On the Unknown Proteins of Eukaryotic Proteomes

Yves-Henri Sanejouand

Received: 30 September 2022 / Accepted: 7 May 2023 / Published online: 23 May 2023
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

Abstract
To study unknown proteins on a large scale, a reference system has been set up for the three better studied eukaryotic kingdoms, built with 36 proteomes as taxonomically diverse as possible. Proteins from 362 other eukaryotic proteomes with no known homologue in this set were then analyzed, focusing noteworthy on singletons, that is, on such proteins with no known homologue in their own proteome. Consistently, for a given species, no more than 12% of the singletons thus found are known at the protein level, according to Uniprot. In addition, since they rely on the information found in the alignment of homologous sequences, predictions of AlphaFold2 for their tridimensional structure are poor. In the case of metazoan species, the number of singletons rarely exceeds 1000 for the species the closest to the reference system (divergence times below 75 Myr). Interestingly, in the cases of viridiplantae and fungi, larger amounts of singletons are found for such species, 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

Introduction
Since the earliest genome sequencing projects proteins with no known homologue have been found in significant amounts (Siew and Fischer 2003; Tautz and Domazet-Lošo 2011). In Escherichia coli, most of them were found to be functional (Daubin and Ochman 2004), some being even found essential for the survival of organisms as complex as Drosophila (Chen et al. 2010; Xia et al. 2021).

In practice, the lack of homologue makes the functional annotation of such proteins difficult. On the other hand, their origin remains a matter of debate, a variety of mechanisms being advocated (Long et al. 2003; Vakirlis et al. 2020) such as gene duplication (Ohno 1970; Ohta 1989), followed by a quick neutral drift of their sequence (King and Jukes 1969; Kimura 1983; Trinquier and Sanejouand 1999; Bershtein et al. 2008), incorporation of viruses (Daubin and Ochman 2004) or transposable elements (Toll-Riera et al. 2009; Carelli et al. 2016), de novo genesis, that is, evolution from random amino acid sequences (White and Jacobs 1993; Knowles and McLysaght 2009; Heinen et al. 2009; Carvunis et al. 2012; Schmitz and Bornberg-Bauer 2017), etc.

With the advent of massive genome sequencing projects (Chain et al. 2009; Cao et al. 2020; Sun et al. 2021), the total number of such unknown proteins is increasing dramatically, with for instance 50% of the expressed genes in open ocean Tara stations that have no match in public sequence databases (Carradec et al. 2018) or 75% of metagenomic viral sequence data from 3042 geographically diverse samples that have no similarity to proteins of previously known viruses (Paez-Espino et al. 2016). As a corollary, a protein can one day be considered as being unknown and, the next day, as being just taxon specific.

Moreover, in the case of eukaryotic species, protein annotation is strongly 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. Thus, for the sake of reproducibility and, noteworthy, to allow for an easier comparison with results of other quantitative studies, a protocol allowing to retrieve proteins coined unknown in a robust way shall prove welcome, especially to perform comparisons between remote, less well studied, species.

Handling editor: Cara Weisman.

Yves-Henri Sanejouand
yves-henri.sanejouand@univ-nantes.fr

1 US2B, UMR 6286 of CNRS, Nantes University, rue de la Houssinière, 44322 Nantes, France

Springer
On the other hand, in a context of an ongoing massive extinction of species (Barnosky et al. 2011; Ceballos et al. 2015), 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 (Crozier 1997; Faith et al. 2010; Small 2011). In particular, molecules of particular importance for human health have been found in various species (Rates 2001; Kinghorn et al. 2016; Lichota and Gwozdzinski 2018). 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 promising characteristics needs to be considered.

Hereafter, in order to identify a significant set of such proteins, proteins with no known homologue were looked for in nearly 400 eukaryotic proteomes. On the other hand, to downsize the weight of model organisms, a specific definition is proposed for unknown proteins, based on the setup of a reference system for the proteomes of the three better studied eukaryotic kingdoms, namely, metazoa, viridiplantae (land plants) and fungi. Note that the use of a reference system is expected, for a given proteome, to yield counts of unknown proteins, hereafter coined specific, that will not change when new proteomes are obtained. However, while with a reference system the status of a given protein can be determined in a robust way, in the case of species far from those chosen for the reference set, lineage-specific proteins (Aravind et al. 2000; Cai and Petrov 2010; Weisman et al. 2020) are more likely to be considered as being unknown. Hereafter, in order assess the significance of this drawback, a set of proteomes of species from other (unicellular) eukaryotic lineages is also analyzed.

In practice, orthologs can be found across eukaryotic species using specialized databases such as OrthoDB (Waterhouse et al. 2011), OMA (Altenhoff et al. 2018) or eggNOG (Hernández-Plaza et al. 2022). However, since the present study focuses on proteins with no known homologue in a dedicated database, they were instead looked for directly.

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, making it difficult to compare counts of unknown proteins per species obtained in different studies. In the present one, noteworthy to address this issue, unknown proteins, hereafter coined specific, are instead defined with respect to a reference system, namely, a set of well-known proteomes as diverse as possible. To ensure diversity, the choice of these proteomes rely on taxonomy. It also rely on a couple of simple criteria, as detailed below.

As a reference system for eukaryotic proteomes, 36 proteomes were selected as follows, among the 398 reference proteomes (UniProt Consortium 2017) with more than 10,000 proteins available in Uniprot, smaller proteomes being ignored in order to focus on proteomes the more likely to be nearly complete. Note that splice isoforms are included in the reference proteomes of Uniprot.

For each of the three better studied eukaryotic kingdoms, namely, metazoa, viridiplantae and fungi, starting from the root, their taxonomic tree, as provided by Uniprot (UniProt Consortium 2007), was scanned up to the first node encountered where at least ten taxa 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.

Note that the reference proteomes of Uniprot are expected to be fairly well annotated (UniProt Consortium 2017). As a matter of fact, among the 398 proteomes considered in the present study, according to the Complete Proteome Detector (UniProt Consortium 2021), nearly 40% are high-value outliers, meaning that they have significantly more identified proteins than closely taxonomically related species. In addition, for these proteomes, complete BUSCO predictions of single-copy orthologs (Simão et al. 2015) are found for 92% of the cases (median value).

Search of Homologues

Homologues in the reference database were looked for using BLAST (Altschul et al. 1997), version 2.6.0+, two proteins being assumed to be homologous when the E-value of their pairwise alignment is lower than $10^{-6}$ (Lobley et al. 2007; Hu et al. 2009; Lucas et al. 2014). Note that, in order to avoid an overestimation of the number of specific proteins, due to the filtering of low-entropy segments, that is, of segments of restricted amino-acid composition, composition-based statistics (Schäffer et al. 2001) was not considered (-comp_based_stats 0).

In the present study, ubiquitous, specific and singleton proteins are defined as follows: ubiquitous proteins have homologues in all 36 eukaryotic proteomes of the reference set; specific proteins have none, singletons being such specific proteins with no homologue in their own proteome.

---

1 Only proteins with at least 50 amino-acid residues were considered.
2 On June 23th, 2020.
3 Most of them are likely to be “housekeeping” proteins.
Divergence times between a given species and the species of the reference system were retrieved from the TimeTree database, version 5 (Kumar et al. 2022), being in practice obtained with the nw_distance command of the Newick utilities package (Junier and Zdobnov 2010), as provided by the Biocnda project (Grüning et al. 2018). Note that for 12% of the species considered herein no divergence time could be obtained.

### Table 1: A reference system for eukaryotic proteomes

| Kingdom | Taxon | Species<sup>a</sup> | 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 floridæ | 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 | Hydatigera 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 |
|         | Caryophyloïdaceae | Phoenix dactylifera | PHODC | 34,033 |
|         | Nelumbonaceae | Nelumbo nucifera | NELNU | 31,582 |
|         | Funariaceae | Physcomitrella patens | PHYPAT | 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 |
|         | Blastocladiomycota | Allomyces macrogynus | ALLM3 | 19,092 |
|         | Pezizomycotina | Ascomycota | ASCIM | 17,778 |
|         | Dothideomycotina | Cortedespora cassicola<sup>b</sup> | CORCC | 17,125 |
|         | Mucorineae | Rhizopus delemar | RHIO9 | 16,971 |
|         | Cunninghamellaceae | Absidia glauca | ABSGL | 14,825 |
|         | Neocallimastigomycota | Piromyces sp | PIRSE | 14,606 |
|         | Eurotiomycotina | Penicillium camemberti<sup>b</sup> | PENCA | 14,390 |
|         | Sordariomycotina | Fusarium poae<sup>b</sup> | FUSPO | 14,048 |
|         | Leotiomycotina | Monilinia fructicola<sup>b</sup> | MONFR | 13,749 |
|         | Syncephalasporaceae | Syncephalastrum racemosum | SYNRA | 11,037 |

For each of the three better studied eukaryotic kingdoms, proteomes were chosen so as to be as taxonomically diverse as possible, among those with more than 10,000 proteins

<sup>a</sup> With the largest proteome of the taxon

<sup>b</sup> Ascomycota
Structure Prediction

Predictions of tridimensional structures were picked in the AlphaFold Protein Structure Database (Varadi et al 2022). These structures were obtained with the AlphaFold2\(^4\) artificial intelligence from DeepMind (Jumper et al 2021), an algorithm whose predictions proved to be the most accurate of all the submissions for over 90% of the targets of the CASP14 blind prediction experiment (Jones and Thornton 2022). Interestingly, AlphaFold2 also provides an estimate of the accuracy of its prediction for the position of each amino-acid residue, coined pLDDT,\(^5\) values over 90% corresponding to a high quality and values below 50% to a poor one (Varadi et al 2022). 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.

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 are ubiquitous, that is, they have homologues in all 36 proteomes of the reference set (see Fig. 1), 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 1000 ubiquitous proteins are found in

\(^4\) The second version of AlphaFold.

\(^5\) Standing for predicted local-distance difference test.
Fig. 2 Number of homologues in their own proteome of specific proteins, that is, of proteins with no homologue in the 36 proteomes of the reference system. Filled squares: metazoa; filled circles: viridiplantae; filled diamonds: fungi; open circles: other eukaryotes

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 (Blake 2015), which probably relies on its host (*Gallus gallus*) for compensating the lack of missing ones.

Interestingly, as suggested by Fig. 1, a significant number of proteins from metazoa and viridiplantae are kingdom—specific (Teakle and Gilmartin 1998; Alam et al. 2007), 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 (there is no obvious peak for eleven proteomes), as if their functional diversity were higher. Note however that there is a small peak for five proteomes, due to the five ascomycota species of the reference system (see Fig. 1, top left, and Table 1). In the case of the unicellular eukaryotes, less peaks are observed (Fig. 1, top right), suggesting that, as expected, their functional diversity is even higher.

**Specific Proteins and Singletons**

As shown in Fig. 1, the percentage of specific proteins, that is, of proteins with no homologue 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 specific proteins were just missed, as a consequence of their high degree of evolutionary divergence.

Interestingly, as shown in Fig. 2, whatever the kingdom, roughly half of specific proteins have homologues within their own proteome. Though such homologues can either be the result of gene-duplication events or just splice isoforms, in both cases it means that such proteins are likely to be older than specific proteins that have none, hereafter called singletons.

**Protein Existence Level**

In Uniprot, the type of evidence that supports the existence of a protein (also called degree of existence) is quantified through a number ranging between one (known at the protein level) and four (predicted). As shown in Fig. 3, only a minority of proteins of our dataset are known at the protein level, 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, model organisms for which extensive omics studies have been performed. 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% of the ubiquitous proteins are known by homology (degree three), meaning that they belong to known protein families.

Such results are in sharp contrast with what is observed for specific and singleton proteins, as they are defined in the present study, that is, with respect to a reference system. In both cases, for a given species, no more 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 (degree three), figures are 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%). For the three first ones, their number of singletons is unusually low (80 at most), as well as their number of specific proteins (138 at most), strongly suggesting that the annotation of these proteomes is incomplete, being biased towards proteins with already known homologues, the case of *Dictyostelium discoideum* illustrating

---

6 The proteome of *Eimeria mitis* is a low-value outlier, according to the Complete Proteome Detector. Note that it does not belong any more to the list of reference proteomes of Uniprot (on February 2023).

7 A small percentage of proteins are also classified as being uncertain (fifth degree).

8 Since this work was performed, the size of the proteomes of *Leptonychotes weddellii* and *Meleagris gallopavo* has increased by a factor of two.
the main drawback of the present approach, namely, that it tends to overestimate the number of singletons in the proteome of organisms far from the reference set.\(^9\)

**Quality of Structural Prediction**

The predicted tridimensional structures of all proteins of ten proteomes considered in the present study are presently\(^10\) available in the AlphaFold Protein Structure Database (Varadi et al 2022), 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 is none from fungi\(^11\).

In this database, the quality of the prediction of the position of each amino-acid residue of a protein by AlphaFold2 is provided as a percentage value (coined pLDDT), values over 90% corresponding to a high quality and values below 50% to a poor one (Varadi et al 2022). As shown in Fig. 4, whatever the eukaryotic kingdom, the quality of the prediction of the structures of ubiquitous proteins is rather high, with a median mean pLDDT value of \(\approx 80\%\), values below 50% being observed in only \(\approx 1\%\) of the cases.

Since the predictions of AlphaFold2 rely partly on the information found in the alignment of homologous sequences (Varadi et al 2022), that is, of sequences just expected to share some partial structural similarity but also of sequences of all known orthologs, the overall quality of the prediction of the structures of specific 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 structures of half of specific and singleton proteins of *Oryza sativa* and *Zea mays* are poorly predicted. This may reflect the fact that the proteomes of viridiplantae

\(\text{Fig. 3} \quad \text{Percentage of singleton (black bars), specific (grey bars) and ubiquitous (light grey bars) proteins with a given degree of existence, according to Uniprot. From left to right: known at protein level, at transcript level, by homology or predicted}\)

\(^9\) *Dictyostelium discoideum* is an amoeba.

\(^10\) As of April 2022 (version 1).

\(^11\) With only 6727 proteins, the proteome of *Saccharomyces cerevisiae* was not considered in the present study.
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\textsuperscript{12} that it does not allow AlphaFold2 to perform well (Varadi et al. 2022). This may also mean that there are more disordered, hard-to-predict proteins (Ruff and Pappu 2021; Tunyasuvunakool et al. 2021), among the specific 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 specific and singleton proteins (see Fig. 4), suggesting that singletons are not an atypical subset of the former. Since, on the other hand, specific proteins with numerous homologues in their own proteome (up to \approx 4000 of them) may do prove atypical, only singleton proteins are considered hereafter.

Interestingly, AlphaFold2 predicts with great confidence (mean 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 Fig. 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 \textit{Drosophila melanogaster}, 7 of \textit{Danio rerio} and a single one of \textit{Mus musculus} and \textit{Rattus norvegicus}. Among these 34 cases, according to Uniprot, 7 are known at the protein level, coming all from \textit{Drosophila melanogaster}.

\section*{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 (Bernstein et al. 1977) 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 \textit{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 previously associated with cilia (RIB21 and RIB30) (Ma et al. 2019). As expected for genuine singletons, they all seem absent in the axoneme of an alveolata, \textit{Tetrahymena thermophila} (Li et al. 2022). However, for three of them (FAP95, FAP107 and FAP143) structural orthologs were found in the cryo-EM electron density map of the axoneme of \textit{Bos taurus} (Gui et al. 2021). Thus, 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 exception being a protein with a CHAD domain found in \textit{Ricinus communis} (PDB 6QV5\textsuperscript{13}). On the other hand, the case of the interleukin 22 of \textit{Danio rerio} (PDB 4O6K; Siupka et al. (2014)) provides a clean example of a sequence drift quick enough so that the homology with its human counterpart is not detected, the E-value of the corresponding alignment (0.02) being over the chosen threshold (10\textsuperscript{-6}). Note, however, that such cases seem rare.

\footnotesize{\textsuperscript{12} Below 30.}

\footnotesize{\textsuperscript{13} Being 76\% identical with an inorganic triphosphatase of a \textit{Duganella} bacterium, this protein is however likely to be a contaminant.}
Evolution of the Number of Singletons

If singletons are added to proteomes from time to time, their number per species should increase as a function of the evolutionary distance from the species of the reference system. Note that a significant scatter is however expected on top of such a trend, as a consequence of the bias of proteome annotation towards proteins with known homologues, which can yield lower estimates for the number of specific proteins per species and, thus, of singletons. However, a trend should be observed for proteomes with the largest number of singletons per species, at a given evolutionary distance from the reference system.

As shown in Fig. 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 as a function of the divergence time between the considered species and the closest species of the reference set. Actually, around 3,000 singletons are found in the proteomes of two species close to the reference system, namely, *Hordeum vulgare* and *Brassica napus*, with divergence times of 10.3 and 20.6 Myr with respect to *Aegilops tauschii* and *Arabidopsis thaliana*, respectively, suggesting that singletons are added to proteomes of viridiplantae on a short timescale (less than 10–20 Myr), their proliferation being kept under control afterwards. However, in order to determine precisely on which timescale such a phenomenon occurs, proteomes closer to the reference system need to be considered.

In the case of metazoa, an evolutionary trend may be there (Fig. 5, top), with at most 1070 singletons found in the proteomes of primates (in the case of *Papio anubis*), and more than 1000 of them in the proteomes of only eight mammal species, among 65 ones, the largest number of singletons (4079) 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, which is an unusually large number, almost four times over the largest value found in viridiplantae, namely, 8875, in the proteome of *Trifolium medium*. Note that the proteomes of these three species are high-value outliers, according to the Complete Proteome Detector.

While, on average, viridiplantae have more ubiquitous proteins than fungi, namely, 3453 ± 3552 (median value: 2519) and 849 ± 351 (median value: 761), respectively, their average number of singletons are similar, namely, 2388 ± 1958 (median value: 1774) and 1922 ± 1077 (median value: 1638), respectively. Moreover, when only high-value outlier proteomes are considered, the average number of singletons is higher, namely, 3143 ± 2245 (median value: 2804) and 2738 ± 715 (median value: 2913), respectively, suggesting that the number of singletons could prove useful for assessing the completeness of a proteome.

On the other hand, singletons from unicellular eukaryotic lineages are also, on average, as numerous, namely, 3315 ± 1950 (median value: 2757), but this is a more likely consequence of the high degree of evolutionary divergence of these proteomes, with respect to the reference system.

Conclusion

In the present study, ubiquitous, specific and singleton proteins were defined as follows: ubiquitous proteins have homologues in all 36 eukaryotic proteomes as taxonomically diverse chosen as a reference set, specific proteins have none, singletons being such specific proteins with no homologue in their own proteome.

Thus, there are at least 1000 ubiquitous proteins in almost all 398 eukaryotic proteomes considered, the main exception being the proteome of *Eimeria mitis*, an api-complexan parasite. Whatever the eukaryotic kingdom, most ubiquitous proteins (40–50% of them) are known by homology (Fig. 3) while, 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* (Domazet-Loso and Tautz 2003; Wang et al. 2004; Palmieri et al. 2014; Lange et al. 2021).

As expected, the tridimensional structures of ubiquitous proteins are predicted by AlphaFold2 with a high level of confidence (Fig. 4), probably because such predictions rely on the information found in the alignment of homologous sequences (Varadi et al. 2022). As a matter of fact, in the case of singletons, the predictions of AlphaFold2 are poor (Fig. 4).

Interestingly, in the case of metazoan species, the number of singletons seems to increase for species away from the reference system (divergence times over 75 Myr). However, no such trend is observed in the cases of viridiplantae or fungi (Fig. 5). Though this phenomenon needs to be confirmed, by considering more proteomes of non-primatetype species close to the reference system, such results suggest that the timescale on which singletons are added to proteomes is slower in metazoa than 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 (Ocaña-Pallarés et al. 2022).

Overall, the number of singletons per species fluctuates by orders of magnitude, even in the case of evolutionary close species (Fig. 5). While low numbers may prove to be the result of incomplete annotation, proteomes of species with large numbers of singletons may deserve to be further scrutinized, like those of *Liparis tanakae* (33,715 of them), *Trifolium medium* (8875 of them) or, in the case of mammals, *Cricetulus griseus* (4079 of them).

Acknowledgements I thank Joseph Parello, for inspiring discussions, and the referees, for their careful and constructive reading of the manuscript.

References

Alam I, Hubbard SJ, Oliver SG et al. (2007) A kingdom-specific protein domain HMM library for improved annotation of fungal genomes. BMC Genomics 8(1):1–12

Altenhoff AM, Glover NM, Train CM et al. (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. Nucl Acids Res 46(D1):D477–D485

Altschul SF, Madden TL, Schäffer AA et al. (1997) Gapped blast and psi-blast: a new generation of protein database search programs. Nucl Acids Res 25(17):3389–3402

Aravind L, Watanabe H, Lipman DJ et al. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes. Proc Natl Acad Sci USA 97(21):11319–11324

Barnosky AD, Matzke N, Tomiya S et al. (2011) Has the earth’s sixth mass extinction already arrived? Nature 471(7336):51–57

Bernstein FC, Koetzle TF, Williams GJB et al. (1977) The protein data bank: a computer-based archival file for macromolecular structures. J Mol Biol 112:535–542

Bershtein S, Goldin K, Tawfik DS (2008) Intense neutral drifts yield robust and evolvable consensus proteins. J Mol Biol 379(5):1029–1044

Blake DP (2015) Eimeria genomics: where are we now and where are we going? Vet Parasitol 212(1–2):68–74

Cai JJ, Petrov DA (2010) Relaxed purifying selection and possibly high rate of adaptation in primate lineage-specific genes. Gen Biol Evol 2:393–409

Cao Y, Li L, Xu M et al. (2020) The ChinaMAP analytics of deep whole genome sequences in 10,588 individuals. Cell Res 30(9):717–731

Carelli FN, Hayakawa T, Go Y et al. (2016) The life history of retrocopies illuminates the evolution of mammalian genes. Genome Res 26(3):301–314

Carradec Q, Pelletier E, Da Silva C et al. (2018) A global ocean atlas of eukaryotic genes. Nat Commun 9(1):373

Carvunis AR, Rolland T, Wapinski I et al. (2012) Proto-genes and de novo gene birth. Nature 487(7407):370–374

Ceballos G, Ehrlich PR, Barnosky AD et al. (2015) Accelerated modern human-induced species losses: entering the sixth mass extinction. Sci Adv 1(5):e1400,253

Chain PSG, Graftham DV, Fulton RS et al. (2009) Genome project standards in a new era of sequencing. Science 326:236–237

Chen S, Zhang YE, Long M (2010) New genes in Drosophila quickly become essential. Science 330(6011):1682–1685

Crozier RH (1997) Preserving the information content of species: genetic diversity, phylogeny, and conservation worth. Annu Rev Ecol Syst 28:243–268

Daubin V, Ochman H (2004) Bacterial genomes as new gene homes: the genealogy of ORFans in E. coli. Genome Res 14(6):1036–1042

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

Faith DP, Magallón S, Hendry AP et al. (2010) Ecosystem services: an evolutionary perspective on the links between biodiversity and human well-being. Curr Opin Environ Sustain 2(1–2):66–74

Grüning B, Dale R, Sjödin A et al. (2018) Bioconda: sustainable and comprehensive software distribution for the life sciences. Nat Methods 15(7):475–476

Gui M, Farley H, Anuwan P et al. (2021) De novo identification of mammalian ciliary motility proteins using cryo-EM. Cell 184(23):5791–5806

Heinen TJ, Staubach F, Häming D et al. (2009) Emergence of a new gene from an intergenic region. Curr Biol 19(18):1527–1531

Hernández-Plaza A, Szklarczyk D, Botas J et al. (2022) eggNOG 6.0: enabling comparative genomics across 12 535 organisms. Nucl Acids Res 51:D389

Hu P, Janga SC, Babu M et al. (2009) Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol 7(4):e1000,096

Jones DT, Thornton JM (2022) The impact of AlphaFold2 one year on. Nat Methods 19(1):15–20

Jumper J, Evans R, Pritzel A et al. (2021) Applying and improving AlphaFold at CASP14. Proteins 89(12):1711–1721

Junier T, Zdobnov EM (2010) The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26(13):1669–1670

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

King JL, Jukes TH (1969) Non-darwinian evolution: most evolutionary change in proteins may be due to neutral mutations and genetic drift. Science 164(3881):788–798

Kinghorn AD, De Blanco EJC, Lucas DM et al. (2016) Discovery of anticancer agents of diverse natural origin. Anticancer Res 36(11):5623–5637

Knowles DG, Mclysaght A (2009) Recent de novo origin of human protein-coding genes. Genome Res 19(10):1752–1759
