Grammar of protein domain architectures

Lijia Yu[a], Deepak Kumar Tanwar[a,1], Emanuel Diego S. Penha[a,2], Yuri I. Wolf[b], Eugene V. Koonin[b,3], and Malay Kumar Basu[a,3]

[a]Department of Pathology, University of Alabama, Birmingham, AL 35249; and [b]National Center for Biotechnology Information, National Institutes of Health, Bethesda, MD 20894

Edited by Clyde A. Hutchison III, J. Craig Venter Institute, La Jolla, CA, and approved January 4, 2019 (received for review August 27, 2018)

From an abstract, informational perspective, protein domains appear analogous to words in natural languages in which the rules of word association are dictated by linguistic rules, or grammar. Such rules exist for protein domains as well, because only a small fraction of all possible domain combinations is viable in evolution. We employ a popular linguistic technique, *n*-gram analysis, to probe the "proteome grammar"—that is, the rules of association of domains that generate various domain architectures of proteins. Comparison of the complexity measures of "protein languages" in major branches of life shows that the relative entropy difference (information gain) between the observed domain architectures and random domain combinations is highly conserved in evolution and is close to being a universal constant, at ∼1.2 bits. Substantial deviations from this constant are observed in only two major groups of organisms: a subset of Archaea that appears to be cells simplified to the limit, and animals that display extreme complexity. We also identify the *n*-grams that represent signatures of the major branches of cellular life. The results of this analysis bolster the analogy between genomes and natural language and show that a "quasi-universal grammar" underlies the evolution of domain architectures in all divisions of cellular life. The nearly universal value of information gain by the domain architectures could reflect the minimum complexity of signal processing that is required to maintain a functioning cell.

n-gram | bigram | protein domain | language | domain architecture

Ever since the inception of the human genome project, the metaphorical expression "book of life," denoting the genome sequence, has captured imaginations of both scientific and lay communities (1–3). Extending the analogy of a genome to a book (a text, or a corpus in linguistics), we can think of amino acid residues as letters, protein domains as words, and proteins as sentences consisting of ordered arrangements of protein domains (domain architectures) (4).

Genomes show remarkable similarities to natural languages. Like all cellular life forms, all natural languages are believed to have descended from a single ancestor (5) and have evolved through mechanisms comparable to biological evolution (6). In a written language, individual letters cannot carry semantic information; the smallest unit of information, therefore, is a word (7, 8). Protein domains are structural, functional, and evolutionary units of proteins (9, 10) and are thus analogous to words. This analogy is reflected in the statistical properties of the domain repertoires of diverse organisms. The frequency distribution of domains encoded in any genome follows a power law (9, 11–13). Power-law distributions have been found in numerous natural and social contexts, including a broad variety of biological systems (14–16). An important variant of power-law distributions is Zipf's law, which describes the frequency distribution of words in natural languages (17). The slope of the curve in Zipf's law for a natural language is approximately −1 (7), which is close to the slope of domain frequency distribution in a genome (1). Additionally, bigram (defined as two consecutive words; nonconsecutive two word combinations that co-occur in a sentence do not count as bigrams) frequency distributions in natural languages also follow power laws; in this case, with a slope of approximately −2 (18). A similar value has been reported for protein domain bigrams (19).

The function of a protein, to a large extent, is determined by the arrangement of its constituent domains—that is, its domain architecture (13, 19). All life forms possess many multidomain proteins, but both the number of unique domains and the fraction of multidomain proteins increase with the organismal complexity (defined as the number of unique cell types in an organism): Eukaryotes have more multidomain proteins than prokaryotes (4, 9, 20–25), and animals have more multidomain proteins than unicellular eukaryotes (26). This trend of increased multidomain protein formation with increasing organismal complexity is known as domain accretion (27) and apparently plays a major role in evolution, particularly in major evolutionary transitions such as the origin of multicellularity (28–33). Of the numerous possible domain combinations, only a limited subset is actually represented in genomes, suggesting that domain architectures are shaped by natural selection (10, 19, 34). It is, therefore, imperative to decipher the rules of association of protein domains.

The smallest information unit in a language is a word, and a grammar is the set of rules regulating the association of words. Given that protein domains are analogous to words, the rules of

Significance

Genomes appear similar to natural language texts, and protein domains can be treated as analogs of words. To investigate the linguistic properties of genomes further, we calculated the complexity of the "protein languages" in all major branches of life and identified a nearly universal value of information gain associated with the transition from a random domain arrangement to the current protein domain architecture. An exploration of the evolutionary relationship of the protein languages identified the domain combinations that discriminate between the major branches of cellular life. We conclude that there exists a "quasi-universal grammar" of protein domains and that the nearly constant information gain we identified corresponds to the minimal complexity required to maintain a functional cell.

Author contributions: M.K.B. designed research; L.Y., D.K.T., E.D.S.P., and M.K.B. performed research; L.Y., D.K.T., Y.I.W., E.V.K., and M.K.B. analyzed data; Y.I.W., E.V.K., and M.K.B. interpreted results; M.K.B. wrote the paper; and E.V.K. and M.K.B. revised and rewrote the manuscript.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: All data and code are available from the GitHub repository, https://github.com/malaybasu/2019-domain_pnas.

[1]Present address: Laboratory of Neuroepigenetics, Medical Faculty of the University of Zürich and Department of Health Sciences and Technology of the Swiss Federal Institute of Technology Zürich, CH-8057 Zürich, Switzerland.

[2]Present address: Department of Nutrition, UniFanor Wyden, 60191-156, Dunas, Fortaleza CE, Brazil.

[3]To whom correspondence may be addressed. Email: malay@uab.edu or koonin@ncbi.nlm.nih.gov.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1814684116/-/DCSupplemental.

Published online February 7, 2019.

association, or the "grammar" of proteins, can be investigated using tools borrowed from linguistics. The simplest way to explore the grammar of an unknown language is to perform an n-gram analysis, a probabilistic language-modeling technique whereby consecutive words in sentences are treated as a unit to identify meaningful word associations. Depending on the number of words (n) in the unit, the analysis can be unigram ($n = 1$), bigram ($n = 2$), trigram ($n = 3$), and so forth. Such modeling allows determination of the conditional probabilities of a word, given the previous word (s). n-Gram language modeling has been widely employed in various text processing applications and speech recognition (7, 8).

Previously, we introduced an informal bigram analysis to explore the evolution of protein domain promiscuity—that is, the tendency of some domains to participate in many different domain architectures. A pair of domains on a protein sequence was considered a bigram, and bigram frequencies were calculated to measure promiscuity of domains in all major branches of the eukaryotic evolutionary tree (13, 19). The concept of bigrams, as applied to protein domains, has since been widely employed in studies on the evolution of protein domain architectures (35–38).

A formal n-gram modeling of domains provides a probabilistic framework for deciphering the rules of domain combination in multidomain protein architectures. Here, we analyze bigram models from all major branches of cellular life and probe the evolutionary characteristics of these models. Such modeling yields an objective measure of genome complexity from the information theory perspective and provides tools to study the evolution of complexity. Domain rearrangements and, in particular, domain accretion made major contributions to evolutionary transitions such as the origin of eukaryotes and, subsequently, the origin of multicellular organisms. Estimation of entropy (a measure of complexity) changes accompanying these events provides a quantitative framework for the analysis of these crucial aspects of evolution. We show that the loss of entropy (information gain) resulting from domain arrangements in genomes is nearly constant across the entire course of cellular life evolution and identify both similarities and dissimilarities between the "language" of proteins and natural languages.

Results and Discussion

Domain *n*-Gram Modeling. Complete proteome data from the three superkingdoms of life (Bacteria, Archaea, and Eukaryota—often called domains of cellular life, a term that we do not use here to denote taxa so as to avoid confusion with protein domains), were downloaded from the UniProt (39) database (see *Methods* and Datasets S1 and S2), and domains were mapped onto the sequence of each protein using HMMER3 (40) and the Pfam database (41). Altogether, we identified about 23 million domains across 4,794 species. The domain maps were filtered (see *Methods*) to generate nonoverlapping orders of domains for each protein. Having constructed these domain maps, we proceeded to generate n-gram models for these genomes.

The n-gram models are made by calculating the conditional probability of one word (domain), given the proximal adjacent words. Depending on the number of words considered (the order), n-gram models are called unigrams (one word), bigrams (two words), trigrams (three words), and so forth. In this work, we used only bigram models, not only because these are the easiest to analyze computationally but also because the fractions of proteins with three or more domains are low in most organisms, particularly in prokaryotes (42). Thus, the higher-order n-gram models would include many missing probabilities and therefore become uninformative. The bigram models were constructed using Eq. **1**, after adding the faux markers "N" and "C" to the beginning and end of each protein sequence, respectively. We also made models without the additional markers to control for the effect of these additions. Unless otherwise stated, the results are from the models with the N and C markers. In addition, we shuffled the domains in each genome and constructed bigram models from these shuffled genomes (see *Methods*).

Domains are present in restricted contexts; only a small subset of the conditional probabilities of all possible bigrams assume nonzero values. More than 95% of all possible bigrams are not found in any genome (Dataset S3). Thus, for the analyses in which normalized data are important, such as phylogenetic tree calculation (see below), we used Good–Turing smoothing (43, 44) (see *Methods*) to assign nonzero probability values to all possible bigrams.

Entropy of *n*-Gram Models and Their Evolutionary Trends. There are ~7,000 languages in the world, divided into 19 linguistic families (45, 46). Notwithstanding the differences, linguistic universals have been identified in the grammar and vocabularies among all these languages (47). In a natural language, the symbols are concatenated under the constraints of syntactic rules of grammar. The resulting sequence shows a balance between order and disorder. A rigorous measure of the degree of order can be obtained by calculating the entropy of such sequences (48).

Entropic concepts are intimately linked to the concept of complexity in biology (49, 50), which has been defined in various ways. Entropic measurements of nucleotide and protein sequences have shown that entropy increases in the course of evolution (51), and this increase has been explained as a by-product of genome evolution via a largely nonadaptive, stochastic process (52). This view of evolution is buttressed by the existence of "universal laws"—that is, conserved patterns of evolutionary change that recur in all major divisions of life and do not appear to represent direct adaptations (53, 54).

We calculated the entropy of the protein language model from each of the analyzed genomes (Dataset S3). We considered unigram models as null models, whereby the entropies are determined by the frequencies of individual domains in the genome. The entropies of the unigram and bigram models were calculated using Eqs. **3** and **4**, respectively. We then investigated the relationships between the obtained entropies and various characteristics of proteomes—such as the number of unique domain families in a genome, domain count, protein count, and amino acid count—in each of the three superkingdoms of cellular life (Fig. 1 and *SI Appendix*, Figs. S1 and S2) and in five major kingdoms (*SI Appendix*, Figs. S3 and S4). The relationships were nonlinear in many cases, most prominently in Bacteria, but converting the values of the proteomic variables to the $\log_{10}$ scale resulted in a significant improvement of the linear regression coefficients (Fig. 1, *SI Appendix*, Fig. S1, and Dataset S4). As expected, for all three superkingdoms, the unigram entropy shows significant positive correlations with the number of domain families, total count of domain, number of proteins, and total amino acid count, each of which can be considered a proxy for the proteome size (and the genome size in prokaryotes) (Fig. 1 and *SI Appendix*, Fig. S1). Similarly, the entropy of natural languages is known to increase with the vocabulary size (8).

For unigram entropies, the slopes of the regression lines are much greater for the number of domain families (or types) than they are for the other variables (Fig. 1 and *SI Appendix*, Fig. S1) because the entropy of the n-gram models, by design, is primarily determined by the diversity of domains in a genome. All regression slopes in Archaea are greater than the respective slopes in Bacteria, suggesting that, in archaeal evolution, increase in the genome size typically leads to a more pronounced innovation in the domain repertoire as compared to bacteria. This effect might be linked to the massive influx of bacterial genes that is thought to have occurred independently in several major groups of Archaea (55). The slope of the regression curve of the unigram entropy on the number of unique domain families in eukaryotes is considerably greater than in prokaryotes, indicating that the growth of the domain repertoire in eukaryotes results in more uniform domain frequency distributions than in either Bacteria or Archaea (Fig. 1).

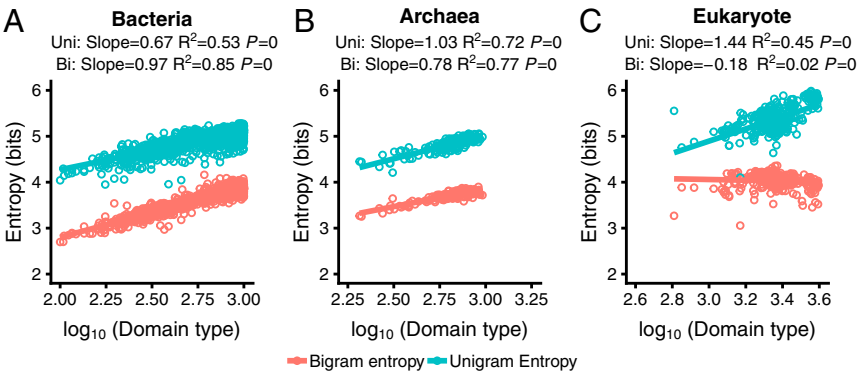

Fig. 1. Log-linear regression of the unigram and bigram entropies with domain families (unique domain types in the genome) in the three superkingdoms of life: Bacteria (*A*), Archaea (*B*), and Eukaryota (*C*). Each point represents a genome. Some points are removed to keep all the figures in the same scale. See *SI Appendix*, Fig. S1 for the full data and regressions with other genomic variables. The *x* axis is converted to the $\log_{10}$ scale. See *SI Appendix*, Fig. S2 for the raw regression data. The slopes of the regression lines, R^2, and *P* values are indicated on top of each plot.

However, all of the other regression lines (*SI Appendix*, Fig. S1) for unigram entropies are notably flatter in eukaryotes than they are in prokaryotes, which is likely to reflect the substantially greater contribution of gene duplication to the increase of the proteome size in eukaryotes compared with prokaryotes (56–60). Analysis of individual phyla of Bacteria and Archaea shows the same trends as the bulk analysis, whereas among the eukaryotic kingdoms, unigram entropies decrease with the increase in genome size both in fungi and in plants (*SI Appendix*, Figs. S3 and S4). This trend is likely to stem from the major contribution of whole-genome duplications that are common in fungi and plants (61).

The bigram entropy regressions show clear differences between the two prokaryotic superkingdoms and eukaryotes (Fig. 1 and *SI Appendix*, Fig. S1). In prokaryotes, the bigram regression lines are roughly parallel to those for the unigrams, indicating that the diversification of domain combinations follows the growth of the domain repertoire which, at least in principle, is compatible with the notion that multidomain architectures evolve through random domain combination (11). In contrast, in eukaryotes (in bulk and in all individual kingdoms; *SI Appendix*, Figs. S1 and S4), the bigram entropy regression slopes are slightly negative, indicating that, with the increasing size of the proteomes, they tend to become more ordered—that is, the distributions of domain architectures become increasingly skewed. Most likely, this pattern is due to the proliferation of favorable domain combinations by gene duplication in complex multicellular eukaryotes, such as animals and plants. A striking example of such a bigram is the extensive amplification of nucleotide-binding and leucine-rich repeat proteins that combine an NTPase domain with an array of leucine-rich repeats (and, in some cases, additional domains), which are essential for both plant and animal innate immunity but, apparently, have evolved independently in plants and animals (62).

Relative Entropy of Protein Language Models. The entropy of a language model indicates how much information is carried by the symbols of a given language in a particular text. The higher the entropy, the more uncertain we are about the information carried by the text (7, 8). The symbols can be alphabets, words, lines, or even the full corpus. A surprising and yet unexplained observation is that all known natural languages possess a nearly constant relative entropy, which is a measure of entropy loss (information gain) between a text written in the given language following the strict rules of grammar and a random sequence of words (63–66). It has been observed that for all natural languages, the information gain is about 3.6 bits, which is compatible with the existence of a universal grammar, despite some distinct, language-specific variations (46, 63).

We compared the relative entropy values (H_g in Eq. **5**) of protein languages across the major prokaryotic and eukaryotic taxa (Dataset S3). Because the unigram entropy is derived from frequencies of individual domains in a genome, it can be considered the entropy of a random disorganized genome. Thus, we calculated the relative entropy (information gain) by subtracting the bigram entropy from the unigram entropy for each genome (H_g in Eq. **5**). The difference between the unigram and bigram entropy measures the amount of information that is gained upon transition from a random collection of domains in the genome (unigrams) to the observed domain architectures (bigrams). This difference in entropy is a measure of the order imposed on the domain architectures by the rules of domain association forced by the biological functions that are relevant for the particular organism—that is, the grammar of the protein language. Clearly, the relative entropy calculated using only bigrams is but an approximation that ignores the information gain from more complex domain architectures (trigrams, tetragrams, etc.). However, given the relatively low fraction of proteins with more than two domains in proteomes (9), these relative entropy values can be expected to accurately reflect proteomic complexity.

In both the unigram and the bigram entropy distributions, the median values increase in the following order: Archaea < Bacteria < Eukaryota (Fig. 2*A* and Dataset S5). This trend is not surprising because archaeal genomes are typically smaller in size and encode fewer domain families than bacterial, let alone eukaryotic genomes (67). The median values of the relative entropy (H_g) follow the same order: Archaea (1.13 bits) < Bacteria (1.21 bits) < Eukaryota (1.28 bits) (Fig. 2*A* and Dataset S5). The differences between these median values of the relative entropy for the three superkingdoms are statistically significant according to the permutation test (Dataset S6). Nevertheless, the three distributions strongly overlap as shown by counting discordant points and calculating Bhattacharyya coefficients (68) for pairs of distributions (Fig. 2*A* and Dataset S6).

For both Archaea and Eukaryota, the distributions of relative entropies are bimodal. The bimodality of the distribution in Archaea is mainly due to the difference between two groups, one of which consists of Euryarchaeota and Nanoarchaeota, and the other consisting of Crenarchaeota and other archaeal taxa. Euryarchaeota and Nanoarchaeota show an information gain value close to that in Bacteria, ∼1.2 bits (Fig. 2*B* and Dataset S5), whereas the rest of the archaea have a lower value of ∼1.04 bits. Thus, these archaea are characterized by anomalously low proteomic complexity. In eukaryotes, the two peaks correspond to plants and fungi (∼1.2 bits) and animals (>1.6 bits) (Fig. 2*B* and Dataset S5). Thus, animals show the highest information gain among the analyzed groups, in accord with the notion that domain architectures in animals are more elaborate and evolve under stronger constraints than those in other organisms (27). In contrast to archaea and eukaryotes, bacterial phyla exhibit remarkable conservation of relative entropy: Except Tenericutes, all analyzed bacteria have similar relative entropy close to ∼1.2 bits.

The above calculations of entropies are based on *n*-gram models with added N and C markers that potentially could bias the entropy calculations, especially for smaller genomes. To

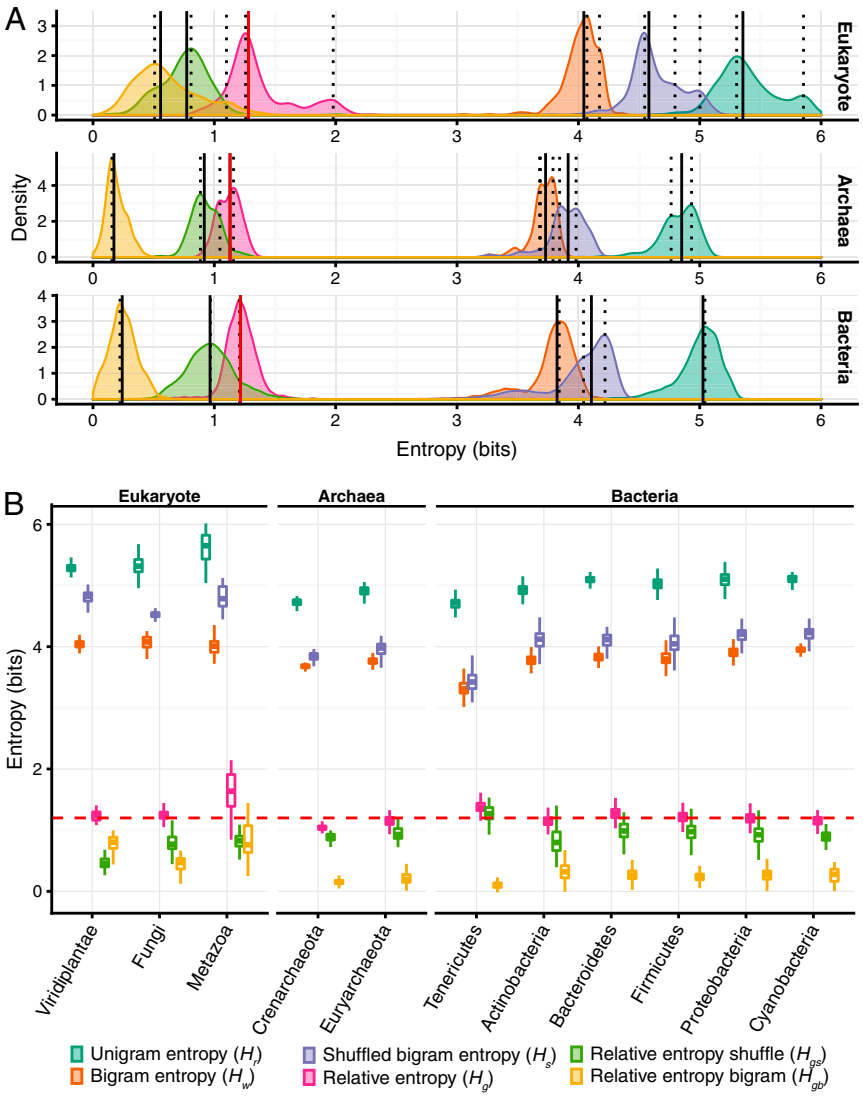

Fig. 2. Distributions of the unigram, bigram, and the three relative entropies. (*A*) Density plot of entropy values for the three superkingdoms. Each panel represents one superkingdom (from top to bottom: Eukaryota, Archaea, and Bacteria). Peaks in distributions are marked with dotted lines. The median values are indicated with solid lines. The median values of relative entropies (H_g) are marked with solid red lines. The *x* axis represents entropy in bits. (*B*) Box plots of the unigram, bigram, and the three relative entropies in three eukaryotic kingdoms (green plants, fungi, animals); two archaeal phyla (Crenarchaeota and Euryarchaeota); and six bacterial phyla (Tenericutes, Actinobacteria, Bacteroidetes, Firmicutes, Proteobacteria, and Cyanobacteria). For the full list of median entropy values in all of the groups, see Dataset S5. The dashed horizontal red line represents the near-universal relative entropy of 1.2 bits.

control for any such effect, we calculated the relative entropies without using the end markers (Dataset S3). As expected, this approach resulted in a substantial increase in unigram, bigram, and relative entropy. Nevertheless, the relative entropy values across all taxa were closely similar, ∼7 bits; moreover, all of the clade-specific trends noticed with the first approach remained valid (*SI Appendix*, Fig. S5).

Thus, apart from the two notable deviations, namely, the low information gain (low complexity) in a subset of Archaea and the high information gain (high complexity) in animals, the median relative entropies lie within a narrow range between ∼1.1 and ∼1.3 bits in many groups of highly diverse organisms. Together, these findings suggest the existence of a "quasi-universal" grammar of protein domain architecture.

The difference in entropy between the unigram and the bigram distributions comes from two constraints on the bigram distributions: first, the genome-specific distribution of proteins by the number of domains (relative frequencies of single- and multidomain proteins); and second, the biologically permissible and preferred domain combinations in the multidomain proteins. To differentiate between these two effects, we shuffled the domains in each genome, keeping constant the number of proteins and the number of domains in each protein (see *Methods*). This shuffling procedure does not change the unigram entropies of the genomes because the frequencies of domains do not change, but the shuffling changes the bigram entropies because the domain combinations are randomized. We then subtracted the shuffled bigram entropy (H_s) from the unigram entropy (H_r) to estimate the relative shuffled entropy (H_{gs} in Eq. **6**) and subtracted the bigram entropy before shuffling (H_w) from the shuffled bigram entropy (H_s) to derive the relative bigram entropy (H_{gb} in Eq. **7**) (Fig. 2 and Dataset S5). The bigram entropies calculated from these shuffled genomes (H_s) are, as expected, lower than the corresponding unigram entropies (H_r) but higher than the empirical bigram entropies (H_w) (Fig. 2). The only exception is Nanoarchaeota, where H_s is equal to or even slightly less than H_w. The bigram entropy gain due to randomization (relative bigram entropy; H_{gb} in Eq. **7**) should be less in smaller genomes with fewer multidomain proteins. This is indeed the case, with Eukaryota having higher H_{gb} (0.56 bits), compared with Archaea (0.17 bits) and Bacteria (0.24 bits) (Fig. 2 and Dataset S5). This difference measures the information gain due to nonrandom, biologically meaningful domain combinations that are maintained by selection. In contrast, the difference between the unigram and shuffled bigram entropies (relative shuffled entropy; H_{gs} in Eq. **6**) reflects the contribution of the global domain architecture—that is, the distribution of domains among the existing number of proteins. We found these values to be lower in Eukaryota (0.77 bits) than in

Archaea (0.92 bits) and Bacteria (0.96 bits). Thus, in complex organisms, the effect of the global domain architecture, although greater than the contribution from specific domain combinations, plays a relatively less important role.

Using Cross-Entropy of Bigram Models to Build an Evolutionary Tree. Several studies, including our own earlier work, have shown that domain frequency as well as domain architectures carry phylogenetic information (19, 69, 70). Therefore, it could be expected that this signal strengthens with the refinement of models of domain architecture. A domain-based phylogeny might be helpful to address certain long-standing questions in evolution, such as finding the root of the eukaryotic tree. Such problems are difficult to solve using traditional phylogenetic methods, and there is considerable interest in harnessing rare genomic changes (RGCs) for this purpose (71), given that they are less prone to various phylogeny-construction artifacts (71–73). Such RGCs could include diagnostic domain architectures—that is, taxon-specific domain combinations that are features for specific taxa.

A probabilistic language model can be used to determine the probability of a given genome by computing the conditional probabilities of all domain combinations it encodes. Given a set of n-gram models, we can calculate which model predicts the highest probability for a given genome. This value can be calculated directly by measuring perplexity, or the cross-entropy (7, 8), of a given genome under all the models. Perplexity is a measure of how well a given n-gram model describes a language. The model that has the maximum prediction power (lowest cross-entropy or lowest perplexity) is considered optimal.

Given two n-gram models generated from two separate genomes and a target genome, the cross-entropy calculation allows one to select the model that has a higher probability, or lower perplexity, for the target genome (see Eq. **8** and *Methods*). Domain models created from phylogenetically closer taxa are expected to possess higher predictive power (lower perplexity) compared with distant taxa. Thus, the pairwise cross-entropies can be represented as distances and, accordingly, can be used to create a whole-genome phylogenetic tree.

We calculated pairwise cross-entropies for 37 selected eukaryotic genomes (Dataset S2; see *Methods* for the selection procedure), and the resulting values were used to build a distance tree (Eq. **9** and Fig. 3). We focused on eukaryotes because of the well-defined topology of the main clades as opposed to the case of prokaryotes. The tree (Fig. 3) exhibits a near-perfect separation of the established major groups of eukaryotes. Depending on the placement of the root, the tree can be interpreted as being compatible with the unikont–bikont topology (74, 75). Under this root placement, the eukaryotic tree consists of three major clades: (*i*) bikonts, which include Archaeplastida (Viridiplantae and Rhodophyta) and Apicomplexa; (*ii*) unikonts (Amoebozoa, fungi, and animals); and (*iii*) Excavata (Diplomonadida and Kinetoplastida) (74, 76). The internal branching in the tree is also compatible with modern phylogenies. Examples include monophyly of mammals (human and mouse), insects (*Apis mellifera*, *Drosophila melanogaster*, and *Anopheles gambei*), fungi (*Aspergillus nidulans*, *Neurospora crassa*, *Schizosaccharomyces pombe*, *Saccharomyces cerevisiae*, *Allomyces macrogynus*, and *Spizellomyces punctatus*), angiosperms (*Oryza sativa*, *Arabidopsis thaliana*, and *Zea mays*), choanoflagellates (*Monosiga brevicollis* and *Salpingoeca rosetta*), and excavates (diplomonads and kinetoplastids). Furthermore, the tree correctly positions Choanozoa and Ichthyosporea at the base of the animal clade.

The congruence of the resulting tree with traditional, sequence-based phylogenies indicates that the language models of domain architectures indeed carry robust phylogenetic information and that the models generally coevolve with the core genes that are used for phylogeny construction. However, the tree also shows some notable deviations from the accepted phylogeny; in particular, the apparent monophyly of Rhodophyta and angiosperms (Fig. 3). Such deviations from the species tree could reflect anomalous changes in domain architectures and, in this particular case, probably in Chlorophyta.

Clade-Specific Signatures of Domain Architecture. Formation of new domain combinations, particularly domain accretion that leads to increased functional complexity, is an important avenue of protein evolution (27). In our previous analyses, we investigated the evolution of eukaryotes based on domain architecture evolution by measuring promiscuity of protein domains (13, 19). The results revealed that the repertoires of promiscuous domains (and especially domain architectures) are specific to major clades of eukaryotes and apparently reflect distinct biological functionalities. To a large extent, the evolution of domain architectures appears to be governed by natural selection (19, 34).

We identified signature bigrams for different major branches of organisms and examined their potential functional implications. To this end, we analyzed bigram language models of major subdivisions of life (bacteria, archaea, green plants, fungi, and animals) by using sparse partial least squares discriminant analysis (sPLS-DA) (77) to identify the bigrams that contributed maximally to the differentiation of each clade from the rest. The bigram entropy values (Eq. **10**) were weighted as features for species of Proteobacteria (n = 1,345), Crenarchaeota (n = 36), Euryarchaeota (n = 111), Viridiplantae (n = 57), Fungi (n = 187), and Metazoa (n = 154), with each proteome containing at least 1,000 proteins.

We used the *splsda* method from the R package mixOmics (78) and performed feature selection after variable tuning. Two methods were employed for feature selection, one using multiclass and the other using binary classes. The first method included a recursive technique by eliminating one clade at a time from the datasets and repeating sPLS-DA on the remaining clades (*SI Appendix*, Fig. S6). Under this method, sPLS-DA was carried out using the multiclass output. In the second method, we used binary classes for each partition and merged classes based on their taxonomic supergroup (Fig. 4). In the first method, the first round of sPLS-DA was run on the entire dataset, with six classes as output (Proteobacteria, Crenarchaeota, Euryarchaeota, Viridiplantae, Fungi, and Metazoa) (*SI Appendix*, Fig. S6*A*). Based on the component showing the maximum difference, the second round of sPLS-DA was run on two datasets, one of which included the three prokaryotic groups (*SI Appendix*, Fig. S6*B*) and the second including the three eukaryotic groups (*SI Appendix*, Fig. S6*D*). The process was iterated until all of the groups selected for the analysis were classified (*SI Appendix*, Fig. S6 *C* and *E*). In the second method, the output classes were always kept binary. In the first round, the sPLS-DA merged the prokaryotes (Bacteria and Archaea) as one combined output class, and eukaryotes (green plants, fungi, and metazoans) as the other (Fig. 4*A*). In subsequent rounds of sPLS-DA, prokaryotes and eukaryotes were split into the corresponding subdivisions: prokaryotes were split into Proteobacteria and Archaea (Fig. 4*B*), and eukaryotes were split into Viridiplantae and Opisthokonta (Fig. 4*D*). An additional round of sPLS-DA was carried out with Archaea (Fig. 4*C*) and Opisthokonta (Fig. 4*E*) to classify all of the groups. These guided, nested sPLS-DA analyses allowed us to identify the domain bigrams that are characteristic in each clade.

In both methods, we visually selected the components that showed maximum separation between the classes and extracted the features that contributed to these components using the *selectVar* function of mixOmics package (Fig. 5 and *SI Appendix*, Fig. S7). In most cases, features selected using multiclass sPLS-DA were identical to or comprised a subset of the features selected using binary classes. In the resulting clustering, prokaryotes are perfectly separated from eukaryotes along component 1 (Fig. 4*A*, *SI Appendix*, Fig. S6*A*, and Dataset S7). We identified 15 bigrams (Fig. 5*A*, *SI Appendix*, Fig. S7*A*, and Dataset S7) that

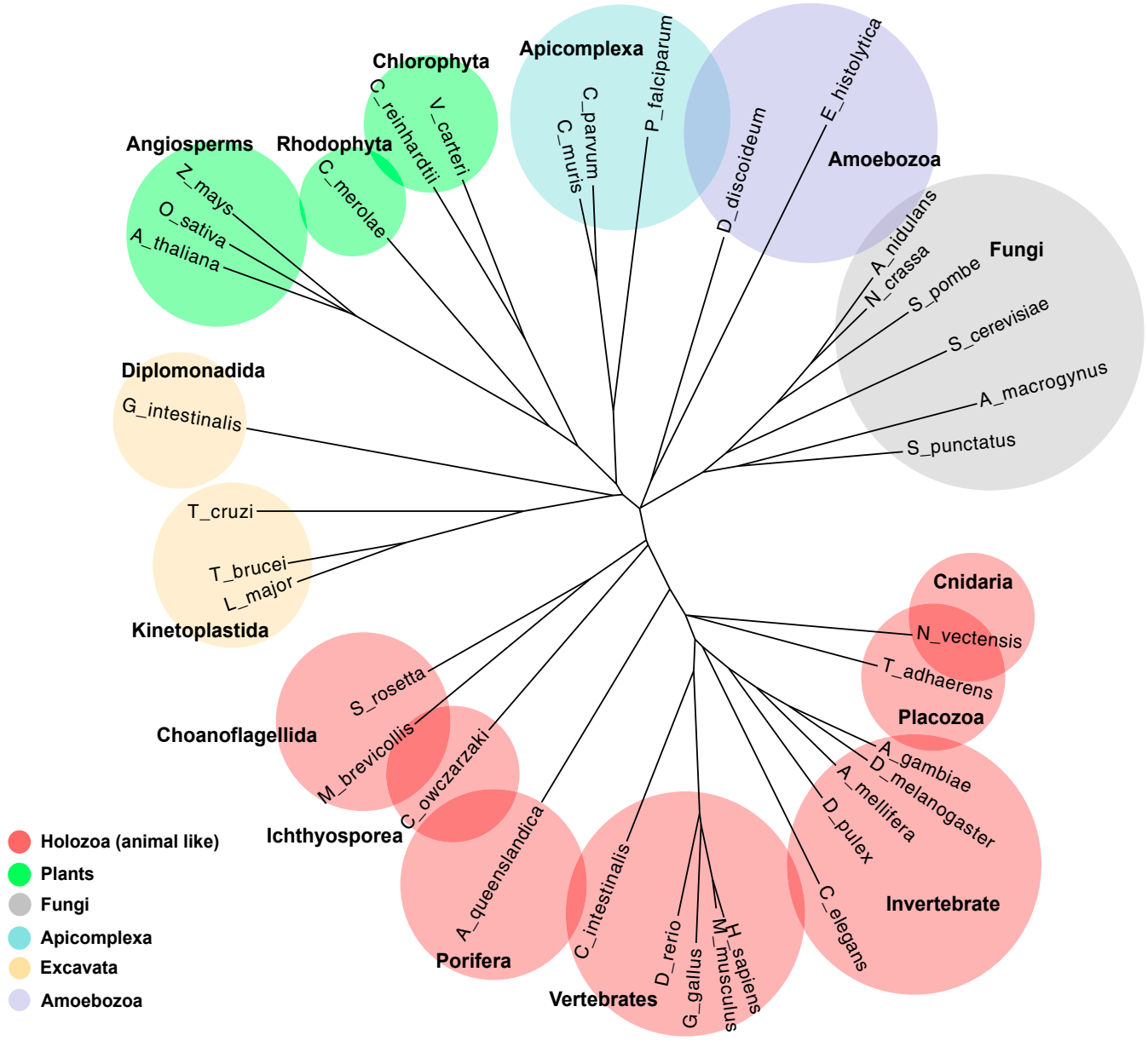

Fig. 3. Phylogenetic tree built from cross-entropy values. Domain bigram models were generated from 37 selected eukaryotic clades (Dataset S2) from the main branches of Eukaryota. The cross-entropies of bigram models were calculated in an all-vs.-all comparison. The entropy values were then normalized to create a distance matrix (see *Methods* for details), and the tree was constructed using the neighbor-joining method. The major groups are colored as shown in the legend.

differentiate the two clusters. All these bigrams are overrepresented in eukaryotes, indicating that eukaryotes and prokaryotes are separated mostly through gain of new domain architectures (and, by inference, new functions) in eukaryotes. The proteins that contain these bigrams are involved in signature eukaryotic functions such as ubiquitin signaling pathways and splicing (Dataset S8). In particular, the three top bigrams are CL0229 (RING)-CL0125 (Peptidase_CA), CL0221 (RRM)-CL0221(RRM), and CL0229 (RING)-CL0229 (RING). The first and the third of these are involved in the ubiquitin network, whereas the second is represented in spliceosome subunits.

Within the prokaryotic cluster (Fig. 4*B* and *SI Appendix*, Fig. S6*B*), Proteobacteria and Archaea are also clearly separated along component 1, with 15 bigrams differentiating the archaeal and bacterial clusters, of which 13 are shared by the binary and multiclass sPLS-DA (Fig. 5, *SI Appendix*, Fig. S7*B*, and Dataset S7). Many of these domain combinations are involved in cell-wall biogenesis (Dataset S8), which is biologically plausible, given the distinct molecular structures of the cell walls of archaea and bacteria (79).

Crenarchaeota and Euryarchaeota also form nonoverlapping clusters separated on component 1 (Fig. 4*C* and *SI Appendix*, Fig. S6*C*), with 15 discriminating bigrams that are common in both the binary and multiclass sPLS-DA (Fig. 5*C*, *SI Appendix*, Fig. S7*C*, and Dataset S7). These domain combinations are mostly involved in ATP hydrolysis, DNA replication, and proteolysis, apparently related to heat-shock response (Datasets S7 and S8).

Comparing green plants (Viridiplantae) with fungi/metazoans (Opisthokonta), we found that the discriminating bigrams are all plant specific, indicating gain of function in plants (Figs. 4*D* and 5*D* and *SI Appendix*, Figs. S6*D* and S7*D*). Altogether, there are 10 such bigrams (Dataset S7), five of which are common to both multiclass

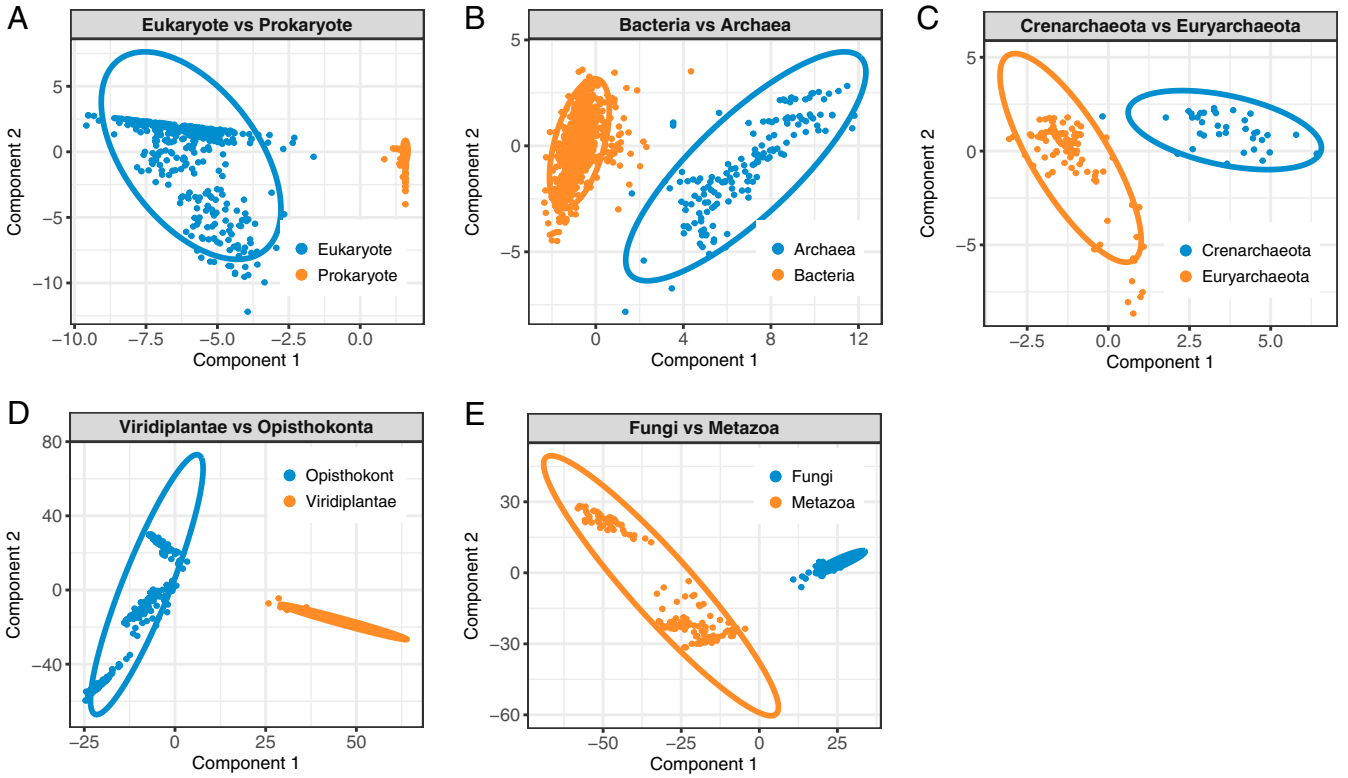

Fig. 4. Bigram feature selection using sPLS-DA. (*A*–*E*) Weighted bigram probabilities were calculated from each species, and sPLS-DA was carried out using binary classes as outcome vectors. For multiclass classification, see *SI Appendix*, Fig. S6. The analyses were carried out in a nested hierarchical manner, beginning with the eukaryotes and prokaryotes at the top level, where all the species were binned into these two divisions (*A*). In the subsequent rounds, each division was split into the respective subdivisions: prokaryotes were split into Bacteria and Archaea (*B*); Archaea was split into Crenarchaeota and Euryarchaeota (*C*); eukaryotes were first split into Viridiplantae (green plants) and Opisthokonta (fungi and animals) (*D*); and finally, Opisthokonta was split into Fungi and Metazoa (animals) (*E*). For each analysis, a biplot with component 1 on the *x* axis and component 2 on the *y* axis is shown. The ellipses represent the 95% confidence area of each cluster.

and binary sPLS-DA. The discriminating domain combinations mostly included NTPases and protein kinases (Dataset S8).

Finally, Fungi and Metazoa are distinguished by 15 bigrams (Fig. 4*E*, *SI Appendix*, Fig. S6*E*, and Dataset S7), most of which are dominant in Metazoa (Fig. 5*E* and *SI Appendix*, Fig. S7*E*). The Metazoa-specific bigrams are related to cell–cell adhesion, including the hedgehog family proteins and various membrane proteins, channels, and kinases—that is, functions mostly associated with metazoan multicellularity (Dataset S8).

Together, these findings indicate that domain bigram models readily and cleanly distinguish between the major divisions of cellular life. Following the linguistic metaphor, the protein languages in different divisions of cellular life are clearly distinct, the quasi-universal grammar notwithstanding.

Conclusions

The similarities between natural languages and genomes are apparent when domains are treated as functional analogs of words in natural languages. Here, we investigated these similarities using modifications of methods employed by computational linguistics. Using domain bigram models, we show that for most groups of organisms (both prokaryotes and eukaryotes), the information gain in the observed domain architectures lies within the narrow interval between 1.1 and 1.3 bits. As shown by our analysis of shuffled-genome entropy, these nearly universal values can be decomposed into distinct contributions from the global domain architecture and the specific domain combinations. This finding implies the existence of a quasi-universal grammar of domain architectures. The nature of the rules that underlie this universal grammar remains to be further investigated. Generally, multidomain architectures are most common among proteins that are involved in signal transduction, regulatory processes, and immune functions (19, 80). Conceivably, the near-universal value of information gain by the genome-wide domain architectures represents the minimum complexity that is required to maintain a functioning cell capable of adequately processing internal and external signals. The significant deviations from the universal value of the information gain in a subset of archaeal phyla seem to represent streamlining under extreme conditions to the lowest limit of complexity sustainable for autonomous cells. Conversely, in animals, the high value of the information gain reflects the exceptional complexity that is incumbent in multicellular organisms with differentiated tissues.

The near-universal information gain relates the protein languages of biology to human natural languages. However, the characteristic values of the information gain are substantially different, namely, ∼1.2 bits in biology and ∼3.6 bits in linguistics. Thus, both the protein and the human (and formal) languages seem to be based on quasi-universal grammars, but the complexity (orderliness) of the latter is substantially greater than that of the former. This difference could be expected because, unlike human languages, all proteomes are rich in single-domain proteins (one-word sentences) (4, 25) and because the role of the stochastic component in the evolution of protein languages appears to be much greater than it is in the evolution of natural languages (9, 11).

We also show that domain bigram models can be used to generate evolutionary trees by measuring the probability of a genome, given the language models from other genomes. This cross-entropy largely accurately reflects the topology of the currently accepted sequence-based trees, showing that domain architectures

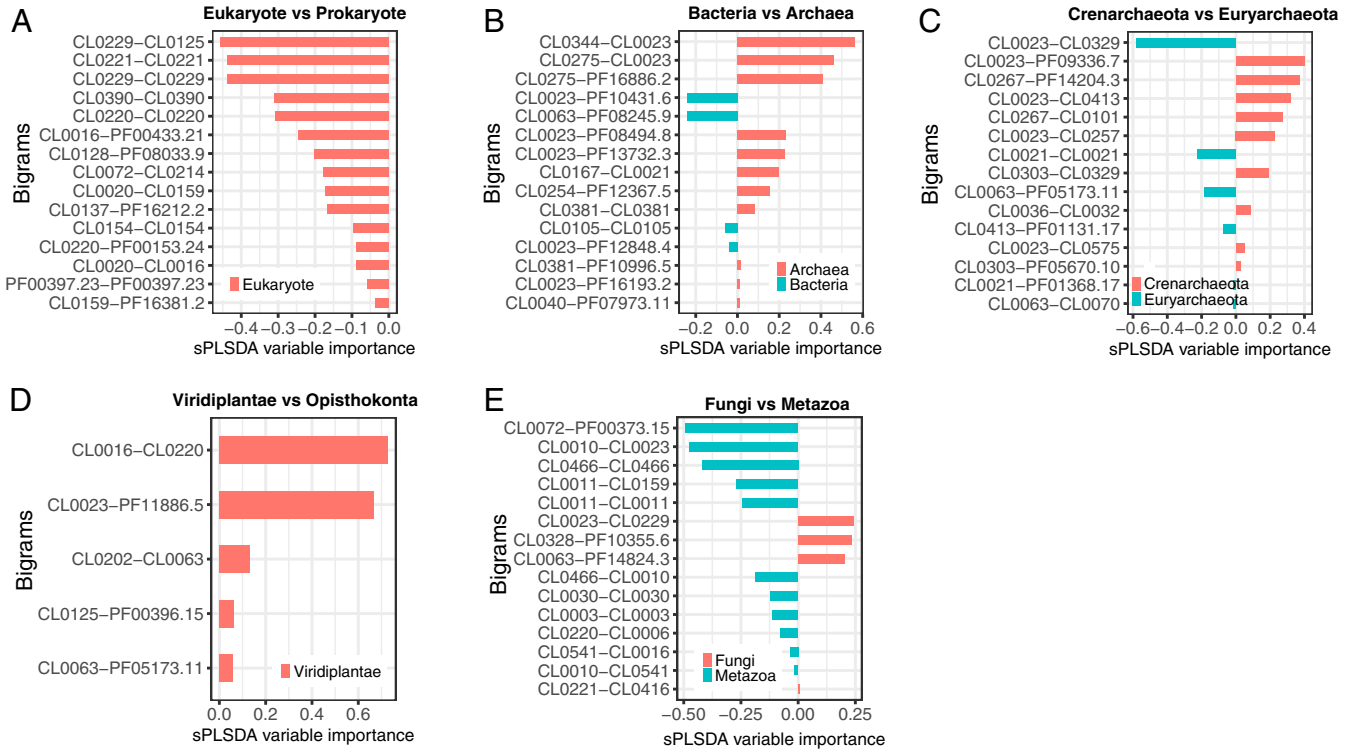

Fig. 5. Discriminating bigrams selected using sPLS-DA. (*A–E*) For each analysis in Fig. 4, the bigram variables were selected for the component that showed maximum separation between the binary classes. For multiclass selection, see *SI Appendix*, Fig. S7. The bar plots show loading values on the *x* axis for each selected bigram on the *y* axis. The color of the bar indicates the group shown in the legend. The bar plots show the selected bigrams and their loading weights for the following binary classes: eukaryotes vs. prokaryotes (*A*), Bacteria vs. Archaea (*B*), Crenarchaeota vs. Euryarchaeota (*C*), Viridiplantae (green plants) vs. Opisthokonta (fungi plus animals) (*D*), and Fungi vs. Metazoa (animals) (*E*).

(languages) mostly coevolve with the core components of genomes. Along similar lines, using feature selection, we identified the sets of bigrams that discriminate between the major clades and indeed seem to correspond to clade-specific functions such as ubiquitin signaling and splicing in eukaryotes. Under the linguistic metaphor, these bigrams are signature phrases of the respective protein languages, and the tree generated using cross-entropy values as distances reflects the evolution of these languages.

Methods

Genomes Used in the Study. Reference proteome FASTA files were downloaded from UniProt (www.uniprot.org/proteomes/) consisting of 4,159 bacterial, 187 archaeal, and 448 eukaryotic genomes. The eukaryotic genomes were a subset selected based on whether the genome has an isoform file. Only genomes with isoforms were used in the analysis ($n = 448$; Dataset S1). However, isoform data were not included in this analysis. Only the canonical proteins were used for all further calculations. A subset of the eukaryotic genomes ($n = 37$) was manually selected for phylogenetic analysis to maximize coverage of the main branches of the eukaryotic tree topology (Dataset S2).

Domain Structure Determination. We used HMMER (v. 3.1b) (40) and Pfam-A database (release 30) (41) to identify domains in each genome. The details are provided in *SI Appendix, Supplementary Methods*.

The n-Gram Model of a Protein Language. Modeling domains under a first-order Markov process, in which the probability of a domain (d_n) depends only on the one preceding domain (d_{n-1}), is called a bigram model, and the domain pair (d_{n-1}, d_n) is called a bigram. We estimated the conditional probability of the domain (d_n), given the preceding domain (d_{n-1}), using maximum likelihood estimation (MLE) (see *SI Appendix* for details):

$$P_{MLE}(d_n|d_{n-1}) = \frac{C(d_{n-1}, d_n)}{C(d_{n-1})}, \quad [1]$$

where $C(d_{n-1}, d_n)$ is the count of the bigram (d_{n-1}, d_n) in the genome, and $C(d_{n-1})$ is the count of all bigrams in which the first domain is d_{n-1}, which is equivalent to the count of domain d_{n-1} in the genome.

In bigram models of natural languages, the beginnings and the ends of sentences are marked with faux word markers (7, 8). Similarly, the start and the end of protein sequences were marked with the two artificial domain markers N and C, respectively. The addition of the end markers allows us to include all domains in the analysis, including those that solely occur in the single-domain proteins; otherwise, bigram models can only be constructed for multidomain proteins. The addition of these markers is a common practice in *n*-gram modeling in linguistics that makes these models probabilistic and generative (7, 8). However, the addition of these markers has large effects on small genomes; therefore, we also analyzed models without these additional markers. In this case, although models are not truly probabilistic, we could compare them against models created without additional markers to measure the effect of these markers on small genomes.

Good–Turing Smoothing. The biggest problem with *n*-gram models is the sparsity of the data. Most of the domains are present in strictly constrained contexts and participate only in a restricted number of domain pairs. Therefore, a large number of the conditional probabilities are 0 in a genome. We used a smoothing technique called Simple Good–Turing (SGT) (43) to assign counts to missing bigrams in the genomes (*SI Appendix*). Once the SGT counts were estimated, these counts were used to calculate the conditional probabilities as shown in Eq. **1**.

Entropy and Relative Entropy of *n*-Gram Language Models. In general, entropy of a probabilistic system, as defined by Shannon (81) is

$$H = -\sum_i P(X_i) \times \log_2 P(X_i), \quad [2]$$

where $P(X_i)$ is the probability of the event X_i. In this equation, entropy is the sum of weighted log probabilities converted to a positive value.

Entropy of a unigram model (a random genome) is calculated by simply replacing $P(X_i)$ in Eq. **2** with the frequency of the domains in a genome, which can be simplified as follows:

$$H_r = -\frac{1}{N}\sum C(d_n) \times \log_2 P(d_n), \quad [3]$$

where d_n is a domain, $C(d_n)$ is its count, $P(d_n)$ is its frequency, and N is the total number of domains in the genome. This entropy calculation effectively gives the entropy of the genome after a random shuffling of all of the domains in the genome, disregarding the protein structures. In other words, it is a "bag of words" model in which the probability of a domain in the genome is proportional to its frequency.

According to Eq. **2**, the entropy of a system is the sum of weighted probabilities of the events. The entropy of a bigram model is defined as a weighted average branching factor of a language (7, 8), where each conditional probability is weighted by the frequency of the particular bigram in the genome. This can be simplified as follows:

$$H_w = -\frac{1}{N_b}\sum C(d_{n-1}, d_n) \times \log_2 \frac{C(d_{n-1}, d_n)}{C(d_{n-1})}, \quad [4]$$

where $C(d_{n-1}, d_n)$ is the count of the bigram, $C(d_{n-1})$ is the count of the first domain in the bigram domain pair, and N_b is the total count of bigrams in the genome.

The loss of entropy resulting from the transition from the completely random version of the genome to the observed domain architectures can, therefore, be calculated as

$$H_g = H_r - H_w, \quad [5]$$

where H_g is the entropy loss (information gain), H_r is the entropy of the unigram (random) model, and H_w is the entropy of the bigram model.

Entropy of a Shuffled Genome. We also estimated the entropy of a genome after shuffling all its constituent domains. The domains encoded in a genome were shuffled in such a way that the total numbers of proteins, domains, and domain families, as well as the number of domains in each protein and the positions of the N and C markers were kept unchanged, but all of the *n*-grams were randomized. We then estimated the bigram entropy from this shuffled genome following the procedure described above. The shuffling was repeated 100 times and the average bigram entropy from these 100 runs was taken as the shuffled bigram entropy (H_s). This entropy value for each genome was then subtracted from the corresponding unigram (H_r) entropy to derive the relative shuffled entropy (H_{gs}):

$$H_{gs} = H_r - H_s \quad [6]$$

The bigram entropy (H_w) before shuffling for each genome were then subtracted from H_s to derive relative bigram entropy (H_{gb}):

$$H_{gb} = H_s - H_w \quad [7]$$

Cross-Entropy (Perplexity). Given that *n*-gram language models are generative, cross-entropy or perplexity is a measure of how well a given language model describes a language (8). Perplexity can be utilized to evaluate the probability of a genome given a bigram model, where the conditional probabilities of the bigrams come from the model, but the weight for each bigram is derived from the genome being evaluated. The better the model describes the genome, the higher the probability and the lower the perplexity (see *SI Appendix* for details). Given a language model from a source genome M, the cross-entropy of a target genome G can be calculated as

$$\begin{aligned} H_w(G, M) &\approx -\frac{1}{N_G}\sum_{i=1}^{N_G} C_G(d_{i-1}, d_i) \times \log_2[P_M(d_i|d_{i-1})] \\ &\approx -\frac{1}{N_G}\sum_{i=1}^{N_G} C_G(d_{i-1}, d_i) \times \log_2 \frac{C_M(d_{i-1}, d_i)}{C_M(d_{i-1})}, \end{aligned} \quad [8]$$

where the function C_G is the count of the bigram (d_{i-1}, d_i) in the target genome G, the function P_M is the conditional probability of the same bigram in the source genome provided by the model M from the source genome, C_M is the count of the bigram in the source genome, and N_G is the number of bigrams in the target genome G.

Phylogeny Construction Using Cross-Entropy Data. A phylogenetic tree was built from the cross-entropy data from all-vs.-all comparison of 37 selected species representing all major divisions of the eukaryotes. Because cross-entropy requires assigning probability to a genome, given the bigram model from another genome, missing bigrams in the target genome have to be taken into consideration. We used SGT-smoothed bigrams for this analysis. We first normalized the pairwise cross-entropy values (Eq. **8**) using the self-entropy (entropy calculated using the models created from the same genome). Given genomes $G_1 \cdots G_i$ with models $M_1 \cdots M_i$ and target genomes $G_1 \cdots G_j$, the distance between the two genomes is calculated as

$$D(G_i, G_j) = \frac{H_w(G_j, M_i)}{H_w(G_i, M_i)} - 1, \quad [9]$$

where $H_w(G_j, M_i)$ is the cross-entropy of genome G_j (under the bigram model M_i), and $H_w(G_i, M_i)$ is the self-entropy (the model is derived from the same genome for which the entropy is being calculated). Because self-entropy is the lowest for any genome, the self-distance $D(G_i, G_j) = 0$, where $i \neq j$. The cross-entropy value is unidirectional; therefore, we took the average of $D(G_i, G_j)$ and $D(G_j, G_i)$ to derive the final distance between the two genomes. The pairwise distance table created in this manner was used to build a neighbor-joining tree using the R package APE (82).

sPLS-DA. After creating the cross-entropy tree, we sought to identify the sets of bigrams defining the major branches in the tree. To this end, supervised feature selection classifying each clade was performed using sPLS-DA (77) as implemented in the R package mixOmics (78). For this analysis, we only used bigrams that were actually present in the genome (N and C markers were removed from the analysis). We also used only species having more than 1,000 proteins in the genomes taken from the following major subdivisions of each superkingdom: Proteobacteria (n = 1,345), Crenarchaeota (n = 36), Euryarchaeota (n = 111), Fungi (n = 187), Viridiplantae (n = 57), and Metazoa (n = 154).

For each genome, all bigram conditional probabilities were weighted by their counts in the genome to calculate entropy values for individual bigrams using the following formula:

$$H(d_{n-1}, d_n) = -\frac{1}{N} C(d_{n-1}, d_n) \times \log_2 P(d_n|d_{n-1}), \quad [10]$$

where $P(d_n|d_{n-1})$ is calculated as in Eq. **1**, and $C(d_{n-1}, d_n)$ is the count of bigram (d_{n-1}, d_n) in the genome. This was divided by the total number of domains (N) in the genome to normalize this value across all species. The formula calculates individual entropy of a given bigram $H(d_{n-1}, d_n)$ measuring the uncertainty that a particular domain d_{n-1} chooses the next domain d_n with the probability $C(d_{n-1}, d_n)/C(d_{n-1})$.

sPLS-DA uses a feature matrix (X) and a group or outcome vector (Y). The feature matrix in this case was the bigram entropy values (Eq. **10**) of each bigram on rows and of each species on columns. Absent bigrams in a genome were all given 0 values. We performed two types of analyses using two types of Y vectors. In the first category, each species in the study was assigned its known clade, and sPLS-DA was carried out using this multiclass outcome vector. In the second analysis, we performed a binary classification using sPLS-DA, whereby each species was preclassified into two supergroup clades.

We tuned (*tune.splsda* function in the mixOmics package) sPLS-DA runs using fivefold cross-validation repeated 10 times to determine the best parameters of the classification (optimum number of components). After each run, we determined the classification accuracy using the area under the receiver operating characteristic curve (AUROC). We visually selected the components that maximally separated each clade from the rest of the clades and selected features (bigrams) using the *selectVar* function from the mixOmics package.

Ontology Analysis of Domains. After the feature selection, we performed ontology analysis of the bigrams to find the functional significance of each bigram. First, we identified all the proteins in the clades that were used in the analysis that contained the identified bigrams. We then searched the UniProt database using the REST application programming interface (https://www.uniprot.org/help/programmatic_access) to identify Gene Ontology (GO) terms corresponding to those proteins. We then calculated the frequency of each GO term in the result and kept only the most frequent GO terms.

ACKNOWLEDGMENTS. The computational analyses were carried out on the Genifx computational facility (genifx.ifx.uab.edu) at the University of Alabama at Birmingham. This work was partially supported by grants from the University of Alabama Health Services Foundation (M.K.B.). Y.I.W. and E.V.K. are supported by intramural funds of the National Institutes of Health.

1. Searls DB (2002) The language of genes. *Nature* 420:211–217.
2. Scaiewicz A, Levitt M (2015) The language of the protein universe. *Curr Opin Genet Dev* 35:50–56.
3. List J-M, Pathmanathan JS, Lopez P, Bapteste E (2016) Unity and disunity in evolutionary sciences: Process-based analogies open common research avenues for biology and linguistics. *Biol Direct* 11:39.
4. Scaiewicz A, Levitt M (2018) Unique function words characterize genomic proteins. *Proc Natl Acad Sci USA* 115:6703–6708.
5. Ruhlen M (1994) *The Origin of Language: Tracing the Evolution of the Mother Tongue* (Wiley, New York).
6. Atkinson QD, Meade A, Venditti C, Greenhill SJ, Pagel M (2008) Languages evolve in punctuational bursts. *Science* 319:588.
7. Manning C, Schütze H (1999) *Foundations of Statistical Natural Language Processing* (MIT Press, Cambridge, MA).
8. Jurafsky D, Martin JH (2008) *Speech and Language Processing* (Prentice Hall, Upper Saddle River, NJ), 2nd Ed.
9. Koonin EV, Wolf YI, Karev GP (2002) The structure of the protein universe and genome evolution. *Nature* 420:218–223.
10. Doolittle RF (1995) The multiplicity of domains in proteins. *Annu Rev Biochem* 64: 287–314.
11. Karev GP, Wolf YI, Rzhetsky AY, Berezovskaya FS, Koonin EV (2002) Birth and death of protein domains: A simple model of evolution explains power law behavior. *BMC Evol Biol* 2:18.
12. Kuznetsov VA (2002) *Computational and Statistical Approaches to Genomics* (Kluwer, Boston).
13. Basu MK, Poliakov E, Rogozin IB (2009) Domain mobility in proteins: Functional and evolutionary implications. *Brief Bioinform* 10:205–216.
14. Luscombe NM, Qian J, Zhang Z, Johnson T, Gerstein M (2002) The dominance of the population by a selected few: Power-law behaviour applies to a wide variety of genomic properties. *Genome Biol* 3:RESEARCH0040.
15. Barabási A-L (2002) *Linked: The New Science of Networks* (Perseus Books Group, New York).
16. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL (2000) The large-scale organization of metabolic networks. *Nature* 407:651–654.
17. Zipf GK (1949) *Human Behaviour and the Principle of Least Effort* (Addison-Wesley, Boston).
18. Krishna M, Hassan A, Liu Y, Radev D (2011) The effect of linguistic constraints on the large scale organization of language. Available at https://arxiv.org/abs/1102.2831. Accessed August 15, 2011.
19. Basu MK, Carmel L, Rogozin IB, Koonin EV (2008) Evolution of protein domain promiscuity in eukaryotes. *Genome Res* 18:449–461.
20. Wolf YI, Brenner SE, Bash PA, Koonin EV (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Res* 9:17–26.
21. Apic G, Gough J, Teichmann SA (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol* 310:311–325.
22. Ekman D, Björklund AK, Elofsson A (2007) Quantification of the elevated rate of domain rearrangements in metazoa. *J Mol Biol* 372:1337–1348.
23. Liu J, Rost B (2004) CHOP proteins into structural domain-like fragments. *Proteins* 55: 678–688.
24. Novozhilov AS, Karev GP, Koonin EV (2006) Biological applications of the theory of birth-and-death processes. *Brief Bioinform* 7:70–85.
25. Levitt M (2009) Nature of the protein universe. *Proc Natl Acad Sci USA* 106: 11079–11084.
26. Tordai H, Nagy A, Farkas K, Bányai L, Patthy L (2005) Modules, multidomain proteins and organismic complexity. *FEBS J* 272:5064–5078.
27. Koonin EV, Aravind L, Kondrashov AS (2000) The impact of comparative genomics on our understanding of evolution. *Cell* 101:573–576.
28. Rokas A (2008) The origins of multicellularity and the early history of the genetic toolkit for animal development. *Annu Rev Genet* 42:235–251.
29. Koonin EV, et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol* 5:R7.
30. Chothia C, Gough J, Vogel C, Teichmann SA (2003) Evolution of the protein repertoire. *Science* 300:1701–1703.
31. Nichols SA, Dirks W, Pearse JS, King N (2006) Early evolution of animal cell signaling and adhesion genes. *Proc Natl Acad Sci USA* 103:12451–12456.
32. Kusserow A, et al. (2005) Unexpected complexity of the Wnt gene family in a sea anemone. *Nature* 433:156–160.
33. Marsh JA, Teichmann SA (2010) How do proteins gain new domains? *Genome Biol* 11:126.
34. Forslund K, Henricson A, Hollich V, Sonnhammer ELL (2008) Domain tree-based analysis of protein architecture evolution. *Mol Biol Evol* 25:254–264.
35. Dong Q, Wang K, Liu X (2016) Identifying the missing proteins in human proteome by biological language model. *BMC Syst Biol* 10:113.
36. Xie X, Jin J, Mao Y (2011) Evolutionary versatility of eukaryotic protein domains revealed by their bigram networks. *BMC Evol Biol* 11:242.
37. Seidl MF, Van den Ackerveken G, Govers F, Snel B (2011) A domain-centric analysis of oomycete plant pathogen genomes reveals unique protein organization. *Plant Physiol* 155:628–644.
38. Weiner J, 3rd, Moore AD, Bornberg-Bauer E (2008) Just how versatile are domains? *BMC Evol Biol* 8:285.
39. Bateman A, et al.; The UniProt Consortium (2017) UniProt: The universal protein knowledgebase. *Nucleic Acids Res* 45:D158–D169.
40. Eddy SR (2011) Accelerated profile HMM searches. *PLoS Comput Biol* 7:e1002195.
41. Finn RD, et al. (2010) The Pfam protein families database. *Nucleic Acids Res* 38:D211–D222.
42. Ekman D, Björklund AK, Frey-Skött J, Elofsson A (2005) Multi-domain proteins in the three kingdoms of life: Orphan domains and other unassigned regions. *J Mol Biol* 348: 231–243.
43. Gale WA, Sampson G (1995) Good-turing frequency estimation without tears. *J Quant Linguist* 2:217–237.
44. Good IJ (1953) The population frequencies of species and the estimation of population parameters. *Biometrika* 40:237–264.
45. Lewis M, ed (2009) *Ethnologue: Languages of the World* (SIL International, Dallas), 16th Ed.
46. Montemurro MA, Zanette DH (2011) Universal entropy of word ordering across linguistic families. *PLoS One* 6:e19875.
47. Greenberg JH (1969) Language universals: A research frontier. *Science* 166:473–478.
48. Shannon CE (1951) Prediction and entropy of printed English. *Bell Syst Tech J* 30: 50–64.
49. Adami C, Ofria C, Collier TC (2000) Evolution of biological complexity. *Proc Natl Acad Sci USA* 97:4463–4468.
50. Adami C (2002) What is complexity? *BioEssays* 24:1085–1094.
51. Koonin EV (2004) A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man. *Cell Cycle* 3:280–285.
52. Lynch M, Conery JS (2003) The origins of genome complexity. *Science* 302:1401–1404.
53. Koonin EV (2011) Are there laws of genome evolution? *PLoS Comput Biol* 7:e1002173.
54. Koonin EV (2011) *The Logic of Chance: The Nature and Origin of Biological Evolution* (FT Press Science, Upper Saddle River, NJ).
55. Nelson-Sathi S, et al. (2015) Origins of major archaeal clades correspond to gene acquisitions from bacteria. *Nature* 517:77–80.
56. Wolfe KH (2001) Yesterday's polyploids and the mystery of diploidization. *Nat Rev Genet* 2:333–341.
57. Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155.
58. Van de Peer Y (2004) Computational approaches to unveiling ancient genome duplications. *Nat Rev Genet* 5:752–763.
59. Treangen TJ, Rocha EPC (2011) Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet* 7:e1001284.
60. Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV (2005) Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res* 33:4626–4638.
61. Zhou X, Lin Z, Ma H (2010) Phylogenetic detection of numerous gene duplications shared by animals, fungi and plants. *Genome Biol* 11:R38.
62. Urbach JM, Ausubel FM (2017) The NBS-LRR architectures of plant R-proteins and metazoan NLRs evolved in independent events. *Proc Natl Acad Sci USA* 114: 1063–1068.
63. Dunn M, Greenhill SJ, Levinson SC, Gray RD (2011) Evolved structure of language shows lineage-specific trends in word-order universals. *Nature* 473:79–82.
64. Rao RPN, et al. (2009) A Markov model of the Indus script. *Proc Natl Acad Sci USA* 106: 13685–13690.
65. Rao RPN, et al. (2009) Entropic evidence for linguistic structure in the Indus script. *Science* 324:1165.
66. Greenberg JH (1963) Some universals of grammar with particular reference to the order of meaningful elements. *Universals of Human Language* (MIT Press, Cambridge, MA).
67. Koonin EV, Wolf YI (2008) Genomics of bacteria and archaea: The emerging dynamic view of the prokaryotic world. *Nucleic Acids Res* 36:6688–6719.
68. Bhattacharyya A (1943) On a measure of divergence between two statistical populations defined by their probability distributions. *Bull Calcutta Math Soc* 35:99–109.
69. Yang S, Doolittle RF, Bourne PE (2005) Phylogeny determined by protein domain content. *Proc Natl Acad Sci USA* 102:373–378.
70. Wang M, Caetano-Anollés G (2006) Global phylogeny determined by the combination of protein domains in proteomes. *Mol Biol Evol* 23:2444–2454.
71. Rogozin IB, Basu MK, Csuros M, Koonin EV (2009) Analysis of rare genomic changes does not support the unikont-bikont phylogeny and suggests cyanobacterial symbiosis as the point of primary radiation of eukaryotes. *Gen Biol Evol* 1:99–113.
72. Luo Y, Fu C, Zhang D-Y, Lin K (2006) Overlapping genes as rare genomic markers: The phylogeny of gamma-Proteobacteria as a case study. *Trends Genet* 22:593–596.
73. Rokas A, Holland PW (2000) Rare genomic changes as a tool for phylogenetics. *Trends Ecol Evol* 15:454–459.
74. Keeling PJ, et al. (2005) The tree of eukaryotes. *Trends Ecol Evol* 20:670–676.
75. Keeling PJ (2007) Genomics. Deep questions in the tree of life. *Science* 317:1875–1876.
76. Adl SM, et al. (2005) The new higher level classification of eukaryotes with emphasis on the taxonomy of protists. *J Eukaryot Microbiol* 52:399–451.
77. Lê Cao K-A, Boitard S, Besse P (2011) Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics* 12:253.
78. Rohart F, Gautier B, Singh A, Lê Cao K-A (2017) mixOmics: An R package for 'omics feature selection and multiple data integration. *PLoS Comput Biol* 13:e1005752.
79. Lombard J (2016) Early evolution of polyisoprenol biosynthesis and the origin of cell walls. *PeerJ* 4:e2626.
80. Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA (2004) Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol* 14:208–216.
81. Shannon CE (1948) A mathematical theory of communication. *Bell Syst Tech J* 27: 379–423.
82. Paradis E, Claude J, Strimmer K (2004) APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* 20:289–290.