



# Origins and Functional Significance of Eukaryotic Protein Folds

Martin Romei<sup>1,2</sup> · Mathilde Carpentier<sup>1</sup> · Jacques Chomilier<sup>2</sup> · Guillaume Lecointre<sup>1</sup>

Received: 27 February 2023 / Accepted: 3 October 2023 / Published online: 7 December 2023  
© Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

Folds are the architecture and topology of a protein domain. Categories of folds are very few compared to the astronomical number of sequences. Eukaryotes have more protein folds than Archaea and Bacteria. These folds are of two types: shared with Archaea and/or Bacteria on one hand and specific to eukaryotic clades on the other hand. The first kind of folds is inherited from the first endosymbiosis and confirms the mixed origin of eukaryotes. In a dataset of 1073 folds whose presence or absence has been evidenced among 210 species equally distributed in the three super-kingdoms, we have identified 28 eukaryotic folds unambiguously inherited from Bacteria and 40 eukaryotic folds unambiguously inherited from Archaea. Compared to previous studies, the repartition of informational function is higher than expected for folds originated from Bacteria and as high as expected for folds inherited from Archaea. The second type of folds is specifically eukaryotic and associated with an increase of new folds within eukaryotes distributed in particular clades. Reconstructed ancestral states coupled with dating of each node on the tree of life provided fold appearance rates. The rate is on average twice higher within Eukaryota than within Bacteria or Archaea. The highest rates are found in the origins of eukaryotes, holozoans, metazoans, metazoans *stricto sensu*, and vertebrates: the roots of these clades correspond to bursts of fold evolution. We could correlate the functions of some of the fold synapomorphies within eukaryotes with significant evolutionary events. Among them, we find evidence for the rise of multicellularity, adaptive immune system, or virus folds which could be linked to an ecological shift made by tetrapods.

**Keywords** Protein structure · Protein folds · Evolution · Informational genes · Operational genes

## Introduction

Folds, as defined by the SCOP classification (Murzin et al. 1995), represent the architecture and topology of a protein domain. Categories of folds are very few compared to the astronomical number of sequences, 1232 reported folds so far (Chandonia et al. 2019), and this number is increasing very slowly. Only 35 new folds have been identified between

2018 and 2022 in the new version of SCOP named SCOPe (Chandonia et al. 2019). The predictive number of folds is estimated to reach 4000 to 10000 according to some authors (Govindarajan et al. 1999; Coulson and Moulton 2002). It has been shown in a number of studies (Caetano-Anollés and Caetano-Anollés 2003; Yang et al. 2005; Caetano-Anollés 2005; Kauko and Lehto 2018) that eukaryotes host many specific folds. These specific folds provide reliable synapomorphies which characterize eukaryotes or clades within eukaryotes (Romei et al. 2022). However, eukaryotes share some folds with the other super-kingdoms, resulting in two types of eukaryotic folds: either specific to eukaryotes or shared with Bacteria or Archaea (Caetano-Anollés 2005; Kauko and Lehto 2018). These latter can be the imprint of the mixed origin of Eukaryota. It is now admitted that the Eukaryota have a dual origin, resulting from the endosymbiosis of an Archaea and a Bacteria (McInerney et al. 2014). What are the inputs of folds of probable Archaea and Bacteria origins in eukaryotic functions? The gene content of Eukaryota is inherited from both super-kingdoms but it has

---

Handling editor: **David Alvarez-Ponce**.

---

Martin Romei and Mathilde Carpentier have contributed equally to this work.

---

✉ Mathilde Carpentier  
mathilde.carpentier@mnhn.fr

<sup>1</sup> Institut Systématique Evolution Biodiversité (ISYEB UMR 7205), Sorbonne Université, MNHN, CNRS, EPHE, UA, Paris, France

<sup>2</sup> IMPMC (UMR 7590), BiBiP, Sorbonne Université, CNRS, MNHN, Paris, France

been pointed out that the repartition of functions is uneven: a greater proportion of informational genes would come from Archaea and a greater proportion of operational genes would come from Bacteria. The dual origin of Eukaryota is quite visible in the fold content (Caetano-Anollés and Caetano-Anollés 2003, p. 20): many folds common to all eukaryotes are also common either with Archaea or with Bacteria. In the literature, the functions inherited from Bacteria and Archaea have been studied at the sequence level using networks (Alvarez-Ponce et al. 2013; Méheust et al. 2018). Several studies (Rivera et al. 1998; Alvarez-Ponce et al. 2013) show that genes inherited from Bacteria and Archaea carry different functions that can be split into two groups: informational functions, which correspond to the information machinery and its regulation (replication, transcription, translation, etc.), and operational functions (metabolism, structures, etc.), which are the other functions. Studies on gene networks (Alvarez-Ponce et al. 2013) and on protein structures (Kauko and Lehto 2018) show that the Archaea ancestor brought a majority of informational functions while the Bacteria ancestor brought a lot of operational functions. However, more recent studies on symbiogenetic genes, which are a fusion of sequences of Archaea and Bacteria origins, show that these separate origins of functions are not so obvious (Cotton and McInerney 2010; Méheust et al. 2018).

Several attempts have been made in order to understand the dynamics of fold emergence, diversification, and disappearance (Winstanley et al. 2005; Wang et al. 2007, 2011). A reconstruction of phylogenetic trees based on folds repartition allows dating the relative age of folds. This relative age of folds has been studied and results in a relative dating of the four SCOP classes: mainly alpha, mainly beta, alpha + beta, alpha, and beta. The reconstructions show that the alpha + beta class is more ancient than the alpha and beta class (Caetano-Anollés 2005; Winstanley et al. 2005). Some studies use this reconstruction and external information to date fold appearances and replace them in a global evolutionary context (Wang et al. 2007, 2011). They use phylogenetic reconstruction based on the abundance of folds to determine a universal molecular clock of protein folds that is calibrated with geological ages by means of folds whose function may be dated.

Another way to date folds is to reconstruct ancestral states on a reference tree, with branch length corresponding to geological time. This reconstruction has two interests: the datation of protein folds can define local rate of evolution on each branch in number of appearances per million years and a global rate of evolution. The ancestral states reconstruction allows linking events of fold innovations to clades. Doing so, folds become synapomorphies.

The aim of this study is to understand eukaryotic folds origins and functions. On a selected sample of 210 species

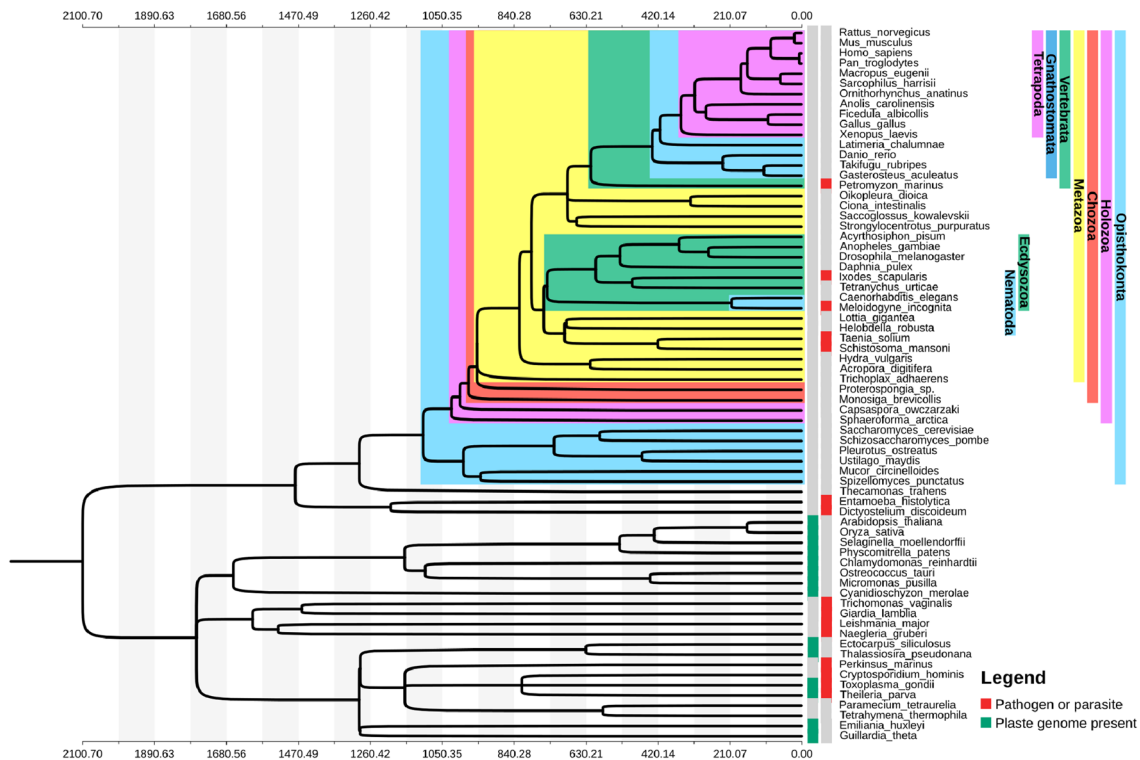
equally spread among the three super-kingdoms (Romei et al. 2022), we will first present the folds within eukaryotes inherited from Bacteria and Archaea. We will also discuss the number of folds inherited and the functions supplied. Then, the numerous folds that appeared during eukaryotic evolution will be explored. We will correlate the functions of some of the fold synapomorphies with significant evolutionary events that occurred at the time of their origin.

## Material and Methods

### Species Selection

The dataset samples 210 species on the tree of life with 70 eukaryotes, 70 Bacteria, and 70 Archaea (Romei et al. 2022). Species are chosen to sample all main lineages while maximizing the divergence time among them, provided their genomes were available. The original reference trees are from Lecointre and Le Guyader (Lecointre and Le Guyader 2017). The trees have been updated by adding Archaea and in particular Asgard Archaea with Hug et al. 2016. The reference phylogeny of the Eukaryotes is presented Fig. 1 and the entire reference tree and the species table list are provided in Supplementary Materials (files S1 and S2, respectively). 13 of the eukaryotes are parasites or have an anaerobic lifestyle (in red in Fig. 1). These characteristics are listed in the Supplementary Table S1 (list of species). We also have retrieved the mitochondrial proteomes from the NCBI database (<https://www.ncbi.nlm.nih.gov/genome/organelle/>) for 52 eukaryotic species. The other 18 eukaryotes were either lacking a mitochondrion or missing from the NCBI database. We have retrieved the proteomes of the chloroplasts for the 12 photosynthetic organisms (in green in Fig. 1) and the proteomes of the 2 apicoplasts (for *Toxoplasma gondii* and *Theileria parva*).

Branch lengths and divergence times were extracted from Time Tree of Life (Kumar et al. 2017) and added to the tree of the 210 species. Time Tree of Life is a searchable tree of life with a real-time scale built from the results of more than a thousand published studies. Only 169 of the 210 species are reported in Time Tree of Life and there are some inconsistencies between Time Tree of life and our reference tree. A Python script has been developed to browse all nodes, check the leaf, and compare them between our reference tree and the tree from Time Tree of Life. For the nodes that are equivalent, the branch lengths extracted from Time Tree of Life are added to our reference tree. If a node from the reference tree has no equivalent in Time Tree of Life (the clade doesn't exist in it), the lengths of the two branches flanking the node are unknown and set to 0 (the node is collapsed).



**Fig. 1** Phylogenetic tree of the 70 eukaryotic species selected for this study (Romei et al. 2022). The complete tree of the 210 species is provided in the Supplementary Materials file S2. The branch lengths are in million year and have been extracted from Time Tree of Life

(see text). The parasitic or anaerobic lifestyle organisms are highlighted in red. The organisms having a chloroplast are in green. The clades in color are those highlighted in Fig. 5 (Color figure online)

## Fold Assignments

The proteins encoded in the complete proteomes and their structural annotation were retrieved from SUPERFAMILY 2.0 database (Gough et al. 2001; Pandurangan et al. 2019) except for the most recent Asgard proteomes which have been scanned with the HMMs via the Superfamily webserver. For each proteome, folds are associated with a binary value which accounts for the presence (1) or absence (0) of the fold when the HMM-associated E-value is, respectively, smaller or greater than  $10^{-4}$  for at least one protein in the proteome. This assignment results in a matrix of 1073 rows (folds) and 210 columns (species). The assignment files and the binary matrix are provided in Supplementary Materials (Supplementary files S3 and S4, respectively) and the R scripts to calculate the matrix are provided in Supplementary files S5 and S6. 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. The folds present in the proteomes of plastids or mitochondria have been also identified and the list is provided in supplementary materials.

## Fold Clustering

The purpose of the study is to delimitate consistent fold groups which can be linked to evolutionary history of Eukaryota. A clustering of the folds has been made by first calculating an Ascendant Hierarchical Clustering with the Ochiai similarity index-derived distance (Gower and Legendre 1986):  $d = \sqrt{1 - S}$  with  $S = \frac{n11}{\sqrt{(n11+n10) \times (n11+n01)}}$  where n11 corresponds to the number of folds present in two species; n10 and n01 correspond to the contingency of different binary data. This distance excludes co-occurring absences (n00) which are desirable because the co-occurring absence of a character is not a rational criterion to group species, and we don't know whether a fold is absent or whether the HMMs didn't allow its identification. Indeed, only  $59\% \pm 10\%$  of the proteins of the Eukaryotic proteomes is annotated with at least one fold. This distance has linear properties and allows the clusters to be separated locally while respecting the global distribution (Hajdu 1981; Gower and Legendre 1986). The ascendant hierarchical clustering produces a tree that is cut at different heights using the Dynamic Tree Cut algorithm in order to define the clusters

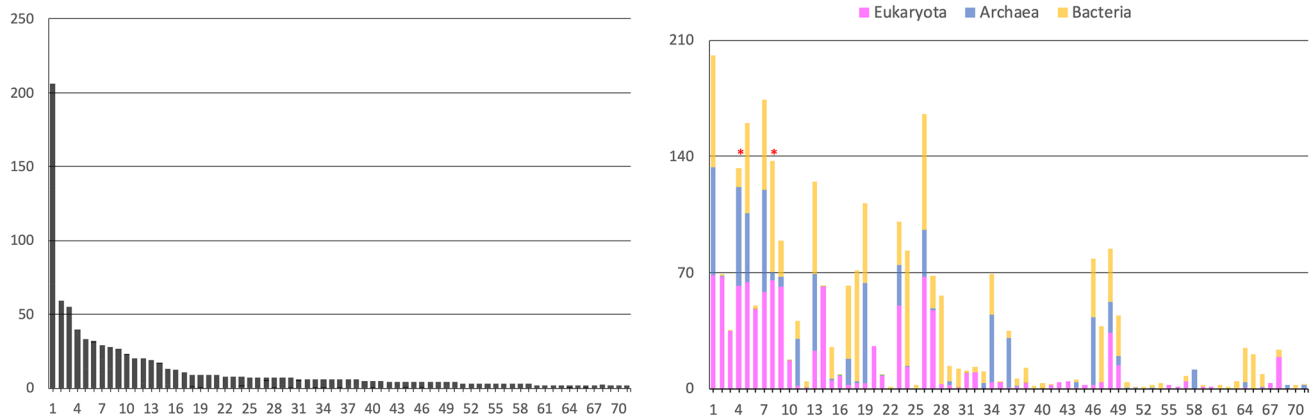
(Langfelder et al. 2008). This method works through iteration of an adaptive process of cluster decomposition and combination until the number of clusters becomes stable. It has four shape criteria to separate clusters: (i) each cluster must have a minimum number of elements; (ii) the elements of each cluster must be close to each other; (iii) the clusters must be separated from other elements; and (iv) the elements of each cluster must be tightly connected. This method allows to define clusters of folds fitting the data without using a cutoff threshold based on the percentage of presence and absence, but it has a parameter influencing the cluster sizes: the deepsplit parameter. As this parameter increases, the clusters become smaller because some clusters are split. We have analyzed two levels of clustering with Dynamic Tree Cut: deepsplit parameter set to 2 or 3. At deepsplit 2, we have 71 clusters, each containing from 2 to 206 folds, and 177 folds are unclassified. The distribution of the cluster sizes is presented in Fig. 2, left. The mean number of folds within the three super-kingdoms is represented on Fig. 2 right. The most populated cluster is the first cluster (206 folds). It contains folds shared by all organisms (more than 60% of the organisms of each super-kingdom, see Fig. 2, right, first bar on the left). We wanted to extract clusters that include eukaryotic folds inherited from Bacteria and clusters with eukaryotic folds inherited from Archaea. The terms Archaea/Eukaryota folds and Bacteria/Eukaryota folds will be used to designate those folds, respectively. The Archaea/Eukaryota folds are shared by most of the 70 eukaryotes and most of the 70 Archaea and are missing in the remaining group. To meet these standards, the mean number of folds must be around 140 (70 Archaea and 70 Bacteria) and they spread across Archaea and eukaryotes. The only cluster that meets these standards is the 4th cluster which contains 40 folds. The same approach results in the selection of cluster

8 which contains 28 Bacteria/Eukaryota folds. Both are highlighted with a red dot above the bar in Fig. 2, right.

With deepsplit level 3, we have selected 9 clusters of eukaryotic folds characterizing 9 clades in the eukaryote phylogeny: Opisthokonta, Holozoa, Chozoa, Metazoa, Vertebrata, Gnathostomata, Tetrapoda, Nematoda, and Ecdysozoa. The associated folds appear as reliable synapomorphies and they are involved in known functions in today's organisms (Romei et al. 2022). All clusters are available in the Table in Supplementary Materials file S7.

## Functional Annotation of Folds

Functional annotations are retrieved from the SUPERFAMILY database (Wilson et al. 2007). There are seven general categories of functions: metabolism, information, intracellular processes, regulation, other, general, and extracellular processes. These general categories are further divided into 50 detailed functions. The analysis of Eukaryota functions inherited from Bacteria and Archaea separates functions in two groups: operational functions and informational functions. The informational functions are linked to translation, ribosomal structure, biogenesis, RNA processing and modification, transcription, DNA replication, recombination and repair, chromatin structure, and dynamic. The operational functions are all the other functions except unknown or unannotated functions (Alvarez-Ponce et al. 2013). Folds may have more than one function assignment. In the SUPERFAMILY database, 10.2% of the folds are associated with several functions. To maintain the same weight for all folds and to perform reliable comparisons, the sum of all the functions associated to each fold is set to 1.



**Fig. 2** Left: Cluster size (in number of folds) distribution for the first level of clustering (deepsplit parameter set to 2). Right: In ordinate: mean number of species where the folds of the cluster are present; the three super-kingdoms are in different colors. In abscissa: cluster ID.

The three horizontal lines are an indication of the number of species in one (70), two (140) or three super-kingdoms (210). Red dots refer to clusters gathering folds shared by Eukaryota and Archaea (cluster 4) and by Eukaryota and Bacteria (cluster 8) (Color figure online)

## Ancestral State Reconstruction

The ancestral states were reconstructed with PastML (Ishikawa et al. 2019) which infers ancestral characters using maximum likelihood (ML) or parsimony. We used the prediction method MPPA for ML and ACCTRAN and DELTRAN optimizations for parsimony. In the analyses, we use the results of ML. Full results are available in Supplementary Materials file S8.

## Results and Discussion

### Folds Shared by Two Super-Kingdoms

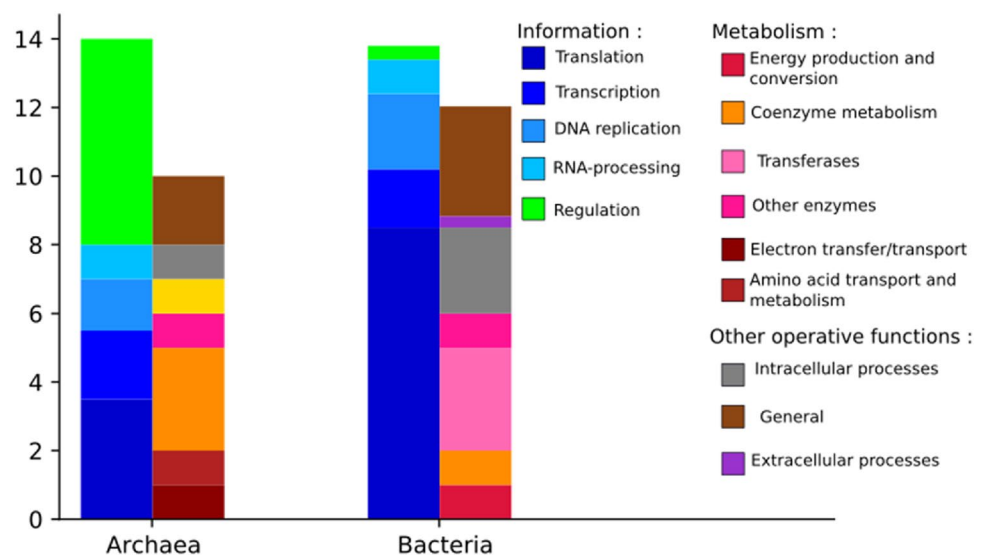
We have delimited a cluster of 28 folds shared by Eukaryota and Bacteria (cluster 8 in Fig. 2) and a cluster of 40 folds shared by Eukaryota and Archaea (cluster 4 in Fig. 2). We have analyzed the distribution of the folds from those clusters across species of each super-kingdom. In the Bacteria/Eukaryota cluster, the folds are present in 56 to 70 Bacteria and in 53 to 70 eukaryotes depending on the considered fold. The Bacteria and the Eukaryota have on average 26.7 and 26.2 of the 28 folds of the cluster, respectively (and the Archaea 1.8). The eukaryotic species missing the most folds are of the parasitic/anaerobic species: *Giardia lamblia* has only 10 of the folds, *Trichomonas vaginalis* has 13, *Cryptosporidium hominis* and *Entamoeba histolytica* have 17, and *Leishmania major* and *Perkinsus marinus* have 24 of the folds. In the Archaea/Eukaryota cluster, the folds are present in 35 to 70 Archaea and in 37 to 70 eukaryotes depending on the considered fold. The Archaea and the Eukaryota have on average 33.8 and 35.5 of the 40 folds of the cluster, respectively (and the Bacteria 6.6). Only the fold a.144 of the Bacteria/Eukaryota cluster is present in a mitochondrion,

that of *Cyanidioschyzon merolae*. This is the L20 protein of the 50S ribosome, which is involved in translation which is one of the information functions. This fold is present in 68 eukaryotes and 70 Bacteria but is not found in the mitochondria except for *Cyanidioschyzon merolae*.

We have looked at the fold functions linked to operational functions and informational functions. The informational functions are all those linked to translation, ribosomal structure, biogenesis, RNA processing and modification, transcription, DNA replication, recombination and repair, chromatin structure, and dynamic. The operational functions are all the other functions except unknown or unannotated functions. Among these two clusters, some folds are not annotated or have an unknown function, 16 within the Archaea/Eukaryota cluster, and 3 within the Bacteria/Eukaryota cluster. The distribution of the functions in the clusters of folds shared by Archaea/Eukaryota and/or shared by Bacteria/Eukaryota is detailed in Fig. 3. The repartition of fold functions is in agreement with other studies (Alvarez-Ponce et al. 2013; Kauko and Lehto 2018; Méheust et al. 2018) for folds shared with Archaea: we have a majority of informational folds, 14, vs 10 operational folds. The informational fold category is divided into regulation function with 6 folds, transcription with 3.5 folds (3 folds and one that has two different functions), translation with 2 folds, DNA replication with 1.5 folds (one fold and one that has two different functions), and RNA processing with one fold.

In the cluster of the folds shared by Bacteria and Eukaryota, Fig. 3 shows that the folds with an informational function are more represented than the operational function. The majority of informational is linked to translation functions (9 out of 13). These folds are not present in all Bacteria and Eukaryota: only 5 of 8.5 folds are present under a constraint of 85%. Operational functions are dominated by transferases (3 folds, present in 85% of the species), operational functions

**Fig. 3** Number of folds linked to informational function in blue and green and operational function in red, pink, and brown for cluster of folds shared between Eukaryota and Archaea (left) and between Eukaryota and Bacteria (right) (clusters 4 and 8, respectively) (Color figure online)



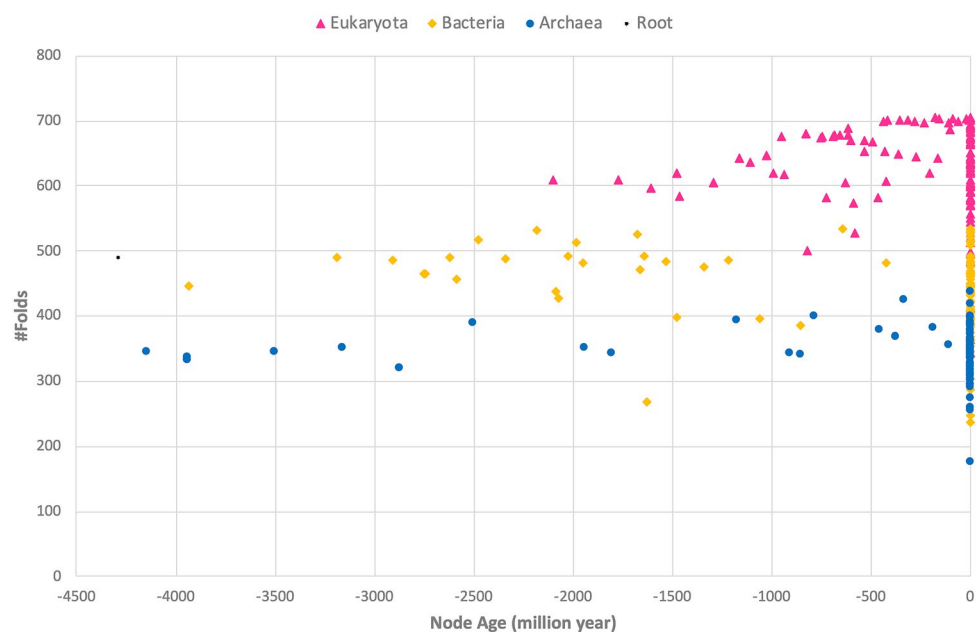
are completed by general (, 3.5 folds, 2 folds at 85%) and intracellular process (3 folds, 2 folds at 85%) (Fig. 4).

We also analyzed all the folds found in mitochondria. We identified 38 folds in mitochondrion. They are listed in Supplementary Materials Table S9. 27 of them are in the first cluster, meaning that they are shared by almost all organisms of the 3 super-kingdoms. 8 folds are from other clusters that group folds present in the 3 super-kingdoms but in less organisms (clusters 5, 7, and 48). The 3 last folds are as follows: the fold a.144 that was discussed earlier; the fold f.18, F1F0 ATP synthase subunit A, that is present in 29 eukaryotes, 58 Bacteria, and 1 Archaea (*Methanosarcina barkeri*); and the fold d.95 (Homing endonuclease like) which is present in 8 eukaryotes, 28 Archaea, and 31 Bacteria. The folds found in the mitochondrion are mainly involved in metabolic functions (19 folds), but also in information (11 folds), regulation (4 folds), intracellular processes (2 folds), and general (onefold). The 42 folds that can be found in plastids are also listed in Supplementary Materials Table S10. 23 of them are in the first cluster, and they are the same as those found in mitochondria. Two informational folds of the Bacteria/Eukaryota cluster are present in plastids (folds a.144 and a.108). One fold of the Archaea/Eukaryota cluster that is involved in regulation, more specifically in signal transduction (b.86, Hedgehog/intein (Hint) domain) can also be found in plastids. All folds annotated as part of the photosystem are also found in plastids (folds g.36, f.29, d.187, d.73, f.31, f.55, f.26). They are shared between all cyanobacteria and most photosynthetic eukaryotes.

The result obtained with Dynamic Tree Cut can be discussed on two points: the repartition of function within the two clusters and the comparison of functions inherited by

eukaryotes from the two clades. For the first point, studies conducted on gene networks (Alvarez-Ponce et al. 2013), on symbiogenetic genes (Méheust et al. 2018), and on folds (Kauko and Lehto 2018) agree with the fact that Archaea and Eukaryota share a majority of informational functions while Bacteria and Eukaryota folds share a majority of operational functions. Our clusters show similar results for Archaea/Eukaryota folds but Bacteria/Eukaryota fold clusters show a number of informational folds higher than operational folds. The selection of folds which have more than 85% of presence within eukaryotes and within Bacteria for our Eukaryota/Bacteria cluster invert the function proportions; however, the number of informational folds is never neglectable. For the second point, the same studies (Alvarez-Ponce et al. 2013; Kauko and Lehto 2018; Méheust et al. 2018) show that operative folds are mainly inherited from Bacteria but Bacteria supply most of the folds linked to informational functions too. The number of folds linked to translation is the highest number of information structures with 24 folds followed by transcription and replication with 15 and 12 structures in the article of Kauko et. A.I (Kauko and Lehto 2018). This repartition of informational functions is similar to our repartition of functions and reflects the supply of informational folds which are not negligible. However, in these other studies, the number of gene/structure inherited from Bacteria is higher than from Archaea. In the study based on folds (Kauko and Lehto 2018), the number of folds in each group is higher than the number of folds in the clusters we analyzed, especially for Bacteria (54 vs 40 for Archaea and 231 vs 28 for Bacteria). In this same study, the selection of folds does not use a criterion of selection on species but only of absence in the third super-kingdom (less

**Fig. 4** Number of folds present at each node vs. the age of the node. 0 is present time and the ordinate is the number of folds present in each of the 210 organisms



than 0.5% of presence). The folds shared by a few species of a super-kingdom are very less likely to originate from the first endosymbiosis and retrace origins within Eukaryota. Indeed, the second endosymbiosis also brought folds from Bacteria and it can contribute to blur results. This bias is highlighted in the study of gene networks (Alvarez-Ponce et al. 2013) where the number of bacterial genes shared with plants is higher than in other eukaryotes.

### Fold Appearance and Distribution

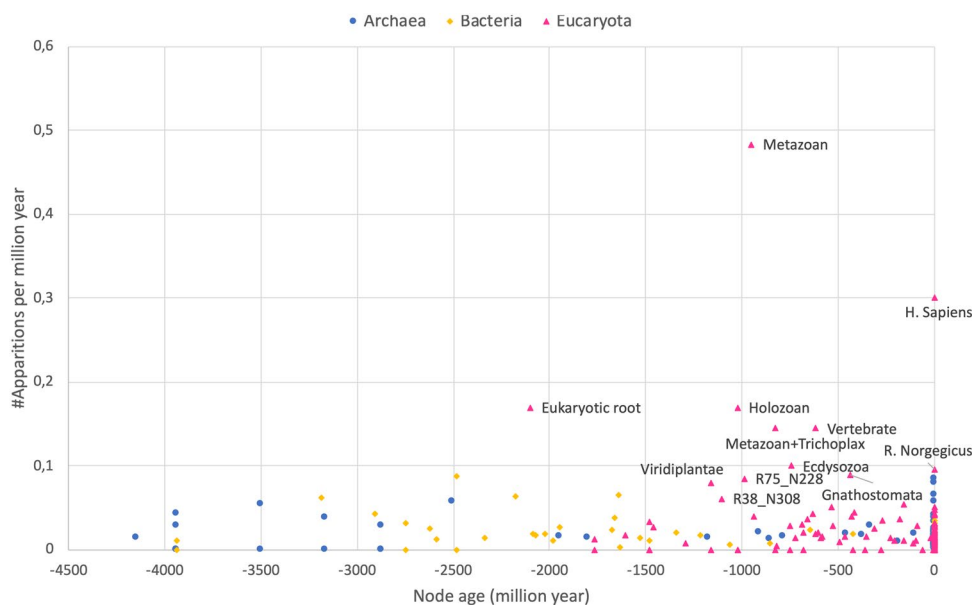
According to the ancestral state reconstruction calculated with Maximum Likelihood, the number of folds present at each node differs among super-kingdoms: there are fewer folds in archaeal organisms (344.5 on average) than in bacterial organisms (443.3 on average) and than in eukaryotic organisms (623.3 on average) (see Fig. 5). These differences in fold contents across the three super-kingdoms are also found in the studies of the fold, superfamily, or family occurrences in the genomes. In 2013 (Bukhari and Caetano-Anollés 2013), Bukhari and Caetano-Anollés counted 1733 SCOP superfamilies distributed as follows: 885 in Archaea, 1312 in Bacteria, and 1508 in Eukaryota. At the SCOP family level, Staley et al. found, in 2018 (Staley and Caetano-Anollés 2018), 703 families in Archaea, 1510 families in Bacteria, and 1696 families in Eukaryota. The methodologies are different but convergence of results is a footprint of the same content variations across the super-kingdoms. In our study, there are 490 folds present at the root but 222 of them have a probability too low to be considered as certain by PASTML (Ishikawa et al. 2019). Indeed, 194 of them disappear on either one of the two next branches of the phylogeny and the others on the following branches. Consequently,

only 268 folds have of high enough probability to be present at the root. In 2016, Weiss et al. reconstructed the ancestral genome of the eukaryotic last common ancestor (Weiss et al. 2016). They identified 355 protein families. Some of these families may share the same fold but this number of families is very similar to our number of folds at the root.

We have calculated the number of fold appearances per million years onto each branch of the tree. The number of fold appearances per million years is very similar for all three kingdoms except for a few eukaryotic nodes. We have a basal rate of appearances of 0.017 for Archaea, 0.016 for Bacteria, and 0.028 for eukaryotes. The rate for the eukaryotic root branch is high, above all other rates but 2 (metazoans and *H. sapiens* in Fig. 5). The highest rate is for Metazoa (36 fold appearances, rate of 0.49), and in decreasing order Eukaryota (312, 0.17), Holozoa (14, 0.17), Metazoa *stricto* sensu (18, 0.15), Vertebrata (10, 0.15), Ecdysozoa (10, 0.1), Gnathostomata (16, 0.09), Fungi (10, 0.09), Viridiplantae (49, 0.09), and Opisthokonta (23, 0.06). The rate of fold appearances is also high for *Homo sapiens* and *Rattus norvegicus* but this is an artefact. Actually, the branches of those two species are very short (6.5 and 20.9 million years, respectively) and there are only two fold appearances for both of them. A supplementary study of the number of appearances according to the time branch length shows that some eukaryotic branches have clearly higher numbers of new folds regardless of the length of the branch (cf. Supplementary Figure S11\_2). Therefore, the distribution of the number of appearances as shown in Fig. 5 is not due to shorter branches within eukaryotes.

We also have calculated the number of fold appearances and disappearances between the eukaryotic root and each eukaryote leaf to measure the impact of the inclusion of

**Fig. 5** Number of fold appearances per million years in ordinate (i.e., the number of appearances on one branch/the branch length) per million years in abscissa. The three colors represent nodes within Eukaryota, Archaea and Bacteria branches (Color figure online)



parasitic or anaerobic organisms. Indeed, parasites classically lose functions, then structures as the parasitic link increases in integration. The results are provided in Supplementary Materials S12. In average in the eukaryotic lineages, we observe 109.9 folds appearances and 118.7 folds losses. *Giardia lamblia*, *Cryptosporidium hominis*, *Theileria parva*, *Entamoeba histolytica*, and *Trichomonas vaginalis* have, as expected, the greatest number of fold losses. They have more than 200 fold losses and less than 42 fold appearances. The same trend, but slightly less pronounced, is observed for *Toxoplasma gondii*, *Leishmania major*, and *Perkinsus marinus* (between 140 and 165 losses). *Taenia solium* and *Schistosoma mansoni* have also a number of losses that are greater than the average (164 and 146, respectively) but they also have a number of appearances that is above the average (134 for both of them). *Ixodes scapularis* has a number of losses close to the average (112), which is not surprising as it is an ectoparasite (the less integrated to its host(s)).

### Functional Analysis of Folds Within Eukaryotic Clades

The 9 clusters selected from our previous study (Romei et al. 2022) correspond to 9 clades in the eukaryote phylogeny: Opisthokonta (45 organisms), Holozoa (39 organisms), Choza (37 organisms), Metazoa (35 organisms), Vertebrata (16 organisms), Gnathostomata (15 organisms), Tetrapoda (11 organisms), Nematoda (2 organisms), and Ecdysozoa (8 organisms). The repartition of the fold functions of these eukaryotic specific clusters is different from the global fold function repartition. There are very few informational and metabolism functions while we observe an over-representation of intracellular and extracellular processes and regulation. The clusters 1 to 4 of Fig. 6 are nested so all folds which appeared in group 1 are still present in species of groups 2, 3, and 4. Groups 5 to 7 are nested and 8 to 9 are nested too. The number of folds appearing at each node is variable. The majority of regulation functions, intracellular functions, and extracellular functions appeared in the four first groups. These folds and their function are therefore present in the majority of eukaryotic species.

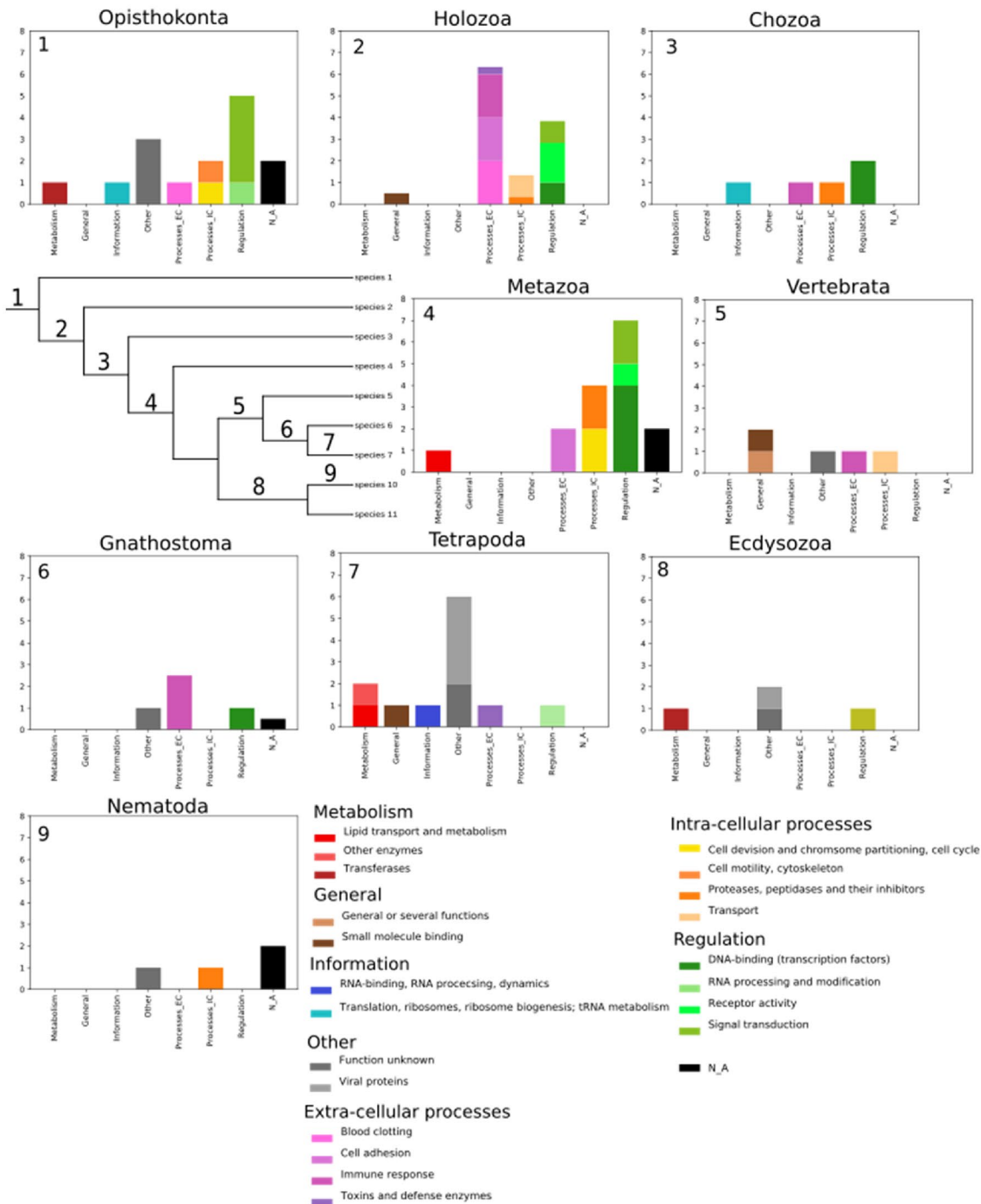
It is interesting to check whether such folds were recruited in functional innovations at the time of common ancestors to the clade for which they constitute a phylogenetic signature. Doing this, we may expect a limited number of correspondences between today's functions of the proteins where the fold is found and the evolutionary innovations of the ancestor of the clades for which the fold is a synapomorphy. Indeed, correspondence between structure and function is flexible during the course of evolution. For instance, the protein called beta-hcg -human chorionic gonadotropin- is used by humans as a signal from the embryo that triggers maternal

recognition of pregnancy, but homologous sequences called "CG" are used for a completely different function in Bacteria (Schwartz 1998). Structures do not always appear in the past with the function it performs today. Exaptation (Gould and Vrba 1982) is the phenomenon where a derived function is recruiting a primitive structure. On the macroscopic scale, one of the best documented examples is the feather that firstly existed as a thermal coat, as an impermeable protection of skin and as an organ of display, long before being used for flight (Benton et al. 2019). So, prudence is needed in assigning today's functions to a structure at the time of its appearance. In spite of these potential difficulties, it is noteworthy to point out the functions that make sense when considering members of the clade for which the involved fold is a synapomorphy.

There are several trends among the Opisthokonta toward pluricellular life. Fungi and animals (Metazoa) exhibit permanent pluricellular forms of life. Choanoflagellates (sister group of the Metazoa) can form colonies. The fold g.52 is implied in "apoptosis inhibition," which is part of the general functions of the cell cycle. Most probably, g.52 was not recruited in "apoptosis inhibition" in a unicellular ancestor of Opisthokonta, but in a function related to cell cycle potentially favoring the rise of various forms of pluricellularity (Ros-Rocher et al. 2021). Fold a.216 (I/LWEQ domain) participates to a general family of functions called "cell motility and cytoskeleton," possibly in line with the cytokinesis of opisthokonts where a contractile actomyosin-based structure specifically plays a central role (Glotzer 2016). The I/LWEQ domain is a 250-residue actin-binding module present in yeast and mammals. Regarding cell motility, one of the synapomorphies of Opisthokonta is the fact that the spermatozoan is pushed forward by the flagellum located at the back and not pulled by flagellae located at the front.

The complete genome of the filasterean *Capsaspora owczarzaki*, the closest known unicellular relative of metazoans besides choanoflagellates, allowed discovering that they had a richer repertoire of proteins involved in cell adhesion and transcriptional regulation (Suga et al. 2013; Lecointre and Guyader 2016) than previously inferred with the available choanoflagellate genome (*Monosiga brevicollis*). Interestingly, folds g.64 and a.215 are synapomorphies of the clade uniting three groups of holozoans (metazoans, choanoflagellates, and filastereans), implied in proteins called today "somatomedin B domain" and "middle domain Talin 1," respectively, which are classified in the "cell adhesion" functional category. In parallel, the fold a.12 is involved in Kix domain of Creb Binding Protein, corresponding to the increase of capacities in transcriptional regulation in holozoans. It is noteworthy that this increase starts at the clade Holozoa but continues to provide synapomorphies to clades nested within holozoans (Ivanovic and Vlaski-Lafarge 2016). Folds involved in transcriptional regulation are found





**Fig. 6** Simplification of our Eukaryota reference tree with only clades with specific folds and function repartition for each of these clades. In all diagrams, the main functions are in the same order as in the leg-

end: Metabolism, general, information, other, extracellular processes, intracellular processes, regulation, and non-attributed

as synapomorphies for Choza (i.e., metazoans + choanoflagellates: g.73 and a.37 (Ivanovic and Vlaski-Lafarge 2016)) and Metazoa (b.54, d.164, g.62, d.217). The fold g.65 is involved in Notch domain, which has to do with intercellular regulation, a function relevant with the trends

of pluricellular life in filastereans, choanoflagellates, and metazoans, even if some of these exhibit pluricellularities only transitory (Ros-Rocher et al.2021).

These considerations hold for holozoans, but in absence of any mesomycetozoans (sister-group to the rest

of holozoans) in the present taxonomic sampling, they could be restricted to their sister group, the filozoans (i.e., filastereans, choanoflagellates, and metazoans). Filozoans are known to be supported by synapomorphies related to cell adhesion (e.g., presence of cadherin domains) and to the rise of components of the Notch signaling pathways (Lecointre and Guyader 2016). However, Ros-Rocher et al. (2021) consider that cadherin domains and Notch pathways have a holozoan origin, which is compatible with our findings and suggesting that the absence of mesomycetozoans in the present sample is not misleading.

Metazoans are permanent pluricellular organisms. Signaling activity among cells is logically expected to increase. Folds g.76, a.123 and g.1 are synapomorphies involved today in receptor activity and signal transduction, respectively. Cell division and then cell cycle has to be regulated, and folds a.77 and g.17 are involved in it. A. 77 is referred to “Death domain,” a domain linked to apoptosis.

Three fold synapomorphies of the Gnathostomata are found in proteins classified under the general function “Immune response”: a.109, d.9, and d.19. This is probably related to the rise of the adaptive immune system in gnathostomes (Bleyzac et al. 2005).

Interestingly, the origin of tetrapods corresponds to a drastic change in living environment. Early tetrapods, some 380 million years ago, were able to cope with underwater life and with aerial terrestrial constraints when their place went dry. This must have corresponded to a renewal of viral parasites. Folds a.206, H.3, a.61, and b.20 are involved in viral proteins, which DNA coding sequences must have been integrated into the tetrapod genome.

Ecdysozoa are grouped with a fold a.163 called “crustacean CHH/MIH/GIH neurohormone,” which is remarkable because CHH (crustacean hyperglycemic hormone) and MIH (molt-inhibiting hormone) are domains both involved in the molting cycle, which is under the control of ecdysteroids. The fold is obviously not restricted to crustaceans (this is the initial name given to these proteins) but concerns all molting animals, for example, nematodes. The clade containing them is the Ecdysozoans, defined by molting activity. The fold a.85 (hemocyanin, N-terminal domain) is the pigment that plays the role of blood dioxygen carrier in arthropods. It is then consistent to find this fold present in arthropods; however, the fold is here a synapomorphy for a wider clade, Ecdysozoa. One has to keep in mind that hemocyanins are involved in several other major physiological processes like immune response and molting. A research perspective to follow is to examine whether the N-terminal domain of hemocyanin could play a role in the molting nematode and/or its immune innate response.

## Conclusion

Eukaryotic folds have two origins: inherited from Bacteria and Archaea on one hand and specific to Eukaryota on the other hand. Folds inherited from Bacteria and Archaea show specific repartition of functions. The groups detailed from hierarchical clustering and Dynamic Tree Cut give slightly different results than previous studies on gene (Alvarez-Ponce et al. 2013; Méheust et al. 2018) and protein structure (Kauko and Lehto 2018). The Archaea/Eukaryota folds are mainly linked to informational function that is split into informational machinery and its regulation. Similarly, the Bacteria/Eukaryota folds show a comparable contribution of informational functions too, higher than expected from previous studies. Those informational functions are mainly linked to translational functions.

The reconstruction of ancestral states coupled with dating of each node on the tree of life enables the possibility of count appearance and disappearance along time and for each super-kingdom. The number of folds is quite stable over time but is different for each super-kingdom. The number of folds within Eukaryota is higher than within Bacteria, which is itself higher than within Archaea. The dating and synapomorphies of each node led to calculate fold appearance rate per million years. This rate is low and stable for Bacteria and Archaea with less than 0.1 appearance by million years. For Eukaryota, the rate is on average twice higher. Moreover, some eukaryotic clades exhibit peak rates such as metazoans (0.49), holozoans (0.17), vertebrates (0.15), and metazoans *stricto sensu* (0.15). To summarize, the eukaryotic origins, the holozoan origins, and the metazoan origins correspond to bursts of fold evolution.

Some of the fold synapomorphies within eukaryotes can be associated with the functions of evolutionary significance. These folds exhibit cell motility functions and early clue of cellular cycle within Opisthokonta, cell adhesion suggesting first clue of multicellularity within Filastera, signaling activity and apoptosis within Metazoa (permanent multicellularity), adaptive immune system folds within Gnathostomata, new function associated to viral proteins corresponding to changes in living environment for Tetrapoda and functions specific to neurohormones and hemocyanins for the Ecdysozoa.

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1007/s00239-023-10136-x>.

**Author Contributions** JC, GL, MC designed the study. MR and MC produced the data. MR and MC analyzed the data. MR, JC, GL, and MC wrote the paper.

## Declarations

**Conflict of interest** The authors declare no conflict of interest.

## References

- Alvarez-Ponce D, Lopez P, Baptiste E, McInerney JO (2013) Gene similarity networks provide tools for understanding eukaryote origins and evolution. *Proc Natl Acad Sci USA* 110:E1594–1603. <https://doi.org/10.1073/pnas.1211371110>
- Benton MJ, Dhouailly D, Jiang B, McNamara M (2019) The early origin of feathers. *Trends Ecol Evol* 34:856–869. <https://doi.org/10.1016/j.tree.2019.04.018>
- Bleyzac P, Exbrayat J-M, Fellah J (2005) Émergence du système immunitaire adaptatif: hypothèses en présence. *Med Sci* 21:210–215
- Bukhari SA, Caetano-Anollés G (2013) Origin and evolution of protein fold designs inferred from phylogenomic analysis of cath domain structures in proteomes. *PLoS Comput Biol* 9:e1003009. <https://doi.org/10.1371/journal.pcbi.1003009>
- Caetano-Anollés D (2005) Universal sharing patterns in proteomes and evolution of protein fold architecture and life. *J Mol Evol* 60:484–498. <https://doi.org/10.1007/s00239-004-0221-6>
- Caetano-Anollés G, Caetano-Anollés D (2003) An evolutionarily structured universe of protein architecture. *Genome Res* 13:1563–1571. <https://doi.org/10.1101/gr.1161903>
- Chandonia J-M, Fox NK, Brenner SE (2019) SCOPe: classification of large macromolecular structures in the structural classification of proteins—extended database. *Nucleic Acids Res* 47:D475–D481. <https://doi.org/10.1093/nar/gky1134>
- Cotton JA, McInerney JO (2010) Eukaryotic genes of archaeobacterial origin are more important than the more numerous eubacterial genes, irrespective of function. *Proc Natl Acad Sci* 107:17252–17255. <https://doi.org/10.1073/pnas.1000265107>
- Coulson AFW, Moulton J (2002) A unifold, mesofold, and superfold model of protein fold use. *Proteins* 46:61–71. <https://doi.org/10.1002/prot.10011>
- Glotzer M (2016) Cytokinesis in metazoa and fungi. *Cold Spring Harb Perspect Biol*. <https://doi.org/10.1101/cshperspect.a022343>
- Gough J, Karplus K, Hughey R, Chothia C (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J Mol Biol* 313:903–919
- Gould SJ, Vrba ES (1982) Exaptation—a missing term in the science of form. *Paleobiology* 8:4–15. <https://doi.org/10.1017/S009483730004310>
- Govindarajan S, Recabarren R, Goldstein RA (1999) Estimating the total number of protein folds. *Proteins* 35:408–414
- Gower JC, Legendre P (1986) Metric and Euclidean properties of dissimilarity coefficients. *J Classif* 3:5–48
- Hajdu LJ (1981) Graphical comparison of resemblance measures in phytosociology. *Vegetatio* 48:47–59
- Hug LA, Baker BJ, Anantharaman K et al (2016) A new view of the tree of life. *Nat Microbiol* 1:1–6. <https://doi.org/10.1038/nmicrobiol.2016.48>
- Ishikawa SA, Zhukova A, Iwasaki W, Gascuel O (2019) A fast likelihood method to reconstruct and visualize ancestral scenarios. *Mol Biol Evol* 36:2069–2085. <https://doi.org/10.1093/molbev/msz131>
- Ivanovic Z, Vlaski-Lafarge M (2016) 11 - metabolic and genetic features of ancestral eukaryotes versus metabolism and “Master Pluripotency Genes” of stem cells. In: Ivanovic Z, Vlaski-Lafarge M (eds) *Anaerobiosis and stemness*. Academic Press, Boston, pp 211–234
- Kauko A, Lehto K (2018) Eukaryote specific folds: part of the whole. *Proteins Struct Funct Bioinform* 86:868–881. <https://doi.org/10.1002/prot.25517>
- Kumar S, Stecher G, Suleski M, Hedges SB (2017) TimeTree: a resource for timelines, timetrees, and divergence times. *Mol Biol Evol* 34:1812–1819. <https://doi.org/10.1093/molbev/msx116>
- 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. <https://doi.org/10.1093/bioinformatics/btm563>
- Lecointre G, Le Guyader H (2016) *Classification phylogénétique du vivant*, t.1. 4e ed. Belin
- Lecointre G, Le Guyader H (2017) *Classification phylogénétique du vivant*, t.2. 4e ed. Belin
- McInerney JO, O’Connell MJ, Pisani D (2014) The hybrid nature of the Eukaryota and a consilient view of life on Earth. *Nat Rev Microbiol* 12:449–455. <https://doi.org/10.1038/nrmicro3271>
- Méheust R, Bhattacharya D, Pathmanathan JS et al (2018) Formation of chimeric genes with essential functions at the origin of eukaryotes. *BMC Biol* 16:30. <https://doi.org/10.1186/s12915-018-0500-0>
- Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247:536–540. [https://doi.org/10.1016/S0022-2836\(05\)80134-2](https://doi.org/10.1016/S0022-2836(05)80134-2)
- Pandurangan AP, Stahlhake J, Oates ME et al (2019) The superfamily 2.0 database: a significant proteome update and a new webserver. *Nucleic Acids Res* 47:D490–D494. <https://doi.org/10.1093/nar/gky1130>
- Rivera MC, Jain R, Moore JE, Lake JA (1998) Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci* 95:6239–6244. <https://doi.org/10.1073/pnas.95.11.6239>
- Romei M, Sapriel G, Imbert P et al (2022) Protein folds as synapomorphies of the tree of life. *Evolution* 76:1706–1719. <https://doi.org/10.1111/evo.14550>
- Ros-Rocher N, Pérez-Posada A, Leger MM, Ruiz-Trillo I (2021) The origin of animals: an ancestral reconstruction of the unicellular-to-multicellular transition. *Open Biol* 11:200359. <https://doi.org/10.1098/rsob.200359>
- Schwartz RS (1998) Can bacteria cause cancer? *Nat Med* 4:1197–1198. <https://doi.org/10.1038/2695>
- Staley JT, Caetano-Anollés G (2018) Archaea-first and the co-evolutionary diversification of domains of life. *BioEssays* 40:1800036. <https://doi.org/10.1002/bies.201800036>
- Suga H, Chen Z, de Mendoza A et al (2013) The Capsaspora genome reveals a complex unicellular prehistory of animals. *Nat Commun* 4:2325. <https://doi.org/10.1038/ncomms3325>
- Wang M, Yafremava LS, Caetano-Anollés D et al (2007) Reductive evolution of architectural repertoires in proteomes and the birth of the tripartite world. *Genome Res* 17:1572–1585. <https://doi.org/10.1101/gr.6454307>
- Wang M, Jiang Y-Y, Kim KM et al (2011) A universal molecular clock of protein folds and its power in tracing the early history of aerobic metabolism and planet oxygenation. *Mol Biol Evol* 28:567–582. <https://doi.org/10.1093/molbev/msq232>
- Weiss MC, Sousa FL, Mrnjavac N et al (2016) The physiology and habitat of the last universal common ancestor. *Nat Microbiol* 1:1–8. <https://doi.org/10.1038/nmicrobiol.2016.116>
- Wilson D, Madera M, Vogel C et al (2007) The superfamily database in 2007: families and functions. *Nucleic Acids Res* 35:D308–D313. <https://doi.org/10.1093/nar/gkl910>
- Winstanley HF, Abeln S, Deane CM (2005) How old is your fold? *Bioinformatics* 21:i449–i458. <https://doi.org/10.1093/bioinformatics/bti1008>
- Yang S, Doolittle RF, Bourne PE (2005) Phylogeny determined by protein domain content. *Proc Natl Acad Sci USA* 102:373–378. <https://doi.org/10.1073/pnas.0408810102>

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.