Are protein folds reliable phylogenetic markers?

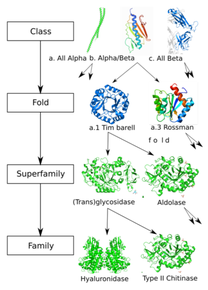
Guillaume Sapriel, Pierre Imbert, Mathilde Carpentier, Jacques Chomilier, Guillaume Lecointre, Martin Romei

Are protein folds reliable phylogenetic markers?

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

[1]Institut de Systématique, Évolution, Biodiversité (ISYEB), Muséum National d'Histoire Naturelle, Sorbonne Université, EPHE, UA, CNRS
[2]Sorbonne Université, CNRS, MNHN, IMPMC (UMR 7590), BiBiP, Paris, France
[3]UFR des sciences de la santé, Université Versailles-St-Quentin, Versailles, France

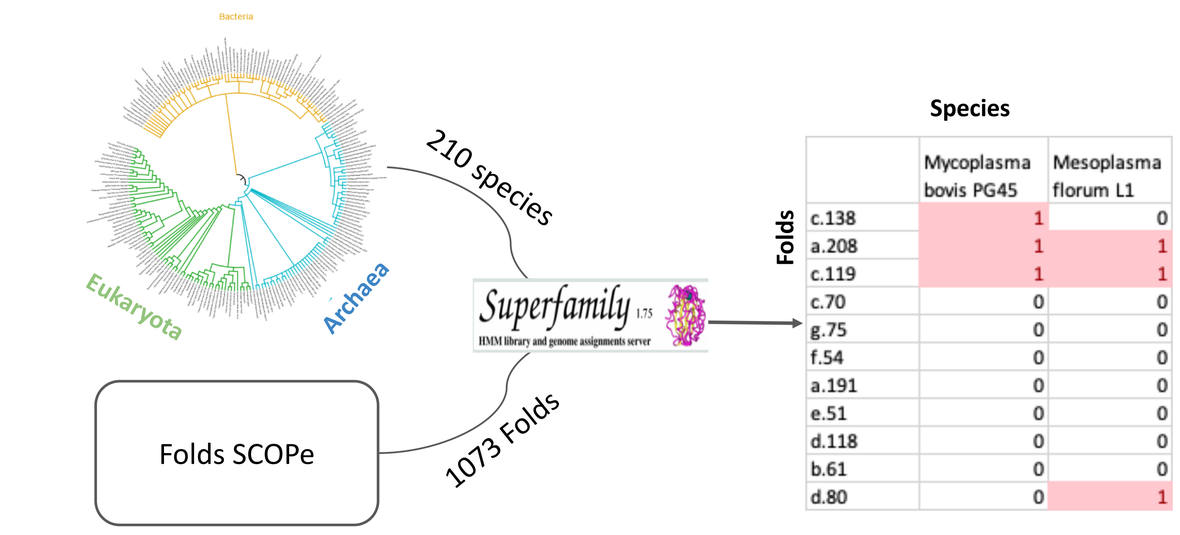
Introduction



Several studies showed that folds (topology of protein secondary structures) distribution in proteomes may be a global proxy to build phylogeny. Many attempts to reconstruct phylogenies from fold content have been made, the first being in the 90s [1, 2]. More recently, phylogenies has been reconstructed from fold abundances (copy-count) or occurrences (binary) using parsimonious reconstruction (see the work of Caetano-Anollés and coll. from 2003 [3] to 2012 [4]) or distance methods [5] or both [6]). None of these approaches were suitable to provide identified synapomorphies. and all these results are controversial and we propose here to confront fold distribution with well-acknowledged phylogenies in order to explore the fold history to understand the sources of these differences.

Material and Methods

Species selection and fold annotation



- Selection of 210 species with complete sequenced proteomes from the reference tree of life from Lecointre & Le Guyader [7], completed by the tree from Hug & al. [8] for bacteria and by the Asgard species from [9]. The balance is maintained among the three superkingdoms with 70 species for each group.
- Fold annotation with the SUPERFAMILY online server [10].
- 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 species are ordered according to the reference phylogeny. The branches are swapped with the package Dendser [11]. The folds are ordered with a hierarchical clustering and the tree is also swapped with Dendser.
- 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 [12].
- The same experiment has also been conducted with CATH v4.3 (level T, the 3rd level of the hierarchy) and ECOD v20220113 (level X, 1st level of the hierarchy).

Retention index

We calculated the retention index (RI) [13] to measure the adequacy between the characters (folds) and our reference phylogeny

$$RI = \frac{g - s}{g + m}$$

with

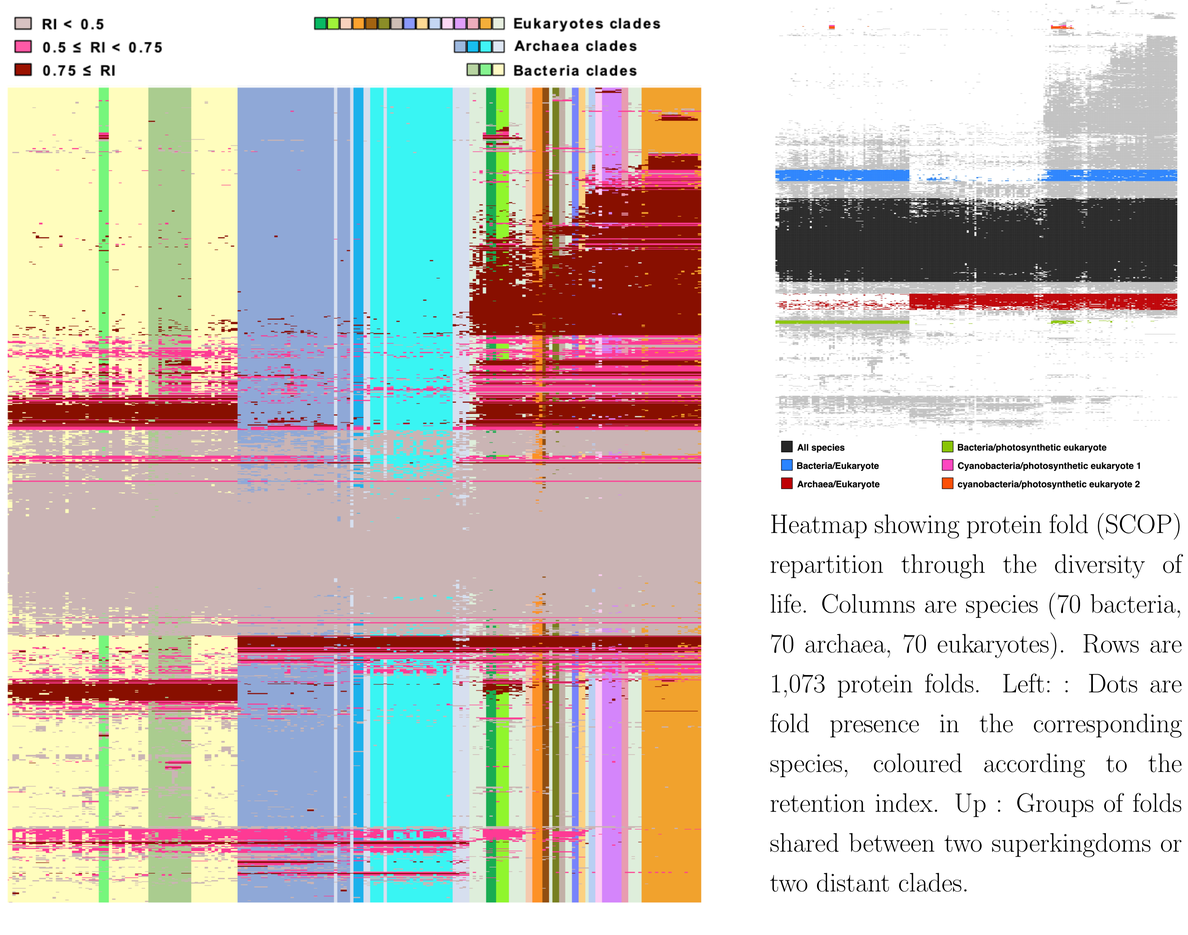
- g: the maximum number of steps, which is the number of changes of a character onto the tree with a single node (star-like tree: all changes being reported onto individual branches);
- s: the number of steps calculated parsimoniously with the considered tree;
- m: the minimum number of steps that the character may have. For a single two-state character, the minimum number of changes is one (the number of character states minus one).

For a character, the retention index 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.

[1] M. Gerstein. "Patterns of protein-fold usage in eight microbial genomes: a comprehensive structural census". In: *Proteins* 33.4 (Dec. 1, 1998), pp. 518–534.
[2] Y I Wolf et al. "Distribution of protein folds in the three superkingdoms of life." In: *Genome research* 9.1 (Jan. 1, 1999), pp. 17–26.
[3] Gustavo Caetano-Anollés et al. "An Evolutionarily Structured Universe of Protein Architecture". In: *Genome Research* 13.7 (July 2003). 00142, pp. 1563–1571.
[4] Kyung Mo Kim et al. "The evolutionary history of protein fold families and proteomes confirms that the archaeal ancestor is more ancient than the ancestors of other superkingdoms". In: *BMC evolutionary biology* 12.1 (Jan. 27, 2012), p. 13.
[5] Song Yang et al. "Phylogeny determined by protein domain content". In: *Proceedings of the National Academy of Sciences of the United States of America* 102.2 (Jan. 11, 2005). tex.ids= yangPhylogenyDeterminedProtein2005a, pp. 373–378.
[6] H F Winstanley et al. "How old is your fold?" In: *Bioinformatics (Oxford, England)* 21 (Suppl 1 June 16, 2005), pp. i449–i458.
[7] Guillaume Lecointre et al. *La classification phylogénétique du vivant. 4e ed.* Belin, 2017.
[8] Laura A. Hug et al. "Phylogenetic Distributions and Histories of Proteins Involved in Anaerobic Pyruvate Metabolism in Eukaryotes". In: *Molecular Biology and Evolution* 27.2 (Feb. 1, 2010). 00072, pp. 311–324.
[9] Panagiotis S Adam et al. "The growing tree of Archaea: new perspectives on their diversity, evolution and ecology". In: *The ISME Journal* 11.11 (Nov. 2017). 00133, pp. 2407–2425.
[10] Derek Wilson et al. "SUPERFAMILY—sophisticated comparative genomics, data mining, visualization and phylogeny". In: *Nucleic Acids Research* 37 (Database issue Jan. 2009). 00301, pp. D380–D386.
[11] Denise Earle et al. "Advances in Dendrogram Seriation for Application to Visualization". In: *Journal of Computational and Graphical Statistics* 24.1 (Jan. 2, 2015). 00013, pp. 1–25.
[12] Peter Langfelder et al. "Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R". In: *Bioinformatics (Oxford, England)* 24.5 (Mar. 1, 2008). 01435, pp. 719–720.
[13] James S. Farris. "The Retention Index and the Rescaled Consistency Index". In: *Cladistics* 5.4 (Dec. 1989). 01832, pp. 417–419.

Results

A global structured repartition of folds

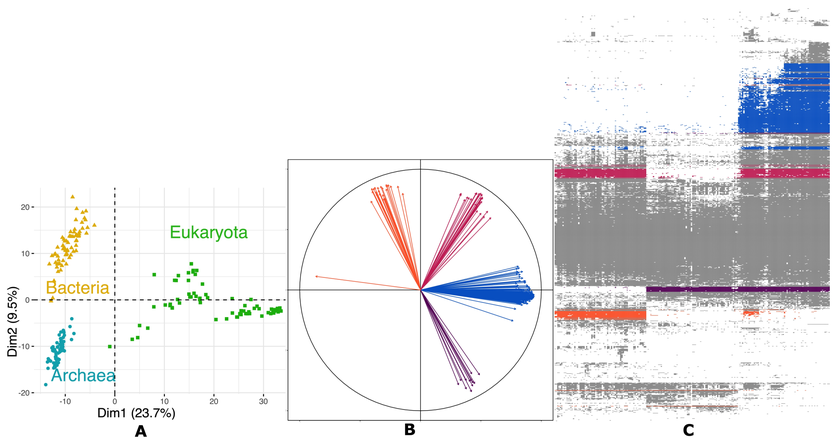


Heatmap showing protein fold (SCOP) repartition through the diversity of life. Columns are species (70 bacteria, 70 archaea, 70 eukaryotes). Rows are 1,073 protein folds. Left: : Dots are fold presence in the corresponding species, coloured according to the retention index. Up : Groups of folds shared between two superkingdoms or two distant clades.

	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

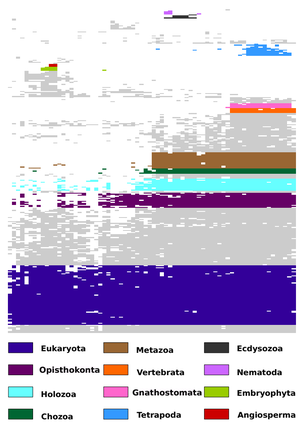
Average Retention Index calculated for all characters with either all organisms or only Bacteria, Eukarya or Archaea (in line). The characters are the predicted presence or absence in the proteomes of SCOP folds, T level architecture of CATH or X level architecture of ECOD (in column).

Fold mosaicism discriminates the three superkingdoms



A. Principal Component Analysis of fold repartition (2 first axes).
B. Protein fold contributions to species repartition; four clusters are identified.
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.

Specific Fold Blocks within Eukaryotes



The clustering and high value of RI allows the identification of 11 clades in Eukaryotes: 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. Within these 11 clades, metabolism and informational functions are clearly under-represented with 6 (7.1%) and 4 (4.7%) folds, whereas regulation and extra-cellular processes are over-represented with 22 (24.5%) and 17 folds (18.6%) respectively. This over representation can be linked to the rise of multicellularity which constrains cells to make junctions and to communicate.

Conclusion and perspectives

Using a bicluster mapping approach we define synapomorphic blocks of folds sharing similar presence/absence patterns. Among the 1,232 folds, 20% are universally present in our TOL, while 54% are reliable synapomorphies. These results are similar with CATH and ECOD databases. Eukaryotes are characterized by a large number of them, and several synapomorphic blocks of folds clearly supported nested eukaryotic clades (divergence times from 1,100 to 380 mya). While 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.

Reference for this work: M. Romei, et al, Protein folds as synapomorphies of the tree of life, Evolution, In press