doi:10.1111/evo.14550

Protein folds as synapomorphies of the tree of life

Martin Romei,[1,2,3] Guillaume Sapriel,[1,4] Pierre Imbert,[1] Théo Jamay,[1] Jacques Chomilier,[2] Guillaume Lecointre,[1] and Mathilde Carpentier[1]

[1] *Institut Systématique Evolution Biodiversité (ISYEB UMR 7205), Sorbonne Université, MNHN, CNRS, EPHE, UA, Paris, France*

[2] *IMPMC (UMR 7590), BiBiP, Sorbonne Université, CNRS, MNHN, Paris, France*

[3] *E-mail: martin.p.romei@gmail.com*

[4] *UFR des sciences de la santé, Université Versailles-St-Quentin, Versailles, France*

Received October 6, 2021
Accepted May 31, 2022

Several studies showed that folds (topology of protein secondary structures) distribution in proteomes may be a global proxy to build phylogeny. Then, some folds should be synapomorphies (derived characters exclusively shared among taxa). However, previous studies used methods that did not allow synapomorphy identification, which requires congruence analysis of folds as individual characters. Here, we map SCOP folds onto a sample of 210 species across the tree of life (TOL). Congruence is assessed using retention index of each fold for the TOL, and principal component analysis for deeper branches. Using a bicluster mapping approach, we define synapomorphic blocks of folds (SBF) sharing similar presence/absence patterns. Among the 1232 folds, 20% are universally present in our TOL, whereas 54% are reliable synapomorphies. These results are similar with CATH and ECOD databases. Eukaryotes are characterized by a large number of them, and several SBFs clearly support nested eukaryotic clades (divergence times from 1100 to 380 mya). Although clearly separated, the three superkingdoms reveal a strong mosaic pattern. This pattern is consistent with the dual origin of eukaryotes and witness secondary endosymbiosis in their phothosynthetic clades. Our study unveils direct analysis of folds synapomorphies as key characters to unravel evolutionary history of species.

KEY WORDS: Phylogeny, protein folds, synapomorphy, tree of life.

To reconstruct the tree of life, phylogeneticists need to identify features shared by a limited number of species in samples of life diversity. Hennig (1950, 1999) distinguished three types of similarities: (i) characters too general for the sample (symplesiomorphies), that is, appearing before the last common ancestor of the sample; (ii) similarities not obtained from a common ancestry (called homoplasies after Lankester (1870)) ; and (iii) similarities exclusively arising from a common ancestry (synapomorphies). Only the latter category of similarities allows the construction of sister-group relationships, which are necessary for phylogenetic inference, and the creation of taxonomic groups. Most subsequent systematists followed these principles using morphological and anatomical characters but in the late 1960s arose molecular systematics (Zuckerkandl and Pauling 1965; Fitch and Margoliash 1967). Similarities were then mostly calculated from alignments of homologous sequences of amino acids or nucleic acids (Hillis et al. 1996). As a result, most of the present phylogenies are based on sequence alignments but the events they account for (substitutions) are usually far less interpretable in terms of functions than the variation of anatomical characters. Some phylogenies combine both types of information (Lecointre and Le Guyader 2017). However, it is possible to use other types of molecular variation by considering the structural organization of amino acid sequences. If we name "anatomy" the 3D spatial organization of proteins, we can say that in the course of the 20th-century history of biology, the anatomical characters used to reconstruct phyletic patterns stopped at the levels of histology and cytology, but never went closer to macromolecules. This article aims to examine whether the anatomy of globular soluble proteins could provide reliable synapomorphies to document relationships of the tree of life.

In this study, our attempt to identify shared features of molecular anatomy among species is based on protein structural domains which is the evolutionary unit of protein 3D structures (Doolittle 1995; Dawson et al. 2017). Structural domains are compact structural units of few hundreds of amino acids (from 50 up to 500). One domain is characterized by the three-dimensional succession of its secondary structures; this topology is called the fold. The question of whether the fold space is discrete (Dokholyan et al. 2002; Taylor 2002; Pascual-García et al. 2009) or continuous (Skolnick et al. 2009; Nepomnyachiy et al. 2017) has been debated but the two points of view can be merged into a physical universe of proteins, which is theoretically continuous but whose populations (the number of proteins actually existing) make it discrete in practice (Sadowski and Taylor 2009). The three major hierarchical classifications of structural domains are CATH (Class Architecture Topology Homology) (Dawson et al. 2017), SCOP (Structural Classification Of Protein) (Murzin et al. 1995), which is now SCOPe (Chandonia et al. 2017), and ECOD (Evolutionary Classification of Protein Domains) (Cheng et al. 2014). These three classifications are hierarchical; the first two are based on structure, sequence, and function, and the third is more evolutionary based. The first four levels of SCOP are as follows: class, fold, superfamily, and family; in CATH, they are class, architecture, topology, and homology, and in ECOD, they are architecture, X-group, H-group, and T-group. In SCOP, class and fold levels are fully structure based, whereas superfamily and family also rely on sequence and/or function similarities; in CATH, the first three levels (CAT) are fully structure-based and in ECOD, the classification is evolutionary based. In SCOP, 12 classes are defined but four of them represent more than 80% of all folds. These four classes are as follows: all-alpha (α), all-beta (β), or mixed alpha-beta (α/β, α+β). In CATH, the classes (α/β, α+β) are merged and the three classes (α, β, αβ) represent 98% of the structural domains. This first level is too general to contain useful evolutionary information (Koonin et al. 2002) and in ECOD, the authors admit that this level has been largely introduced for convenience of users and does not directly correspond to evolutionary grouping (Cheng et al. 2014). Concerning the other levels, it has been established that proteins of the same SCOP superfamily or family derive from a common ancestor, but the common origin of the folds is still discussed (Koonin et al. 2002). Even if there are some discrepancies between the classifications, the best correspondence for the fold level of SCOP is the T level in CATH (Csaba et al. 2009) and the X-level in ECOD (Cheng et al. 2014). The number of recorded folds is limited to a few thousands (Zhang 1997; Wolf et al. 2000), and is almost constant since 2013 according to CATH and SCOP. This number is surprisingly low, several orders of magnitude less than the number of sequences in the biosphere (Landenmark et al. 2015). On the one hand, it is possible that biological or physical constraints significantly limit the repertoire of folds (Mannige 2014; Rose et al. 2015), and that structural convergence is frequent. In this case, fold structural similarities across lineages of the tree of life should be homoplasies, that is, they do not come from a common ancestor. On the other hand, it is also possible that folds represent intrinsically stable characters even in cases of high divergence of sequences (Bajaj and Blundell 1984; Illergård et al. 2009). Because of the low number of different superfamilies within a given fold, the conservation of function, and sequence similarities, a consensus that most protein folds are monophyletic is emerging, with the possible exception of some of the most diverse superfolds (Koonin et al. 2002). It has also been shown that convergent evolution is rare in terms of structural domain combination (Gough 2005). It is expected that, due to the slow dynamics of fold change, folds could highlight the deepest sister-group relationships of the tree of life, and could be useful for large-scale classification of species. Therefore, folds should appear to be suitable character for phylogenetic investigation; however, trees constructed from them are not consistent with the phylogeny of organisms (Williams et al. 2020). Consequently, there is a need to evaluate the propensity of protein folds to be reliable synapomorphies.

Many attempts to reconstruct phylogenies from fold content have been made, the first being in the 1990s (Gerstein 1998; Gerstein and Hegyi 1998; Wolf et al. 1999; Lin and Gerstein 2000). The number of genomes was then limited (13 at most) and most of them were bacterial. The number of experimentally resolved protein structures was also limited. Nevertheless, the trees were globally congruent with current knowledge and the results were promising. In the beginning of this century, the number of complete genomes and resolved structures grew exponentially. Besides, the development of sensitive and efficient methods based on hidden Markov models (HMM) for scanning sequence databases have significantly increased the capacity to retrieve distantly related sequences (Levitt 2009). The HMMs were used to assign folds to the proteins of whole proteomes and phylogenies were reconstructed from fold abundances (copy-count) or occurrences (binary) using parsimonious reconstruction (Caetano-Anollés and Caetano-Anollés 2003, 2005; Wang et al. 2006, 2007, 2011; Caetano-Anollés et al. 2009; Wang and Caetano-Anollés 2009; Kim and Caetano-Anollés 2011, 2012) or distance methods (Yang et al. 2005) or both (Winstanley et al. 2005). None of these approaches was suitable to provide identified synapomorphies. Caetano-Anollés and his group mostly used fold abundances, whereas Yang et al. and Winstanley et al. found that fold abundances are not reliable enough to infer robust trees, arguing that abundance is greatly affected by gene and chromosome duplication, which is not a uniform process (Yang et al. 2005). Caetano-Anollés et al. conducted numerous studies and produced a lot of phylogenies with thousands of organisms. They rooted their phylogenies by choosing the branch of the tree exhibiting the largest numbers of ancestral (plesiomorphic) character states (Nasir et al. 2014). They also calculated phylogenies from RNA structures (Caetano-Anollés 2002). They deduced from these rooted phylogenies that the world of organisms is tripartite, and the phylogenetic relationships between the three superkingdoms is [archaea, [bacteria, eukaryotes]], archaea being at the root and paraphyletic, and bacteria and eukaryotes being monophyletic (Kim and Caetano-Anollés 2011). These results are in contradiction with present phylogenies where eukaryotes are very close to archaea (Hug et al. 2016) or even branch within them (Williams et al. 2020). These apparent contradictions are appealing: why protein structure and DNA do not tell the same story? Cateano-Anollés et al. argue that the use of molecular sequences is problematic for phylogeny reconstruction on many grounds: mutational saturation, definition of homology of sites in sequence alignments, taxon sampling and tree imbalance, and different historical signatures in domains of multidomain proteins (Caetano-Anollés and Nasir 2012; Caetano-Anollés et al. 2014), whereas structural domains diversify mostly by vertical descent and are better suited to reconstruct phylogenies (Wang and Caetano-Anollés 2009; Kim and Caetano-Anollés 2012). Several fold phylogenies have also been derived from fold distributions in complete proteomes (the folds are the tree leaves and the proteomes are the characters). The resulting trees represent more the temporal order of fold diversification and it is possible to calculate the age of the folds from the distance between the root of the tree and its corresponding terminal leaf (Caetano-Anollés and Caetano-Anollés 2005; Winstanley et al. 2005). This distance has been used as a relative timescale to study the order of FSF appearance in evolutionary history (Wang et al. 2007; Caetano-Anollés et al. 2011). They found a linear correlation with geological time and defined a global molecular clock of protein folds (Wang et al. 2011). They elaborated a history of life in five phases, where the metabolic domains evolved earlier than informational domains involved in translation and transcription, supporting the metabolism-first hypothesis rather than the RNA world scenario (Kim and Caetano-Anollés 2012).

All these results are controversial and we propose here to confront fold distribution with well-acknowledged phylogenies to explore the fold history to understand the sources of these differences. Also, these studies were not specifically conducted to search for synapomorphies to document relationships among taxa; they were more focused on the timing and process of fold evolution itself. In the present article, instead of reconstructing a tree from folds like in (Caetano-Anollés and Caetano-Anollés 2003), we map folds onto a consensual phylogenetic tree of life to look for potential fold synapomorphies and to study the dynamics of fold changes, to check for their functions, and timescales of fold synapomorphies occurrence.

Materials and Methods

DATA

We have selected 210 species with complete sequenced proteomes. The reference tree of life is taken from Lecointre and Le Guyader (2017), completed by the tree from Hug et al. (2016) for bacteria and by the Asgard species from Adam et al. (2017). Species were chosen to sample all main lineages while maximizing the scope of lineage divergence times among them. The balance is maintained among the three superkingdoms with 70 species for each group. The reference tree is provided in Figure S1 and the species and main lineages list is in Table S1.

Fold annotation is performed using the SUPERFAMILY online server (Wilson et al. 2009). SUPERFAMILY hidden Markov models (HMMs) have been calculated for each superfamily and family of the SCOP 1.75 classification (Murzin et al. 1995). Each protein sequence sharing less than 95% of identity with other members of the same superfamily is used as seed. For each seed, an HMM is iteratively calculated by scanning the NRDB90 (nonredundant sequence database with less than 90% of identity) and aligning the retrieved sequences. Complete proteomes retrieved from NCBI database are scanned with those HMMs. If a protein has a significant score (*E*-value below 10^{-4}) with several HMMs, only the HMM with the best *E*-value is selected. In addition to the SUPERFAMILY database, supplementary archaeal proteomes have been scanned with the SUPERFAMILY webserver with the same *E*-value threshold. Fold annotations are extrapolated from the superfamily level. Each fold branched out into superfamilies, therefore a fold annotation is the clustering of assignment for each of its superfamily. A binary matrix is created with folds in row and species in column. The matrix contains 1 when the fold is present within a species and 0 otherwise. The same experiment has also been conducted with CATH version 4.3 (level T, the third level of the hierarchy) and ECOD version 20220113 (level X, first level of the hierarchy), which both also provide collections of HMMs. We chose these levels because other levels include nonstructural information, such as sequence similarity and function. In the case of ECOD, evolutionary information is also taken into account because this classification does not merely rely on structural information at level X. Nevertheless, we included it in our study to check the robustness of the results. The thresholds are the same for the three databases and we used CATH-resolve-hits (Lewis et al. 2019) to resolve overlapping HMM matches. The procedure to calculate the heatmaps and RI (see below) is identical for the three databases.

A fold is considered specific of a superkingdom when 100% of its presence is in one superkingdom. Some protein chains have multidomains and consequently may contain several folds. For each proteome, we have also retrieved all pairs of folds localized in the same protein. We wanted to identify the lines of the heatmap that are similar due to the systematic presence of two folds in the same protein. Folds may be found several times in a proteome, and the evaluation of these redundancies allows the estimation of the bias due to the multidomain chains. If two folds are found once in one given protein, they are tagged as co-occurring folds. We identified only 11,641 co-occurring folds present in the same proteome, among the 575,128 possible pairs ((1073 × 1072)/2). These co-occurring pairs are composed of 902 of the 1073 folds. Note that 629 co-occurring pairs are twice the same fold.

HEATMAP CALCULATION

The heatmap is generated by R with the heatmap.2 package from the ggplot2 package (Wickham 2016). To appreciate fold repartition and to search for synapomorphies, we developed a two-way analysis inspired by a former algorithm seeking to group objects and their attributes simultaneously (Birks et al. 2012). The ordering procedure is different for the folds (lines) and the species (columns). The species are ordered according to the reference phylogeny of organisms and the branches are swapped with the package Dendser (Earle and Hurley 2015) to improve the visualization (Gruvaeus and Wainer 1972; Eisen et al. 1998). The folds are ordered with a hierarchical clustering and the tree is also swapped with Dendser. Hierarchical clustering is calculated with the Mcquitty method (Sokal and Michener 1958) and the Ochiai similarity index derived distance (Gower and Legendre 1986) : $d\ = \sqrt{1 - S}$ with $\mathrm{S}\ = \frac{n11}{\sqrt{(n11+n10)*(n11+n01)}}$, where $n11$ corresponds to the contingency of the same binary data (i.e., the number of folds present in two species); $n10$ and $n01$ correspond to the contingency of different binary data. This distance excludes negative matches ($n00$). It follows metric rules, and has good resolution and linear properties, thus allowing both local cluster separation and fidelity to global distribution patterns (Hajdu 1981; Gower and Legendre 1986).

EXTRACTION OF FOLD CLUSTERS

Fold clusters are extracted by cutting the fold dendrogram resulting from their hierarchical clustering at different heights with the Dynamic Tree Cut algorithm, hybrid version (Langfelder et al. 2008). Four shape criteria are used to separate clusters: (i) a minimum number of elements, (ii) elements in the same branch must also be close to each other, (iii) clusters must be separated from other elements, and (iv) elements must be tightly connected. Two levels of clustering are used in the analyses: the first to analyze the deepest divergences of the tree of life and the second for more recent divergences within eukaryotes. This cluster of folds is used as a template to extract interesting groups of folds consistent with the reference tree of life at these two cutting levels.

FOLD CO-OCCURRENCES

We calculated the number of co-occurring domains in a given protein. In such a case, closeness of the two folds in the heatmap may be due to nonindependent characters. We have extracted the pairs of co-occurring folds in each protein. For a pair of co-occurring folds AB, we have calculated the proportion of presence of AB relative to either A or B (occurrences(AB)/occurrences(A) or occurrences(AB)/occurrences (B)).

RETENTION INDEX

The number of different structural folds is remarkably low and it is possible that fold structural similarities across lineages of the tree of life are homoplasies, that is, they do not come from the common ancestors of existing lineages. Many events of endosymbiosis and horizontal transfers may blur the phylogenetic signals. We calculated the retention index (RI) (Farris 1989) to measure the adequacy between the characters (folds) and our reference phylogeny. We do not know if the appearance and disappearance of folds are equiprobable, and there is no satisfactory evolutionary model for structures suggesting that simple distances—based on the presence or number of occurrences of folds in the proteomes—used for these reconstructions are probably strongly biased. In the absence of such models, standard parsimony is suitable (Farris 1983), either for phylogenetic inference or for mapping characters onto a pre-existing tree. The RI is "the fraction of apparent synapomorphy in the characters that is retained as synapomorphy on the tree" (Farris 1989). It can be calculated for a single character or for any set of characters. If the characters are discrete, a change of state is named a step. In our case, the folds (characters) have two states: present (1) or absent (0). The RI is calculated for each fold as $\mathrm{RI}\ = \frac{g-s}{g-m}$ with g: the maximum number of steps, which is the number of changes of a character onto the tree with a single node (star-like tree: all changes being reported onto individual branches); s: the number of steps calculated parsimoniously with the considered tree; m: the minimum number of steps that the character may have. For a single two-state character, the minimum number of changes is one (the number of character states minus one).

The RI has the advantage over the consistency index to be less sensitive (less biased) to the number of species (the number of terminals). For a character, the RI value is 1 if it perfectly fits the tree and 0 if it fits the tree as poorly as possible. If the character is uninformative for the tree (having a single state for all taxa), the value will be 0. This calculation is performed using the mesquite software (Maddison and Maddison 2019) with our data matrix and our tree of life.

We calculated the RI for all folds and for each superkingdom separately to determine if the folds are, on average, reliable phylogenetic markers within each superkingdom. Thereafter, we calculated the average RI for blocks of folds grouping clades of interest to determine if those specific folds are reliable as clade synapomorphies, that is, congruent with the tree of life.

To express the RI value with a statistical index, a *p*-value is calculated by referring to random RI calculations. We randomly simulated presence and absence of folds in same quantities across the tree 1 million times. For each replicate, a new RI is calculated. The ranking of actual character RI among random RIs of this character results in a *p*-value.

FUNCTION ANALYSIS

The function analysis is based on the SUPERFAMILY database annotation (Wilson et al. 2007). The superfamilies are associated with seven general functions (Table S2). The functions in our dataset are, in order of decreasing frequency, as follows: metabolism, other, intracellular processes, regulation, information, general and extracellular processes. For each group of interest, we count the number of superfamilies associated with a function. However, some folds are subdivided into many superfamilies with different functions. Therefore, to maintain the same weight for each fold, the sum of all the functions for one given fold is set to 1.

PRINCIPAL COMPONENT ANALYSIS

To extract the folds that separate the species, we use a principal component analysis with folds as variables. The PCA is an exploratory method that consists of a summary of many dimensions by exhibiting the dimensions that explain the widest variance (Lebart et al. 2000). The contribution of folds to the separation or grouping of species is characterized on the first two dimensions by projected vectors on the first two axis; the norm of those vectors represents the global contribution on these two axes. We select folds with vectors with high norms in the first two dimensions. The threshold for the norm value is fixed at 0.8, as determined by the distribution of the fold norm. The observation of variables reveals four groups that can be separated according to the sign of their coordinates. This fold selection is highlighted in the heatmap.

Results

DATA

This study focuses on the fold level of SCOP because we want to test the evolutionary signal in protein structural information, not in sequence information, thus excluding the other levels (superfamily and family) that contain sequence information. However, the fold and superfamily levels are very close: 87% of the folds have only one superfamily (934/1073) and 94.2% (1011) have less than five superfamilies. To assert the robustness of our results, we have also used the X-group of ECOD and the T level of CATH, which are the most similar levels in these two classifications to the fold level of SCOP (Csaba et al. 2009; Cheng et al. 2014). We have chosen to analyze only the presence/absence of folds to focus on the dynamics of their appearance and disappearance, but not their duplication, which is a different process. Previous studies have been mostly based on the number of occurrences of folds in proteomes, which may add information for phylogenetic reconstructions but is out of the scope of the present analysis. On average in the proteomes, 64% of the proteins are annotated with at least one SCOP fold (66% with CATH and 67% with ECOD). Eukaryotes are slightly less annotated than archaea and bacteria (59 ± 10%, 65 ± 4%, and 67 ± 6%, respectively). These annotation proportions are similar with CATH and ECOD (Table S5). The remaining unannotated proteins may be false negatives (the HMM missed a known fold), unstructured proteins, transmembrane proteins, or yet unknown folds.

A GLOBAL STRUCTURED REPARTITION OF FOLDS

The ordered and serialized heatmap exhibits the presence of folds across species (Fig. 1). The proteomes/species (in columns) are sorted according to the reference phylogenetic tree and the folds (in rows) are sorted according to the hierarchical clustering conducted on the folds (see Materials and Methods). The RI is calculated for each fold (character) according to the reference phylogenetic tree. In Figure 1, when a given fold is present in a given proteome, the cell is in dark red (RI > 0.75), pink ($0.75 \geq$ RI > 0.5) or gray-pink (RI ≤ 0.5). In the background, the colored columns highlight the different phylum of the selected organisms (see Fig. S2 for the detailed legend). We have calculated the RI for 10^6 random distributions of the folds on the same reference tree (see Materials and Methods) resulting in a mean RI value of 0.12. The average RI of all folds is 0.56 with an empirical *p*-value below 10^{-4}, indicating that the fold taxonomic distribution significantly fits the tree and that folds contain phylogenetic information. Indeed, 813 (76%) folds have a significantly high RI (5% threshold, Fig. S3), 351 (33%) have an RI > 0.75 and 578 (54%) have an RI > 0.5. The global RI are quite similar with ECOD or CATH (0.54 and 0.53, respectively, see Table 1).

The heatmap—the visual representation of fold repartition—is strongly structured. This structure is also visible with CATH and ECOD (see Figs. S6 and S7). We tried to quantify the part of this structuring due to multidomain proteins to check for possible bias due to the statistical dependence of fold characters due

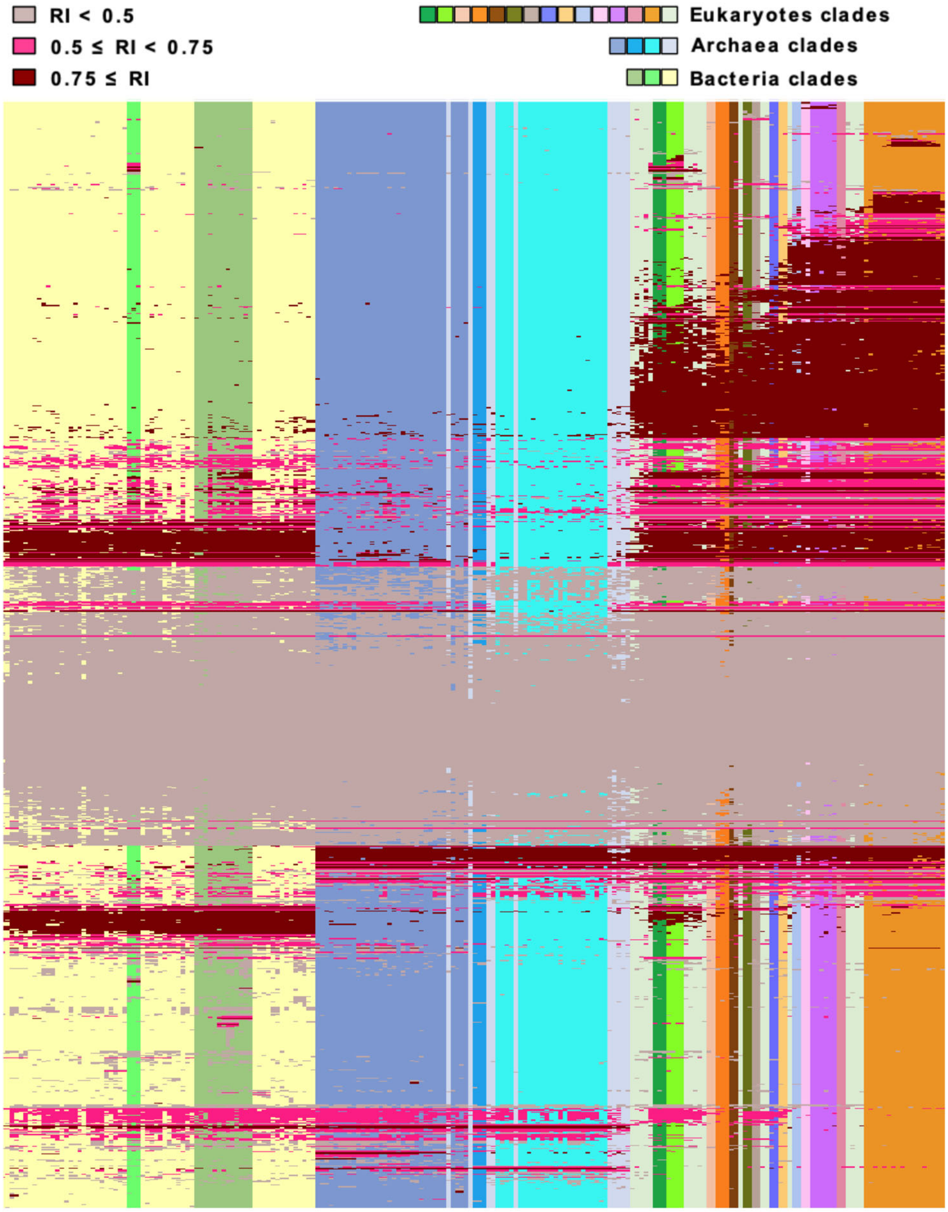

Figure 1. Heatmap showing protein fold repartition through the diversity of life. Columns are species ordered according to the reference phylogenetic tree. The 210 species from left to right are as follows: 70 bacteria, 70 archaea, 70 eukaryotes. For convenience, column colors exhibit taxonomic groups according to NCBI nomenclature. For bacteria, we chose to exhibit only two phyla of interest. Rows are 1073 protein folds as extracted from SCOP. Dots are fold presence in the corresponding species, colored according to the retention index as calculated for the fold repartition onto the reference phylogenetic tree. Darker red dots refer to folds that can be interpreted as reliable taxonomic markers (i.e., group synapomorphies). An interactive version of this heatmap is given in the Supporting Information.

to multidomain proteins. We have calculated the number of fold co-occurrences in a given protein (see Materials and Methods). Globally, the great majority of the folds close in the heatmap do not co-occur in the same protein. First, the co-occurrences of folds are rare: the co-occurrences of the folds represent less than 5% of each fold occurrences. Second, the folds co-occurring in the same proteins are mainly found in the central strip of ubiquitous folds (in light pink in Fig. 1). Outside this central strip, we find only 27 pairs of folds close in the heatmap (rank ±20), mostly corresponding to folds specific to eukaryotes (see Fig. S4).

The folds present in all organisms are localized in the central grey-pink strip in the heatmap in Figure 1. They are of no interest for phylogenetic inference, although interesting for investigating the origin of life. Note that 54 folds (4.8%) are present in all organisms, and 181 folds (16.9%) are found in 90% of

Table 1. Average retention index calculated for all characters with either all organisms or only bacteria, Eukarya or Archaea (in line). The characters are the predicted presence or absence in the proteomes of SCOP folds, T level architecture of CATH or X level architecture of ECOD (in column).

	SCOP	CATH	ECOD
All	0.56	0.53	0.54
Bacteria	0.29	0.26	0.27
Eukaryotes	0.44	0.43	0.47
Archaea	0.27	0.27	0.27

the organisms. These "ubiquitous" folds are from all structural classes of SCOP and their distribution does not statistically differ from the class distributions in all folds, although the mixed alpha and beta class, which has been found as the most ancestral (Caetano-Anollés and Caetano-Anollés 2003; Winstanley et al. 2005), is slightly more represented (Table S3). They have no obvious structural specificity. The superfolds (Orengo et al. 1994) such as TIM barrels (fold c.1) or Rossmann Fold (fold c.2) are found in all proteomes. For those ubiquitous folds, we observe an overrepresentation of functions related to metabolism (44.6% vs. 28.3%) (Table S2).

Considering the three superkingdoms, more folds are identified in eukaryotes (915) than in bacteria (804) or archaea (627). The average RI calculated on all folds is different for the three superkingdoms: it is worth 0.44 within eukaryotes, 0.27 within archaea, and 0.29 within bacteria (see Table 1). The global mean RI (0.56) is unrelated to any of the three superkingdom mean RIs because these later ones are calculated on distinct sub-trees. Its higher value indicates a better tendency to describe the total distribution compared to the three fold distributions in each of the superkindoms. The heatmap is indeed separated into three distinct parts. The "eukaryotes part" contains 224 specific folds that are clustered in nested blocks (Fig. S5, green lines). Their RI is high (0.59). The archaea and bacteria have fewer specific folds (19 and 89, respectively). The RI values are 0.68 for archaea and 0.33 for bacteria that do not show such a structure (Fig. S5, cyan and yellow folds). The unique fold common to all archaea and all bacteria (f.58 MetI-like, RI > 0.75), the few folds present in several archaea and bacteria but not present in eukaryotes and a few folds shared by archaea, bacteria and two eukaryotic clades are also in this zone and the RI is low for most of them (see Fig. 1).

The lower RI values obtained for archaea and bacteria and theirm few specific folds indicate that the reconstruction of the evolutionary history of organisms from folds is rather risky within bacteria and within archaea. However, the high value of RI, the high proportion of specific folds, and the step shape of eukaryotes on the heatmap surprisingly reveal that folds can distinctly separate recent clades within eukaryotes.

FOLD MOSAICISM DISCRIMINATES THE THREE SUPERKINGDOMS

The separation between the three superkingdoms seems clear in the heatmap and we identified the folds allowing this separation by calculating a principal component analysis (PCA). Several other PCA have been computed in the literature from the fold repartition among species and a clear separation results from the projection on the two first axes (Kurland and Harish 2015), but the folds implied in the axes have not been explored. Here, we calculated such a PCA to identify the folds allowing the separation of the three superkingdoms (see Fig. 2 and Material and Methods for details). The three superkingdoms are indeed clearly separated on the projection in the first two dimensions at the origin of 32.2% of the diversity (Fig. 2a). Only four groups of folds mainly participate to the two first axes (Fig. 2b) and these folds are grouped together in the heatmap (see Fig. 2c). The blue folds are eukaryote-specific folds, separating eukaryotes from archaea and bacteria. The pink and purple folds are shared between eukaryotes and bacteria and eukaryotes and archaea, respectively. The orange folds are shared between bacteria and photosynthetic eukaryotes, which pull the photosynthetic organisms together and above the other eukaryotes closer to the bacteria height (Fig. 2a).

The PCA results show that the grouping of archaea and bacteria according to the two first axes is less due to specific folds which are a few, than to folds also shared with eukaryotes. There are only 19 folds specific to archaea and 89 folds specific to bacteria. The separation of the three superkingdoms clearly reflects the primary endosymbiosis of the mitochondria and the secondary endosymbiosis of chloroplast. Therefore, mosaicism in fold repartition (i.e., folds shared between distant clades) significantly contributes to discriminate the three superkingdoms of the tree of life, in spite of the well-documented paraphyly of archaea (Hug et al. 2016).

FOLDS SHARED BY EUKARYOTES AND EITHER BACTERIA OR ARCHAEA

Some consistent blocks of folds are shared exclusively between eukaryotes and bacteria or eukaryotes and archaea (Fig. 3). These common blocks reflect the origin of eukaryotes and the first endosymbiosis. In these blocks, Archaea and eukaryotes share 40 folds (3.7%); the mean RI equals 0.72. Only 28 folds (2.6%) are shared between bacteria and eukaryotes; their RI is also of 0.72.

According to recent studies, eukaryotes find their origin in archaea and more specifically in the *Asgard* lineage (Zaremba-Niedzwiedzka et al. 2017). The folds exclusively shared between eukaryotes and archaea would therefore be the evidence of this common origin. The folds exclusively shared between eukaryotes and bacteria may originate from the alpha-proteobacterial origin of the mitochondria (Fan et al. 2020). Within eukaryotes,

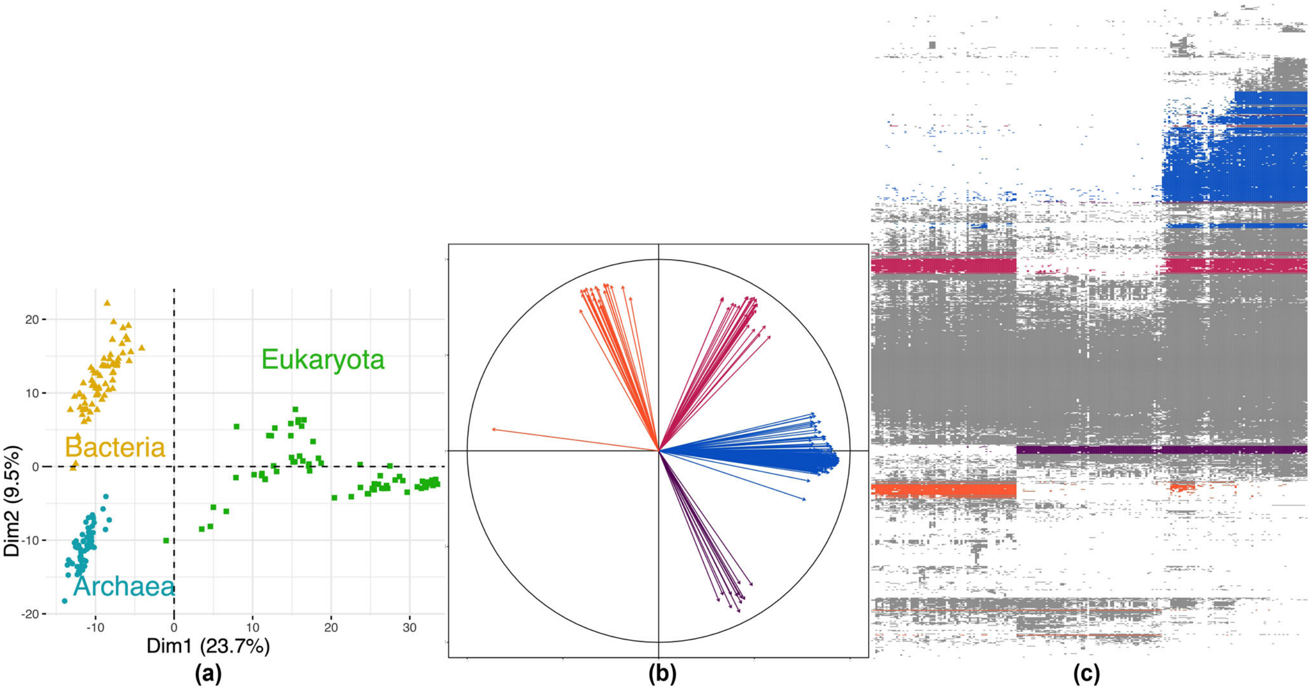

Figure 2. (a) Projection of species in the two first dimensions of a principal component analysis of fold repartition. Colors refer to the three superkingdoms of life: yellow refers to bacteria, green refers to eukaryotes, and blue refers to archaea. (b) Protein fold contributions to species repartition in the previous two dimensions of the principal component analysis. Arrows show the most contributive folds to this species repartition, in which length shows the strength of the contribution. Four clusters are distinguished with the following color code: blue refers to fold repartitions discriminating eukaryotes, and pink and purple refer to folds discriminating archaea and bacteria, orange refers to folds discriminating bacteria and archaea too and photosynthetic eukaryotes from other eukaryotes. (c) Same clusters of folds spread onto the heatmap. It shows that blue folds are markedly distributed among eukaryotes, pink folds are markedly shared by eukaryotes and bacteria, purple folds by eukaryotes and archaea, and orange folds by bacteria and photosynthetic eukaryotes.

determining the bacterial or archaeal origins of the folds common to the three kingdoms would require other information regarding the proteins themselves (sequence, detailed function). The functions of archaea and eukaryotes shared folds are mainly information and regulation with 8 (20.0%) and 7 folds (17.5%), respectively (Table S2, line 5). However, the number of folds with a metabolic function is high with six folds (15.0%). Moreover, the functions of the folds shared by eukaryotes and bacteria are mainly informational with 14 folds (47.9%) and metabolism with six folds (21.4%) (Table S2, line 6). Other studies show that the majority of eukaryotic metabolism genes came from bacteria, whereas most of informational genes came from archaea (Alvarez-Ponce et al. 2013).

The second endosymbiosis (chloroplast) is highlighted by two groups of folds (Fig. 3, lines in orange and in red). The first group (Fig. 3, lines in orange) contains seven folds with photosynthetic eukaryotes and cyanobacteria, mainly metabolic (six folds, 66.7%), and more specifically associated with photosynthetic functions (Table S2, line 8). The second group of eight folds (Fig. 3, in red) is shared between photosynthetic eukaryotes and all bacteria. This second group highlights the transfer of cyanobacteria nonspecific folds through the second endosymbiosis. These folds are involved in information function and intracellular processes with four folds (43.8%) and two folds (25.0%) (Table S2, line 7).

The repartition of folds within the three superkingdoms highlights the mixed origin of eukaryota. An equal amount of eukaryotes folds is inherited from bacteria and archaea. The traditional separation of information/metabolism function for archaea/bacteria is not observed. The repartition of folds highlights the second endosymbiosis with folds inherited from cyanobacteria, which are photosynthetic linked folds and folds from all bacteria with more informational function.

SPECIFIC FOLD BLOCKS WITHIN EUKARYOTES

Many folds are specific to eukaryotes (224, 20.9% of the folds), evidencing the fold innovations that took place in eukaryotes. The serialization of the heatmap allows the visualization of blocks representing specific and consistent groups of folds present in groups of species (i.e., clades), which are circumscribed by Dynamic Tree Cut on the hierarchical clustering tree of the folds (see Materials and Methods). Within eukaryotes, 12 blocks are

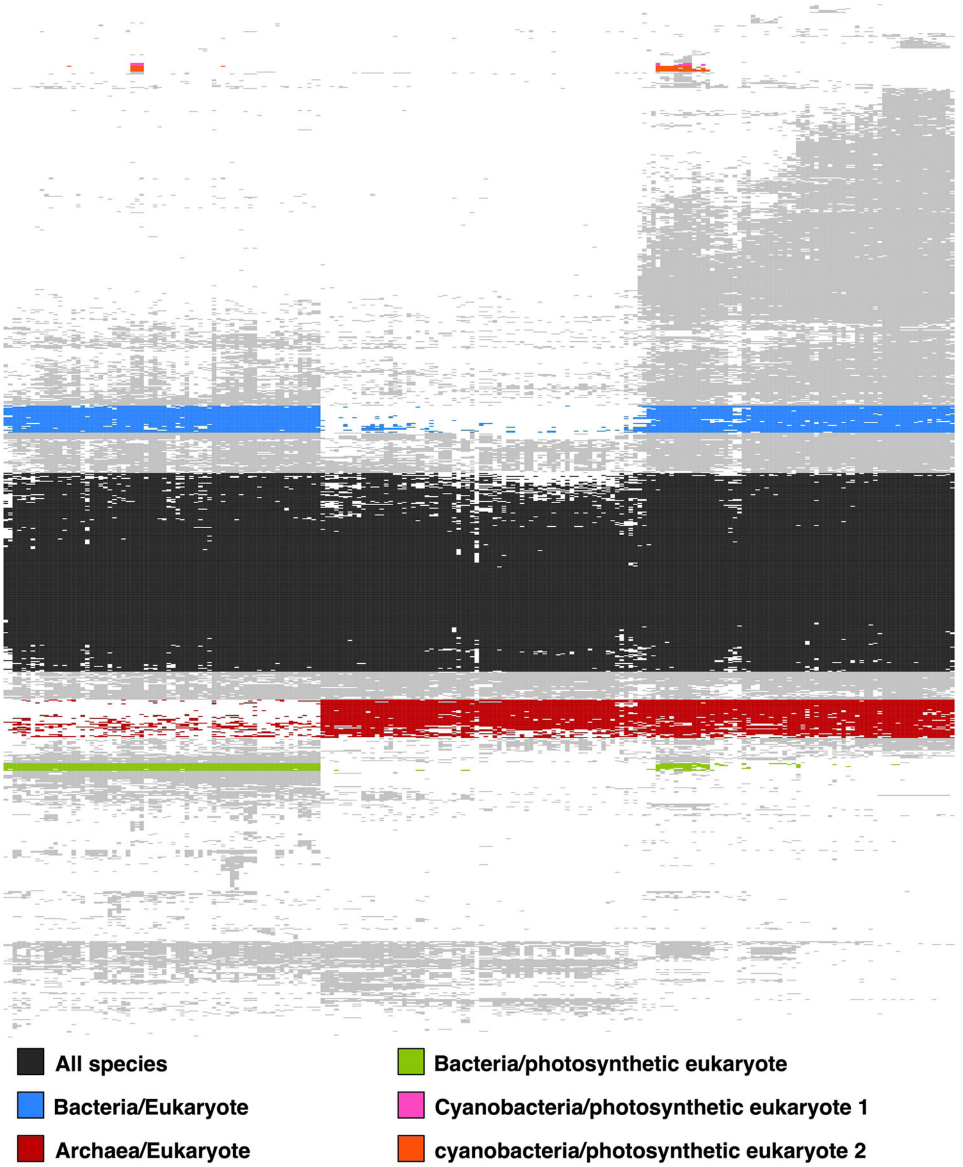

Figure 3. Heatmap of folds and species as in Figure 1, with a color code showing groups of folds shared between two superkingdoms or two distant clades. Three types of groups are extracted with the Dynamic Tree Cut algorithm. The black groups are folds shared within all species. The blue and red groups are folds shared between eukaryotes and one of the two other superkingdoms. The green and orange groups are shared between bacteria and photosynthetic eukaryote groups.

specifically associated with clades (monophyletic taxonomic groups) and are reliably supported by at least three folds. They are highlighted in Figure 4.

The first block encompasses all eukaryotes with 59 folds (5.5%). We observe an overrepresentation of functions related to regulation (19 folds, 26.5%) and intracellular processes (17 folds, 24.0%) (see Table S2, line 4), which is consistent with previous studies of eukaryote specific folds (Kauko and Lehto 2018).

The clustering and high value of RI allows the identification of 11 clades: Opisthokonta, Holozoa, Chozoa, Metazoa, Vertebrata, Gnasthostomata, Tetrapoda, Ecdysozoa, Nematoda, Embryophyta, and Angiospermae. The first seven clades are nested. The two last groups are photosynthetic clades. The folds that constitute synapomorphies for each of these clades are listed in Table 2 and Table S4 with their RI. Within these 11 clades, metabolism and informational functions are clearly underrepresented with 6 (7.1%) and 4 (4.7%) folds, whereas regulation and extracellular processes are overrepresented with 22 (24.5%) and 17 folds (18.6%), respectively (Table S2, line 2). This overrepresentation can be linked to the rise of multicellularity, which constrains cells to make junctions and to communicate.

Table 2. List of putative fold synapomorphies found within eukaryotes: for the 11 blocks specifically associated with clades (monophyletic taxonomic groups) and are reliably supported by at least three folds and with high RI.

Clade	Folds (from SCOP)
Nematoda	e.76 (Viral glycoprotein ectodomain-like), d.62 (pepsin inhibitor-3), a.226 (Her-1), b.169 (MFPT repeat-like)
Ecdysozoa	a.260 (Rhabdovirus nucleoprotein-like), b.102 (Methuselah ectodomain), a.85 (hemocyanin, N-terminal domain), a.163 (crustacean CHH/MIH/GIH neurohormone)
Tetrapoda	a.206 (P40 nucleoprotein), h.3 (Stalk segment of viral fusion proteins), a.61 (retroviral matrix proteins), b.20 (ENV polyprotein, receptor-binding domain), h.6 (apolipoprotein A-II), g.77 (resistin), g.9 (defensin-like), b.63 (oncogene products), d.234 (proguanylin), a.101 (uteroglobin-like), a.212 (KRAB domain [Kruppel-associated box]), d.5 (RNase A-like)
Gnathostomata	a.109 (Class II MHC-associated invariant chain ectoplasmic trimerization domain), d.6 (prion-like), d.9 (IL8-like), d.19 (MHC antigen-recognition domain), d.288 (GTF2I-like repeat)
Vertebrata	h.7 (Synuclein), g.25 (heparin-binding domain from vascular endothelial growth factor), f.50 (Connexin43), a.126 (serum albumin-like), a.26 (4-helical cytokines)
Metazoa	b.54 (Core binding factor beta, CBF), d.200 (integrin beta tail domain), g.1 (insulin-like), a.77 (DEATH domain), g.28 (thyroglobulin type-1 domain), g.27 (FnI-like domain), d.164 (SMAD MH1 domain), g.62 (cysteine-rich DNA binding domain, (DM domain)), g.17 (cystine-knot cytokines), a.277 (TAFH domain-like), g.76 (hormone receptor domain), g.22 (serine protease inhibitors), a.123 (nuclear receptor ligand-binding domain), a.271 (SOCS box-like), f.7 (lipovitellin-phosvitin complex), beta-sheet shell regions), d.217 (SAND domain-like)
Chozoa	b.22 (TNF-like), g.73 (CCHHC domain), g.8 (BPTI-like), a.194 (L27 domain), a.37 (A DNA-binding domain in eukaryotic transcription factors)
Holozoa	g.64 (Somatomedin B domain), d.171 (fibrinogen C-terminal domain-like), d.170 (SRCR-like), a.12 (Kix domain of CBP (creb binding protein)), a.135 (tetraspanin), a.215 (a middle domain of Talin 1), g.16 (Trefoil/Plexin domain-like), g.12 (LDL receptor-like module), a.256 (RUN domain-like), g.65 (Notch domain), g.18 (complement control module/SCR domain), g.14 (Kringle-like)
Opisthokonta	a.83 (Guanido kinase N-terminal domain), d.246 (mRNA decapping enzyme DcpS N-terminal domain), a.68 (Wiscott–Aldrich syndrome protein, WASP, C-terminal domain), d.370 (BTG domain-like), d.332 (RGC domain-like), f.52 (ATP synthase B chain-like), a.216 (I/LWEQ domain), g.20 (blood coagulation inhibitor (disintegrin)), g.52 (inhibitor of apoptosis [IAP] repeat), e.55 (Rap/Ran-GAP), a.117 (Ras GEF), a.87 (DBL homology domain [DH-domain]), a.205 (Hsp90 co-chaperone CDC37), a.141 (Frizzled cysteine-rich domain), a.221 (Lissencephaly-1 protein [Lis-1, PAF-AH alpha] N-terminal domain)
Angiospermae	a.220 (Hypothetical protein At3g22680), g.13 (crambin-like), g.88 (intrinsically disordered proteins)
Embryophyta	g.69 (Plant proteinase inhibitors), a.52 (bifunctional inhibitor/lipid-transfer protein/seed storage 2S albumin), b.162 (At5g01610-like), b.143 (NAC domain)

Discussion

We have explored the protein fold universe and its repartition among living organisms. We developed a new approach, taking advantage of the knowledge of the phylogeny of species, a phylogenetic coherence index and by using matrices double seriation and clustering methods, which allows rendering the results visually and meaningfully on a heatmap linking fold distribution to species.

We limited the sample to 210 species, that is, 70 bacteria, 70 archaea, and 70 eukaryotes, to keep our results intelligible, and we chose to sample the different depths of the tree of life. In terms of folds, the total number of different folds is currently stable, probably a sign that all common folds are known, even if it is probable that some rare folds are still unknown (Koonin et al. 2002). As they would be present only in a few species, they should presumably not drastically change the heatmap. Moreover, proteomes are not fully annotated, particularly in eukaryotes. Consequently, we need to keep in mind that the absence of folds may not be real absences, but may originate from undetected folds (both unknown folds and undetected known folds). Nevertheless, the detected (present) folds represent a significant part of the fold universe and their distribution in living organisms can therefore be considered as reliable.

Exploring this universe, we have shown that protein folds witness various aspects of the evolutionary history of taxa that bear them, emphasizing the fact that information derived from compared protein structures, which we call comparative molecular anatomy, are reasonable candidates for the description of the tree of life. They bear the trace of vertical transmissions but also of important introgressions.

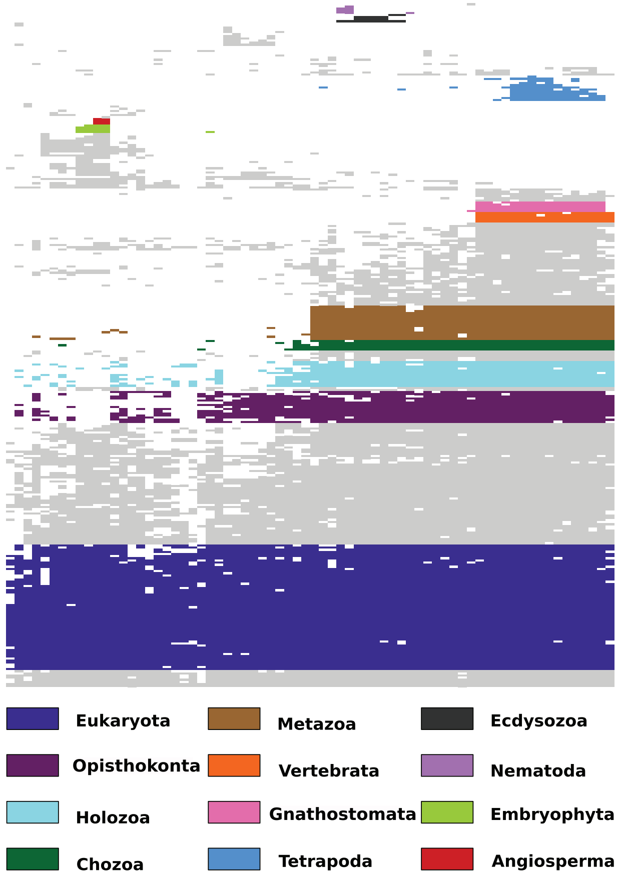

Figure 4. Heatmap of groups of eukaryotic folds. Each group extracted with the Dynamic Tree Cut algorithm matches with a eukaryotic clade. The dark blue groups are folds specific to all eukaryotes. The light green and red are folds specific to photosynthetic clades. The other colors are imbricate clades from Opisthokonta to Nematoda.

First, our results confirm the hypothesis of the dual origin of eukaryotes, as in several other approaches (Rivera and Lake 2004; Pisani et al. 2007; Alvarez-Ponce et al. 2013). Some folds are only shared by bacteria and eukaryotes, whereas others are only shared by archaea and eukaryotes. The number of shared folds is of the same order of magnitude in both cases, which seems in contradiction with the latest work (Brueckner and Martin 2020), but the folds shared by the three kingdoms may also have an archaeal and/or bacterial origin in eukaryotes. Therefore, taking these folds into account, our results are more in line with (Brueckner and Martin 2020). To identify the archaeal or bacterial origin of the folds, it will be necessary to add information, for example by comparing the structures more thoroughly, or by looking for additional similarities of functions or sequences. The functions of the folds shared by two kingdoms do not strictly follow the traditional informational/operational cleavage for archaeal/bacterial contributions to eukaryogenesis (Alvarez-Ponce et al. 2013). Actually, we do find a majority of informational folds in the shared archaea/eukaryotic folds but we also find many folds carrying informational functions in bacteria/eukaryotic folds as well. These results can also be explained because we do not take into account the folds shared by the three kingdoms, which indeed contain many folds with metabolic function. However, it is interesting to note that some bacterial specific informational folds have been maintained in eukaryotic organisms.

We detect synapomorphies for several eukaryotic clades at remarkably fine levels of the tree hierarchy, a sign of the continuous fold innovation process in eukaryotes. These folds are overrepresented within regulation and extracellular or intracellular processes. The number of folds specific to eukaryotes but common to all them is important. It seems possible that the rhythm of fold innovation is not even and has encountered some bursts at the moment of major evolutionary innovations such as the origin of eukaryotes or at the origin of metazoans. This is consistent with the evolution of spliceosomal introns where the moment of substantial gain of introns is concomitant with these events (Rogozin et al. 2012).

For the bacteria and archaea, the number of folds is smaller and their distribution is less in line with our species phylogeny, which has a low resolution for these two kingdoms. In the case of bacteria, it is interesting to note that groups of folds seem characteristic of certain groups of species but many horizontal transfers have taken place. For the archaea, the results should be tempered with the still limited knowledge of this kingdom.

This work brings insight into why phylogenies calculated by standard parsimony with fold data are sometimes contradictory. Indeed, a fully dichotomous tree is inadequate to represent the dual origin of eukaryotes. The numbers of folds shared by eukaryotes and bacteria and by eukaryotes and archaea are quite close. Depending on the species and fold selections made by authors, this number varies slightly and therefore the eukaryotes may by chance be closer to archaea or to bacteria. The same inadequacy of dichotomic representations occurs with the endosymbiosis of the chloroplast. A more progressive approach, for example by calculating separate phylogenies for archaea and eukaryotes and for bacteria and eukaryotes (Raymann et al. 2015), or signal stripping phylogenies as in Pisani et al. (2007), may yield promising results.

Our initial question was whether the folds are good phylogenetic markers. They are definitely good markers, for both ancient and recent events: they are stable over time but also appear regularly enough. However, horizontal transfers are visible and a more detailed study of the distribution of folds in bacteria or archaea will probably allow us to identify them in these organisms better. We also believe that such an approach, relying heavily on current knowledge of the phylogeny of species, will also allow a better understanding of the mechanisms of evolution of folds and the appearance of new folds. Finally, we hope that this study will help to give the reader a clearer and more accurate view of the universe of protein folds and the distribution of folds in living organisms.

AUTHOR CONTRIBUTIONS

JC, GL, and MC designed the study. MR, MC, PI, TJ, and GS produced the data. MR, GS, and MC analyzed the data. MR, JC, GL, GS, and MC wrote the paper.

ACKNOWLEDGMENTS

MR has benefited from a thesis grant from the IPV program of Sorbonne University.

DATA ARCHIVING

Data files are available at https://github.com/MartinRomei/Protein_structure_phylogenetic_marker.

LITERATURE CITED

Adam, P.S., Borrel, G., Brochier-Armanet, C. & Gribaldo, S. (2017) The growing tree of Archaea: New perspectives on their diversity, evolution and ecology. *The ISME Journal*, 11, 2407–2425.

Alvarez-Ponce, D., Lopez, P., Bapteste, E. & McInerney, J.O. (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proceedings of the National Academy of Sciences of the United States of America*,, 110, E1594–E1603.

Bajaj, M. & Blundell, T. (1984) Evolution and the tertiary structure of proteins. *Annual Review of Biophysics & Bioengineering*, 13, 453–492.

Birks, H.J.B., Lotter, A.F., Juggins, S. & Smol, J.P. (2012) Clustering and partitioning. *in* Tracking environmental change using lake sediments: data handling and numerical techniques. Springer Science & Business Media.

Brueckner, J. & Martin, W.F. (2020) Bacterial genes outnumber archaeal genes in eukaryotic genomes. *Genome Biology and Evolution*, 12, 282–292.

Caetano-Anollés, D., Kim, K.M., Mittenthal, J.E. & Caetano-Anollés, G. (2011) Proteome evolution and the metabolic origins of translation and cellular life. *Journal of Molecular Evolution*, 72, 14–33.

Caetano-Anollés, G. (2002) Evolved RNA secondary structure and the rooting of the universal tree of life. *Journal of Molecular Evolution*, 54, 333–345.

Caetano-Anollés, G. & Caetano-Anollés, D. (2003) An evolutionarily structured universe of protein architecture. *Genome Research*, 13, 1563–1571.

Caetano-Anollés, G. & Caetano-Anollés, D. (2005) Universal sharing patterns in proteomes and evolution of protein fold architecture and life. *Journal of Molecular Evolution*, 60, 484–498.

Caetano-Anollés, G. & Nasir, A. (2012) Benefits of using molecular structure and abundance in phylogenomic analysis. *Frontiers in Genetics*, 3, 172.

Caetano-Anollés, G., Wang, M., Caetano-Anollés, D. & Mittenthal, J.E. (2009) The origin, evolution and structure of the protein world. *Biochemical Journal*, 417, 621–637.

Caetano-Anollés, G., Nasir, A., Zhou, K., Caetano-Anollés, D., Mittenthal, J.E., Sun, F.-J. & Kim, K.M. (2014) Review article archaea: the first domain of diversified life. *Archaea*,, 1–26.

Chandonia, J.-.M., Fox, N.K. & Brenner, S.E. (2017) SCOPe: manual curation and artifact removal in the structural classification of proteins - extended database. *Journal of Molecular Biology*, 429, 348–355.

Cheng, H., Schaeffer, R.D., Liao, Y., Kinch, L.N., Pei, J., Shi, S., Kim, B.-.H. & Grishin, N.V. (2014) ECOD: an evolutionary classification of protein domains. *PLOS Computational Biology*, 10, e1003926-18.

Csaba, G., Birzele, F. & Zimmer, R. (2009) Systematic comparison of SCOP and CATH: a new gold standard for protein structure analysis. *BMC Structural Biology*, 9, 23.

Dawson, N.L., Lewis, T.E., Das, S., Lees, J.G., Lee, D., Ashford, P., Orengo, C.A. & Sillitoe, I. (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research*, 45, D289–D295.

Dokholyan, N.V., Shakhnovich, B. & Shakhnovich, E.I. (2002) Expanding protein universe and its origin from the biological Big Bang. *Proceedings of the National Academy of Sciences of the United States of America*, 99, 14132–14136.

Doolittle, R.F. (1995) The multiplicity of domains in proteins. *Annual Review of Biochemistry*, 64, 287–314.

Earle, D. & Hurley, C.B. (2015) Advances in dendrogram seriation for application to visualization. *Journal of Computational and Graphical Statistics*, 24, 1–25.

Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95, 14863–14868.

Fan, L., Wu, D., Goremykin, V., Xiao, J., Xu, Y., Garg, S., Zhang, C., Martin, W.F. & Zhu, R. (2020) Phylogenetic analyses with systematic taxon sampling show that mitochondria branch within Alphaproteobacteria. *Nature Ecology and Evolution*, 4, 1213–1219.

Farris, J. (1983) The logical basis of phylogenetic analysis. In Advances in Cladistics, proceedings of the second meeting of the Willi Hennig Society, Platnick, N.I. and Funk, V.A. (eds.), pp. 1–36. Columbia University Press, New York Botanical Garden, New York.

Farris, J.S. (1989) The retention index and the rescaled consistency index. *Cladistics*, 5, 417–419.

Fitch, W.M. & Margoliash, E. (1967) Construction of phylogenetic trees. *Science*, 155, 279–284.

Gerstein, M. (1998) Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census. *Proteins*, 33, 518–534.

Gerstein, M. & Hegyi, H. (1998) Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiology Reviews*,

Gough, J. (2005) Convergent evolution of domain architectures (is rare). *Bioinformatics*, 21, 1464–1471.

Gower, J.C. & Legendre, P. (1986) Metric and Euclidean properties of dissimilarity coefficients. *J. Classification*, 3, 5–48.

Gruvaeus, G. & Wainer, H. (1972) Two additions to hierarchical cluster analysis. *British Journal of Mathematical and Statistical Psychology*, 25, 200–206.

Hajdu, L.J. (1981) Graphical comparison of resemblance measures in phytosociology. *Vegetatio*, 48, 47–59.

Hennig, W. (1950) Grundzuge einer Theorie der phylogenetischen Systematik. Deutscher zentralverlag.

Hennig, W. (1999) Phylogenetic systematics. Univ. of Illinois Press.

Hillis, D.M., Moritz, C., Mable, B.K. & Olmstead, R.G. (1996) Molecular systematics. Sinauer Associates, Sunderland, MA.

Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield, C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016) A new view of the tree of life. *Nature Microbiology*, 1:16048.

Illergård, K., Ardell, D.H. & Elofsson, A. (2009) Structure is three to ten times more conserved than sequence–a study of structural response in protein cores. *Proteins*, 77, 499–508.

Kauko, A. & Lehto, K. (2018) Eukaryote specific folds: part of the whole. *Proteins Structure, Function, and Bioinformatics*, 86, 868–881.

Kim, K.M. & Caetano-Anollés, G. (2011) The proteomic complexity and rise of the primordial ancestor of diversified life. *BMC Evolutionary Biology*, 11, 140.

Kim, K.M. & Caetano-Anollés, G. (2012) The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms. *BMC Evolutionary Biology*, 12, 13.

Koonin, E.V., Wolf, Y.I. & Karev, G.P. (2002) The structure of the protein universe and genome evolution. *Nature*, 420, 218–223.

Kurland, C.G. & Harish, A. (2015) The phylogenomics of protein structures: the backstory. *Biochimie*, 119, 284–302.

Landenmark, H.K.E., Forgan, D.H. & Cockell, C.S. (2015) An estimate of the total DNA in the biosphere. *PLOS Biology*, 13:e1002168.

Langfelder, P., Zhang, B. & Horvath, S. (2008) Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R. *Bioinformatics*, 24, 719–720.

Lankester, E.R. (1870) II.—On the use of the term homology in modern zoology, and the distinction between homogenetic and homoplastic agreements. *Ann. Mag. Nat. Hist.*, 6, 34–43.

Lebart, L., Morineau, A. & Piron, M. (2000) Statistique exploratoire multidimensionnelle. Dunod.

Lecointre, G. & Guyader, HLe. (2017) La classification phylogénétique du vivant. 4e ed. Belin.

Levitt, M. (2009) Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 11079–11084.

Lewis, T.E., Sillitoe, I. & Lees, J.G. (2019) cath-resolve-hits: a new tool that resolves domain matches suspiciously quickly. *Bioinformatics*, 35, 1766–1767.

Lin, J. & Gerstein, M. (2000) Whole-genome trees based on the occurrence of folds and orthologs: implications for comparing genomes on different levels. *Genome Research*, 10, 808–818.

Maddison, W.P. & Maddison, D.R. (2019) Mesquite: a modular system for evolutionary analysis. https://www.mesquiteproject.org

Mannige, R. (2014) Origination of the protein fold repertoire from oily pluripotent peptides. *Proteomes*, 2, 154–168.

Murzin, A.G., Brenner, S.E., Hubbard, T. & Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247, 536–540.

Nasir, A., Kim, K.M. & Caetano-Anollés, G. (2014) Global patterns of protein domain gain and loss in superkingdoms. *PLOS Computational Biology*, 10, e1003452-18.

Nepomnyachiy, S., Ben-Tal, N. & Kolodny, R. (2017) Complex evolutionary footprints revealed in an analysis of reused protein segments of diverse lengths. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 11703–11708.

Orengo, C.A., Jones, D.T. & Thornton, J.M. (1994) Protein superfamilies and domain superfolds. *Nature*, 372, 631–634.

Pascual-García, A., Abia, D., Ortiz, A.R. & Bastolla, U. (2009) Cross-over between discrete and continuous protein structure space: insights into automatic classification and networks of protein structures. *PLOS Computational Biology*, 5, e1000331-20.

Pisani, D., Cotton, J.A. & McInerney, J.O. (2007) Supertrees disentangle the chimerical origin of eukaryotic genomes. *Molecular Biology and Evolution*, 24, 1752–1760.

Raymann, K., Brochier-Armanet, C. & Gribaldo, S. (2015) The two-domain tree of life is linked to a new root for the Archaea. *Proceedings of the National Academy of Sciences of the United States of America*, 112, 6670–6675.

Rivera, M.C. & Lake, J.A. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes. *Nature*, 431, 152–155.

Rogozin, I.B., Carmel, L., Csuros, M. & Koonin, E.V. (2012) Origin and evolution of spliceosomal introns. *Biology Direct*, 7, 11.

Rose, P.W., Prlić, A., Bi, C., Bluhm, W.F., Christie, C.H., Dutta, S., Green, R.K., Goodsell, D.S., Westbrook, J.D., Woo, J., et al. (2015) The RCSB Protein Data Bank: views of structural biology for basic and applied research and education. *Nucleic Acids Research*, 43, D345-D356.

Sadowski, M.I. & Taylor, W.R. (2009) Protein structures, folds and fold spaces. *Journal of Physics: Condensed Matter*, 22:033103.

Skolnick, J., Arakaki, A.K., Lee, S.Y. & Brylinski, M. (2009) The continuity of protein structure space is an intrinsic property of proteins. *Proceedings of the National Academy of Sciences of the United States of America*, 106, 15690–15695.

Sokal, R.R. & Michener, C.D. (1958) A statistical method for evaluating systematic relationships. University of Kansas.

Taylor, W.R. (2002) A "periodic table" for protein structures. *Nature*, 416, 657–660.

Wang, M. & Caetano-Anollés, G. (2009) The evolutionary mechanics of domain organization in proteomes and the rise of modularity in the protein world. *Structure*, 17, 66–78.

Wang, M., Boca, S.M., Kalelkar, R., Mittenthal, J.E. & Caetano-Anollés, G. (2006) A phylogenomic reconstruction of the protein world based on a genomic census of protein fold architecture. *Complex*, 12, 27–40.

Wang, M., Yafremava, L.S., Caetano-Anollés, D., Mittenthal, J.E. & Caetano-Anollés, G. (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Research*, 17, 1572–1585.

Wang, M., Jiang, Y.-.Y., Kim, K.M., Qu, G., Ji, H.-.F., Mittenthal, J.E., Zhang, H.-.Y. & Caetano-Anollés, G. (2011) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Molecular Biology and Evolution*, 28, 567–582.

Wickham, H. (2016) ggplot2 - elegant graphics for data analysis. Springer International Publishing, Cham.

Williams, T.A., Cox, C.J., Foster, P.G., Szöllősi, G.J. & Embley, T.M. (2020) Phylogenomics provides robust support for a two-domains tree of life. *Nature Ecology and Evolution*, 4, 138–147.

Wilson, D., Madera, M., Vogel, C., Chothia, C. & Gough, J. (2007) The SUPERFAMILY database in 2007: families and functions. *Nucleic Acids Res.*, 35, D308–D313.

Wilson, D., Pethica, R., Zhou, Y., Talbot, C., Vogel, C., Madera, M., Chothia, C. & Gough, J. (2009) SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Research*, 37, D380–D386.

Winstanley, H.F., Abeln, S. & Deane, C.M. (2005) How old is your fold? *Bioinformatics*, 21, i449–i458.

Wolf, Y.I., Brenner, S.E., Bash, P.A. & Koonin, E.V. (1999) Distribution of protein folds in the three superkingdoms of life. *Genome Research*, 9, 17–26.

Wolf, Y.I., Grishin, N.V. & Koonin, E.V. (2000) Estimating the number of protein folds and families from complete genome data. *Journal of Molecular Biology*, 299, 897–905.

Yang, S., Doolittle, R.F. & Bourne, P.E. (2005) Phylogeny determined by protein domain content. *Proceedings of the National Academy of Sciences of the United States of America*, 102, 373–378.

Zaremba-Niedzwiedzka, K., Caceres, E.F., Saw, J.H., Bäckström, D., Juzokaite, L., Vancaester, E., Seitz, K.W., Anantharaman, K., Starnawski, P., Kjeldsen, K.U., et al. (2017) Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature*, 541, 353–358.

Zhang, C.T. (1997) Relations of the numbers of protein sequences, families and folds. *Protein Engineering, Design and Selection*, 10, 757–761.

Zuckerkandl, E. & Pauling, L. (1965) Molecules as documents of evolutionary history. *Journal of Theoretical Biology*, 8, 357–366.

Associate Editor: S. Charlat
Handling Editor: A. McAdam

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Supporting Information
Figure S1. Reference phylogenetic tree from Lecointre & Le Guyader (Lecointre and Guyader 2017) and Hug et al. (Hug et al. 2016).
Figure S2. This figure is a supplementary legend for Figure 1, which specifies all background colours distributed in the three superkingdoms.
Figure S3. Heatmap of species and folds sorted like in Figure 1.
Figure S4. Highlights of co-occuring folds in a given protein.
Figure S5. Heatmap of species and folds sorted like in Figure 1.
Figure S6. Heatmaps showing protein CATH Topologies (left) and ECOD X-levels (rigth) repartition through the diversity of life.
Figure S7. Heatmaps showing protein CATH Topologies (left) and ECOD X-levels (rigth) repartition through the diversity of life.
Table S1. List of the 210 selected organisms, their phylum, superkingdom, taxid (from NCBI) and their associated color in the heatmap.
Table S2. Repartition of fold main functions across taxonomic groups or pairs of groups.
Table S3. Repartition of SCOP class level within all folds and within folds present in 90 % of species.
Table S4. List of putative fold synapomorphies found within eukaryotes.
Table S5. Among the 210 proteomes, proportion of proteins with at least one significant hit with HMMS in either CATH, SCOP or ECOD databases.