ADVANCED BIOLOGY
www.advanced-bio.com

Unique Patterns in Amino Acid Sequences of Aging-Related Proteins

*Eszter Zita Szatmári, Attila Csordás, and Csaba Kerepesi**

Aging has strong genetic components and the list of genes that may regulate the aging process is collected in the GenAge database. There may be characteristic patterns in the amino acid sequences of aging-related proteins that distinguish them from other proteins and this information will lead to a better understanding of the aging process. To test this hypothesis, human protein sequences are extracted from the UniProt database and the relative frequency of every amino acid residue in aging-related proteins and the remaining proteins is calculated. The main observation is that the mean relative frequency of aspartic acid (D) is consistently higher, while the mean relative frequencies of tryptophan (W) and leucine (L) are consistently lower in aging-related proteins compared to the non-aging-related proteins for the human and four examined model organisms. It is also observed that the mean relative frequency of aspartic acid is higher, while the mean relative frequency of tryptophan is lower in pro-longevity proteins compared to anti-longevity proteins in model organisms. Finally, it is found that aging-related proteins tend to be longer than non-aging-related proteins. It is hoped that this analysis initiates further computational and experimental research to explore the underlying mechanisms of these findings.

1. Introduction

A better understanding of human aging can help achieve a higher quality of life and healthier aging, potentially contributing to the fight against aging-related diseases. However, aging is still a poorly understood process.[1] In 1993 it was possible to duplicate the lifespan of the roundworm (*Caenorhabditis elegans*) by knocking out just one gene, Daf-2.[2] Today it is possible to extend lifespan and slow aging in other model organisms, even in mice[3] leading to the conclusion that aging has strong genetic components.[4,5] The list of human genes that may regulate the human aging process is collected in the manually-curated GenAge database.[1,6] The current version of GenAge (v20) consists of 307 human aging-related genes that include members that are directly linked to human aging (e.g., progeroid syndrome genes), or genes that are linked to aging by evidence in model organisms (the list of all criteria at http://genomics.senescence.info/help.html#genage).

Although individual aging-related genes have been intensively studied, their bulk analysis is still limited. Previously, using machine learning, we predicted and characterized aging-related proteins based on 36 Gene Ontology and protein-protein interaction network features.[7] We did not utilize features extracted from the amino acid sequences in that study. In the current work, we hypothesize that there may be characteristic patterns in the amino acid sequences of aging-related proteins that distinguish them from other human proteins and this information would lead to a better understanding of the human aging process. To test this hypothesis, we extracted human protein sequences from the UniProt database[8] and calculated the relative frequency of the amino acid residues in aging-related proteins and the remaining human proteins (non-aging-related proteins, hereafter).

2. Results

We observed that the relative frequency of certain amino acids was significantly different in the aging-related set compared to the non-aging-related set of proteins (**Figure 1**A). The most remarkable differences were in the case of aspartic acid, tryptophan and leucine (p-value < 0.0035, FDR corrected p-value < 0.01314, Data S1, Supporting Information). We examined if these phenomena are evolutionarily conserved. The relative frequency of tryptophan and leucine was consistently lower while the relative frequency of aspartic acid was consistently higher in aging-related proteins compared to the non-aging-related proteins for humans and four examined model organisms such as *C. elegans*, *D. melanogaster*, *M. musculus*, and *S. cerevisiae* (Figure 1B–D).

E. Z. Szatmári, C. Kerepesi
Institute for Computer Science and Control (SZTAKI)
Hungarian Research Network (HUN-REN)
Budapest 1111, Hungary
E-mail: kerepesi@sztaki.hu
E. Z. Szatmári
Department of Applied Analysis and Computational Mathematics
Eötvös Loránd University (ELTE)
Pázmány Péter sétány 1/C, Budapest 1117, Hungary
A. Csordás
AgeCurve Limited
Cambridge CB21SD, UK

The ORCID identification number(s) for the author(s) of this article can be found under https://doi.org/10.1002/adbi.202300436

DOI: 10.1002/adbi.202300436

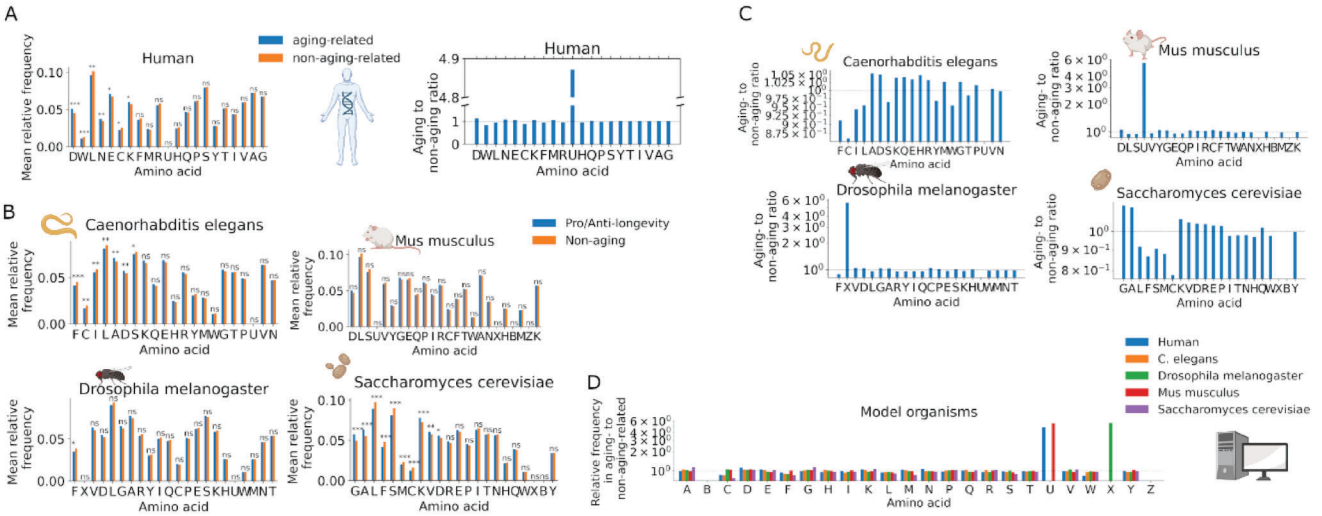

Figure 1. A) The mean relative frequency of each amino acid (in percentage) in aging-related proteins and non-aging-related proteins (left panel) and the ratio of the mean relative frequencies in the aging-related- and non-aging-related proteins in the human (right panel). B) The mean relative frequency of each amino acid (in percentage) in aging-related proteins and non-aging-related proteins in model organisms. C) The ratio of the mean relative frequencies in the aging-related- and non-aging-related proteins in model organisms. D) The ratio of the mean relative frequencies in the aging-related- and non-aging-related proteins in human and different model organisms.

We also examined the ratio of relative frequencies in pro-longevity proteins (products of genes whose decreased expression reduced lifespan and/or overexpression extended lifespan) to anti-longevity proteins (products of genes whose decreased expression extended lifespan and/or overexpression reduced lifespan) in the four model organisms (**Figure 2**). The mean relative frequency of aspartic acid was consistently higher while the mean relative frequency of tryptophan was consistently lower in pro-longevity proteins compared to anti-longevity proteins.

We also examined the length of amino acids in regarding our analysis. We found that the aging-related proteins tend to be longer than the non-aging-related proteins for human and model organisms (**Figure 3**). However, there was no significant difference between the mean length of pro-longevity proteins and anti-longevity proteins for any examined model organisms.

3. Discussion

We discuss the possible role of the observed unique patterns of the amino acid sequences of aging-related proteins in the human aging process. As aging is often described as a damage accumulation it is possible that some amino acids are more susceptible to some forms of molecular damage then other amino acids and this is associated with the longevity effect of a protein.

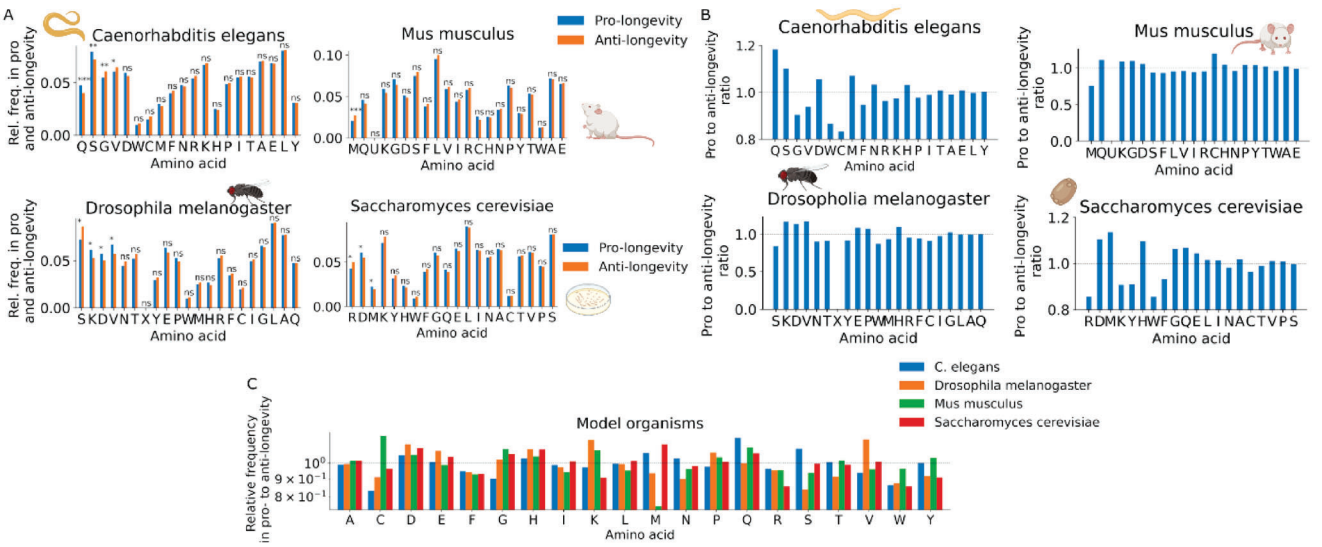

Figure 2. A) The mean relative frequency of each amino acid (in percentage) in pro-longevity proteins and anti-longevity proteins in model organisms. B) The ratio of the mean relative frequencies in the pro-longevity- and anti-longevity proteins in model organisms. C) The ratio of the mean relative frequencies in the pro-longevity- and anti-longevity proteins in different model organisms.

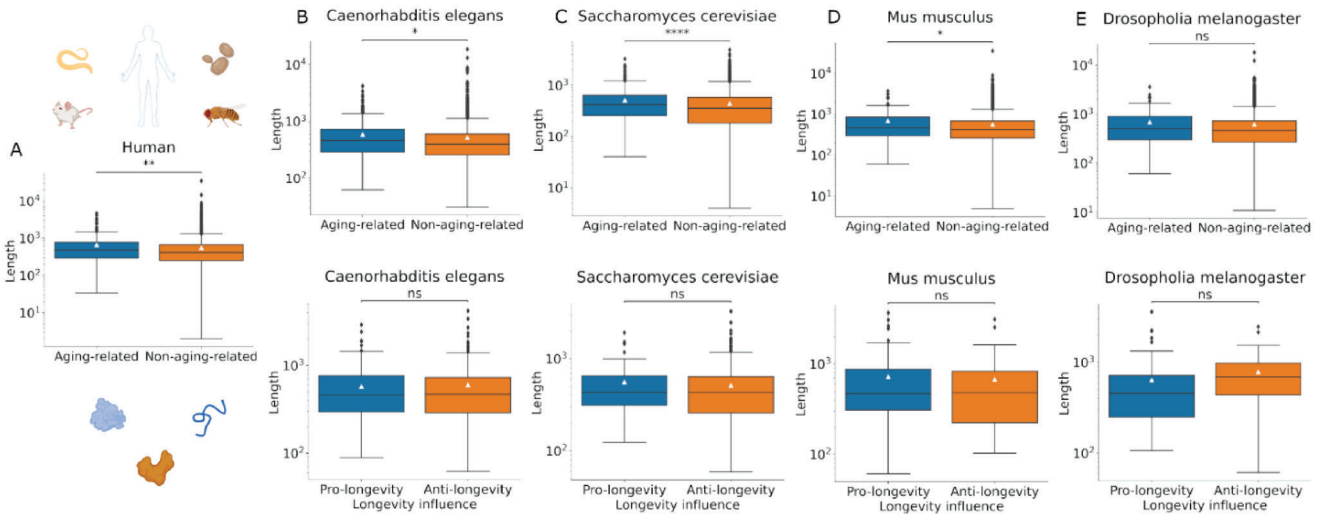

Figure 3. A) Comparing the length of aging-related proteins and non-aging-related proteins for humans. B–E) Length differences between aging-related and non-aging-related proteins as well as between pro-longevity and anti-longevity proteins for *C. elegans*, *S. cerevisiae*, *M. musculus*, and *D. melanogaster*, respectively.

There are several lines of distinct, but overlapping evidence for the inclusion of a human gene into GenAge (see inclusion criteria list consisting of 9 categories here: https://genomics.senescence.info/help.html#genage). Evidence can directly link gene products to aging in humans, mammalian or non-mammalian model organisms, in vitro cell cultures, human longevity, upstream gene regulation control elements, functional pathways or mechanisms, or downstream elements amongst others. Below we propose a mechanism that could explain some of our findings. It requires further investigation to establish how this mechanism can be related to some of these evidence categories.

We found that the mean relative frequency of aspartic acid was consistently higher in the aging-related proteins compared to the non-aging-related proteins. Aspartic acid can undergo spontaneous, non-enzymatic isomerization to form isoaspartic acid. Eventually, the equilibrium favors the formation of isoaspartate because it is a less strained ring structure, which is more energetically favorable. This modification can affect protein structure and function and has been implicated in aging and diseases like Alzheimer's.[9] Aging-related proteins may be more prone to isoaspartate accumulation (as damage) due to their higher relative mean aspartate acid frequency than the non-aging-related protein set. If pro-longevity proteins have more pronounced isoaspartate accumulation with age this can block their effects in older cohorts. If anti-longevity proteins accumulate isoaspartate with age spontaneously this might have pro-longevity effects. In the future it would be possible to test for the reversibility of these effects via enzymes known as protein isoaspartate methyltransferases (PIMT) that can recognize and repair isoaspartate residues, converting them back to aspartate in an effort to maintain protein integrity.

Most amino acids are coded by multiple codons while tryptophan is coded by a single one. Therefore a single point mutation in any of the nucleotides in the tryptophan encoding triplet will result in a different amino acid. Remarkably, we observed a strong difference in relative frequencies between aging-related proteins and non-aging-related proteins in the case of tryptophan. It is possible that tryptophan is sensitive to somatic mutations and this can be linked to its underrepresentation in aging-related proteins for human and all model organisms.

Amino acid imbalance has roles in the regulation of aging and aging-related diseases.[10] Tryptophan through some final products produced through its metabolisms (MTOR, IIS, and Kynurenic acid) can expand lifespan.[11] The increase of leucine availability upregulates the mRNA translation, thereby increases muscle protein synthesis, which, in turn, leads to greater net muscle protein accretion.[12] Racemization of aspartyl residues in certain tissues in teeth has been shown to occur at a rate that corresponds to an enrichment in the D-aspartic acid content of 0.1% per year, meaning this enrichment indicates age.[13]

It was shown previously, that human genes related to aging have longer transcript lengths when compared with the rest of protein-coding genes.[14] Also, the protein products of longevity genes were shown to be in general longer than the rest of the *C. elegans* proteins.[15] Consistently, here we found that the aging-related proteins tend to be longer than the non-aging-related proteins for the human and four model organisms. A recent research found that in humans and mice the genes with the longest transcripts enriched for lifespan-extending genes reported to extend lifespan, while genes with the shortest transcripts enriched for genes reported to shorten lifespan.[16] In contrast, we found no significant difference between the mean length of pro-longevity proteins and anti-longevity proteins for any examined model organisms.

4. Conclusion

We found that certain amino acids are overrepresented while others are underrepresented in aging-related proteins and that the relative frequencies of certain amino acids are different in pro-longevity proteins compared to anti-longevity proteins. Altogether our simple data analysis suggests a link between the amino acid pattern of a protein and its capability to regulate the

aging process. We speculate that this may be (at least partly) because some amino acids are more prone to molecular damage than others, however, future experimental research needs to reveal the exact mechanisms underlying these findings.

5. Experimental Section

Databases: The Uniprot[8] and the GenAge[1] database were analyzed. UniProt is a collaboration between the European Bioinformatics Institute (EMBL-EBI), the SIB Swiss Institute of Bioinformatics, and the Protein Information Resource (PIR). GenAge database is part of the Human Ageing Genomic Resources that collected data mining tools and contributes to exploring the genetic background of human aging. Uniprot contains amino acid sequences and identifiers of all human and model organism proteins. GenAge contains the identifiers of aging-related proteins.

Identifying Aging-Related Proteins in Swiss-Prot: The reviewed version of Uniprot (i.e., Swiss-Prot) for the human containing 20401 human proteins was downloaded. The human section of GenAge (genage_human.csv) containing 307 aging-related genes coding 306 proteins (only TERC, telomerase RNA component, did not code any protein) was also downloaded. Then, a protein in Swiss-Prot as aging-related was considered if its gene name appeared in the human section of the GenAge database. All 306 aging-related proteins were identified in this way. Then, the reviewed (Swiss-Prot) sequences of UniProt in fasta format were downloaded for *Caenorhabditis elegans (C. elegans)*, *Drosophila melanogaster (D. melanogaster)*, *Mus musculus (M. musculus)*, and *Saccharomyces cerevisiae (S. cerevisiae)*. The fasta files contained 4429, 3708, 17 148, and 7913 protein sequences, respectively. For each protein, its gene name was searched in the "symbol" column of the model organism section of the GeneAge database (genage_models.csv) after filtering for the given organism. If the gene name was found, the same label was assigned as appeared in the "longevity influence" column (i.e., Pro-longevity, Anti-Longevity, Unannotated, or Unclear), otherwise, the protein was labeled as Non-longevity. 58/265/3986, 61/34/3610, 76/43/17024, and 55/382/6865 proteins were labeled as Pro-Longveity/Anti-Longevity/Non-longevity for *C. elegans*, *D. melanogaster*, *M. musculus*, and *S. cerevisiae*, respectively. The resulting labeling is available in Data S1 (Supporting Information).

Calculation of Amino Acid Frequencies: The frequency (number of occurrences) for each amino acid (i.e., character in the amino acid sequence) that appeared in each protein was counted. The relative frequency was the frequency divided by the length of the amino acid sequence. Finally, the mean (i.e., average) frequency and relative frequency for each amino acid and organism (including the human) were calculated. The resulting data are available in Data S1 (Supporting Information).

Statistical Analysis: The Python package SciPy (v1.3.1) was used for statistical analysis. Two-sided *t*-tests were calculated for comparing two groups: ns, $p > 0.05$; *, $0.01 < p \leq 0.05$; **, $0.001 < p \leq 0.01$; ***, $0.0001 < p \leq 0.001$; **** $p \leq 0.0001$. Asterisks and the term "p-values" referred to uncorrected *p*-values while the term "FDR" corrected *p* values' referred to *p*-values corrected for multiple hypothesis testing by using Benjamini/Hochberg method (see Data S1, Supporting Information).

Supporting Information

Supporting Information is available from the Wiley Online Library or from the author.

Acknowledgements

S.E.Z. and C.K. were supported by the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory, Hungary. The authors thank Zsófia Dobolyi, Ádám Sturm, Ádám Lendvai, and João Pedro Magalhães for discussion. The authors used BioRender for the figures.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

C.K. conceived and supervised the study. S.E.Z. and C.K. acquired data and performed data analysis. All authors interpreted the results. All authors wrote the manuscript.

Data Availability Statement

The data that support the findings of this study are available in the supplementary material of this article

Keywords

aging, aging-related proteins, amino acid sequence, aspartic acid, tryptophan

Received: August 21, 2023
Revised: October 14, 2023
Published online:

[1] R. Tacutu, D. Thornton, E. Johnson, A. Budovsky, D. Barardo, T. Craig, E. Diana, G. Lehmann, D. Toren, J. Wang, V. E. Fraifeld, J. P. DeMagalhães, *Nucleic Acids Res.* **2018**, *46*, D1083.
[2] C. Kenyon, J. Chang, E. Gensch, A. Rudner, R. A. C. Tabtiang, *Nature* **1993**, *366*, 461.
[3] L. Y. Sun, A. Spong, W. R. Swindell, Y. Fang, C. Hill, J. A. Huber, J. D. Boehm, R. Westbrook, R. Salvatori, A. Bartke, *eLife* **2013**, *2*, 01098.
[4] J. P. De Magalhães, *Biogerontology* **2003**, *4*, 119.
[5] C. J. Kenyon, *Nature* **2010**, *464*, 504.
[6] J. P. De Magalhães, O. Toussaint, *FEBS Lett.* **2004**, *571*, 243.
[7] C. Kerepesi, B. Daróczy, Á. Sturm, T. Vellai, A. Benczúr, *Sci. Rep.* **2018**, *8*, 4094.
[8] A. Bateman, *Nucleic Acids Res.* **2019**, *47*, D506.
[9] J. Wang, S. Mukherjee, R. A. Zubarev, *Aging* **2022**, *14*, 8882.
[10] C. A. Canfield, P. C. Bradshaw, *Transl. Med. Aging* **2019**, *3*, 70.
[11] A. T. Van Der Goot, E. A. A. Nollen, *Trends Mol. Med.* **2013**, *19*, 336.
[12] M. Leenders, L. J Van Loon, *Nutr. Rev.* **2011**, *69*, 675.
[13] P. M. Helfman, J. L. Bada, M. Y. Shou, *Gerontology* **1977**, *23*, 419.
[14] I. Lopes, G. Altab, P. Raina, J. P. De Magalhães, *Front. Genet.* **2021**, *12*, 559998.
[15] Y. H. Li, M. Q. Dong, Z. Guo, *Mech. Ageing Dev.* **2010**, *131*, 700.
[16] T. Stoeger, R. A. Grant, A. C. Mcquattie-Pimentel, K. R. Anekalla, S. S. Liu, H. Tejedor-Navarro, B. D. Singer, H. Abdala-Valencia, M. Schwake, M. P. Tetreault, H. Perlman, W. E. Balch, N. S. Chandel, K. M. Ridge, J. I. Sznajder, R. I. Morimoto, A. V. Misharin, G. R. S. Budinger, L. A. Nunes Amaral, *Nat. Aging* **2022**, *2*, 1191.