Journal of Molecular Evolution (2024) 92:363–370
https://doi.org/10.1007/s00239-024-10174-z

ORIGINAL ARTICLE

Are Most Human-Specific Proteins Encoded by Long Noncoding RNAs?

Yves-Henri Sanejouand[1]

Received: 17 November 2023 / Accepted: 3 May 2024 / Published online: 25 June 2024

Abstract

By looking for a lack of homologs in a reference database of 27 well-annotated proteomes of primates and 52 well-annotated proteomes of other mammals, 170 putative human-specific proteins were identified. While most of them are deemed uncertain, 2 are known at the protein level and 23 at the transcript level, according to UniProt. Interestingly, 23 of these 25 proteins are found to be encoded or to have close homologs in an open reading frame of a long noncoding human RNA. However, half of them are predicted to be at least 80% globular, with a single structural domain, according to IUPred, and with at least 80% of ordered residues, according to flDPnn. Strikingly, there is a near-complete lack of structural knowledge about these proteins, with no tertiary structure presently available in the Protein Data Bank and a fair prediction for one of them in the AlphaFold Protein Structure Database. Moreover, knowledge about the function of these possibly key proteins remains scarce.

Keywords Tertiary structure · Globularity · UniProt · RNAcentral · IUPred · FlDPnn · AlphaFold

Introduction

Each time a new genome is sequenced, genes coding for proteins with no known homolog are found, even when genomes of closely related species are available, as is the case of primates (Toll-Riera et al. 2009; Cai and Petrov 2010; Ruiz-Orera et al. 2015; Sandmann et al. 2023) or *Drosophila* (Domazet-Loso and Tautz 2003; Wang et al. 2004; Heames et al. 2020; Grandchamp et al. 2023).

In the former case, the hypothesis that human-specific proteins may prove to be involved in the behavioral or anatomical peculiarities of the human species (Vakirlis et al. 2022; Papadopoulos and Albà 2023), such as the size of its brain (Rich and Carvunis 2023) or its functions (Li et al. 2010; Duffy et al. 2022; An et al. 2023), seems worthy of consideration. In the present study, taking advantage of both the high quality of the annotation of the human proteome (Amaral et al. 2023) and the availability of a significant number of well-annotated primate proteomes (Marques-Bonet et al. 2009; Juan et al. 2023), an extensive search of human-specific proteins was undertaken.

To achieve this, as in a previous study (Sanejouand 2023), a reference database was set up. Then, information about the tertiary structure of the identified putative human-specific proteins was gathered, with the idea that such knowledge could provide hints about their function or origin. No such information based on experimental data was found, so advantage was taken of recent progress in structure prediction methods (Kryshtafovych et al. 2021; Necci et al. 2021; Liu et al. 2023). Since, as is noteworthy in the case of proteins with no known homolog, predictions can vary significantly from one method to another (Monzon et al. 2022; Aubel et al. 2023; Middendorf and Eicholt 2023), three prediction methods based on different approaches were considered: first, IUPred (Dosztányi 2018), which provides a qualitative prediction stating whether a polypeptide is expected to adopt a globular fold; then, flDPnn (Hu et al. 2021), a neural network that provides structure disorder predictions and was ranked among the top methods in the recent Critical Assessment of protein Intrinsic Disorder (CAID) prediction experiment (Necci et al. 2021); and finally, AlphaFold (Jumper et al. 2021), which has proved able to predict the tertiary structure of proteins at an atomic level of detail (Kryshtafovych et al. 2021; Jones and Thornton 2022).

Handling editor: **Erich Bornberg-Bauer**.

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

[1] US2B, UMR 6286 of CNRS, Nantes University, 2 rue de la Houssinière, Nantes 44322, Pays de la Loire, France

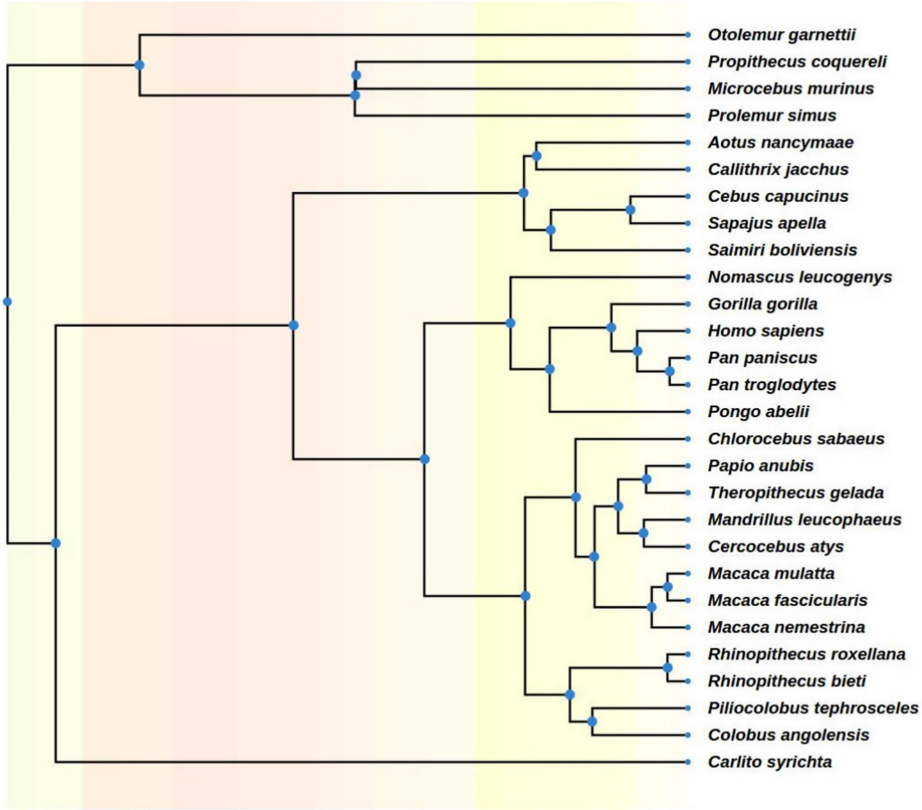

Fig. 1 The phylogenetic tree of the primate species with well-annotated proteomes considered herein

Methods

Choice of a Reference Database

The definition of what a species-specific protein is depends upon what is known about the proteomes of the closest species at a given time. For studying species-specific proteins, as well as for the sake of reproducibility, it is thus important to choose a well-defined reference system (Sanejouand 2023), that is, a set of reference proteomes. Moreover, since the quality of the annotation of a proteome can vary significantly from one proteome to another, these proteomes have to be chosen from among the best annotated ones.

To do this, 27 UniProt reference proteomes (UniProt Consortium 2017) of primates were picked, that is, all those that have at least a standard level of annotation,[1] according to the Complete Proteome Detector (UniProt Consortium 2021). Note that ten of these are high-value outliers, namely those of *Callithrix jacchus*, *Cercocebus atys*, *Gorilla gorilla*, *Macaca fascicularis*, *Macaca mulatta*, *Macaca nemestrina*, *Pan troglodytes*, *Papio anubis*, *Rhinopithecus roxellana*, and *Sapajus apella*, meaning that they have significantly more identified proteins than closely taxonomically related species.

For these 27 reference proteomes, complete BUSCO predictions of single-copy orthologs (Simão et al. 2015) are found for 94% (median value) of the cases. However, the percentage of short (less than 50 amino acid residues long) proteins varies widely from one proteome to another, ranging between 0.1% (*Sapajus apella*) and 3.5%, being over 1% in the case of only four primate species, namely *Pongo abelii*, *Pan troglodytes*, *Macaca fascicularis*, and *Homo sapiens*.

Overall, there are between 19,229 (*Chlorocebus sabaeus*) and 50,207 (*Macaca fascicularis*) proteins per proteome (40,000 ± 7500, on average) in our reference database, with a total of 1,083,746 proteins, including known isoforms.

Figure 1 shows the phylogenetic tree of the primate species considered in the present study, according to the TimeTree webserver version 5 (Kumar et al. 2022).

Search for Homologs

For each of the 20,449 human proteins that were at least 30 amino acid residues long, associated to a given gene, as found in UniProt (in February 2023), homologs in the reference database were sought using BLAST (Altschul et al. 1997) version 2.6.0+, assuming that two proteins are

[1] In June 2023.

homologous when the E-value of their pairwise alignment is lower than 10^{-6} (Lobley et al. 2007; Lucas et al. 2014; Sanejouand 2023). Note that, to avoid an overestimation of the number of specific proteins due to the filtering of low-entropy segments, that is, of segments with restricted amino acid composition, composition-based statistics (Schäffer et al. 2001) were not considered (-comp_based_stats 0).

Noncoding RNAs

For each human-specific protein found, that is, for each human protein with no homolog in the reference database, its possible encoding by a noncoding RNA (ncRNA), as found in the RNAcentral database (The RNAcentral Consortium 2015) version 22 was checked. To do this, its sequence was compared with those obtained by translating all nonoverlapping open reading frames at least 90 nucleotides long. Note that it was assumed that both start and stop codons are standard ones, while it is known that peptides encoded by ncRNAs can have atypical stop codons (Dragomir et al. 2020).

Protein Globularity

The degree of globularity of each human-specific protein found, that is, the percentage of the protein length predicted to be globular, as well as the number of globular domains, was estimated using version 1 (Dosztányi et al. 2005) of the standalone version of IUPred (Dosztányi 2018; Pajkos et al. 2023). Note that IUPred performs its predictions by considering the local sequential environment of each amino acid residue within 2–100 residues in either direction. Note also that, at variance with most recent methods (Necci et al. 2021), IUPred does not make use of evolutionary data, which are expected to be lacking in the case of species-specific proteins.

Ordered Residues

Binary predictions of ordered/disordered residues were obtained using the flDPnn neural network (Hu et al. 2021), as implemented in the eponymous webserver.[2]

In this study, high percentages, namely over 80%, of ordered residues are assumed to indicate that the considered protein is globular, meaning that it can adopt a stable tertiary structure. Note that the results obtained herein depend slightly upon the threshold chosen to indicate that a protein is predicted to be globular.

Structure Prediction

Predictions of tertiary structure were picked from the AlphaFold Protein Structure Database (Varadi et al. 2022), except for PACMP, which was included in UniProt after the release of the fourth version of the database. In this case, the prediction was performed using the standalone version of AlphaFold2, version 2.3, as available on the GitHub webserver.[3]

AlphaFold2 can also be used for predicting whether a protein has disordered segments. Indeed, AlphaFold2 provides an estimate of the accuracy of its prediction for the position of each amino acid residue, which is coined pLDDT,[4] with values over 90% corresponding to high quality, meaning that residue positions can be trusted, while for values below 50% they should not (Varadi et al. 2022). In the latter case, this can mean that the residues belong to disordered segments (Ruff and Pappu 2021; Pajkos et al. 2023), but it can also be interpreted as a possible lack of homologs, including remote ones, in the sequence databases available at the time of training the network (Varadi et al. 2022).

Hereafter, 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 ($\langle$pLDDT$\rangle$).

Results

How Many Human-Specific Proteins?

The requirement that the reference database be large enough (Vakirlis and McLysaght 2019) was assessed as follows: When proteins of *Homo sapiens* not found in the proteome of its closest relative, namely *Pan troglodytes*, are sought, 347 are identified. Note that this number is lower than a previous estimate obtained 10 years ago, namely 634 (Ruiz-Orera et al. 2015), maybe as a result of the improvement of the annotation of the human proteome (Amaral et al. 2023). Indeed, when the search was performed the other way around (Sanejouand 2023), 1036 chimpanzee-specific proteins[5] were found.

In fact, as shown in Fig. 2, when the number of proteomes in the reference database is increased, by adding proteomes one after another starting from the proteomes closest to the human species, the number of human-specific proteins drops

[2] http://biomine.cs.vcu.edu/servers/flDPnn/.

[3] https://github.com/google-deepmind/alphafold.

[4] Standing for predicted local-distance difference test.

[5] 847 singletons.

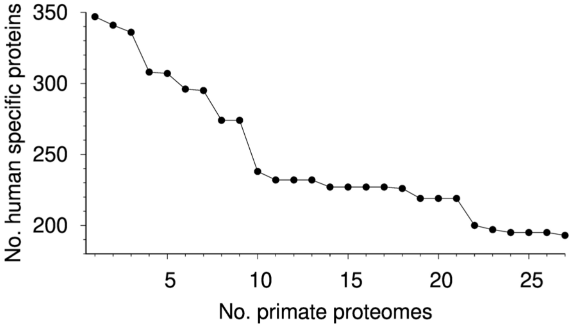

Fig. 2 Number of human-specific proteins as a function of the number of primate proteomes in which homologs of the human proteins were sought. Primate proteomes were added one by one according to the time of divergence between the primate and the human species, as provided in the TimeTree database, *Pan troglodytes* being added first (left) and *Protolemur simus* the 27th (right)

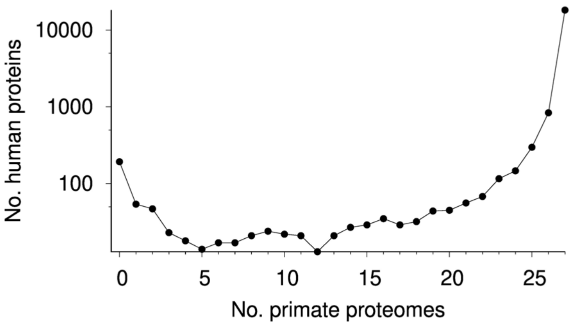

Fig. 3 Number of human proteins with homologs found in a given number of primate proteomes; 193 human proteins have no homolog in the proteomes of the 27 other primates considered

from 347 (left) to 193 (right). Interestingly, a few species make significantly higher contributions to this reduction, like the fourth (*Pongo abelii*) and tenth (*Macaca fascicularis*) (the two sharpest drops in Fig. 2), further suggesting that their proteomes are more complete than the others. However, while the proteome of *Macaca fascicularis* is indeed the largest in our reference database, the size of the proteome of *Pongo abelii*, with 39,491 proteins, is slightly below the average. Indeed, its level of annotation is only considered standard, according to the Complete Proteome Detector (UniProt Consortium 2021).

As shown in Fig. 3, while 89% of human proteins have homologs in all 27 proteomes in our reference database, 836 of them have homologs in all but one, 298 in all but two, etc., suggesting that the annotation of several reference proteomes is far from being complete. Of course, if the annotation of the 27 proteomes considered were improved or if more primate proteomes were added, the number of proteins found to be specific to the human species would continue to drop. To partially take this expected trend into account, homologs of the 193 proteins found above were sought in 52 UniProt reference proteomes of other mammalian species, these other proteomes being chosen on the basis of their high level of annotation, being all high-value outliers,[6] according to the Complete Proteome Detector (UniProt Consortium 2021).

Homologs were indeed found for 23 (12%) of them. However, in more than half of these cases, they were found in a single mammalian species *only*, as if the annotation of these proteins was intrinsically difficult. A possible reason is that these proteins are often short, with an average length of 106 ± 54 amino acid residues.

A total of 170 putative human-specific proteins were identified above, but as suggested, note that this number is expected to drop year on year as a consequence of the ongoing progress of proteome annotation. Note, however, that the protocol used in the present study was designed to be easy to reproduce, allowing for independent updates.

How Many Well-Known Ones?

In UniProt, the degree of knowledge, that is, the type of evidence that supports the existence of a protein, is quantified through a number ranging between one (known at the protein level) and five (uncertain).

Among the 170 putative human-specific proteins identified above, only 2 are known at the protein level (top of Table 1) according to UniProt, namely PACMP, the poly-ADP-ribosylation-amplifying and CtIP-maintaining micropeptide (Zhang et al. 2022), and SDIM1, the stress-responsive DNAJB4-interacting membrane protein 1 (Lei et al. 2011). Such a result is in sharp contrast to the fact that 90% of the human proteome is nowadays known at this level (Adhikari et al. 2020). Note that PACMP is short (44 residues) and, as such, could have escaped annotation in the proteomes considered above. In fact, PACMP was included in UniProt quite recently.[7]

On the other hand, while 23 of these proteins are known at the transcript level (Table 1), 23 others are just predicted. Strikingly, the 122 others (72%) are deemed uncertain in UniProt, being annotated as dubious CDS or gene predictions, possible pseudogenes, etc. This means that, according to UniProt, although a few of them may prove to be actual proteins, this is unlikely for the vast majority of them.

Actually, among the 25 human-specific proteins known at either the protein or the transcript level, except PACMP, SDIM1, CATR1, and HCP5, all of them are considered to be uncharacterized, meaning that they do not have any known

[6] In February 2023.

[7] In October 2023.

Table 1 The 25 human-specific proteins known at either the protein (top) or the transcript (bottom) level

UniProt Identifier	Protein length	RNA Identifier[a]	Start basis	Ordered Residues[b]	Globularity[c] (domains)	⟨pLDDT⟩
PACMP	44	URS0002617ABC	139	0	0 (0)	57
SDIM1	146	**No**	–	100	100 (1)	29
Q68DW6	58	URS00008B55A1	409	90	100 (1)	61
CATR1	79	**No**[d]	–	100	100 (1)	83
CF195	127	URS00025F035E	356	88	97 (1)	29
YV004	128	**No**	–	81	83 (1)	41
A4AS1	129	URS000233F457	1298	0	0 (0)	49
YS039	131	URS0002546BBE	2824	55	100 (1)	36
HCP5	132	URS00025BFBD4	372	91	100 (1)	39
YP023	132	URS000259F674	8	65	61 (2)	39
CA220	134	URS00025BEBB1	359	99	100 (1)	39
CL036	138	URS00025D0BC1	219	98	100 (1)	49
FEAS1	138	URS00009BBF20	179	73	100 (1)	40
YI001	140	URS00025BBEEA	315	26	0 (0)	38
YP033	140	URS0002573B74	1427	88	100 (1)	31
YK004	145	**No**[e]	–	87	100 (1)	38
YS045	151	URS0002339A91	77	80	93 (1)	41
YF010	163	URS00025B1F38	1769	69	60 (2)	34
YB035	174	URS00025A88CD	314	94	100 (1)	28
YV008	177	URS000258F499	37	51	42 (1)	26
IDAS1	188	URS0000A912F7	1	71	99 (1)	36
YT009	211	URS0000E60B4D	490	59	30 (1)	41
CQ077	243	URS00021231B2	527	88	91 (1)	25
CK072	251	URS00025C6C2F	242	86	87 (1)	27
YU004	302	URS0000EBC2BD	89	14	23 (1)	26

Most of these are found in an open reading frame of a long noncoding human RNA. Proteins predicted to be at least 80% globular, by IUPred, and with more than 80% of ordered residues, according to flDPnn, are shown in bold. The only protein with a fair quality of three-dimensional (3D) structure prediction, according to AlphaFold, is underlined

[a] In the RNAcentral database

[b] Percentage predicted to be ordered, according to flDPnn

[c] Percentage of protein length predicted to be globular, according to IUPred

[d] Putative peptides 72% identical are found in human RNA URS0001BF6EDB and URS00008B3362

[e] Putative peptides 90% identical, but less than 90 residues long, are found in human antisense RNA URS0001BF7FF4, URS0000E9D220, and URS000233D552

function. On the other hand, as stated in Table 1, 21 of them are found to be encoded by an open reading frame of a long noncoding human RNA (lncRNA), while 2 others, namely CATR1 and YK004, have close RNA-encoded homologs.

Interestingly, these 25 proteins, except SDIM1, CATR1, YV004, and YS039, also have close homologs encoded in the open reading frames of RNAs of other primate species, meaning that, at the RNA level, their sequences are not human specific. Since no transcript is known for any of them in UniProt, this raises the possibility that, in the human species, these RNAs have acquired the ability to be recognized as messenger ones. Of course, they may also just have been missed so far in species other than humans, at both the protein and the transcript level. Note that the growth of the number of reported noncoding RNA genes has been rapid, suggesting that primate catalogs may, in this respect, prove rather incomplete (Amaral et al. 2023).

How Many Globular Ones?

From the results above, it is tempting to speculate that most human-specific genes do not code for proteins and may instead be involved, like many lncRNAs (Statello et al. 2021), in the regulation of gene expression (Nahon 2003). However, it has recently been shown that translation is widespread at many annotated lncRNA transcripts (Patraquim et al. 2020, 2022; Mudge et al. 2022; Broeils et al. 2023), with up to 3330 human lncRNAs found bound

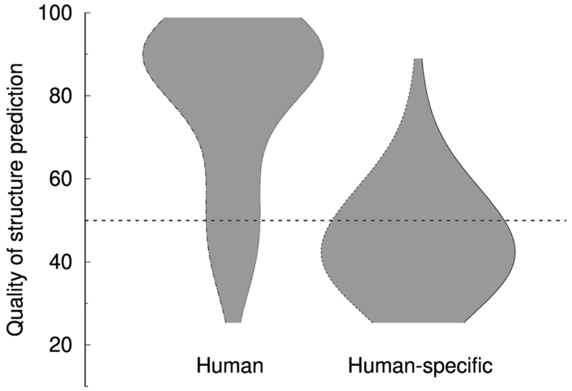

Fig. 4 Quality of tertiary structure prediction, according to AlphaFold2, for the whole human proteome (left) and for the putative human-specific proteins identified herein (right). The dashed line indicates the quality threshold below which the confidence in the models is very low (⟨pLDDT⟩ < 50)

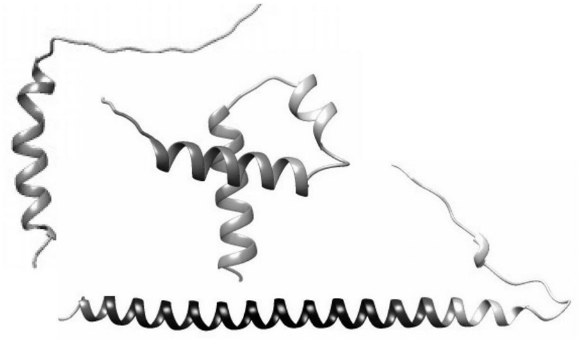

Fig. 5 The best predicted structures of human-specific proteins, according to AlphaFold2. Left: PACMP, the poly-ADP-ribosylation-amplifying and CtIP-maintaining micropeptide (⟨pLDDT⟩=57). Middle: Q68DW6, an uncharacterized protein (⟨pLDDT⟩=61). Bottom: CATR1, the CATR tumorigenic conversion 1 protein (⟨pLDDT⟩=83). The darker the color, the higher the level of confidence (pLDDT). For entries Q13166 (CATR1) and Q68DW6, colored representations can be found at https://www.uniprot.org/uniprotkb. Drawn with Chimera (Pettersen et al. 2004)

to ribosomes with active translation elongation (Lu et al. 2019).

Actually, lncRNAs often show coding potential and sequence constraints similar to evolutionarily young protein coding sequences (Ruiz-Orera et al. 2014). It is thus necessary to assess the coding potential of lncRNAs. A straightforward way to do this is to predict how globular the encoded proteins are expected to be (Papadopoulos et al. 2021; Peng and Zhao 2024). As shown in Table 1, ten human-specific proteins known at the transcript level and encoded by an lncRNA (50% of them) are predicted to be at least 80% globular, by IUPred, with a single structural domain and more than 80% of ordered residues, according to flDPnn. Note that the predictions of IUPred and flDPnn are similar. In fact, they differ by more than 20% for five cases only, namely YS049, FEAS1, YI001, IDAS1, and YT009.

Note also that the two human-specific proteins that are *not* known to be encoded or to have homologs encoded by a human lncRNA, namely SDIM1 and YV004, are predicted to be at least 83% globular, by IUPred, also with a single structural domain, and to have more than 80% of ordered residues, according to flDPnn.

On the other hand, since nearly 30% of regions within the proteome are expected to be disordered (Ruff and Pappu 2021), other human-specific lncRNAs could also encode genuine, though disordered, proteins.

What about Their Structure?

No homolog was found in the Protein Data Bank (Kouranov et al. 2006) for the 25 human-specific proteins identified above. However, thanks to machine learning algorithms, major progress has recently been witnessed in the field of protein structure prediction (Jumper et al. 2021; Jones and Thornton 2022). Moreover, such predictions have been performed on a large scale. Furthermore, they are nowadays available in public databases (Varadi et al. 2022).

As illustrated in Fig. 4, the tertiary structure of most human proteins has been predicted with a high level of confidence by AlphaFold2 (Varadi et al. 2022), the average pLDDT being over 90 for nearly 40% of them, and over 70 for more than 70% of them. Note that the structures of only 14% of human proteins are predicted with low confidence (⟨pLDDT⟩ below 50).

However, in the case of the putative human-specific proteins identified above, up to 70% of them are predicted with such low confidence (Fig. 4). As specified in Table 1, among the 25 human-specific proteins known at either the protein or the transcript level, AlphaFold2 is able to make a fair prediction in the case of one of them (⟨pLDDT⟩ = 83) only, namely CATR1, the CATR tumorigenic conversion 1 protein (Li et al. 1995, 1998). However, as shown in Fig. 5, its predicted structure is fairly simple, with a single, long α-helical segment.

The tendency of AlphaFold to predict helical structures for short proteins with no known homolog has already been noted (Monzon et al. 2022). In fact, the two other rather confident predictions of AlphaFold2 (⟨pLDDT⟩ > 50; see Table 1) are also for small proteins with very simple topologies (Fig. 5).

Note that the structure of most proteins predicted to be globular, by IUPred, and to have more than 80% of ordered residues, by flDPnn, is predicted with little confidence by AlphaFold2 (Table 1). This result further suggests that AlphaFold2 is of little help for predicting the structure of proteins with no known homolog (Varadi et al. 2022;

Monzon et al. 2022; Middendorf and Eicholt 2023), as for instance illustrated in a previous study of 362 eukaryotic proteomes (Sanejouand 2023). On the other hand, such discrepancies could also indicate conditional folding of intrinsically disordered regions (Alderson et al. 2023).

Conclusions

By looking for a lack of homologs in a reference database of 27 well-annotated proteomes of primates and 52 well-annotated proteomes of other mammals, 170 putative human-specific proteins were identified. However, most of these are deemed uncertain in UniProt, casting doubts on the 23 that are deemed to be predicted. Indeed, given the efforts made to complete the annotation of the human proteome (Amaral et al. 2023), it becomes less and less likely to have human proteins that are not known at least at the transcript level (Adhikari et al. 2020).

On the other hand, 23 of the 25 human-specific proteins known at either the protein or the transcript level are found to be encoded or to have close homologs in an open reading frame of a human lncRNA (Table 1). While one of them, namely PACMP, the poly-ADP-ribosylation-amplifying and CtIP-maintaining micropeptide, is known at the protein level (Zhang et al. 2022), 12 others are predicted to be at least 80% globular, with a single structural domain, and to have more than 80% of ordered residues, suggesting that a majority of human-specific proteins may prove to be encoded by lncRNAs.

In fact, de novo proteins have already been found to have an lncRNA origin (Ruiz-Orera et al. 2014, 2020). Such lncRNAs could come from RNAs with a former regulatory function or from intergenic open reading frames (Papadopoulos et al. 2021), which in turn may appear randomly, becoming new functional proteins when they happen to confer selective advantages (Ruiz-Orera et al. 2020).

Acknowledgements I thank the referees, for their constructive suggestions, Philippe Minard, for a careful reading of the manuscript, Johann Hendrickx, for performing the structure prediction for PACMP, and Aymeric Blondel, for making AlphaFold available on the Centre de Calcul Intensif des Pays de la Loire.

Declarations

Conflict of interest The author declares no conflicts of interest.

References

Adhikari S, Nice EC, Deutsch EW et al (2020) A high-stringency blueprint of the human proteome. Nat Comm 11(1):5301

Alderson TR, Pritišanac I, Kolarić D et al (2023) Systematic identification of conditionally folded intrinsically disordered regions by AlphaFold2. Proc Natl Acad Sci USA 120(44):e2304302120

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

Amaral P, Carbonell-Sala S, De La Vega FM et al (2023) The status of the human gene catalogue. Nature 622(7981):41–47

An NA, Zhang J, Mo F et al (2023) De novo genes with an lncRNA origin encode unique human brain developmental functionality. Nat Ecol Evol 7(2):264–278

Aubel M, Eicholt L, Bornberg-Bauer E (2023) Assessing structure and disorder prediction tools for de novo emerged proteins in the age of machine learning. F1000 Research 12:347

Broeils LA, Ruiz-Orera J, Snel B et al (2023) Evolution and implications of de novo genes in humans. Nat Ecol Evol 7:804–815

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

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

Dosztányi Z (2018) Prediction of protein disorder based on IUPred. Protein Sci 27(1):331–340

Dosztányi Z, Csizmok V, Tompa P et al (2005) IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. Bioinformatics 21(16):3433–3434

Dragomir MP, Manyam GC, Ott LF et al (2020) FuncPEP: a database of functional peptides encoded by non-coding RNAs. Non-coding RNA 6(4):41

Duffy EE, Finander B, Choi G et al (2022) Developmental dynamics of RNA translation in the human brain. Nat Neurosci 25(10):1353–1365

Grandchamp A, Kühl L, Lebherz M et al (2023) Population genomics reveals mechanisms and dynamics of de novo expressed open reading frame emergence in *Drosophila melanogaster*. Genome Res 33(6):872–890

Heames B, Schmitz J, Bornberg-Bauer E (2020) A continuum of evolving de novo genes drives protein-coding novelty in *Drosophila*. J Mol Evol 88(4):382–398

Hu G, Katuwawala A, Wang K et al (2021) flDPnn: accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Comm 12(1):4438

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

Juan D, Santpere G, Kelley JL et al (2023) Current advances in primate genomics: novel approaches for understanding evolution and disease. Nat Rev Genet 24(5):314–331

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

Kouranov A, Xie L, De La Cruz J et al (2006) The RCSB PDB information portal for structural genomics. Nucl Acid Res 34:D302–D305

Kryshtafovych A, Schwede T, Topf M et al (2021) Critical assessment of methods of protein structure prediction (CASP)-Round XIV. Proteins Struct Funct Bioinf 89(12):1607–1617

Kumar S, Suleski M, Craig JM et al (2022) TimeTree 5: an expanded resource for species divergence times. Mol Biol Evol 39(8):msac174

Lei JX, Cassone CG, Luebbert C et al (2011) A novel neuron-enriched protein SDIM1 is down regulated in Alzheimer's brains and attenuates cell death induced by DNAJB4 over-expression in neuro-progenitor cells. Mol Neurodegener 6(1):1–16

Li D, Noyes I, Shuler C et al (1995) Cloning and sequencing of CATR1.3, a human gene associated with tumorigenic conversion. Proc Natl Acad Sci USA 92(14):6409–6413

Li D, Sun XL, Casto B et al (1998) Epstein-Barr virus growth-transformed cells are converted to malignancy following transfection of a 1.3-kb CATR1 antisense construct independent of a change in the level of c-Myc expression followed by a 8;14 chromosomal translocation. Proc Natl Acad Sci USA 95(9):4894–4899

Li CY, Zhang Y, Wang Z et al (2010) A human-specific de novo protein-coding gene associated with human brain functions. PLoS Comput Biol 6(3):e1000734

Liu J, Yuan R, Shao W et al (2023) Do "Newly Born'' orphan proteins resemble "Never Born'' proteins? A study using three deep learning algorithms. Proteins Struct Funct Bioinf 91(8):1097–1115

Lobley A, Swindells MB, Orengo CA et al (2007) Inferring function using patterns of native disorder in proteins. PLoS Comput Biol 3(8):e162

Lu S, Zhang J, Lian X et al (2019) A hidden human proteome encoded by "non-coding'' genes. Nucleic Acids Res 47(15):8111–8125

Lucas SJ, Akpınar BA, Šimková H et al (2014) Next-generation sequencing of flow-sorted wheat chromosome 5D reveals lineage-specific translocations and widespread gene duplications. BMC Genomics 15(1):1–18

Marques-Bonet T, Ryder OA, Eichler EE (2009) Sequencing primate genomes: what have we learned? Annu Rev Genomics Hum Genet 10:355–386

Middendorf L, Eicholt LA (2023) Random, de novo and conserved proteins: how structure and disorder predictors perform differently. bioRxiv 07.18:549582

Monzon V, Haft DH, Bateman A (2022) Folding the unfoldable: using AlphaFold to explore spurious proteins. Bioinf Adv 2(1):vbab043

Mudge JM, Ruiz-Orera J, Prensner JR et al (2022) Standardized annotation of translated open reading frames. Nat Biotechnol 40(7):994–999

Nahon JL (2003) Birth of "human-specific'' genes during primate evolution. Genetica 118:193–208

Necci M, Piovesan D, Tosatto SC (2021) Critical assessment of protein intrinsic disorder prediction. Nat Methods 18(5):472–481

Pajkos M, Erdős G, Dosztányi Z (2023) The origin of discrepancies between predictions and annotations in intrinsically disordered proteins. Biomolecules 13(10):1442

Papadopoulos C, Albà MM (2023) Newly evolved genes in the human lineage are functional. Trends Genet 39(4):235–236

Papadopoulos C, Callebaut I, Gelly JC et al (2021) Intergenic ORFs as elementary structural modules of de novo gene birth and protein evolution. Genome Res 31(12):2303–2315

Patraquim P, Mumtaz MAS, Pueyo JI et al (2020) Developmental regulation of canonical and small ORF translation from mRNAs. Gen Biol Evol 21(1):1–26

Patraquim P, Magny EG, Pueyo JI et al (2022) Translation and natural selection of micropeptides from long non-canonical RNAs. Nat Comm 13(1):6515

Peng J, Zhao L (2024) The origin and structural evolution of de novo genes in *Drosophila*. Nat Comm 15:810

Pettersen EF, Goddard TD, Huang CC et al (2004) UCSF chimera: a visualization system for exploratory research and analysis. J Comput Chem 25(13):1605–1612

Rich A, Carvunis AR (2023) De novo gene increases brain size. Nat Ecol Evol 7(2):180–181

Ruff KM, Pappu RV (2021) AlphaFold and implications for intrinsically disordered proteins. J Mol Biol 433(20):167208

Ruiz-Orera J, Messeguer X, Subirana JA et al (2014) Long non-coding RNAs as a source of new peptides. elife 3:e03523

Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C et al (2015) Origins of de novo genes in human and chimpanzee. PLoS Genet 11(12):e1005721

Ruiz-Orera J, Villanueva-Cañas JL, Albà MM (2020) Evolution of new proteins from translated sORFs in long non-coding RNAs. Exp Cell Res 391(1):111940

Sandmann CL, Schulz JF, Ruiz-Orera J et al (2023) Evolutionary origins and interactomes of human, young microproteins and small peptides translated from short open reading frames. Mol Cell 83(6):994–1011

Sanejouand YH (2023) On the unknown proteins of eukaryotic proteomes. J Mol Evol 91:492–501

Schäffer AA, Aravind L, Madden TL et al (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. Nucleic Acids Res 29(14):2994–3005

Simão FA, Waterhouse RM, Ioannidis P et al (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31(19):3210–3212

Statello L, Guo CJ, Chen LL et al (2021) Gene regulation by long non-coding RNAs and its biological functions. Nat Rev Mol Cell Biol 22(2):96–118

The RNAcentral Consortium (2015) RNAcentral: an international database of ncRNA sequences. Nucleic Acids Res 43:D123–D129

Toll-Riera M, Bosch N, Bellora N et al (2009) Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol 26(3):603–612

UniProt Consortium (2017) Uniprot: the universal protein knowledgebase. Nucleic Acids Res 45(D1):D158–D169

UniProt Consortium (2021) Uniprot: the universal protein knowledgebase in 2021. Nucleic Acids Res 49(D1):D480–D489

Vakirlis N, McLysaght A (2019) Computational prediction of de novo emerged protein-coding genes. Meth Mol Biol 1851:63–81

Vakirlis N, Vance Z, Duggan KM et al (2022) De novo birth of functional microproteins in the human lineage. Cell Rep 41(12):111808

Varadi M, Anyango S, Deshpande M 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

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

Zhang C, Zhou B, Gu F et al (2022) Micropeptide PACMP inhibition elicits synthetic lethal effects by decreasing CtIP and poly(ADP-ribosyl)ation. Mol Cell 82(7):1297–1312