On the unknown proteins of eukaryotic proteomes

Yves-Henri Sanejouand

On the unknown proteins of eukaryotic proteomes

Yves-Henri Sanejouand*

US2B, UMR 6286 of CNRS, Nantes University, France.

September 21^{th}, 2022

Abstract

In order to study unknown proteins on a large scale, a reference system has been set up for the three major eukaryotic lineages, built with 36 proteomes as taxonomically diverse as possible. Proteins from 362 eukaryotic proteomes with no known homologue in this set were then analyzed, focusing noteworthy on singletons, that is, on unknown proteins with no known homologue in their own proteome.

Consistently, according to Uniprot, for a given species, no more than 12% of the singletons thus found are known at the protein level. Also, since they rely on the information found in the alignment of homologous sequences, predictions of AlphaFold2 for their tridimensional structure are usually poor.

In the case of metazoan species, the number of singletons seems to increase as a function of the evolutionary distance from the reference system. Interestingly, no such trend is found in the cases of viridiplantae and fungi, as if the timescale on which singletons are added to proteomes were different in metazoa and in other eukaryotic kingdoms. In order to confirm this phenomenon, further studies of proteomes closer to those of the reference system are however needed.

Keywords: Ubiquitous proteins, Singletons, Metazoa, Viridiplantae, Fungi, Evolutionary distance, Uniprot, AlphaFold2.

*yves-henri.sanejouand@univ-nantes.fr

Introduction

Since the earliest genome sequencing projects proteins with no known homologue have been found in significant amounts [1, 2]. Various origins have been proposed for such unknown proteins [3, 4]: gene duplication [5, 6], followed by a neutral drift of their sequence [7, 8, 9, 10], incorporation of transposable elements [11, 12], *de novo* genesis, that is, evolution from random amino acid sequences [13, 14, 15, 16, 17], *etc.*

With the advent of massive genome sequencing projects [18, 19, 20], the total number of unknown proteins is expected to increase dramatically. On the other hand, the definition of what is an unknown protein relies on the information already available which, in the case of eukaryotic species, is biased towards what is known for a small set of model species, as well as for species close to, or the most useful for the human one. As a consequence, a robust definition is needed for unknown proteins, especially in order to perform quantitative comparisons between species.

On the other hand, in a context of an ongoing massive extinction of species [21, 22], it could prove worthwhile focusing the efforts on preserving those (their genomes, at least) that are the more likely to prove useful for humans in the future [23, 24, 25]. In particular, molecules of particular importance for our health have been found in various species [26, 27, 28]. Since such molecules are synthesized by enzymes with original specificities (or functions), the hypothesis that species hosting a lot of unknown proteins may prove more likely to yield enzymes with promizing characteristics needs to be considered.

Herein, as a first step towards this end, an instru-

mental definition is proposed for unknown proteins, based on the setup of a reference system for the proteomes of the three major eukaryotic kingdoms, namely, metazoa, viridiplantae (land plants) and fungi. While with such a definition the status of a given protein can be determined in a robust way, in the case of species far from the reference system, lineage-specific proteins [29, 30, 31] are more likely to be considered as being unknown. Hereafter, in order to cope with this drawback, a set of proteomes of species from other (unicellular) eukaryotic lineages is also analyzed.

Methods

Choice of a reference system

Unknown proteins are usually defined through the fact that they do not share any significant homology with other known proteins. As a consequence, the status of a given protein may change each time a new proteome is unraveled. In the present study, in order to address this issue, unknown proteins are instead defined with respect to a reference system, namely, a set of well-known proteomes as taxonomically diverse as possible.

As a a reference system for eukaryotic proteomes, 36 proteomes were selected as follows, among the 398 reference proteomes [32] with more than 10,000 proteins available in Uniprot.[1] For each of the three better studied eukaryotic kingdoms, namely, metazoa, viridiplantae and fungi, their taxonomic tree, as provided by Uniprot [33], was scanned down to the node where at least ten taxons with proteomes of more than 10,000 proteins could be found, retaining for each taxon the proteome with the largest number of proteins. This protocol yielded 15, 10 and 11 proteomes for metazoa, viridiplantae and fungi, respectively (see Table 1), corresponding to a total number of 1,174,474 reference sequences.

Search of homologues

Homologues in this reference database were looked for using BLAST [34] version 2.6.0+, two proteins being assumed to be homologous when the E-value of their pairwise alignment is lower than 10^{-6} [35, 36, 37]. Note that, in order to avoid an overestimation of the number of unknown proteins, due to the filtering of low-entropy segments, that is, of segments of restricted amino-acid composition, composition-based statistics [38] was not considered (-comp_based_stats 0).

Evolutionary distance

The evolutionary distance between two species has long been estimated by comparing the sequences found in both species for a given protein, such as myoglobin or cytochrome c [39] or, for more distantly related species, highly conserved biomolecules like the ribosomal RNA [40, 41, 42]. With the advent of whole genome sequencing projects, it is nowadays possible to estimate this distance using a large set of common proteins [43, 44]. Hereafter, the evolutionary distance between a species and the 36 species of the reference system (Table 1) is estimated by comparing their ubiquitous proteins, that is, proteins that have homologues in all 36 proteomes of the reference set.[2] In practice, for each ubiquitous protein of a given species, the closest protein in the reference database is picked and the percentage of differences in their alignment is recorded, the evolutionary distance being the corresponding average over all ubiquitous proteins of the species. For this measure, ubiquitous proteins with more than ten homologues in the considered proteome were not taken into account.

Results

Ubiquitous proteins

Homologues in the reference database were identified for each protein of the 398 eukaryotic proteomes with more than 10,000 known proteins, that is, 189, 83, 99, 27 proteomes from metazoa, viridiplantae, fungi and other eukaryotic lineages, respectively. On average, whatever the kingdom, 10–15% of the proteins have homologues in all 36 proteomes of the reference set, the largest numbers of them being found in three viridiplantae, namely, *Triticum turgidum* (37,911), *Aegilops tauschii* (31,733) and *Hordeum vulgare* (29,499). On the other hand, at least 1,000 such ubiquitous

[1] On June 23^{th}, 2020.

[2] Most of them are likely to be "housekeeping" proteins.

Table 1: A reference system for eukaryotic proteomes. For each of the three major eukaryotic kingdoms, proteomes were chosen so as to be as taxonomically diverse as possible, among those with more than 10,000 proteins.

Kingdom	Taxon	Species[a]	Uniprot Id.	Proteins
Metazoa	Arthropoda	Portunus trituberculatus	PORTR	99,420
	Craniata	Homo sapiens	HUMAN	75,004
	Rotifera	Brachionus plicatilis	BRAPC	52,387
	Demospongiae	Amphimedon queenslandica	AMPQE	43,437
	Nematoda	Caenorhabditis japonica	CAEJA	35,024
	Brachiopoda	Lingula unguis	LINUN	34,415
	Eleutherozoa	Stichopus japonicus	STIJA	30,032
	Cephalochordata	Branchiostoma floridae	BRAFL	28,544
	Mollusca	Crassostrea gigas	CRAGI	25,997
	Anthozoa	Nematostella vectensis	NEMVE	24,435
	Annelida	Helobdella robusta	HELRO	23,328
	Tunicata	Ciona savignyi	CIOSA	20,004
	Trematoda	Opisthorchis felineus	OPIFE	18,330
	Tardigrada	Hypsibius dujardini	HYPDU	14,867
	Cestoda	Hydatigena taeniaeformis	HYDTA	11,591
Viridiplantae	Poaceae	Aegilops tauschii	AEGTS	214,162
	Musaceae	Ensete ventricosum	ENSVE	58,382
	Papaveraceae	Papaver somniferum	PAPSO	41,351
	Pentapetalae	Arabidopsis thaliana	ARATH	39,353
	Coryphoideae	Phoenix dactylifera	PHODC	34,033
	Nelumbonaceae	Nelumbo nucifera	NELNU	31,582
	Funariaceae	Physcomitrella patens	PHYPA	30,858
	Amborellaceae	Amborella trichopoda	AMBTC	27,371
	Asparagaceae	Asparagus officinalis	ASPOF	24,059
	Bromeliaceae	Ananas comosus	ANACO	23,408
Fungi	Agaricomycetes	Armillaria gallica	ARMGA	25,522
	Blastocladiaceae	Allomyces macrogynus	ALLM3	19,092
	Pezizomycetes	Ascobolus immersus[b]	ASCIM	17,778
	Dothideomycetes	Corynespora cassiicola[b]	CORCC	17,125
	Mucorineae	Rhizopus delemar	RHIO9	16,971
	Cunninghamellaceae	Absidia glauca	ABSGL	14,825
	Neocallimastigaceae	Piromyces sp.	PIRSE	14,606
	Eurotiomycetes	Penicillium camemberti[b]	PENCA	14,390
	Sordariomycetes	Fusarium poae[b]	FUSPO	14,048
	Leotiomycetes	Monilinia fructicola[b]	MONFR	13,749
	Syncephalastraceae	Syncephalastrum racemosum	SYNRA	11,037

[a]With the largest proteome of the taxon.
[b]Ascomycota.

proteins were found in all eukaryotic proteomes considered herein, except in the cases of *Megaselia scalaris* (864 of them) and *Eimeria mitis* (336), the later being an apicomplexan parasite [45], which probably relies on its host (*Gallus gallus*) for compensating the lack of missing ones.

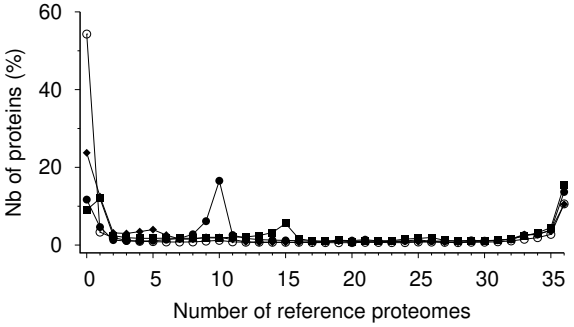

Figure 1: Percentage of proteins as a function of the number of proteomes of the reference system in which their homologues are found. Filled squares: proteins from metazoa; filled circles: from viridiplantae; filled diamonds: from fungi; open circles: from other eukaryotes.

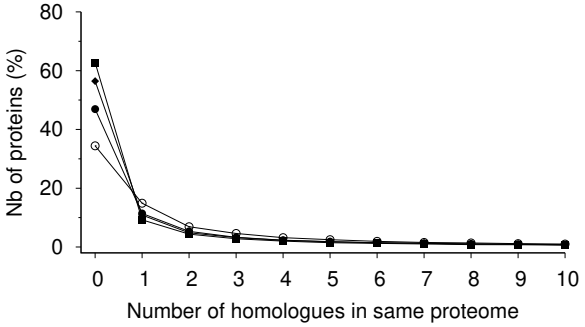

Figure 2: Number of homologues in their own proteome of unknown proteins, that is, those with no homologue in the 36 proteomes of the reference system. Filled squares: metazoa; filled circles: viridiplantae; filled diamonds: fungi; open circles: other eukaryotes.

Interestingly, as suggested by Figure 1, a significant number of proteins from metazoa and viridiplantae are kingdom–specific [46, 47], that is, they only have homologues in the proteomes of the reference system coming from their own kingdom (15 and 10 proteomes, respectively). On the other hand, fungi do not seem to have a significant number of them (no peak in the case of eleven proteomes), as if their functional diversity were higher. Note however that there is a peak for five proteomes, due to the five ascomycota species of the reference system (see Figure 1 and Table 1).

Unknown proteins and singletons

As shown in Figure 1, the percentage of unknown proteins, that is, of proteins not found in the reference database, is around 10%, on average, in the case of proteomes from metazoa and viridiplantae, around 25%, in the case of fungi, and as high as 54%, in the case of the 27 proteomes from other eukaryotic lineages. This later result makes sense if it is assumed that, in this case, homologues of a large amount of unknown proteins were just missed, as a consequence of their high degree of evolutionary divergence.

Interestingly, as shown in Figure 2, whatever the kingdom, roughly half of the unknown proteins have homologues within their own proteome. Note that such proteins are likely to be older than unknown proteins that have none, hereafter called singletons.

Degree of existence

In Uniprot, the degree of knowledge about a protein (the so-called degree of existence) is quantified through a number ranging between one (known at the protein level) and four (predicted).[3] As shown in Figure 3, only a minority of proteins of our dataset are well-characterized ones (degree one), even in the case of the ubiquitous proteins of metazoa (top left). Specifically, no more than 3% of them, except in the cases of *Homo sapiens* (92%), *Mus musculus* (84%), *Rattus norvegicus* (61%), *Drosophila melanogaster* (59%), *Danio rerio* (41%) and *Sus scrofa* (40%), that is, a few model organisms. Likewise, more than 5% of the ubiquitous proteins are known at the transcript level (degree two) in the cases of 13 metazoa, 4 viridiplantae and 1 species from an unicellular eukaryotic lineage.

However, whatever the eukaryotic kingdom, around 40–50%, on average, of the ubiquitous proteins are known by homology (degree three), meaning that they belong to known protein families.

As expected, such results are in sharp contrast with what is observed for unknown and singleton proteins. In both cases, for a given species, no more

[3] A few proteins are also classified as being uncertain (fifth degree).

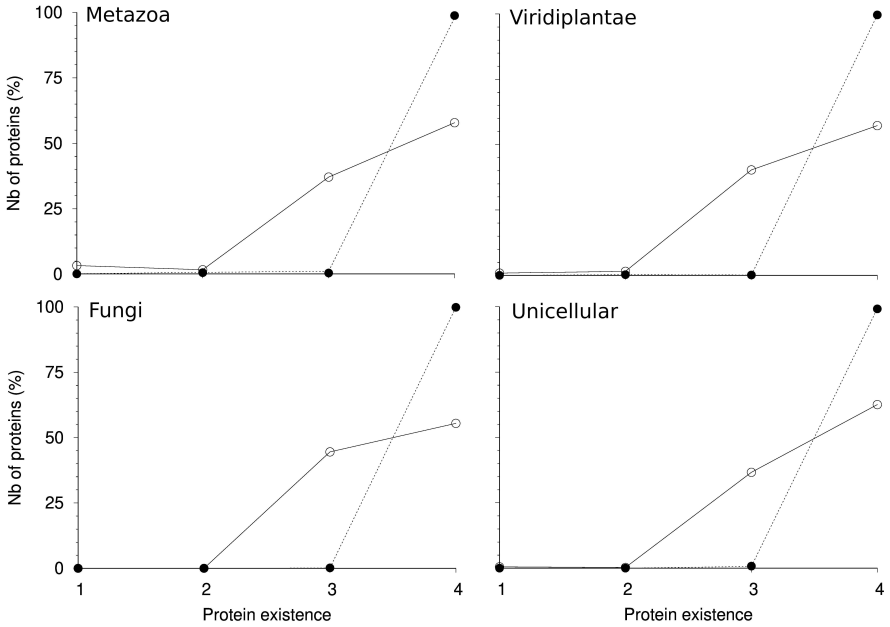

Figure 3: Number of proteins as a function of their degree of existence, according to Uniprot, for each eukaryotic kingdom. 1 means: known at the protein level; 2: at the transcript level; 3: by homology; 4: predicted. Open circles and plain lines: ubiquitous proteins; dotted lines: unknown proteins; filled circles: singletons.

than 12% of them are known at the protein level. Actually, more than 1% of the singletons are known at the protein level in the cases of eight species *only*, all of them belonging to the metazoan kingdom. For singletons that are, according to Uniprot, actually known by homology, figures are however a bit higher. As a matter of fact, they represent more than 5% of the singletons of a proteome in the cases of four species, namely, *Lipotes vexillifer* (16%), *Leptonychotes weddellii* (10%), *Meleagris gallopavo* (6%), *Beauveria bassiana* (6%) and *Dictyostelium discoideum* (6%). However, for the three first ones, their number of singletons is unusually low (80 at most), as well as their number of unknown proteins (138 at most), strongly suggesting that the annotation of these proteomes is incomplete, being biased towards proteins with already known homologues.

Quality of structural prediction

The predicted tridimensional structures of all proteins of ten proteomes considered in the present study are presently[4] available in the AlphaFold Protein Structure Database [48], namely, two proteomes of the reference system, *Homo sapiens* and *Arabidopsis thaliana*, four from other metazoa, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster* and *Danio rerio*, two from other viridiplantae, *Oryza sativa* and *Zea mays*, and two from other eukaryotic lineages, *Dictyostelium discoideum* and *Trypanosoma cruzi*. Note that there was none from fungi.

In this database, the quality of the prediction of the position of each amino-acid residue of a protein by AlphaFold2[5] is provided as a percentage value (coined pLDDT[6]), values over 90% corresponding to a high quality and values below 50% to a poor one [48]. Herein, the overall quality of the prediction of the structure of a protein is assumed to be given by the average of the quality of the prediction of the position of its residues.

As shown in Figure 4, whatever the eukaryotic

[4] As of April 2022.
[5] The second version of AlphaFold.
[6] Standing for predicted local-distance difference test.

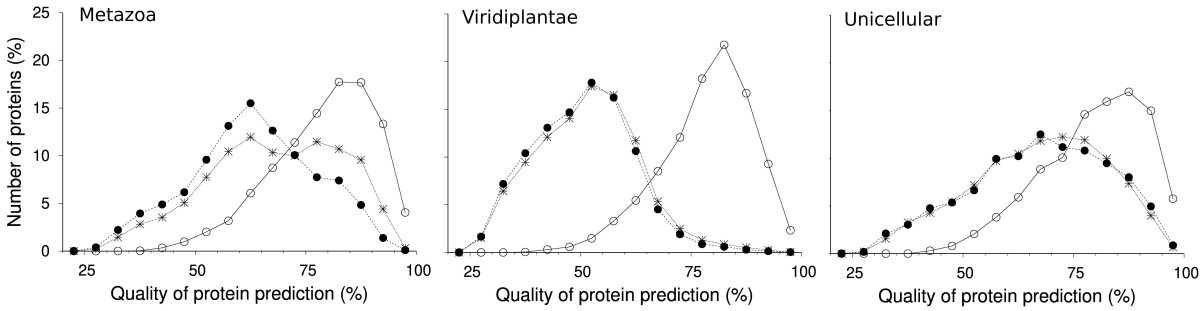

Figure 4: Number of proteins with a given quality of structural prediction (pLDDT), according to AlphaFold2, for three eukaryotic kingdoms. Over 90% means: high quality; below 50%: low one. Open circles: ubiquitous proteins; stars: unknown proteins; filled circles: singletons.

kingdom, the overall quality of the prediction of the structures of ubiquitous proteins is rather high, with a median value of $\approx$ 80%, values below 50% being observed in only $\approx$ 1% of the cases.

Since the predictions of AlphaFold2 are partly based on the information found in the alignment of homologous sequences [48], the overall quality of the prediction of the structures of unknown and singleton proteins is expected to be significantly lower. Indeed, whatever the eukaryotic kingdom, the median pLDDT value is below 70%. In the case of viridiplantae, it is as low as $\approx$ 50%, meaning that the structure of half of the unknown and singleton proteins of *Oryza sativa* and *Zea mays* are poorly predicted. This may reflect the fact that the proteomes of viridiplantae are less extensively studied than other eukaryotic proteomes, with the consequence that there may have more proteins with not enough known homologues, that is, with a number of homologues so low[7] that it does not allow AlphaFold2 to perform well [48]. This may also mean that there are more disordered, hard-to-predict proteins [49, 50], among the unknown and singleton proteins of these two viridiplantae.

At a more general level, note that, whatever the eukaryotic kingdom, the quality of the prediction of a structure by AlphaFold2 is similar for unknown and singleton proteins (see Figure 4), suggesting that the later sequence set is a genuine subset of the former (see also Figure 3). This is the reason why only singleton proteins are considered hereafter.

Interestingly, AlphaFold2 predicts with great confidence (average pLDDT over 90%) the tridimensional structure of 192 singletons, among the 13,141 ones (1.5% of them) found in the AlphaFold Protein Structure Database. As suggested by Figure 1, only a few (16) come from viridiplantae, most of them (142) coming from unicellular eukaryotic lineages. In the case of metazoan species, accurate structures are predicted for 25 singletons of *Drosophila melanogaster*, 7 of *Danio rerio* and a single one of *Mus musculus* and *Rattus norvegicus*. Among the later 34 cases, according to Uniprot, 7 are known at the protein level, coming all from *Drosophila melanogaster*.

Singletons with known 3D structure

If structures predicted with AlphaFold2 were considered above it is, essentially, because too few singletons are known at the protein level. As a matter of fact, there is a structure in the Protein Data Bank [51] for only 29 of them, among the 679,509 singletons (0.004% of them) found in the 362 eukaryotic proteomes considered.

Among these 29 singletons with a known tridimensional structure, seven come from five metazoan species, fifteen from two viridiplantae, and seven from unicellular eukaryotic lineages. Amazingly, twelve of them belong to the axoneme of *Chlamydomonas reinhardtii*, whose structure of the 48-nm repeat is an assembly of 38 different proteins (PDB 6U42), seven singletons being known as flagellar associated proteins (FAP68, FAP85, FAP95, FAP107, FAP143, FAP222, FAP273). Two others were identified during the determination of the structure of the doublet-microtubule, being not

[7] Below 30.

previously associated with cilia (RIB21 and RIB30) [52]. As expected for genuine singletons, they all seem absent in the axoneme of an alveolata, *Tetrahymena thermophila* [53]. However, for three of them (FAP95, FAP107 and FAP143) structural orthologs were found in the cryo-EM electron density map of the axoneme of *Bos taurus* [54]. So, these three singletons are likely to have orthologs in the human species, their structure and function being well conserved while their sequences are not.

The ten other singletons from viridiplantae or metazoa with a known tridimensional structure seem also to have a known function, the only obvious orphan being a protein with a CHAD domain found in *Ricinus communis* (PDB 6QV5 [55]).[8] On the other hand, the interleukin 22 of *Danio rerio* (PDB 4O6K [56]) looks like a clean example of a sequence drift quick enough so that homology with its human counterpart is hardly detected at the sequence level, while both display a typical class II cytokine architecture.

Evolution of the number of singletons

If singletons are randomly added to proteomes from time to time, their number should increase regularly as a function of the evolutionary distance from the reference system. Note that this may not be true for average values since, as mentioned above, proteome annotation may sometimes prove biased towards proteins with known homologues. However, such a trend would be expected for proteomes with the largest number of singletons, at a given evolutionary distance from the reference system.

As shown in Figure 5 (bottom), no such trend is found in the cases of viridiplantae and fungi. Indeed, in both kingdoms, the number of singletons per proteome does not seem to vary significantly as a function of the evolutionary distance. Actually, 2,979 and 865 singletons were found in the proteomes of the two species the closest to the reference system, namely, *Hordeum vulgare* and *Triticum turgidum*, suggesting that singletons are added to the proteomes of viridiplantae on a short timescale, their proliferation being kept under control afterwards. However, in order to determine on which timescale such a phenomenon occurs, proteomes closer to the reference system need to be considered.

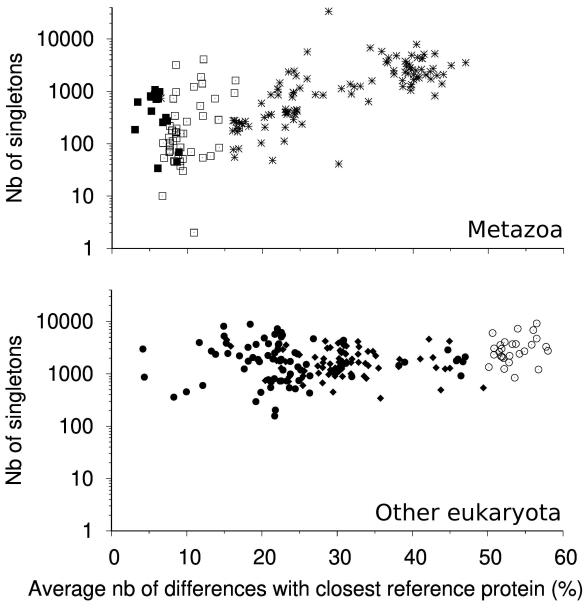

Figure 5: Number of singletons per proteome, as a function of the evolutionary distance between the species and the reference system. Top, filled squares: primates; open squares: other mammals; stars: other metazoa. Bottom, filled circles: viridiplantae; diamonds: fungi; open circles: unicellular eukaryotic lineages.

In the case of metazoa, an evolutionary trend seems there (Fig. 5, top), with at most 1,070 singletons found in the proteomes of primates (in the case of *Papio anubis*), and more than 1,000 of them in the proteomes of only eight mammal species, among 65 ones, the largest number of singletons (4,079) being found in the proteome of *Cricetulus griseus*. For other metazoa, the largest number of singletons is found in the proteome of *Liparis tanakae*, namely, 33,715. Note that this is an unusually large number, almost four times over the largest value found in viridiplantae, namely, 8,875, in the proteome of *Trifolium medium*.

Of course, the observed trend could be a consequence of the increasing evolutionary distance from the reference system, the homology of more and more proteins being not recognized as a conse-

[8] Being 76% identical with an inorganic triphosphatase of a *Duganella* bacterium, this protein is however likely to be a contaminant.

quence of a too quick neutral sequence drift [7, 8], like in the case of the interleukin 22 mentioned above. However, since such a drift is not observed for eukaryotic species other than metazoan ones (bottom of Fig. 5), it would anyway mean that singleton dynamics is different in metazoan species.

Note that while, on average, viridiplantae have more ubiquitous proteins than fungi, namely, 3,453 $\pm$ 3,552 (median value: 2,519) and 849 $\pm$ 351 (median value: 761), respectively, their average numbers of singletons are similar, namely, 2,388 $\pm$ 1,958 (median value: 1,774) and 1,922 $\pm$ 1,077 (median value: 1,638), respectively. On the other hand, singletons from unicellular eukaryotic lineages are, on average, more numerous, namely, 3,315 $\pm$ 1,950 (median value: 2,757). However, this is also a likely consequence of the high degree of evolutionary divergence of these proteomes, with respect to the reference system.

Conclusion

Defining ubiquitous, unknown and singleton proteins with respect to a set of 36 eukaryotic proteomes as taxonomically diverse as possible, the following results have been obtained.

For instance, a significant number of proteins from metazoa and viridiplantae are kingdom–specific, that is, they only have homologues in the proteomes of the reference system coming from their own kingdom. However, in the case of fungi, such proteins seem rare (Fig. 1).

Noteworthy, roughly half of the unknown proteins are singletons, that is, no homologue could be found for them, either in the reference system or in their own proteome (Fig. 2). On the other hand, there are at least 1,000 ubiquitous proteins in nearly all 398 eukaryotic proteomes considered, the main exception being the proteome of *Eimeria mitis*, an api-complexan parasite.

According to Uniprot, no more than 3% of the proteins of a given eukaryotic proteome are known at the protein level, except in the case of the ubiquitous proteins of six metazoan species. As a matter of fact, whatever the eukaryotic kingdom, most ubiquitous proteins (40–50% of them) are only known by homology (Fig. 3). In the case of singletons, 1% of them are known at the protein level in eight metazoan species *only*. Note that this figure rises up to 12% (157 among 1342 singletons), in the case of the proteome of *Drosophila melanogaster*, likely as a result of the number of studies dedicated to new genes found in *Drosophila* [57, 58, 59, 60].

As expected, in the case of ubiquitous proteins, tridimensional structures are predicted by AlphaFold2 with a high level of confidence (Fig. 4), probably because such predictions are based on the information found in the alignment of homologous sequences [48]. As a matter of fact, in the case of singletons, the predictions of AlphaFold2 are poor, in particular in the case of viridiplantae (Fig. 4).

Interestingly, in the case of metazoan species, the number of singletons seems to increase as a function of the evolutionary distance from the reference system, as if they were added to their proteomes rather regularly (Fig. 5, top). On the other hand, no such trend is found in the cases of viridiplantae or fungi (Fig. 5, bottom). Though this phenomenon needs to be confirmed, by considering proteomes closer to the reference system, such results suggest that the timescale on which singletons are added to proteomes is different in metazoa and in other eukaryotic kingdoms. It could also mean that the dominant underlying mechanisms are not the same, as already assessed in the case of fungi [61].

References

[1] Siew, N. & Fischer, D. (2003). Analysis of singleton ORFans in fully sequenced microbial genomes. *Proteins: Struct., Funct., Bioinf.* **53**(2), 241–251.

[2] Tautz, D. & Domazet-Lošo, T. (2011). The evolutionary origin of orphan genes. *Nat. Rev. Genet.* **12**(10), 692–702.

[3] Long, M., Betrán, E., Thornton, K. & Wang, W. (2003). The origin of new genes: glimpses from the young and old. *Nat. Rev. Genet.* **4**(11), 865–875.

[4] Vakirlis, N., Carvunis, A.R. & McLysaght, A. (2020). Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. *Elife* **9**, e53500.

[5] Ohno, S. (1970). *Evolution by gene duplication.* Springer, Berlin.

[6] Ohta, T. (1989). Role of gene duplication in evolution. *Genome* **31**(1), 304–310.

[7] King, J.L. & Jukes, T.H. (1969). Non-darwinian evolution: Most evolutionary change in proteins may be due to neutral mutations and genetic drift. *Science* **164**(3881), 788–798.

[8] Kimura, M. (1983). *The neutral theory of molecular evolution*. Cambridge University Press.

[9] Trinquier, G. & Sanejouand, Y.H. (1999). New protein-like properties of cubic lattice models. *Phys. Rev. E* **59**(1), 942–946.

[10] Bershtein, S., Goldin, K. & Tawfik, D.S. (2008). Intense neutral drifts yield robust and evolvable consensus proteins. *J. Mol. Biol.* **379**(5), 1029–1044.

[11] Toll-Riera, M., Bosch, N., Bellora, N., Castelo, R., Armengol, L., Estivill, X. & Mar Alba, M. (2009). Origin of primate orphan genes: a comparative genomics approach. *Mol. Biol. Evol.* **26**(3), 603–612.

[12] Carelli, F.N., Hayakawa, T., Go, Y., Imai, H., Warnefors, M. & Kaessmann, H. (2016). The life history of retrocopies illuminates the evolution of new mammalian genes. *Genome Res.* **26**(3), 301–314.

[13] White, S.H. & Jacobs, R.E. (1993). The evolution of proteins from random amino acid sequences. I. Evidence from the lengthwise distribution of amino acids in modern protein sequences. *J. Mol. Evol.* **36**(1), 79–95.

[14] Knowles, D.G. & McLysaght, A. (2009). Recent de novo origin of human protein-coding genes. *Genome Res.* **19**(10), 1752–1759.

[15] Heinen, T.J., Staubach, F., Häming, D. & Tautz, D. (2009). Emergence of a new gene from an intergenic region. *Curr. Biol.* **19**(18), 1527–1531.

[16] Carvunis, A.R., Rolland, T., Wapinski, I., Calderwood, M.A., Yildirim, M.A., Simonis, N., Charloteaux, B., Hidalgo, C.A., Barbette, J., Santhanam, B. *et al.* (2012). Proto-genes and de novo gene birth. *Nature* **487**(7407), 370–374.

[17] Schmitz, J.F. & Bornberg-Bauer, E. (2017). Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000 Research* **6**.

[18] Chain, P.S.G., Grafham, D.V., Fulton, R.S., Fitzgerald, M.G., Hostetler, J., Muzny, D., Ali, J., Birren, B., Bruce, D.C., Buhay, C. *et al.* (2009). Genome project standards in a new era of sequencing. *Science* **326**(5950), 236–237.

[19] Cao, Y., Li, L., Xu, M., Feng, Z., Sun, X., Lu, J., Xu, Y., Du, P., Wang, T., Hu, R. *et al.* (2020). The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. *Cell research* **30**(9), 717–731.

[20] Sun, Y., Shang, L., Zhu, Q.H., Fan, L. & Guo, L. (2021). Twenty years of plant genome sequencing: Achievements and challenges. *Trends Plant Sci.* pages S1360–1385.

[21] Barnosky, A.D., Matzke, N., Tomiya, S., Wogan, G.O., Swartz, B., Quental, T.B., Marshall, C., McGuire, J.L., Lindsey, E.L., Maguire, K.C. *et al.* (2011). Has the earth's sixth mass extinction already arrived? *Nature* **471**(7336), 51–57.

[22] Ceballos, G., Ehrlich, P.R., Barnosky, A.D., García, A., Pringle, R.M. & Palmer, T.M. (2015). Accelerated modern human–induced species losses: Entering the sixth mass extinction. *Sci. Adv.* **1**(5), e1400253.

[23] Crozier, R.H. (1997). Preserving the information content of species: genetic diversity, phylogeny, and conservation worth. *Annu. Rev. Ecol. Syst.* **28**, 243–268.

[24] Faith, D.P., Magallón, S., Hendry, A.P., Conti, E., Yahara, T. & Donoghue, M.J. (2010). Evosystem services: an evolutionary perspective on the links between biodiversity and human well-being. *Curr. Op. Env. Sust.* **2**(1-2), 66–74.

[25] Small, E. (2011). The new Noah's Ark: beautiful and useful species only. Part 1. Biodiversity

conservation issues and priorities. *Biodiversity* **12**(4), 232–247.

[26] Rates, S.M.K. (2001). Plants as source of drugs. *Toxicon* **39**(5), 603–613.

[27] Kinghorn, A.D., De Blanco, E.J.C., Lucas, D.M., Rakotondraibe, H.L., Orjala, J., Soejarto, D.D., Oberlies, N.H., Pearce, C.J., Wani, M.C., Stockwell, B.R. *et al.* (2016). Discovery of anticancer agents of diverse natural origin. *Anticancer Res.* **36**(11), 5623–5637.

[28] Lichota, A. & Gwozdzinski, K. (2018). Anticancer activity of natural compounds from plant and marine environment. *International journal of molecular sciences* **19**(11), 3533.

[29] Aravind, L., Watanabe, H., Lipman, D.J. & Koonin, E.V. (2000). Lineage-specific loss and divergence of functionally linked genes in eukaryotes. *Proc. Natl. Acad. Sci. USA* **97**(21), 11319–11324.

[30] Cai, J.J. & Petrov, D.A. (2010). Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. *Gen. Biol. Evol.* **2**, 393–409.

[31] Weisman, C.M., Murray, A.W. & Eddy, S.R. (2020). Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **18**(11), e3000862.

[32] UniProt Consortium (2014). Activities at the universal protein resource (UniProt). *Nucleic Acids Res.* **42**, D191–D198.

[33] UniProt Consortium (2007). The universal protein resource (UniProt). *Nucleic Acids Res.* **36**(suppl_1), D190–D195.

[34] Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res.* **25**(17), 3389–3402.

[35] Lobley, A., Swindells, M.B., Orengo, C.A. & Jones, D.T. (2007). Inferring function using patterns of native disorder in proteins. *PLoS Comput. Biol.* **3**(8), e162.

[36] Hu, P., Janga, S.C., Babu, M., Díaz-Mejía, J.J., Butland, G., Yang, W., Pogoutse, O., Guo, X., Phanse, S., Wong, P. *et al.* (2009). Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. *PLoS Biol.* **7**(4), e1000096.

[37] Lucas, S.J., Akpınar, B.A., Šimková, H., Kubaláková, M., Doležel, J. & Budak, H. (2014). Next-generation sequencing of flow-sorted wheat chromosome 5D reveals lineage-specific translocations and widespread gene duplications. *BMC genomics* **15**(1), 1–18.

[38] Schäffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. & Altschul, S.F. (2001). Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* **29**(14), 2994–3005.

[39] Zuckerkandl, E. & Pauling, L. (1965). Evolutionary divergence and convergence in proteins. In V. Bryson & H. Vogel, eds., *Evolving Genes and Proteins*, pages 97–166. Academic Press, New York.

[40] Aris-Brosou, S. & Yang, Z. (2002). Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18S ribosomal RNA phylogeny. *Syst. Biol.* **51**(5), 703–714.

[41] Chun, J., Lee, J.H., Jung, Y., Kim, M., Kim, S., Kim, B.K. & Lim, Y.W. (2007). EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences. *Int. J. Syst. Evol. Micr.* **57**(10), 2259–2261.

[42] Del Campo, J., Kolisko, M., Boscaro, V., Santoferrara, L.F., Nenarokov, S., Massana, R., Guillou, L., Simpson, A., Berney, C., de Vargas, C. *et al.* (2018). EukRef: phylogenetic curation of ribosomal RNA to enhance understanding of eukaryotic diversity and distribution. *PLoS Biol.* **16**(9), e2005849.

[43] Tekaia, F., Lazcano, A. & Dujon, B. (1999). The genomic tree as revealed from whole proteome comparisons. *Genome Res.* **9**(6), 550–557.

[44] Zimmer, A., Lang, D., Richardt, S., Frank, W., Reski, R. & Rensing, S.A. (2007). Dating the early evolution of plants: detection and molecular clock analyses of orthologs. *Mol. Genet. Genomics* **278**(4), 393–402.

[45] Blake, D.P. (2015). Eimeria genomics: where are we now and where are we going? *Veterinary Parasitology* **212**(1-2), 68–74.

[46] Teakle, G.R. & Gilmartin, P.M. (1998). Two forms of type IV zinc-finger motif and their kingdom-specific distribution between the flora, fauna and fungi. *Trends Biochem. Sci.* **23**(3), 100–102.

[47] Alam, I., Hubbard, S.J., Oliver, S.G. & Rattray, M. (2007). A kingdom-specific protein domain HMM library for improved annotation of fungal genomes. *BMC genomics* **8**(1), 1–12.

[48] Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A. *et al.* (2022). AlphaFold Protein Structure Database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**(D1), D439–D444.

[49] Ruff, K.M. & Pappu, R.V. (2021). AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.* **433**(20), 167208.

[50] Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., Meyer, C., Laydon, A. *et al.* (2021). Highly accurate protein structure prediction for the human proteome. *Nature* **596**(7873), 590–596.

[51] Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *J. Mol. Biol.* **112**, 535–542.

[52] Ma, M., Stoyanova, M., Rademacher, G., Dutcher, S.K., Brown, A. & Zhang, R. (2019). Structure of the decorated ciliary doublet microtubule. *Cell* **179**(4), 909–922.

[53] Li, S., Fernandez, J.J., Fabritius, A.S., Agard, D.A. & Winey, M. (2022). Electron cryotomography structure of axonemal doublet microtubule from Tetrahymena thermophila. *Life Sci. Alliance* **5**(3).

[54] Gui, M., Farley, H., Anujan, P., Anderson, J.R., Maxwell, D.W., Whitchurch, J.B., Botsch, J.J., Qiu, T., Meleppattu, S., Singh, S.K. *et al.* (2021). De novo identification of mammalian ciliary motility proteins using cryo-EM. *Cell* **184**(23), 5791–5806.

[55] Lorenzo-Orts, L., Hohmann, U., Zhu, J. & Hothorn, M. (2019). Molecular characterization of CHAD domains as inorganic polyphosphate-binding modules. *Life Sci. Alliance* **2**(3).

[56] Siupka, P., Hamming, O.J., Frétaud, M., Luftalla, G., Levraud, J.P. & Hartmann, R. (2014). The crystal structure of zebrafish IL-22 reveals an evolutionary, conserved structure highly similar to that of human IL-22. *Gen. Immun.* **15**(5), 293–302.

[57] Domazet-Loso, T. & Tautz, D. (2003). An evolutionary analysis of orphan genes in Drosophila. *Genome Res.* **13**(10), 2213–2219.

[58] Wang, W., Yu, H. & Long, M. (2004). Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. *Nat. Genet.* **36**(5), 523–527.

[59] Palmieri, N., Kosiol, C. & Schlötterer, C. (2014). The life cycle of drosophila orphan genes. *elife* **3**.

[60] Lange, A., Patel, P.H., Heames, B., Damry, A.M., Saenger, T., Jackson, C.J., Findlay, G.D. & Bornberg-Bauer, E. (2021). Structural and functional characterization of a putative de novo gene in drosophila. *Nat. Comm.* **12**(1), 1–13.

[61] Ocaña-Pallarès, E., Williams, T.A., López-Escardó, D., Arroyo, A.S., Pathmanathan, J.S., Bapteste, E., Tikhonenkov, D.V., Keeling, P.J., Szöllősi, G.J. & Ruiz-Trillo, I. (2022). Divergent genomic trajectories predate the origin of animals and fungi. *Nature* **609**, 1–7.