is not particularly high. This observation can be attributed to the nearly error-free duplication events, aligning with case (1) discussed earlier, and the lengths of these proteins, which are not particularly long.

The second example is another extreme, the ancient protein titin, which is used to emphasize the effects of mutations along protein elongation. Titin serves as a structural support in muscles and is of immense length (e.g., TITIN_HUMAN contains 364 exons) [59]. This gigantic protein consists of numerous domains, some of which belong to the PEVK region, which is rich in highly repetitive sequences (this PEVK region forms a distinct structure in the center of the protein, functioning as an entropic spring [60]). Nevertheless, the η values of titin are not very high and exhibit variations among different species [61,62], as shown in Table IV. This suggests that the effects of duplications, which can increase the hierarchical and nested structures of sequences, have been largely counteracted by long-term and consistent mutations.

The third example is an emerging family at the evolutionary scale, *neuroblastoma breakpoint family* (NBPF), which lies in between the two extremes mentioned above. NBPF is known for its members having varying numbers of Olduvai repeats, with approximately 20 members in humans, playing a certain role in human brain development and cognition [63,64]. These young proteins seem to be predominantly found in proteomes of primates, whereas in nonprimate mammals, their counterparts exist as single-copy Olduvai. As an amplicon, Olduvai has undergone a significant gene amplification within a relatively short time span [63,65]. Thus, η and κ increased significantly, and mutations had not had enough time to largely lower η and κ to counteract the effect of duplication (Table V). Therefore, from Fig. 6(a), we can observe that the NBPF family members form a clear pattern, exhibiting their evolutionary trajectory.

The aforementioned classes of proteins illustrate the elongation seen in ancient proteins and the emergent core duplication found in longer proteins, which can be metaphorically described as an Odyssey-like journey. These examples suggest that, over a long duration, there is a certain degree of synchronization between size expansion and increased complexity, while between expansion events, complexity tends to decrease. Thus, the evolution of sequence complexity appears to follow a zigzag pattern. Most long proteins do not exhibit the same level of extremity as ubiquitin and titin, but instead

TABLE V. Ladderpath-associated indicators of the gene family NBPF.

Protein (entry name)	Organism	S	λ	ω	η	κ
NBPF5_HUMAN	*H. sapiens* (human)	351	268	83	0.082	22.07
NBPF7_HUMAN	*H. sapiens* (human)	421	321	100	0.065	20.84
NBPF3_HUMAN	*H. sapiens* (human)	633	438	195	0.114	49.93
NBPF4_HUMAN	*H. sapiens* (human)	638	464	174	0.066	30.82
NBPF6_HUMAN	*H. sapiens* (human)	638	466	172	0.062	29.01
NBPFF_HUMAN	*H. sapiens* (human)	670	445	225	0.145	64.48
NBPFB_HUMAN	*H. sapiens* (human)	865	479	386	0.268	128.20
NBPF8_HUMAN	*H. sapiens* (human)	869	492	377	0.251	123.33
NBPFP_HUMAN	*H. sapiens* (human)	902	419	483	0.386	161.61
NBPFE_HUMAN	*H. sapiens* (human)	921	420	501	0.396	166.41
NBPF9_HUMAN	*H. sapiens* (human)	1111	522	589	0.362	188.89
NBPF1_HUMAN	*H. sapiens* (human)	1214	523	691	0.410	214.31
NBPFC_HUMAN	*H. sapiens* (human)	1457	532	925	0.489	260.13
NBPFA_HUMAN	*H. sapiens* (human)	3795	584	3211	0.760	444.00
NBPFJ_HUMAN	*H. sapiens* (human)	3843	491	3352	0.801	393.28
NBPFK_HUMAN	*H. sapiens* (human)	5207	248	4959	0.927	229.84

fall somewhere in between, e.g., NBPF. For more examples of such proteins, refer to SM [40] Sec. 8.

III. DISCUSSION

A. On definitions

The ladderpath approach aims to decode the information concealed within the hierarchical and nested relationships among the recurring subsequences found in a specified set of target sequences. It achieves this by iteratively identifying recurring subsequences (termed the *ladderons*) and rearranging them into a tree-like hierarchical structure (termed the *laddergraph*), which distills and encodes the evolutionary information. In the context of biological sequences, these recurring subsequences, or ladderons, could represent motifs, domains, or signify transposable elements, satellite DNA, microduplications within genome scale, and the like. To better encapsulate the tree-like hierarchical structure, two indices were derived. The first is the order-rate η, which, in a normalized manner, quantitatively measures the orderliness of a sequence, ranging from close to 0 [completely disordered, as illustrated in Fig. 2(a)] to 1 [fully ordered, as illustrated in Fig. 2(c)]. When η sits centrally, the structure exhibits significant order while the ladderons display intricate overlaps and nested relationships [as illustrated in Fig. 2(b)]. At this point the other derived index, κ, reaches its maximum, signifying the utmost complexity. The ladderpath-complexity κ gauges complexity by factoring in both orderliness and the length. While sequence length does contribute to complexity, longer does not necessarily equate to more complex.

Ladderpath differs from Shannon entropy in that the latter primarily focuses on the statistics of individual letters, although extensions, such as the k-mer method [4,5], can be adapted to consider substructures. These approaches did not factor in the intricate hierarchical relationships among these substructures. Thus, our order rate η shows a correlation with R_k, an index derived from the k-mer method, but this correlation varies with different internal sequential patterns. Further, Shannon entropy and its variants operate under a strong assumption that the sequence in question represents a realization of a random variable, implying that the sequence should be infinitely long. However, in reality, amino acid sequences invariably have finite lengths. On the other hand, finite lengths mean that methods like the Lempel-Ziv lossless compression (a compression method also under the umbrella of AIT) cannot achieve their optimal or shortest description [19]. This introduces significant variability when trying to deduce genuine evolutionary histories. In contrast, ladderpath does not rely on assumptions of infinite length.

B. Statistical observations on sequential orderliness and complexity

The first statistical observation, based on our examination of the ladderpath-complexity of proteomes, reveals differing complexity distributions among species. Among the six species analyzed, humans, followed by mice, possess relatively more sequences of high complexity that exhibit richer hierarchical and nested structures. We also confirmed using shuffling methods that this complexity does not stem from content differences but arises from internal sequential patterns. From the perspective of protein structure, studies have shown that species with higher complexity possess more proteins with larger radii of gyration (signifying increased flexibility) and a higher degree of modularity [66]. On the other hand, our analysis from the sequential perspective implies that the more complex a species is, the higher the tendency for sequence complexity.

Nevertheless, as we mentioned, there is currently no consensus on the definition of species complexity. Initially, using genome size as a measure led to the well-known C-value enigma [67], as some plants and protozoa have larger genomes than humans. The discovery of noncoding DNA in the early 1970s addressed many of these questions, and now genome size is no longer directly linked to species complexity. Subsequently, using proteome size as a measure introduced the G-value paradox [68], where, for example, *C. elegans* and humans have nearly identical proteome sizes. Additionally, researchers also use other metrics such as the total number of cell types [42,43] or the total number of interacting proteins [41] to gauge species complexity. Indeed, it is conceivable that there may never be a single standard or definitive truth for species complexity. Therefore, our indicator is meant to propose an alternative method from the perspective of sequence analysis, emphasizing the internal structure of sequences. Further research involving more data and comprehensive analysis is likely necessary to explore the relationship between species complexity and sequence complexity.

Another statistical observation is that proteins with a significant proportion of IDRs tend to exhibit higher order rate η with statistical significance. It is crucial to note that these elevated η values usually do not exceed 0.1 [as shown in Fig. 5(e)]. Within this range, a higher η invariably indicates richer hierarchical and nested structures, akin to transitioning from proteins exemplified in Fig. 2(a) to Fig. 2(b) [not possible to Fig. 2(c) since such an extreme hierarchical relationship requires an η of approximately 0.8 or higher]. Thus, our findings suggest that, at the sequence level, proteins with IDRs tend to have richer hierarchical and nested structures compared to typical proteins. This correlation between sequence orderliness and structural uncertainty aligns with previous studies, particularly regarding tandem repeats [52–54] and segmental duplications [55]. However, the underlying mechanism for this structural tendency remains unclear, which warrants further study.

On the other hand, understanding that a higher η often originates from segment duplication, another intriguing question arises: Does the evolution of IDPs involve more duplication events? Lastly, building upon the earlier point that more complex species have more proteins with higher modularity, a bold idea might be developed: Could IDPs be an essential stage in the evolutionary journey toward increasing protein modularity? Specifically, an amino acid sequence "core," through occasional duplication events, generates repetitive subsequences along elongation (which naturally leads to an increase in η), resulting in structures becoming more disordered and flexible, facilitating the exploration of various interactions, and ultimately leading to the fixation of structural modules. Although there is no direct evidence suggesting such evolutionary processes, there are some hints.

For instance, Ref. [54] states that internal repeat regions (such as microsatellites and minisatellites) always retain basic functions and exhibit interspecies variation and polymorphism, providing a foundation for their further evolution into new genetic materials. Similarly, Ref. [53] expresses a compatible view, noting that amino acid repeats tend to evolve very rapidly compared to other parts, which are crucial for the rapid adaptation of species.

C. On evolution

Our results suggest that as the protein elongates, its complexity follows a zigzag pattern, originating from the interplay of duplication and mutation (the latter refers to processes such as substitution and insertion). Duplication results in a sharp increase in sequence orderliness and length, while mutation leads to a decline in orderliness, with the length remaining more or less unchanged, together leading to a significant diversity in the internal patterns of sequences. Owing to the interplay of these mechanisms and their varying occurrence rates, the internal structure of the sequence can become highly hierarchical and interlaced. This might result in proteins having distinct values of κ, η, and S (e.g., leading to different distributions between long and short proteins), potentially promoting a range of structures and functions. Statistically speaking, we did observe that η distributions diverge when protein length exceeds 2000 AA [Fig. 6(c)]. This hints that various species, or those in varied environments, might adopt different elongation strategies or, in other words, different "tinkering" processes. For instance, the trend of evolving into multidomains is more pronounced in eukaryotic proteins than in prokaryotes [69], suggesting that distinct biological elongation dynamics might be at play. The evolution of human-specific segmental duplications (HSDs) seems to exhibit varied patterns across different periods. During the human-chimpanzee divergence, there was a period of relative quiescence, succeeded by a spike in HSD occurrences and the emergence of new genes [70–72]. The previously mentioned NBPF experienced rapid, widespread duplications. The Olduvai domain, in particular, stands out as one of the most extreme and fastest copy number expansions in the human genome (with humans having about 300 copies, great apes 90–120, monkeys 30–40, and single or a few copies in nonprimate mammals, while being absent in nonmammals), which has been strongly linked to human brain evolution and cognitive function [73]. Variations in elongation mechanisms, especially under diverse or rapidly changing environmental stresses, might be advantageous for quick adaptation [27,74], potentially accelerating the emergence of new structures or inducing dose-dependent effects [75,76], among other outcomes.

In Fig. 6(c), more detailed analysis and intriguing insights can be observed. The length of the *E. coli* proteome is interrupted around 2000 AA. Beyond this length, yeast and mouse-ear cress (as species with cell walls) show no increase in order rate. Meanwhile, for the mouse and *C. elegans*, both multicellular species without cell walls, there is an evident rise in order rate, with their trends aligning closely. At lengths greater than 3000 AA, the order rate of human proteome experiences a sudden and significant surge. Based on these observations, we make the following speculations: (1) The juncture at which the *E. coli* length halts could be a pivotal point in the shift from prokaryotes to eukaryotes. Eukaryotes might have developed additional tools for sequence expansion and tools that augment the hierarchical and nested structure of sequences, especially those facilitating intragenomic duplication. This transition could be a landmark event differentiating eukaryotes from prokaryotes. (2) The patterns observed in yeast and mouse-ear cress indicate that being unicellular or multicellular may not be a key factor affecting proteome orderliness, while having a cell wall might pose an obstruction to increasing the order rate. We speculate that the cell wall might hinder horizontal gene transfer between species, preventing elements with capabilities such as translocation and duplication from integrating and emerging as evolutionary tools. (3) The spike in order rate seen in humans, relative to other eukaryotic species, raises the question: have humans undergone certain critical events or acquired novel genetic tools? If a stark contrast remains when comparing humans at this point with other nonhuman primates (taking the example of NBPF, as previously discussed), it might explain the profound impact of social development within human evolution. Collectively, these findings suggest a deeper exploration of evolutionary data using this approach or similar methodologies. It also underscores that the ladderpath approach could harbor significant potential for more in-depth applications in evolutionary biology.

On the other hand, the simulated evolutionary process obtained through alternating segmental duplication and mutation provides a better fit to actual evolutionary data than considering mutations alone. This phenomenon poses a significant challenge to the neutral theory [77] and constructive neutral evolution [78]. At the very least, it suggests that from the time of Darwin to current evolutionary biology theories [79], there has been an overemphasis on the role of mutations, neglecting the effects of gene duplication and transfer. As inferred from the sudden shifts shown above, gene duplication and transfer are likely the main ingredients for significant evolutionary leaps. This resonates with the endosymbiotic theory [80] and horizontal gene transfer [81] applied to explain genome expansion. Such observations hint at two important applications of the ladderpath approach: (1) One application is identifying critical shifts and branching points in the entire evolutionary tree, examining whether new gene modules have been added, and pinpointing which of these modules have undergone extensive duplication and transfer in subsequent evolutionary bursts. The ladderpath approach may address the problem of phylogenetic lineages that are obscured by chimeric, symbiotic, or reticulate evolutionary events, which may provide crucial insights into phenomena like the Cambrian explosion [82]. (2) In fields such as synthetic biology and enzyme engineering [83], as well as pharmaceutical engineering [84], the practice of directed evolution is mainly based on point mutations and mutation libraries. There is limited application of strategies involving extensive gene segmental duplication. Yet, by using the laddergraph (in this case, we are not limited to indicators such as the order rate but the laddergraph that has more comprehensive information), we can identify duplicated ladderons merged within these hierarchical and nested relationships, hinting at which ladderons are important

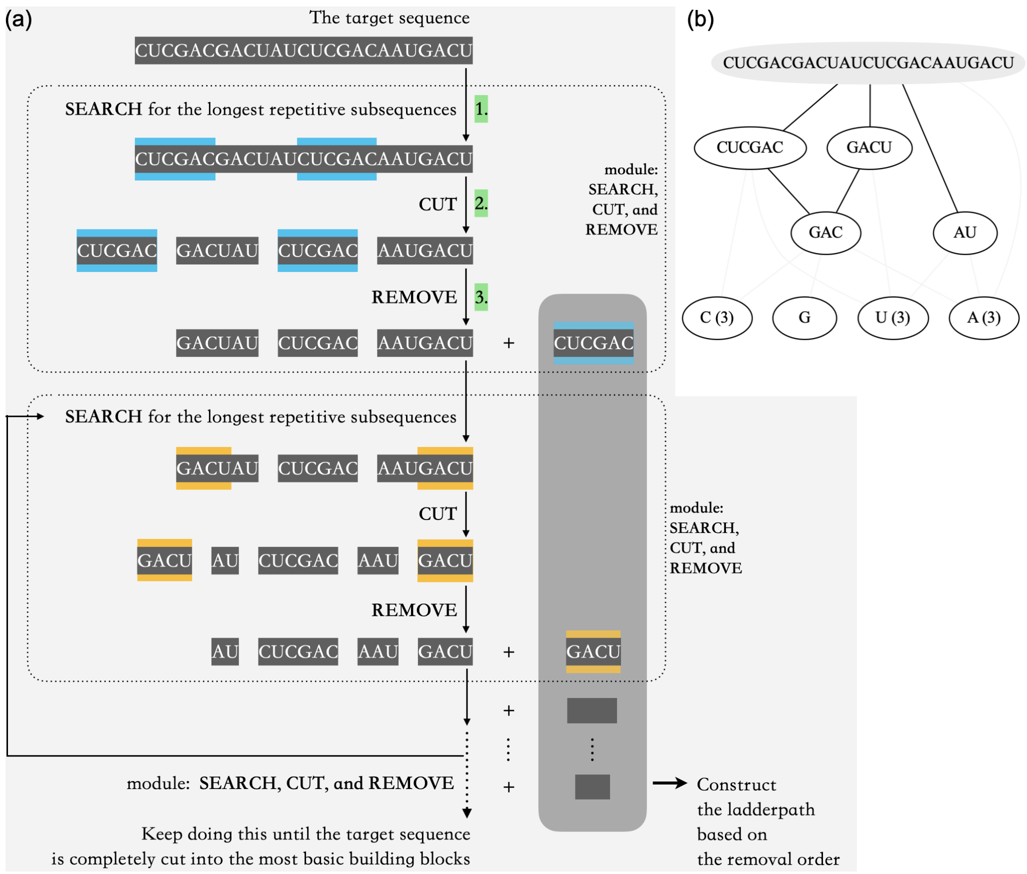
(a)
The target sequence
CUCGACGACUAUCUCGACAAUGACU
SEARCH for the longest repetitive subsequences
1.
CUCGACGACUAUCUCGACAAUGACU
CUT
2.
CUCGAC
GACUAU
CUCGAC
AAUGACU
REMOVE
3.
GACUAU
CUCGAC
AAUGACU
+
CUCGAC
module: SEARCH, CUT, and REMOVE
SEARCH for the longest repetitive subsequences
GACUAU
CUCGAC
AAUGACU
CUT
GACU
AU
CUCGAC
AAU
GACU
REMOVE
AU
CUCGAC
AAU
GACU
+
GACU
module: SEARCH, CUT, and REMOVE
+
+
module: SEARCH, CUT, and REMOVE
Keep doing this until the target sequence
is completely cut into the most basic building blocks
Construct the ladderpath based on the removal order
(b)
CUCGACGACUAUCUCGACAAUGACU
CUCGAC
GACU
GAC
AU
C (3)
G
U (3)
A (3)

FIG. 7. The algorithm for computing ladderpath-associated information. (a) Flowchart illustrating the algorithm with a specific example. (b) The laddergraph of the exemplified target sequence, calculated by this algorithm.

and possibly meaningful modules; and then we can use these information to perform mutations, e.g., replacing point mutations with modular mutations in directed evolution, potentially accelerating the evolutionary process significantly. Furthermore, while gene duplication has found many applications in plant and animal breeding, issues such as the adaptability of inserted duplicate fragments and their loss in subsequent generations have consistently hampered successful breeding rates [85]. Using the ladderpath approach to determine the optimal ratio and strategy for duplication and mutation might offer improved tools for targeted breeding [86] and related biotechnological endeavors.

The ladderpath approach provides a useful tool and a specific computational method to quantitatively describe the complexity of target objects, such as sequences. It focuses on "how objects are generated" rather than on emphasizing uncertainty, as in the case of Shannon entropy, or the efficiency of compression, as seen in lossless compression algorithms like Lempel-Ziv. This approach embodies the evolutionary tinkering process, highlighting the importance of "reuse" and "modularity." While this paper demonstrates the usefulness of derived indicators such as order rate and ladderpath-complexity, it is even more crucial to note that comprehensive information is stored in the laddergraph, which depicts the hierarchical and nested relationships among recurring subsequences, resulting from the evolutionary tinkering process. In practice, we can learn from the tinkering mechanisms of innovation that nature employs (along with sophisticated and powerful reductionist-like innovation) to help us construct complex targets or systems from simpler ones, e.g., peptide drug design (to be discussed in an upcoming paper) and synthetic biology. Using the ladderpath approach as a tool to reverse engineer species evolution might also offer valuable insights, facilitating the design of more effective directed evolution strategies, which could then be applied to fields such as crop breeding and even the design of bioprocesses.

IV. METHODS

A. Algorithm for computing ladderpath-associated information

Here, we show how the algorithm works by taking a target sequence CUCGACGACUAUCUCGACAAUGACU as an example [Fig. 7(a)]. Firstly, we search for the longest repetitive subsequence in the target sequence and find CUCGAC, marked in blue. Secondly, we cut the target sequence into pieces so that the repetitive subsequences are isolated. As a result, we obtain a set of shorter sequences: [CUCGAC, GACUAU, CUCGAC, AAUGACU]. In the third step, we place one CUCGAC into a separate bag, which will then be used to construct the ladderpath. After this step, we have a set of sequences [GACUAU, CUCGAC, AAUGACU] remaining. These three steps constitute the module which we call "SEARCH, CUT, and REMOVE", marked in green in Fig. 7(a).

Next, we treat the remaining set of sequences [GACUAU, CUCGAC, AAUGACU] as a "target" sequence and apply the "SEARCH, CUT, and REMOVE" module to this target. From this, we obtain another longest repetitive subsequence, GACU, which we place into the separate bag. We continue to apply the module until the original target sequence is completely segmented into its most basic building blocks. Finally, based on the order of removal, we construct the ladderpath as {G, A(3), C(3), U(3) // AU, GAC // CUCGAC, GACU }. It characterizes the hierarchical and interlaced relationships within the original target sequence, and has a one-to-one correspondence with a laddergraph shown in Fig. 7(b).

The source code for this algorithm is available on GitHub [87]. Note that for sequences below 10 000 AA, the code can handle everything efficiently (all sequences mentioned in this paper fall within this range, except for the six sequences listed in Table IV). For sequences extending beyond this but below 40 000 AA (in the entire dataset, only six sequences in Table IV have lengths between 10 000 and 40 000 AA), the running times in all aspects are still tolerable, except for determining the order rate η, as computing the accurate value of $\omega_0(S)$ for $S > 10\,000$ AA requires significant computational power. Nonetheless, we have developed a technique to compute $\omega_0(S)$ effectively for large S. The idea is that highly disordered long sequences can be segmented and collectively processed to calculate ω (with further details available in SM [40] Sec. 9).

B. Identifying IDRs

For Fig. 5(d), if for a protein sequence the ratio of the consensus region to the total length is over 25%, we say that this protein contains a significant proportion of IDRs. For Fig. 5(e), we applied the disorder predictor software METAPREDICTOR on the proteomes of the six species *H. sapiens* (human), *M. musculus* (mouse), *A. thaliana* (mouse-ear cress), *C. elegans*, *S. cerevisiae* (yeast), and *E. coli*. For each protein sequence, we used the command-line tools of METAPREDICTOR and obtained the disorder scores for each amino acid. Those amino acids were labeled as "disordered" if the score was over 0.5 (the default value from the software). Then, if the ratio of disordered amino acids is over 25%, we say that this protein contains a significant proportion of IDRs.

The codes and datasets generated and analyzed during the current study are available on GitHub [87]. All protein sequence data were obtained from the UniProt database. The six proteomes (release 2023_04) used in this paper were indexed by the following proteome IDs and taxonomy IDs: UP000000589_10090, UP000000625_83333, UP000001940_6239, UP000002311_559292, UP000005640_9606, UP000006548_3702. The data associated with the IDR analysis were obtained from the DisProt database version 9.4 (release 2023_06).

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (Grants No. 12205012, No. 12305052, and No. 32000720), the Early Career Scheme (ECS) from Research Grants Council (RGC) of Hong Kong (Grant No. 22302723), Beijing Normal University via the Youth Talent Strategic Program (Grant No. 28705-310432106), and Guangdong Scientific Research Projects for the Higher-Educational Institution (Grant No. 2021ZDZX1011). We thank the Self-Generation Reading Group and the Life-Complexity Reading Group for their support, both organized by the Swarma Club. We also thank Philip Gerlee, Fan Jin, Yoshitsugu Oono, and Da Zhou for their fruitful discussions and insightful comments.

Author contributions were as follows. Conceptualization, Y.L. and Z.Z; methodology, Y.L. and Z.Z.; software, Y.L., Z.Z., and Y.Z.; validation, Z.Z., C.L., and Y.Z.; formal analysis, Z.Z., C.L., and L.P.; investigation, Z.Z., Y.L., C.L., and W.Q.; resources, Y.L., H.L., K.Z., and Z.D.; data curation, Z.Z., C.L., and L.P.; writing—original draft preparation, Z.Z., Y.L., C.L., and W.Q.; writing—review and editing, Z.Z., Y.L., C.L., L.P., W.Q., Q.T., H.L., K.Z., and Z.D.; visualization, Z.Z., Y.L., and C.L.; supervision, Y.L., W.Q., Q.T., and Z.D.; project administration, Y.L. and Z.D.; funding acquisition, Y.L., H.L., K.Z., and Z.D.

[1] L. Yu, D. K. Tanwar, E. D. S. Penha, Y. I. Wolf, E. V. Koonin, and M. K. Basu, Grammar of protein domain architectures, Proc. Natl. Acad. Sci. USA **116**, 3636 (2019).

[2] J. E. Garb and C. Y. Hayashi, Modular evolution of egg case silk genes across orb-weaving spider superfamilies, Proc. Natl. Acad. Sci. USA **102**, 11379 (2005).

[3] Å. K. Björklund, D. Ekman, and A. Elofsson, Expansion of protein domain repeats, PLoS Comput. Biol. **2**, e114 (2006).

[4] J. S. Richman and J. R. Moorman, Physiological time-series analysis using approximate entropy and sample entropy, Am. J. Physiol.-Heart Crculatory Phys. **278**, H2039 (2000).

[5] J. Derr, M. L. Manapat, S. Rajamani, K. Leu, R. Xulvi-Brunet, I. Joseph, M. A. Nowak, and I. A. Chen, Prebiotically plausible mechanisms increase compositional diversity of nucleic acid sequences, Nucleic Acids Res. **40**, 4711 (2012).

[6] J. C. Wootton and S. Federhen, Statistics of local complexity in amino acid sequences and sequence databases, Comput. Chem. **17**, 149 (1993).

[7] S. Vinga, Information theory applications for biological sequence analysis, Brief. Bioinform. **15**, 376 (2014).

[8] C. Adami, The use of information theory in evolutionary biology, Ann. N.Y. Acad. Sci. **1256**, 49 (2012).

[9] S. H. Bertz, On the complexity of graphs and molecules, Bull. Math. Biol. **45**, 849 (1983).

[10] K. Motomura, T. Fujita, M. Tsutsumi, S. Kikuzato, M. Nakamura, and J. M. Otaki, Word decoding of protein amino acid sequences with availability analysis: A linguistic approach, PLoS ONE **7**, e50039 (2012).

[11] K. Shahzad, J. E. Mittenthal, and G. Caetano-Anollés, The organization of domains in proteins obeys Menzerath-Altmann's law of language, BMC Syst. Biol. **9**, 44 (2015).

[12] C. W. Coopmans, K. Kaushik, and A. E. Martin, Hierarchical structure in language and action: A formal comparison, Psychological Rev. **130**, 935 (2023).

[13] A. N. Kolmogorov, Three approaches to the quantitative definition of information, Int. J. Comput. Math. **2**, 157 (1968).

[14] G. J. Chaitin, On the length of programs for computing finite binary sequences, J. ACM **13**, 547 (1966).

[15] H. Zenil, S. Hernández-Orozco, N. A. Kiani, F. Soler-Toscano, A. Rueda-Toicen, and J. Tegnér, A decomposition method for global evaluation of Shannon entropy and local estimations of algorithmic complexity, Entropy **20**, 605 (2018).

[16] H. Zenil, F. Soler-Toscano, K. Dingle, and A. A. Louis, Correlation of automorphism group size and topological properties with program-size complexity evaluations of graphs and complex networks, Physica A **404**, 341 (2014).

[17] S. Hernández-Orozco, N. A. Kiani, and H. Zenil, Algorithmically probable mutations reproduce aspects of evolution, such as convergence rate, genetic memory and modularity, R. Soc. Open Sci. **5**, 180399 (2018).

[18] H. Zenil, N. A. Kiani, F. Marabita, Y. Deng, S. Elias, A. Schmidt, G. Ball, and J. Tegnér, An algorithmic information calculus for causal discovery and reprogramming systems, iScience **19**, 1160 (2019).

[19] J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, IEEE Trans. Inf. Theory **24**, 530 (1978).

[20] R. Cilibrasi and P. M. Vitanyi, Clustering by compression, IEEE Trans. Inf. Theory **51**, 1523 (2005).

[21] S. E. Ahnert, I. G. Johnston, T. M. A. Fink, J. P. K. Doye, and A. A. Louis, Self-assembly, modularity, and physical complexity, Phys. Rev. E **82**, 026117 (2010).

[22] I. G. Johnston, K. Dingle, S. F. Greenbury, C. Q. Camargo, J. P. K. Doye, S. E. Ahnert, and A. A. Louis, Symmetry and simplicity spontaneously emerge from the algorithmic nature of evolution, Proc. Natl. Acad. Sci. USA **119**, e2113883119 (2022).

[23] F. Jacob, Evolution and tinkering, Science **196**, 1161 (1977).

[24] R. Solé and S. Valverde, Evolving complexity: How tinkering shapes cells, software and ecological networks, Philos. Trans. R. Soc. B **375**, 20190325 (2020).

[25] S. J. Hoyt *et al.*, From telomere to telomere: The transcriptional and epigenetic state of human repeat elements, Science **376**, eabk3112 (2022).

[26] T. Marques-Bonet and E. Eichler, The evolution of human segmental duplications and the core duplicon hypothesis, *Cold Spring Harbor Symposia on Quantitative Biology* (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, 2009), Vol. 74, pp. 355–362.

[27] S. K. Garushyants, I. B. Rogozin, and E. V. Koonin, Template switching and duplications in SARS-CoV-2 genomes give rise to insertion variants that merit monitoring, Commun. Biol. **4**, 1343 (2021).

[28] T. Marques-Bonet, S. Girirajan, and E. E. Eichler, The origins and impact of primate segmental duplications, Trends Genet. **25**, 443 (2009).

[29] M. Lynch and J. S. Conery, The origins of genome complexity, Science **302**, 1401 (2003).

[30] M. L. Pedulla *et al.*, Origins of highly mosaic mycobacteriophage genomes, Cell **113**, 171 (2003).

[31] L. Journet, C. Agrain, P. Broz, and G. R. Cornelis, The needle length of bacterial injectisomes is determined by a molecular ruler, Science **302**, 1757 (2003).

[32] L. Chen, A. L. DeVries, and C.-H. Cheng, Evolution of antifreeze glycoprotein gene from a trypsinogen gene in antarctic notothenioid fish, Proc. Natl. Acad. Sci. USA **94**, 3811 (1997).

[33] G. Andrews *et al.*, Mammalian evolution of human cis-regulatory elements and transcription factor binding sites, Science **380**, eabn7930 (2023).

[34] C. Bekpen and D. Tautz, Human core duplicon gene families: Game changers or game players? Briefings Funct. Genomics **18**, 402 (2019).

[35] Y. Liu, Z. Di, and P. Gerlee, Ladderpath approach: How tinkering and reuse increase complexity and information, Entropy **24**, 1082 (2022).

[36] D. Knuth,Evaluation of powers, in *Art of Computer Programming, Volume 2: Seminumericnl Algorithms*, 3rd ed. (Addison-Wesle Professional, Boston, 1997), pp. 75–81.

[37] S. M. Marshall, A. R. Murray, and L. Cronin, A probabilistic framework for identifying biosignatures using pathway complexity, Philos. Trans. R. Soc. A **375**, 20160342 (2017).

[38] Y. Liu, C. Mathis, M. D. Bajczyk, S. M. Marshall, L. Wilbraham, and L. Cronin, Exploring and mapping chemical space with molecular assembly trees, Sci. Adv. **7**, eabj2465 (2021).

[39] S. A. Kauffman, *A World Beyond Physics: The Emergence and Evolution of Life* (Oxford University, New York, 2019).

[40] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevResearch.6.023215 for the supplemental material contains nine sections, including details on mathematics and data, with references made at relevant points in the main text.

[41] E. Schad, P. Tompa, and H. Hegyi, The relationship between proteome size, structural disorder and organism complexity, Genome Biol. **12**, R120 (2011).

[42] S. B. Carroll, Chance and necessity: The evolution of morphological complexity and diversity, Nature (London) **409**, 1102 (2001).

[43] M. C. McCarthy and B. J. Enquist, Organismal size, metabolism and the evolution of complexity in metazoans, Evol. Ecology Res. **7**, 681 (2005).

[44] M. Pellegrini, M. E. Renda, and A. Vecchio, *Ab initio* detection of fuzzy amino acid tandem repeats in protein sequences, in *BMC Bioinformatics*, (BioMed Central, New York, 2012), Vol. 13, pp. 1–13.

[45] J. Mollenhauer, U. Holmskov, S. Wiemann, I. Krebs, S. Herbertz, J. Madsen, P. Kioschis, J. Coy, and A. Poustka, The genomic structure of the DMBT1 gene: Evidence for a region with susceptibility to genomic instability, Oncogene **18**, 6233 (1999).

[46] J. Watari, Y. Takata, M. Ogawa, H. Sahara, S. Koshino, M.-L. Onnela, U. Airaksinen, R. Jaatinen, M. Penttilä, and S. Keränen, Molecular cloning and analysis of the yeast flocculation gene FLO1, Yeast **10**, 211 (1994).

[47] J. C. Byrd and R. S. Bresalier, Mucins and mucin binding proteins in colorectal cancer, Cancer Metastasis Rev. **23**, 77 (2004).

[48] A. V. Kajava, Tandem repeats in proteins: From sequence to structure, J. Struct. Biol. **179**, 279 (2012).

[49] S. J. Caldwell, I. C. Haydon, N. Piperidou, P.-S. Huang, M. J. Bick, H. S. Sjöström, D. Hilvert, D. Baker, and C. Zeymer, Tight and specific lanthanide binding in a de novo TIM barrel with a large internal cavity designed by symmetric domain fusion, Proc. Natl. Acad. Sci. USA **117**, 30362 (2020).

[50] F. Quaglia *et al.*, Disprot in 2022: Improved quality and accessibility of protein intrinsic disorder annotation, Nucl. Acids Res. **50**, D480 (2022).

[51] R. J. Emenecker, D. Griffith, and A. S. Holehouse, Metapredict: A fast, accurate, and easy-to-use predictor of consensus disorder and structure, Biophys. J. **120**, 4312 (2021).

[52] J. Jorda, B. Xue, V. N. Uversky, and A. V. Kajava, Protein tandem repeats–The more perfect, the less structured, FEBS J. **277**, 2673 (2010).

[53] M. Simon and J. M. Hancock, Tandem and cryptic amino acid repeats accumulate in disordered regions of proteins, Genome Biol. **10**, R59 (2009).

[54] P. Tompa, Intrinsically unstructured proteins evolve by repeat expansion, BioEssays **25**, 847 (2003).

[55] N. W. Van Bibber, C. Haerle, R. Khalife, G. W. I. Dayhoff, and V. N. Uversky, Intrinsic disorder in human proteins encoded by core duplicon gene families, J. Phys. Chem. B **124**, 8050 (2020).

[56] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker, Intrinsic disorder and functional proteomics, Biophys. J. **92**, 1439 (2007).

[57] V. Vacic, V. N. Uversky, A. K. Dunker, and S. Lonardi, Composition profiler: A tool for discovery and visualization of amino acid composition differences, BMC Bioinf. **8**, 211 (2007).

[58] S. Martín-Villanueva, G. Gutiérrez, D. Kressler, and J. de la Cruz, Ubiquitin and ubiquitin-like proteins and domains in ribosome production and function: Chance or necessity? Int. J. Mol. Sci. **22**, 4359 (2021).

[59] M.-L. Bang *et al.*, The complete gene sequence of titin, expression of an unusual ≈700-kDa titin isoform, and its interaction with obscurin identify a novel Z-line to I-band linking system, Circ. Res. **89**, 1065 (2001).

[60] W. A. Linke, M. Ivemeyer, P. Mundel, M. R. Stockmeier, and B. Kolmerer, Nature of PEVK-titin elasticity in skeletal muscle, Proc. Natl. Acad. Sci. USA **95**, 8052 (1998).

[61] W. Guo, S. J. Bharmal, K. Esbona, and M. L. Greaser, Titin diversity—Alternative splicing gone wild, BioMed Res. Int. **2010**, 753675 (2010).

[62] K. Muenzen, J. Monroy, and F. R. Finseth, Evolution of the highly repetitive PEVK region of titin across mammals, G3: Genes, Genomes, Genetics **9**, 1103 (2019).

[63] M. S. O'Bleness, C. M. Dickens, L. J. Dumas, H. Kehrer-Sawatzki, G. J. Wyckoff, and J. M. Sikela, Evolutionary history and genome organization of DUF1220 protein domains, G3: Genes, Genomes, Genetics **2**, 977 (2012).

[64] L. J. Dumas *et al.*, DUF1220-domain copy number implicated in human brain-size pathology and evolution, Am. J. Human Genet. **91**, 444 (2012).

[65] S. G. Gregory *et al.*, The DNA sequence and biological annotation of human chromosome 1, Nature (London) **441**, 315 (2006).

[66] Q.-Y. Tang, W. Ren, J. Wang, and K. Kaneko, The statistical trends of protein evolution: A lesson from alphafold database, Mol. Biol. Evol. **39**, msac197 (2022).

[67] T. R. Gregory, A bird's-eye view of the c-value enigma: Genome size, cell size, and metabolic rate in the class aves, Evolution **56**, 121 (2002).

[68] J. Mattick and P. Amaral, *RNA, the Epicenter of Genetic Information: A New Understanding of Molecular Biology* (Taylor & Francis, London, 2022), Ch. 7, pp. 77–90.

[69] C. Chothia, J. Gough, C. Vogel, and S. A. Teichmann, Evolution of the protein repertoire, Science **300**, 1701 (2003).

[70] M. Y. Dennis *et al.*, The evolution and population diversity of human-specific segmental duplications, Nat. Ecol. Evol. **1**, 0069 (2017).

[71] M. R. Vollger *et al.*, Segmental duplications and their variation in a complete human genome, Science **376**, eabj6965 (2022).

[72] Y. He and Y. Mao, Exploring the primate genome: Unraveling the mysteries of evolution and human disease, Innovation **4**, 100467 (2023).

[73] J. M. Sikela and F. Van Roy, Changing the name of the NBPF/DUF1220 domain to the Olduvai domain, F1000Research **6**, 2185 (2018).

[74] L. M. Field and A. L. Devonshire, Evidence that the E4 and FE4 esterase genes responsible for insecticide resistance in the aphid Myzus persicae (Sulzer) are part of a gene family, Biochem. J. **330**, 169 (1998).

[75] D. W. Loehlin and S. B. Carroll, Expression of tandem gene duplicates is often greater than twofold, Proc. Natl. Acad. Sci. USA **113**, 5988 (2016).

[76] C. Shao *et al.*, The enormous repetitive Antarctic krill genome reveals environmental adaptations and population insights, Cell **186**, 1279 (2023).

[77] M. Kimura and T. Ohta, Stepwise mutation model and distribution of allelic frequencies in a finite population, Proc. Natl. Acad. Sci. USA **75**, 2868 (1978).

[78] A. Stoltzfus, On the possibility of constructive neutral evolution, J. Mol. Evol. **49**, 169 (1999).

[79] T. C. Scott-Phillips, K. N. Laland, D. M. Shuker, T. E. Dickins, and S. A. West, The niche construction perspective: A critical appraisal, Evolution **68**, 1231 (2014).

[80] R. Belshaw, V. Pereira, A. Katzourakis, G. Talbot, J. Pačes, A. Burt, and M. Tristem, Long-term reinfection of the human genome by endogenous retroviruses, Proc. Natl. Acad. Sci. USA **101**, 4894 (2004).

[81] H. Ochman, J. G. Lawrence, and E. A. Groisman, Lateral gene transfer and the nature of bacterial innovation, Nature (London) **405**, 299 (2000).

[82] J. Spring, Major transitions in evolution by genome fusions: From prokaryotes to eukaryotes, metazoans, bilaterians and vertebrates, J. Struct. Funct. Genomics **3**, 19 (2003).

[83] M. S. Newton, V. L. Arcus, M. L. Gerth, and W. M. Patrick, Enzyme evolution: innovation is easy, optimization is complicated, Curr. Opin. Struct. Biol. **48**, 110 (2018).

[84] D. P. Fewer and M. Metsä-Ketelä, A pharmaceutical model for the molecular evolution of microbial natural products, FEBS J. **287**, 1429 (2020).

[85] L. Comai and E. H. Tan, Haploid induction and genome instability, Trends Genet. **35**, 791 (2019).

[86] M. S. Dar, B. B. Dholakia, A. P. Kulkarni, P. S. Oak, D. Shanmugam, V. S. Gupta, and A. P. Giri, Influence of domestication on specialized metabolic pathways in fruit crops, Planta **253**, 61 (2021).

[87] https://github.com/yuernestliu/LadderpathCalculator.