Prediction and characterization of human ageing-related proteins by using machine learning

Csaba Kerepesi1, Bálint Daróczy1, Ádám Sturm2,3, Tibor Vellai2,3 & András Benczúr1

Ageing has a huge impact on human health and economy, but its molecular basis – regulation and mechanism – is still poorly understood. By today, more than three hundred genes (almost all of them function as protein-coding genes) have been related to human ageing. Although individual ageing-related genes or some small subsets of these genes have been intensively studied, their analysis as a whole has been highly limited. To fill this gap, for each human protein we extracted 21000 protein features from various databases, and using these data as an input to state-of-the-art machine learning methods, we classified human proteins as ageing-related or non-ageing-related. We found a simple classification model based on only 36 protein features, such as the “number of ageing-related interaction partners”, “response to oxidative stress”, “damaged DNA binding”, “rhythmic process” and “extracellular region”. Predicted values of the model quantify the relevance of a given protein in the regulation or mechanisms of the human ageing process. Furthermore, we identified new candidate proteins having strong computational evidence of their important role in ageing. Some of them, like Cytochrome b-245 light chain (CY24A) and Endoribonuclease ZC3H12A (ZC12A) have no previous ageing-associated annotations.

Genetic analysis of mortality rate has clearly revealed that ageing has strong genetic components1–3. Previously identified ageing-related genes are summarized in GenAge, a high quality, manually curated database4. The human section of GenAge (version 18) consists of 305 ageing-related genes. This set of genes includes a few members that are directly linked to human ageing, as well as the best candidate genes are supported by evidence from model organisms, using cellular experiments and functional analyses (see http://genomics.senescence.info/help.html#genage).

Features that distinguish ageing-related genes from the set of remaining human genes (hereafter referred to as “non-ageing-related” genes) may help us better understand the mechanism and regulation of the human ageing process as a whole. It was shown that ageing-related proteins, compared to non-ageing-related ones, tend to have (i) more protein-protein interaction (PPI) partners, (ii) higher K-core values (K-core is a network centrality measure defined in the Methods section), (iii) more ageing-related protein-protein interaction partners, and (iv) higher co-expression coefficients with other genes6.

In the present study, we analyzed not only the co-expression and protein-protein interaction features but also thousands of other protein features. Moreover, we searched not only one-variable differences between ageing-related proteins and non-ageing-related proteins but, using machine learning, we found a multi-variable model that explains what makes a protein ageing-related.

Machine learning is a rapidly growing field of computer science, in which we construct algorithms that can learn from and make predictions on data. Machine learning has many applications for science and technology7, including genetics and genomics8. Here, we applied supervised machine learning to fit a classification model of the protein features to the set of known ageing-related and non-ageing-related proteins, in order to predict ageing-related proteins and, at the same time, to understand ageing-related properties of the proteins. A few dozen ageing studies have applied supervised machine learning methods9, some of them based on the GenAge database (as in the present study). Support-vector machine (SVM), k-nearest neighbour (KNN), and decision tree classifiers were used for predicting ageing-related genes of the nematode (Caenorhabditis elegans),

1Institute for Computer Science and Control (MTA SZTAKI), Hungarian Academy of Sciences, Budapest, Hungary.
2Department of Genetics, Eötvös Loránd University, Budapest, Hungary. 3MTA-ELTE Genetics Research Group, Eötvös Loránd University, Budapest, Hungary. Correspondence and requests for materials should be addressed to C.K. (email: kerepesi@sztaki.hu)

Received: 5 December 2017
Accepted: 19 February 2018
Published online: 06 March 2018
fruit fly (Drosophila melanogaster), and mouse (Mus musculus) genomes\textsuperscript{10-12}. Furthermore, a new feature selection method was constructed for the Bayesian network classifier and applied for predicting pro- or anti-longevity effects of genes of the most important model organisms\textsuperscript{9,13}.

For human genes, naive Bayes classifier and J48 decision tree were used to classify human DNA repair genes as ageing-related or non-ageing-related\textsuperscript{14}. To our knowledge, only one study applied supervised machine learning using the whole set of human protein-coding genes\textsuperscript{6}. Here we made several improvements on the methods of that pioneering study. For example, as databases have been extended in the last 7 years, we could use 304 ageing-related genes (from GenAge) instead of 140. We extracted not only 5 but 21000 protein features, and applied not only 280 but all of the 20183 proteins for every single training. Hence, our improved methodology has yielded new insights for ageing-related proteins.

We applied three state-of-the-art machine learning tools, XGBoost (a scalable tree boosting system\textsuperscript{15}), logistic regression (a regression analysis of binary sequences\textsuperscript{16}), and support-vector machine (a binary classifier for training data that are linearly non-separable\textsuperscript{17}), to classify human proteins as ageing-related or non-ageing-related. The models are built based on 21000 protein features extracted from different databases (UniProt\textsuperscript{18}, Gene Ontology\textsuperscript{19} and GeneFriends\textsuperscript{20}), and fit to known ageing-related human proteins (extracted from GenAge\textsuperscript{5}). The models are built from the full set of human proteins in Swiss-Prot, using the proteins included in the GenAge database as instances of the ageing-related class and all other human proteins in Swiss-Prot as the instances of the non-ageing-related class. Through this process, we uncovered the characteristic ageing-related features of human ageing-related proteins and quantified the relevance of a given protein in the regulation of the human ageing process as well as we predicted new ageing-related protein candidates.

We trained and tested our predictive methods as follows. First, we labelled every protein as an ageing-related or non-ageing-related protein on the basis of existing annotation in GenAge. Second, we selected a machine learning algorithm along with a fixed parameter setting. We then applied a 5-fold cross-validation, in which we split the data into 5 random parts and in each fold (round), used 4 parts to train the machine learning method and evaluated the prediction on the fifth one. Prediction for a protein is a real number between 0 and 1. At the end of the 5 fold-cross-validation, we have predicted values for the entire set of proteins, which ranks the proteins from weakest to strongest expected aging-relatedness. Then we compared the predicted values to the labels to assess prediction accuracy. Based on statistical accuracy measurements, we may compare the combination of algorithms and parameters to select the best performing method. The final prediction used to quantify the relevance of a given protein in the regulation of the human ageing process as well as to identify new ageing-related protein candidates. For more details of our method see the Methods section.

**Results**

A simple model to classify human proteins into ageing-related or non-ageing-related classes. One of our main results is a simple model with a high prediction performance that applies only 36 protein features (listed in Table 1). The model was built by using gradient boosted trees\textsuperscript{5,13}, for feature selection and training, as described in the Methods section. This simple model shows the most important features of the classification and provides an insight into the role of the individual protein features in the regulation of the ageing process.

The model (Table 1) contains only binary (true or false) features. For each human protein, we can compute the predicted relevance of ageing as follows: for each row of the table, we check whether the given feature is true for the protein, and then we add up the corresponding scores. The larger the final sum, the more important the protein is in the human ageing process by the model. Only the features that are listed in Table 1 can increase or decrease the ageing relevance score, hence these are the most important features in the human ageing-process by the model.

The results of Table 1 can be interpreted as follows. In general, the most important types of features are the features representing information about the number of ageing-related neighbours in the PPI network, which is consistent with earlier findings demonstrating that human ageing-related proteins tend to interact with other ageing-related proteins\textsuperscript{8}. We note that degree (number of neighbours, regardless of whether or not they are ageing-related) is not among the most important features of Table 1, because in our machine learning predictions, degree had no additional prediction power when used together with the number of ageing-related neighbours.

There are twenty-one important Gene Ontology features of the biological process (BP) category (e.g. “regulation of insulin-like growth factor receptor signaling pathway” or “response to oxidative stress”), four important Gene Ontology features of the cellular component (CC) category, “extracellular region”, “chromosomal part”, “mitochondrion” and “nucleoplasm”, and six important Gene Ontology features of the molecular function category, “damaged DNA binding”, “organic cyclic compound binding”, “enzyme binding”, “growth factor binding”, “protein binding” and “chromatin binding”. The fact that all of the molecular function features are binding type is consistent with the importance of the number of ageing-related neighbours.

Table 1 also shows that most of the features (32 of the 36) have a positive score, hence their existence in proteins indicates ageing-relatedness. Contrary, the existence for other features (4 of the 36 with negative scores: “ageing_n_0”, “ageing_n_1”, “ageing_n_2”, “ageing_n_3_4”) is an indicator of the non-ageing-related class.

**Human proteins with the highest predicted relevance in ageing.** Sorting human proteins by predicted relevance in the regulation of the ageing process can help find the most promising targets for pharmacological or other interventions to extend human healthy lifespan. Table 2 shows the 20 most relevant ageing-related proteins we obtained by performing 20 predictions for each, by applying three different methods (XGBoost, SVM and logistic regression – see the Methods section) on the final feature set that was selected by XGBoost and sorted by the average of the predicted scores. The process is described in detail in the Methods section. Supplementary Table S1 displays a more detailed list of the predicted ageing relevance of all human proteins.
17 out of the 20 proteins in Table 2 have a record in the GenAge database with a detailed evidence of why it is selected in the database as an ageing-related member. For example, there are experimental evidence for the ageing-association of the homologues of human “forkhead box protein O1” (FOXO1) in worms, fruit flies, and mice. Another example is the serine/threonine protein kinase (MTOR_HUMAN), the role of which in the ageing process was demonstrated in each of the main ageing models (C. elegans, Drosophila, yeast, and mouse).

| Feature ID | Description of the Feature | Category | Score | Relative Frequency in ageing/non-ageing |
|------------|----------------------------|----------|-------|----------------------------------------|
| ageing_n_0 | number of ageing-related neighbours = 0 | Net | -2.896 | 38.8/92.1 |
| ageing_n_1 | number of ageing-related neighbours = 1 | Net | -2.275 | 15.8/5.6 |
| ageing_n_2 | number of ageing-related neighbours = 2 | Net | -1.168 | 15.1/1.4 |
| ageing_n_3_4 | number of ageing-related neighbours = 3,4 | Net | -0.744 | 12.8/0.6 |
| GO:0043567 | regulation of insulin-like growth factor receptor signaling pathway | BP | 1.327 | 2.6/0.1 |
| GO:0006979 | response to oxidative stress | BP | 0.9 | 21.7/1.4 |
| GO:0003684 | damaged DNA binding | MF | 0.837 | 8.6/0.2 |
| GO:0009987 | cellular process | BP | 0.805 | 99.3/70.0 |
| GO:0005576 | extracellular region | CC | 0.636 | 21.7/8.8 |
| GO:0065008 | regulation of biological quality | BP | 0.563 | 60.2/14.9 |
| GO:0051276 | chromosome organization | BP | 0.515 | 14.5/1.6 |
| GO:0032502 | developmental process | BP | 0.497 | 69.4/22.5 |
| GO:0043066 | negative regulation of apoptotic process | BP | 0.474 | 32.9/3.5 |
| GO:0009628 | response to abiotic stimulus | BP | 0.441 | 38.2/4.4 |
| GO:0007169 | transmembrane receptor protein tyrosine kinase signaling pathway | BP | 0.413 | 19.1/2.1 |
| GO:0010332 | response to gamma radiation | BP | 0.411 | 8.6/0.1 |
| GO:0019838 | growth factor binding | MF | 0.405 | 53.0/4.4 |
| GO:0044710 | single-organism metabolic process | BP | 0.388 | 42.1/15.4 |
| GO:0031325 | positive regulation of cellular metabolic proc | BP | 0.331 | 64.8/12.8 |
| GO:0050896 | response to stimulus | BP | 0.288 | 77.3/22.8 |
| GO:0031667 | response to nutrient levels | BP | 0.285 | 16.8/1.5 |
| GO:0005515 | protein binding | MF | 0.271 | 75.7/24.4 |
| GO:2000377 | regulation of reactive oxygen species metabolic process | BP | 0.259 | 13.8/0.6 |
| GO:0051716 | cellular response to stimulus | BP | 0.257 | 62.2/11.1 |
| GO:0005654 | nucleoplasm | CC | 0.235 | 49.7/14.1 |
| GO:0080135 | regulation of cellular response to stress | BP | 0.225 | 27.3/2.6 |
| GO:0048511 | rhythmic process | BP | 0.224 | 15.1/1.2 |
| GO:0044427 | chromosomal part | CC | 0.197 | 24.0/3.4 |
| ageing_n_5+ | number of ageing-related neighbours ≥ 5 | Net | 0.192 | 17.4/0.2 |
| GO:003682 | chromatin binding | MF | 0.171 | 17.1/2.1 |
| GO:0006974 | cellular response to DNA damage stimulus | BP | 0.167 | 27.6/3.1 |
| GO:0097159 | organic cyclic compound binding | MF | 0.166 | 62.8/28.8 |
| GO:0005739 | mitochondrial | CC | 0.16 | 20.4/6.1 |
| GO:0019899 | enzyme binding | MF | 0.128 | 39.8/6.8 |
| GO:0009894 | regulation of catabolic process | BP | 0.125 | 25.7/3.4 |

Table 1. A simple model, produced by tree boosting (XGBoost), to classify human proteins as ageing-related or non-ageing-related. Features are listed by ID and description. Feature category can take values "Net" (Network), "MF" (Molecular Function), "CC" (Cellular Component), or "BP" (Biological Process). The table consists of only binary (true or false) features. For each protein we can compute the predicted relevance of ageing as follows: for each row of the table, we check whether the given feature is true for the protein and then we add up the corresponding scores. The larger the final sum, the more important role of a protein is predicted in the human ageing process. For example, suppose that a protein has 3 ageing-related neighbours and their UniProt record contains only two GO terms, "response to oxidative stress", and "regulation of growth". Then the predicted ageing relevance of that protein is $\sum_{i=1}^{3} \text{score}_i = -0.744 + 0.9 + 0.398 = 0.554$. Predicted scores produced by the above summation method are presented in the "Table1_pred" column of Supplementary Table S1. Scores obtained by summation are not necessarily bounded by 1. The actual output of XGBoost, which we used in the rest of the paper, was normalized to take values in [0…1]. In fact, we use the average of normalized predicted values made by several models (see the Methods). The relative frequency of features in the ageing-related and the non-ageing-related sets of proteins, a value independent of our particular model, is displayed in the last column.
first shown by one of the authors of this paper. Finally, we note that "Werner syndrome ATP-dependent helicase" (WRN_HUMAN) is one of the strongest candidates for proteins influencing human ageing with direct evidence as mutation of WRN gene leads to Werner syndrome, which is characterized by premature ageing (progeria)39.

Whether or not a gene is annotated with the GO term "aging" (GO:0007568) is also displayed in Table 2; however, this term and its descendant terms are not used for modelling, we just display it as extra information. Interestingly, some proteins with a relatively high predicted score are not assigned to the GO term "aging", showing the difference between the set of ageing-related proteins of GenAge and the set of proteins annotated with GO term "aging".

New candidates of ageing-related human proteins predicted by machine learning. Models we found here predict new candidates of ageing-related proteins that were previously not annotated as ageing-related in GenAge database. The 20 most promising new ageing-related candidates are listed in Table 3, and sorted by their average predicted values. The list was obtained from Supplementary Table S1 by selecting the 20 highest scored (average predicted value) proteins with no GenAge annotation. They can be considered as proteins having strong computational evidence of their regulator role in the human ageing process. Proteins highlighted in the following part of this section can be good candidates to expand GenAge database with them in the future.

Table 2. Human proteins with the highest predicted relevances in ageing. The 20 highest scored proteins considered the entire set of human proteins (regardless of whether or not the protein is included in the GenAge database), sorted by decreasing predicted relevance in ageing (average predicted value). Each row consists of an ID of the given protein ("Uniprot ID"), a description ("recommended name in UniProt"), the number of ageing-related protein neighbours of the given protein in the protein-protein interaction network ("ageing neighbours"), a statement about its assignment to the GO term "aging" ("aging GO"), a statement about its inclusion in GenAge ("GenAge"), and the average predicted value of 20 predictions of three machine learning methods each (XGBoost, SVM and LR) by using the final feature set selected by XGBoost ("average predicted value"). Average predicted values close to one indicate very strong predicted relevance for the human ageing process. Supplementary Table S1 is a more detailed list with all of the human proteins.
Table 3. New candidates of ageing-related human proteins predicted by machine learning. The 20 highest scored proteins with no ageing-related GenAge annotation, sorted by decreasing predicted relevance in ageing (average predicted value). The columns have the same meanings as in Table 2.

| Uniprot ID | recommended name | ageing neighbours | ageing GO | GenAge | average predicted value |
|-----------|------------------|------------------|----------|--------|-------------------------|
| SIR2_HUMAN | NAD-dependent protein deacetylase sirtuin-2 | 2 | no | no | 0.909 |
| BECN1_HUMAN | Beclin-1 | 3 | yes | no | 0.827 |
| SYUA_HUMAN | Alpha-synuclein | 2 | yes | no | 0.801 |
| CAV1_HUMAN | Caveolin-1 | 4 | no | no | 0.745 |
| LRRK2_HUMAN | Leucine-rich repeat serine/threonine-protein kinase 2 | 6 | no | no | 0.734 |
| BAD_HUMAN | Bcl2-associated agonist of cell death | 3 | no | no | 0.721 |
| PARK7_HUMAN | Protein DJ-1 | 2 | no | no | 0.711 |
| HS90B_HUMAN | Heat shock protein HSP 90-beta | 8 | no | no | 0.709 |
| SMAD3_HUMAN | Mothers against decapentaplegic homolog 3 | 2 | no | no | 0.662 |
| KDM1A_HUMAN | Lysine-specific histone demethylase 1A | 2 | no | no | 0.66 |
| ERBB4_HUMAN | Receptor tyrosine-protein kinase erbB-4 | 3 | no | no | 0.633 |
| HDAC6_HUMAN | Histone deacetylase 6 | 2 | no | no | 0.606 |
| FACL2_HUMAN | Fanconi anemia group D2 protein | 2 | no | no | 0.585 |
| RARA_HUMAN | Retinoic acid receptor alpha | 5 | no | no | 0.567 |
| XRCC1_HUMAN | DNA repair protein XRCC1 | 4 | no | no | 0.567 |
| CY24A_HUMAN | Cytochrome b-245 light chain | 0 | no | no | 0.562 |
| SRC_HUMAN | Proto-oncogene tyrosine-protein kinase Src | 10 | no | no | 0.562 |
| CBL_HUMAN | E3 ubiquitin-protein ligase CBL | 5 | no | no | 0.561 |
| XBPI_HUMAN | X-box-binding protein 1 | 0 | no | no | 0.551 |
| FYN_HUMAN | Tyrosine-protein kinase Pyn | 3 | no | no | 0.543 |

accelerates the ageing process in mice\textsuperscript{38}, it can be assumed that Caveolin-1 may have a cell protective, anti-ageing function.

LRRK2\textsubscript{HUMAN} is a member of the leucine-rich repeat kinase family. Mutations in \textit{LRKK2} gene are implicated in the development of Parkinson's disease\textsuperscript{39}. While loss-of-function mutations in \textit{LRRK2} cause age-dependent neurodegeneration in \textit{Drosophila}\textsuperscript{40}, gain-of-function mutations in the gene confer resistance to age-related motor decline in mice, possibly via enhancement of LRRK2 kinase activity\textsuperscript{41}. So, it can be assumed that LRRK2 may also have a potential neuroprotective, anti-ageing function.

Histone deacetylases (HDACs) are primarily involved in the deacetylation of histones but some HDACs, such as HDAC6\textsubscript{HUMAN}, can also affect the function of cytoplasmic non-histone proteins. HDAC6 overexpression correlates with tumorigenesis, and improves the survival of cancer cells, which presupposes a cell protective function\textsuperscript{42}. Indeed, the reduced expression of HDAC6 contributes to a decline in stem cell numbers\textsuperscript{43} and brain function\textsuperscript{44} during ageing. Furthermore, HDAC6 overexpression in transgenic mice increases the reproductive lifespan of animals\textsuperscript{45}.

Additionally, we found a few proteins that have high predicted relevance in ageing but have no ageing-related annotation in GenAge, nor in the whole literature. Such proteins are Cytochrome b-245 light chain (\textit{CY24A}\textsubscript{HUMAN}) and Endoribonuclease ZC3H12A (\textit{ZC12A}\textsubscript{HUMAN}). \textit{CY24A}\textsubscript{HUMAN} is the 64th most relevant protein in ageing by our predictions (Table 3, Supplementary Table S1), and \textit{ZC12A}\textsubscript{HUMAN} is the 78th most relevant protein in ageing by our predictions (Supplementary Table S1). Neither of these proteins have ageing-related neighbours but both have 16 GO features of the 31 GO features of Table 1. The 16 ageing-related predictor features for each of these two proteins are listed in Supplementary Table S2.

Figure 1 shows how the new candidates interact with each other and with human ageing-related proteins of GenAge. To evaluate the final prediction, we plotted the receiver operating characteristic curve (ROC, Fig. 2a). The performance of the model was 0.9322, a result we obtained by measuring the area under the curve of the receiver operating characteristic curve (ROC AUC). It is shown that ROC AUC (shortly: AUC) is the probability that a randomly chosen positive example is predicted with a higher score than a randomly chosen negative example\textsuperscript{46}, hence AUC is independent of the class imbalance.

To compare our prediction to the ageing-related proteins of GenAge, and the set of proteins annotated with the GO term "ageing", we chose a threshold (0.24) for the predicted relevance in ageing ("avg pred" in Supplementary Table S1) (Fig. 2c); a protein is predicted as ageing-related by the models if its predicted relevance in ageing is at least 0.24. We selected this threshold because at this point, there is a relatively high \textit{true positive rate} (0.4638) and, at the same time, a relatively low \textit{false positive rate} (0.0081) and maximal \textit{F1 score} (0.46548) and maximal MCC (0.45641) are reached at this point (\textit{TP} = 162, \textit{FP} = 141, \textit{FN} = 163, \textit{TN} = 19717, \textit{precision} = 0.46535, \textit{recall} = 0.46382, \textit{accuracy} = 0.98390). Evaluation measures for more threshold values are available in Supplementary Table S3, and displayed in Fig. 2b. For definitions of the evaluation measures see the Methods section.
Discussion

In this study, we ordered the human proteins on the basis how (to which extent) machine learning algorithms, which automatically build a classifier by learning from a set of labelled data, predict their importance in the regulation or mechanism of the ageing process. The results we obtained have at least two important relevancies. First, they may help identify the ageing-related proteins that have a particularly prominent role in the human ageing process (quantifying the importance of ageing-related proteins in the process). Second, the results may help uncover novel proteins with an ageing function (the role of these proteins in ageing has not been recognized previously). Furthermore, we created a simple, biologically easily interpretable model, based on only 36 protein features that may help to understand better the human ageing process.

Ageing is driven by the progressive accumulation of unrepaired cellular damage\textsuperscript{4,47}. Such damages mainly include oxidized, aggregated and misfolded proteins that are generated by mutations, environmental factors (e.g. heat stress) and metabolic agents (e.g. reactive oxygen species produced by mitochondrial respiration), and act...
as cellular toxins often causing the loss of the affected cells. At advanced ages, massive levels of cell death can lead to the development of an age-associated degenerative disease (tissue dysfunction), and eventually organismal death. Prior to this life period, cellular damages are effectively degraded (i.e. eliminated) by the repair and maintenance processes and mechanisms including autophagy (cellular self-eating) being the most significant form of breaking down cytoplasmic materials, the ubiquitin-proteasome system and molecular chaperons, also called heat-shock proteins, as well as the DNA repair pathways. These processes and mechanisms, however, display a gradual decline in their capacity as the organism ages. In the present study, BCL2 (antiapoptotic B cell lymphoma protein), FOXO1 (Fork head box O transcription factor) and ERCC1 (DNA excision repair protein) were identified as proteins with the highest predicted relevance in human ageing (Table 2). Indeed, BCL2 protects cells from undergoing apoptosis (programmed cell death), and, in both nematodes and human cells, also interacts with the autophagic process through binding the core autophagy protein BECN1 (Beclin – Bcl2-interacting). Genes, however, that modulate (primarily limit) lifespan independently of the ageing process can be identified by machine learning. Boosted trees may be widely used in further analysis of this field.

Supervised machine learning methods are especially effective when they are used on a large set of examples. Earlier machine learning studies on human proteins applied only a few hundred features of a few hundred proteins for each training. By using extensive computational power, here we analyzed all the human proteins, and performed feature selection from 21000 protein features. In 2016, a novel machine learning system was developed, XGBoost, which allows an effective feature selection even in case of a huge number of correlating features. XGBoost is applied widely by data scientists for example at data mining challenges. However, according to our knowledge, we are the first who apply it for ageing research. Boosted trees may be widely used in further analysis of this field.

We used the GenAge database to assign the human proteins into “ageing-related” or “non-ageing-related” classes in the following way: the 304 proteins of GenAge served as “ageing-related” instances and the remaining 19879 human proteins served as “non-ageing-related” instances. These classes then served as labels for training the classifiers. We applied GenAge because it focuses on the ageing process when selecting genes (see http://genomics.senescence.info/help.html#genage). Genes, however, that modulate (primarily limit) lifespan independently of the ageing process are omitted from this database. Such genes are involved in human pathologies or their activity is altered in case of extreme longevity. In addition, several other related supervised machine learning studies also rely on GenAge.

One may ask why the “aging” GO annotation was not used in the process of labelling the proteins for training the classifiers. We used only GenAge for labelling for several reasons. First, GenAge has a more detailed explanation and references than the “aging” GO annotations. Second, we could find no study related to machine learning based on “aging” GO terms. Third, it seems that the “aging” GO assignment process does not focus on the regulation of the ageing process. For example, “aging” GO assignments of the proteins KRA43, KRA45, KRA47, KRA48, KRA49, K1C14, K1C16, KRT83 and KT33B are based on the single evidence that keratin and keratin-associated proteins in white hair are upregulated in comparison with black hair in microarray experiments. However, using both GenAge and “aging” GO annotations would give a wider perspective of ageing. So, we performed a supplementary analysis based on a labelling where a given protein was assigned to the ageing-related class if it is included in GenAge or annotated with the “aging GO” term or its descendants. The results, methods and discussion sections of the supplementary analysis can be found in Supplementary Information, Supplementary Tables S4–S6.
It is important to emphasize that the vast majority of human ageing-related proteins, including those listed in GenAge, have not been validated experimentally for a regulator role in human ageing. Relevant results have been obtained mostly from genetic model systems and assumed that they operate in an evolutionarily conserved way. As an example, defects in the transmembrane receptor for insulin/IGF-1 signalling have been shown to double lifespan in nematodes (C. elegans) but there is no evidence for a gene/protein that can extend human lifespan in such an extreme manner. Some degree of ageing regulator evidence exists only for a few human proteins. WRN, for example, which encodes a RecQ helicase involved in DNA repair, when is mutated, leads to Werner syndrome, the pleiotropic phenotype of which is characterized by extreme progeria. Prominent or novel ageing proteins we identified in this work may become promising drug targets for further efforts in order to extend healthy lifespan in humans, which is a central focus in current pharmacological research.

Despite its medical and social significance, our present knowledge on the biological basis of the (human) ageing process is rather limited. As Cynthia Kenyon wrote in one of her review articles on ageing, genetic factors that primarily cause ageing (i.e. the progressive, lifelong accumulation of cellular damage) remain unexplored. Recent theoretical considerations have tried to identify a novel class and high copy number of genes, mobile genetic elements, as primary genetic determinants of ageing, but a relevant direct experimental evidence is still missing to support this assumption. In the light of these facts and as databases are being improved considerably, our present ageing-related ordering (Supplementary Table S1) may be modified in the future.

Here we ignored an ageing-related gene, telomerase reverse transcriptase (TERT), because it does not code for a protein. An interesting future direction would be to predict not only ageing-related proteins but ageing-related non-coding RNAs. Such a work could be based on results of the computational prediction and characterization of disease-associated human microRNAs, and long non-coding RNAs.

Conclusion

Although single ageing-related proteins have been intensively studied, their analysis as a whole has been largely limited. To fill this gap, in the present work, we applied three state-of-the-art machine learning tools to classify human proteins as ageing-related or non-ageing-related. The classification models are built on all human proteins and 21000 protein features, and fit to known ageing-related human proteins of the GenAge database. The models were built from the full set of human proteins in Swiss-Prot, using the proteins included in the GenAge database as instances of the ageing-related class and all other human proteins in Swiss-Prot as the instances of the non-ageing-related class. The final prediction was used to quantify the relevance of a given protein in the regulation of the human ageing process as well as to identify new ageing-related protein candidates.

Methods

We start this section by describing the source of known ageing-related proteins. We continue by describing the Gene Ontology features, the protein-protein interaction (PPI) network features and the co-expression feature. Then we detail how gradient boosted trees were applied for selecting the most relevant features. The main steps are shown in Fig. 3. We close this section by describing the best performing machine learning methods.

Ageing-related data (labels of the classification).

All the 20183 human Swiss-Prot (manually annotated and reviewed UniProt) entries were downloaded from the UniProtKB protein database on April 6, 2017. In the human section of GenAge database (Build 18), we found 305 candidates of human ageing-related genes. With the exception of a single gene TERT, all of these genes are included in Swiss-Prot. Hence, the target variable (labels) of the classification has value "1" for the 304 proteins of GenAge ("ageing-related class") and value "0" for the remaining 19879 human proteins ("non-ageing-related class").

Gene Ontology features. We compute Gene Ontology (GO) features in a similar way as Freitas et al., but by also using the GO categories "cellular component" and "molecular function". For each human Swiss-Prot protein entry we extract the associated GO terms, all of which are binary, being either yes or no. The difficulty of this task is that the GO assignments of the UniProt entries are not complete: an entry is associated with a given GO term but not all the ancestors containing the given GO term. For example, the ANKE1_HUMAN protein entry has only the GO term "calcium ion binding" but does not have the ancestor GO terms "metal ion binding", "cation binding", "ion binding", "binding", "molecular function". To handle this problem, we downloaded the basic version of the Gene Ontology database (with the database filename "go-basic.obo") and by walking upward in the GO hierarchy, we added all of the ancestor GO terms to the corresponding proteins. Note that "go-basic.obo" is guaranteed to be acyclic, and annotations can be propagated up the (directed) graph. The final feature table contains 20183 proteins and 21019 features. Although the notion of ageing-relatedness of the GenAge database is far from being identical to that of the Gene Ontology database (see Fig. 2), we removed the GO terms, along with their descendants, that contain "aging", "senescence" or "age-related" as substring (these terms are also used by Chautard et al.).

PPI network features. Protein-protein interactions (PPIs) are included in the Swiss-Prot database. In our PPI network of 20183 nodes and 18784 edges, we only kept bidirectional and non-self interactions.

For each protein, we computed the following features based on the constructed network and the ageing-related data. In terms of interaction count statistics, we computed the number of neighbours, the number of ageing-related neighbours, and the ratio of the two. We also computed the K-core value of a node by using the “coreness” function of the R package igraph. A K-core of a graph is a maximal subgraph in which each vertex has degree at least K. The K-core or coreness value of a node is the maximal value of K such that the node is in a K-core. We extracted further network features by Cytoscape, including "Average Shortest Path
Length”, “Betweenness Centrality”, “Closeness Centrality”, “Clustering Coefficient”, “Eccentricity”, “Neighborhood Connectivity”, “Radiality”, “Stress” and “Topological Coefficient”.

**Co-expression feature.** For each human protein-coding gene, we computed its gene co-expression with the set of ageing-related genes using the GeneFriends database. Co-expression is the number of human ageing-related genes of GenAge that increase or decrease in expression simultaneously in the RNAseq datasets processed by GeneFriends.

**Feature selection with XGBoost.** Gradient boosted tree algorithms are capable of selecting the most important uncorrelated features by building small decision trees of a few of the most important features and gradually refining the small models by adding new trees. We used the XGBoost implementation for feature selection with the parameters shown in Table 4. We evaluated the generated models by 5-fold cross-validation and measured the area under the curve of the receiver operating characteristic curve (ROC AUC). For every feature set, we repeated this process 20 times. The average and standard deviation of the 20 predictions are shown in Table 4. In the first steps of the feature selection process we selected the most important Gene Ontology features except the GO terms related to ageing. Original Gene Ontology (GO) terms with the ageing-related terms produced an AUC of 0.8787 and 16820 features. Original Gene Ontology (GO) terms without the ageing-related terms produced an AUC of 0.8729 and 16800 features. The explanation for this surprisingly low increasing is the large difference between the set of ageing-related proteins of GenAge and the set of proteins annotated with GO term “aging” (as Fig. 2c showed). GO ancestor calculation has a considerable added value, reaching an AUC of 0.9086 and 21000 features.

We used feature selection started from this set of 21000 GO features in two passes. First, we used XGBoost for selecting the GO features by computing the importance of features and selecting those with value greater than 0. We reached an AUC of 0.9187 (improvement by 0.0101) with only 373 GO features left from the initial 21000. By the second filter, XGBoost selected the GO features that have feature importance values greater than 0.004. We reached an AUC of 0.9219 with only 65 GO features left from the initial 373.
Given the 65 GO features selected in two passes by XGBoost, we continued feature selection by adding network and co-expression features. All these features produced an AUC of 0.9294, showing a considerable increase. However, we found that the filtered GO features with the addition of a single feature, the number of ageing-related neighbours ("ageing_n") produced a slight increase in AUC (0.9314). Since simpler models usually generalize better, we kept 66 features with the 65 GO features and the number of ageing-related neighbours.

In the last step of feature selection we applied a third filter, where XGBoost (with 50 trees and maximal depth 1) selected features with importance greater than 0. At this point, we reduced the XGBoost parameter depth of tree to achieve a simple, well interpretable model (at the same time we needed to increase the number of trees to reach the same performance).

Only 32 features left from the initial 66, and we reached a final AUC of 0.9322. This final feature set was used for the predictions in the results section and it is shared at https://github.com/kerepesi/aging_ml along with codes to reproduce the results.

### Predictions via SVM and LR on the feature set selected by XGBoost.

Besides XGBoost, we performed 20 predictions of 5 fold cross-validations (5 fold CV is repeated 20 times) with support vector machine (SVM)

\[ L^2 \text{ normalized feature space produced an average AUC of 0.9321 (std dev 0.0015). Average predicted values of each method are presented in Supplementary Table S1.} \]

### Performance of various machine learning algorithms.

We compared performance of XGBoost (learning rate = 0.3, depth of trees = 6, number of trees = 20) with various machine learning algorithms (with the default settings of scikit-learn, version 0.19.0⁶⁹): k-nearest neighbour, decision tree, naïve Bayes, logistic regression, and support-vector machine with linear kernel function. Most of them appeared in related studies. We applied the algorithms on the whole set of features without selection (GO, UniNet, CoExp), as well as, on a feature set containing only the GO features that occur in at least 100 proteins (idea of occurrence threshold is inspired by Freitas et al.¹⁵).

For each algorithm and feature set the average and standard deviation of AUC values generated by predictions of 5-fold cross-validation are presented in Table 5. XGBoost outperformed the remaining methods.

### Evaluation measures for binary classification.

TP (true positive) is the number of positives that are predicted as positives. TN (true negative) is the number of negatives that are predicted as negatives. FP (false positive) is the number of negatives that are predicted as positives. FN (false negative) is the number of positives that are predicted as negatives. In our context "positive" means "ageing-related", "negative" means "non-aging-related".

\[ \text{Precision} \overset{\text{def}}{=} \frac{TP}{TP + FP}, \quad \text{accuracy} \overset{\text{def}}{=} \frac{TP + TN}{TP + TN + FP + FN}, \quad \text{recall} \overset{\text{def}}{=} \frac{TP}{TP + FN}, \quad \text{fall-out} \overset{\text{def}}{=} \frac{FP}{TN + FP} \]

#### Table 4. Feature selection process driven by performance of XGBoost on different feature sets. Performance of various machine learning algorithms.

| short description of the feature set | number of features | depth of trees | number of trees | number of predictions | AUC average | std dev |
|-------------------------------------|-------------------|---------------|----------------|-----------------------|-------------|--------|
| GO w/o ancestors, with ageing GOs  | 16820             | 6             | 20             | 20                    | 0.8787      | 0.0061 |
| GO w/o ancestors                   | 16800             | 6             | 20             | 20                    | 0.8729      | 0.0050 |
| GO                                  | 21000             | 6             | 20             | 20                    | 0.9086      | 0.0049 |
| GO XGBoost one pass filter          | 373               | 6             | 20             | 20                    | 0.9187      | 0.0042 |
| GO XGBoost two pass filter          | 65                | 6             | 20             | 20                    | 0.9219      | 0.0033 |
| GO XGBoost two pass filter UniNet, CoExp | 79          | 6             | 20             | 20                    | 0.9294      | 0.0034 |
| GO XGBoost two pass filter UniNet   | 78                | 6             | 20             | 20                    | 0.9293      | 0.0036 |
| GO XGBoost two pass filter, degree  | 66                | 6             | 20             | 20                    | 0.9283      | 0.0027 |
| GO XGBoost two pass filter, ageing_n | 66            | 6             | 20             | 20                    | 0.9314      | 0.0029 |
| GO XGBoost three pass filter, ageing_n | 32             | 1             | 50             | 20                    | 0.9322      | 0.0011 |
We note that—in a binary classification task—there are at least one positive sample (i.e. TP + FN ≥ 1) and at least one negative sample (i.e. TN + FP ≥ 1), hence the denominator of the formula of recall, fall-out and accuracy can never be equal to zero.

ROC curve (Receiver Operating Characteristic Curve) is defined by the point pairs of true positive rates and false positive rates at different threshold settings. ROC AUC (shortly AUC) is calculated as the area under the ROC curve.

### Data and code availability.

Tables and codes of the final results are available at [https://github.com/kerep-esi/aging_ml](https://github.com/kerep-esi/aging_ml). Other intermediate data and codes of this study are available from the corresponding author upon reasonable request.

### References

1. de Magalhães, J. P. Is mammalian aging genetically controlled? *Biogerontology* **4**, 119–120 (2003).
2. de Magalhães, J. P., Cabral, J. A. & Magalhães, D. The influence of genes on the aging process of mice. *Genetics* **169**, 265–274 (2005).
3. Kenyon, C. J. The genetics of aging. *Nature* **464**, 504–512 (2010).
4. Vellai, T., Takács-Vellai, K., Sass, M. & Klionsky, D. J. The regulation of aging: does autophagy underlie longevity? *Trends in cell biology* **19**, 487–494 (2009).
5. Tacutu, R. et al. Human Ageing Genomic Resources: Integrated databases and tools for the biology and genetics of ageing. *Nucleic acids research* **gks1155** (2012).
6. Li, Y.-H., Zhang, G.-G. & Guo, Z. Computational Prediction of Aging Genes in Human. In *Biomedical Engineering and Computer Science (ICBEC), 2010 International Conference on*, 1–4 (IEEE 2010).
7. Jordan, M. I. & Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science* **349**, 255–260 (2015).
8. Libbrecht, M. W. & Noble, W. S. Machine learning in genetics and genomics. *Nature Reviews. Genetics* **16**, 321 (2015).
9. Fabris, E., de Magalhães, J. P. & Freitas, A. A. A review of supervised machine learning applied to ageing research. *Biogerontology* **1**, 1–8 (2017).
10. Li, Y.-H., Dong, M.-Q. & Guo, Z. Systematic analysis and prediction of longevity genes in *Caenorhabditis elegans*. *Mechanisms of ageing and development* **131**, 700–709 (2010).
11. Song, X., Zhou, Y.-C., Feng, K., Li, Y.-H. & Li, J.-h. Discovering aging-genes by topological features in *Drosophila melanogaster* protein–protein interaction network. In *Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on*, 94–98 (IEEE 2012).
12. Fang, K. et al. Topological analysis and prediction of aging genes in *Mus musculus*. In *Systems and Informatics (ICSAI), 2012 International Conference on*, 2268–2271 (IEEE 2012).
13. Wan, C., Freitas, A. A. & De Magalhães, J. P. Predicting the pro-longevity or anti-longevity effect of model organism genes with new hierarchical feature selection methods. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)* **12**, 262–275 (2015).
36. Zou, H., Stoppani, E., Volonte, D. & Galbiati, F. Caveolin-1, cellular senescence and age-related diseases. *Nature* **426**, 620–620 (2003).

35. Vellai, T. et al. Genetics: influence of TOR kinase on lifespan in *C. elegans*. *Nature* **426**, 885–890 (2004).

34. Burnett, C. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* **215**–242 (1958).

33. Chuck, T. & de Magalhães, J. P. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic acids research* **43**, D1124–D1132 (2014).

32. Tissenbaum, H. A. & Guarente, L. Increased dosage of a sir-2 gene extends lifespan in *Caenorhabditis elegans*. *Science* **305**, 60–63 (2004).

31. Ogg, S. et al. The fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in *C. elegans*. *Nature* **389**, 994–999 (1997).

30. Meléndez, A. J. & Garcia, A. M. Hierarchical deterioration of body systems in Werner’s syndrome: implications for normal ageing. *Mechanisms of ageing and development* **98**, 239–254 (1997).

29. Goto, M. Autophagy genes are essential for dauer development and life-span extension in *C. elegans*. *Science* **301**, 1387–1391 (2003).

28. Wu, J. J. et al. Loss of LRRK2/PARK8 induces degeneration of dopaminergic neurons in Drosophila. *Nature* **479**, 913–918 (2011).

27. Ogg, S. et al. The fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in *C. elegans*. *Nature* **389**, 994–999 (1997).

26. Chen, X. et al. The Piwi-piRNA pathway: road to immortality. *Cellular and molecular life sciences* **62**, 215–242 (2005).

25. Vellai, T. & Takács-Vellai, K. Regulation of protein turnover by longevity pathways. In *Protein Metabolism and Homeostasis in Aging*. Springer (2010).

24. Kamei, Y. & Ogg, S. The NAD+/sirtuin pathway modulates longevity through activation of mitochondrial UPR and FOXO signaling. *Cell* **154**, 430–441 (2013).

23. Giannakou, M. E. & Ogg, S. Gene Ontology Consortium: going forward. *Nature genetics* **43**, 913–920 (2011).

22. Ogg, S. et al. Increased mammalian lifespan and a segmental and tissue-specific slowing of aging after genetic reduction of mTOR expression. *Cell reports* **4**, 913–920 (2013).

21. Vellai, T. & Takács-Vellai, K. Regulation of protein turnover by longevity pathways. In *Protein Metabolism and Homeostasis in Aging*. Springer (2010).

20. van Dam, S., Craig, T. & de Magalhães, J. P. GeneFriends: a human RNA-seq-based gene and transcript co-expression database. *Nucleic acids research* **43**, D1124–D1132 (2014).

19. Consortium, G. O. et al. Hair greying is associated with active hair growth. *Nature* **436**, 994–999 (1997).

18. Consortium, U. et al. Hair greying is associated with active hair growth. *Nature* **436**, 994–999 (1997).

17. Bernard, B. & de Magalhães, J. P. Machine learning and development. *Nature* **436**, 994–999 (1997).

16. Cox, D. R. The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological)* **215**–242 (1958).

15. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794 (ACM, 2016).

14. Freitas, A. A., Vasieva, O. & de Magalhães, J. P. A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related. *BMC genomics* **12**, 1 (2011).

13. Chen, X. miREFRWR: a novel disease-related microRNA-environmental factor interactions prediction method. *Briefings in bioinformatics* **bbw060** (2016).

12. Apweiler, R. et al. UniProt: the universal protein knowledgebase. *Nucleic acids research* **32**, D115–D119 (2004).

11. Sturm, Á., Ivics, Z. & Vellai, T. The mechanism of ageing: primary role of transposable elements in genome disintegration. *Cellular and molecular life sciences* **62**, 215–242 (2005).

10. Hanley, J. A. & McNeil, B. J. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Annals of statistics* **13**, 2617–2624 (2004).

9. Barna, J. et al. Heat shock factor-1 interwines insulin/IGF-1, TGF-β/cGMP signaling to control development and aging. *BioSystems* **13**, 2617–2624 (2004).

8. Choi, H. I. et al. Hair greying is associated with active hair growth. *Nature* **436**, 994–999 (1997).

7. Zhang, X. et al. Hair greying is associated with active hair growth. *Nature* **436**, 994–999 (1997).

6. Lee, S. B., Kim, W. & Chung, J. Loss of LRRK2/PARK8 induces degeneration of dopaminergic neurons in Drosophila. *Biochemical and biophysical research communications* **358**, 534–539 (2007).

5. Vellai, T. et al. Heat shock factor-1 interwines insulin/IGF-1, TGF-β/cGMP signaling to control development and aging. *BMC developmental biology* **12**, 32 (2012).

4. Wu, J. J. et al. Increased mammalian lifespan and a segmental and tissue-specific slowing of aging after genetic reduction of mTOR expression. *Cell reports* **4**, 913–920 (2013).

3. Goto, M. Hierarchical deterioration of body systems in Werner's syndrome: implications for normal ageing. *Mechanisms of ageing and development* **98**, 239–254 (1997).

2. Ogg, S. et al. The fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in *C. elegans*. *Nature* **389**, 994–999 (1997).

1. Lee, S. B., Kim, W., Lee, S. & Chung, J. Loss of LRRK2/PARK8 induces degeneration of dopaminergic neurons in Drosophila. *Biochemical and biophysical research communications* **358**, 534–539 (2007).

0. Barna, J. et al. Heat shock factor-1 interwines insulin/IGF-1, TGF-β/cGMP signaling to control development and aging. *BMC developmental biology* **12**, 32 (2012).
63. Chautard, E., Thierry-Mieg, N. & Ricard-Blum, S. Interaction networks as a tool to investigate the mechanisms of aging. *Biogerontology* **11**, 463–473 (2010).
64. Dorogovtsev, S. N., Goltsev, A. V. & Mendes, J. F. F. K-core organization of complex networks. *Physical review letters* **96**, 040601 (2006).
65. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695**, 1–9 (2006).
66. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498–2504 (2003).
67. Assenov, Y., Ramirez, F., Schelhorn, S.-E., Lengauer, T. & Albrecht, M. Computing topological parameters of biological networks. *Bioinformatics* **24**, 282–284 (2007).
68. Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

**Acknowledgements**

C.K. and A.B. were supported by the Momentum Grant of the Hungarian Academy of Sciences (LP2012-19/2012). This work was supported by the grants OTKA (Hungarian Scientific Research Fund) NK78012, MEDinPROT Protein Science Research Synergy Program (provided by the Hungarian Academy of Sciences; HAS) and VEKOP (No. VEKOP-2.3.2-16-2017-00014). A.S. and T.V. are supported by the MTA-ELTE Genetics Research Group (01062).

**Author Contributions**

C.K., D.B., A.B. conceived the study. C.K. designed and performed the feature generation and the feature selection. C.K. and B.D. designed and performed the machine learning predictions. All authors interpreted the results. Biological interpretation of the results was done by T.V., Á.S., C.K. All authors drafted the manuscript. A.B. supervised the study, A.B. and T.V. acquired funding for the study.

**Additional Information**

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-22240-w.

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note:** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2018