Entropy Rates of Physiological Aging on Microscopy

Tuan D. Pham
Department of Biomedical Engineering
Linkoping University
Linkoping 581-83, Sweden
E-mail: tuan.pham@liu.se

***Abstract*—This paper presents a method for computing entropy rates of images by modeling a stationary Markov chain constructed from a weighted graph. The proposed method was applied to the quantification of the complex behavior of the growing rates of physiological aging of *Caenorhabditis elegans* (*C. elegans*) on microscopic images, which has been considered as one of the most challenging problems in the search for metrics that can be used for identifying differences among stages in high-throughput and high-content images of physiological aging.**

***Keywords*—Entropy rate, Markov chain, Kullback-Leibler divergence, physiological aging, microscopic images.**

I. Introduction

The computerized quantification of complex patterns in big medical and life-science data is an important and challenging area of interdisciplinary biomedical research. Advanced developments in this study is demanding because they can serve as useful tools that enable fast high-throughput screening and timely scientific experimentation with the use of high-content microscopic images for validating new hypotheses as well as testing a large number of drug-like compounds. Most information-theoretic developments for distinguishing nonlinear behavior in data are concerned with time series of dynamical systems. Relatively little effort has been spent on the study to analyze image data of complex biological systems.

Until now the notion of image complexity is diverse and not well-defined [1]. It has been so because different methods introduce different standpoints for measuring image complexity. More generally, the Shannon entropy of an image intensity histogram has been considered as a measure of image complexity, but this formulation does not take into account the spatial distribution of pixels in images [2]. Other methods address the issue of image complexity in the context of image compression. The association between complexity and compression stems from the concept of the Kolmogorov complexity introduced in algorithmic information theory. The Kolmogorov complexity is known as the descriptive complexity of an object being equivalent to the length of the shortest computer program that produces the object as the output from basic elements [3]. However, in addition to the difficulty of the computation of the Kolmogorov complexity, it has been known that the Kolmogorov complexity is not associated with the underlying nature of visual appearance in images [1], [4].

This paper presents a method for measuring image complexity for quantifying the pattern of regularity in image data, with a particular reference to differentiating stages of physiological aging. The method works by using the image histogram to convey statistical information about the image intensity. The weights between image rows or columns can then be computed using the Kullback-Leibler divergence (KLD) [3] to construct a weighted graph of the image. By imposing constraints on the graph of the image as a stationary Markov chain, the entropy rate of the Markov chain, which is the weighted transition entropy, can be computed. Such an entropy rate has been proved to be equal to the regularity statistic obtained from the approximate entropy [5]. The formulation introduced here derives a weighted graph whose vertices are either rows or columns of an image and edge weights are measures of differences between two probability distributions of the corresponding pairs of image rows or columns, and both separate sources of image information are combined as the entropy sum to represent the image regularity statistic.

The rest of this paper is organized as follows. Section II describes how an image can be induced to a weighted graph using the KLD and the image histogram. Section III shows the computation of the entropy rate of a Markov chain of a KLD-weighted graph of an image, which can be used as a tool for quantifying the regularity or complexity of image data. In Section IV, the proposed method was applied for studying the complex phenomena of the growing rate of physiological aging of *C. elegans* on microscopic images. Finally, Section V is the concluding remarks of the research findings.

II. A Markovian KLD-Weighted Graph

To model an image as a discrete stochastic process, let the sequence of random variables $\{X_i\}$, $i = 1, \ldots, N$, represent the sequence of N rows or N columns of the image. Such a sequence of random variables of the image is assumed to be stationary with respect to shifts in the space index, that is $Pr(X_1 = x_1, X_2 = x_2, \ldots, X_N = x_N) = Pr(X_{1+s} = x_{1+s}, X_{2+s} = x_{2+s}, \ldots, X_{N+s} = x_{N+s})$, $\forall N$, $\forall s$, and $\forall x_i \in \mathcal{X}$. The discrete stochastic process of the image is also a Markov chain, in which each X_i depends only on its immediately preceeding random variable and conditionally independent of all other preceding random variables, that is $Pr(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_1 = x_1\} = Pr(X_{n+1} = x_{n+1} | X_n = x_n)$.

The stationary Markov process of an image can be constructed as a weighted graph G that consists of two finite sets: a set of vertices $V(G)$ and a set of edges $D(G)$, where each edge is associated with a pair of vertices called its endpoints. A weight $w_{ij} \geq 0$ is assigned on the edge joining its endpoints from i to j (for an undirected graph, $w_{ij} = w_{ji}$), which can be obtained as the Kullback-Leibler divergence of the two probability distributions of image rows or columns i and j as follows.

Let $g = (g_1, g_2, \ldots, g_M)$ and $h = (h_1, h_2, \ldots, h_M)$ be two discrete probability distributions of two corresponding image rows or columns i and j of length M, respectively, obtained from the image histogram. A weight w_{ij} of graph G of an image can be computed by the KLD as a statistical measure of the departure of the candidate distribution h from the reference model g as follows [3]:

$$w_{ij} = D(g\|h) + D(h\|g), \tag{1}$$

where

$$D(g\|h) = \sum_{k=1}^{M} g_k \log \frac{g_k}{h_k}, \tag{2}$$

$$D(h\|g) = \sum_{k=1}^{M} h_k \log \frac{h_k}{g_k}, \tag{3}$$

where $0\log(0/0) = 0$, $0\log(0/q) = 0$, and $p\log(p/0) = \infty$.

Based on the properties of the KLD, $w_{ij} \geq 0$.

It should be now pointed out that using the KLD to determine the weights of an image-based graph edges is based on several theoretical aspects. In comparison with other metric functions such as the Euclidean distance, the KLD conveys statistical meaning and is geometrically important, because its asymmetric property exists for a manifold of probability distributions while no distance functions can take on this measure [6]. Moreover, the KLD has three particular properties that make them important in information processing [7]: 1) an information-theoretic function that satisfies the data processing inequality, 2) being an exponential rate of optimal classifier performance probabilities, and 3) its Hessian matrix is proportional to the Fisher information matrix. It is also proved that the KLD is the only dissimilarity measure of two probability distributions, which satisfies the characterization of entropy [8].

III. Entropy Rate of a Markov Chain of an Image

After modeling an image as a KLD-weighted graph, the entropy rate of the Markov chain of a weighted graph can be readily computed as follows. Let $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$ be a sequence of N random variables. The entropy rate for the stochastic process $\mathbf{X}$, denoted as $H(\mathbf{X})$, which grows with N, is derived in [3], which is known as the average transition entropy in terms of the entropy of the stationary distribution and the total number of edges in the weighted graph. Because the edge weights of the graph are the KLD values, the derivations of the probabilities of connecting the edges and the state probabilities are taken as the complement. The entropy rate for a stochastic process is mathematically expressed as [3]

$$H(\mathbf{X}) = \lim_{N\to\infty} \frac{1}{N} H(X_1, X_2, \ldots, X_N), \tag{4}$$

when the limit exists.

Furthermore, the conditional entropy rate of $\mathbf{X}$, denoted as $H^*(\mathbf{X})$, is defined as

$$H^*(\mathbf{X}) = \lim_{N\to\infty} \frac{1}{N} H(X_N | X_{N-1}, X_{N-2}, \ldots, X_1), \tag{5}$$

when the limit exists.

For a stationary stochastic process, the limit expressed in Equations (4) and (5) exists and $H(\mathbf{X}) = H^*(\mathbf{X})$. Thus, for a stationary Markov chain, the entropy rate can be calculated as [3]

$$H(\mathbf{X}) = H^*(\mathbf{X}) = H(X_2|X_1) = -\sum_i \mu_i \sum_j p_{ij} \log p_{ij}, \tag{6}$$

where

$$p_{ij} = 1 - \frac{w_{ij}}{c_i}, \tag{7}$$

which is the probability of the edge connecting node i to node j, and c_i is the total weight of edges emitting from node i:

$$c_i = \sum_j w_{ij}, \tag{8}$$

and μ_i is the stationary probability of state i, which is proportional to the weight of all edges emitting from node i, expressed as follows

$$\mu_i = 1 - \frac{c_i}{2C}, \tag{9}$$

where

$$C = \sum_i \sum_{j>i} w_{ij}. \tag{10}$$

IV. Entropy Rates of Physiological Aging of *C. elegans* on Microscopic Images

Modern biological theories of aging are categorized into two major schools of thought, known as the programmed theory and the error theory [9]. However, because of the complex processes and many underlying biological, biochemical, and environmental-exposed mechanisms of aging [10], there are no theories of aging that can adequately explain principle questions such as why and how human age, and what human aging markers are. Answers to these questions will have biological significance because lifespan can be healthily promoted and prolonged. Identifying biomarkers that are predictive of aging is difficult but its successful finding is important because it can address critical issues of the largest risk factor for many human diseases and facilitate the development of effective measures for sustaining older-aging health [11], [12]. It is reported in [13] that the American Federation for Aging Research defines a criterion for a biomarker of aging as a metric that can predict the rate of aging, which is considered to be a better predictive value of life span than chronological age. This definition was the motivation of this paper that applied the entropy rate of an image as a regularity statistic for quantifying the rate of complex aging of *C. elegans* on microscopic images.

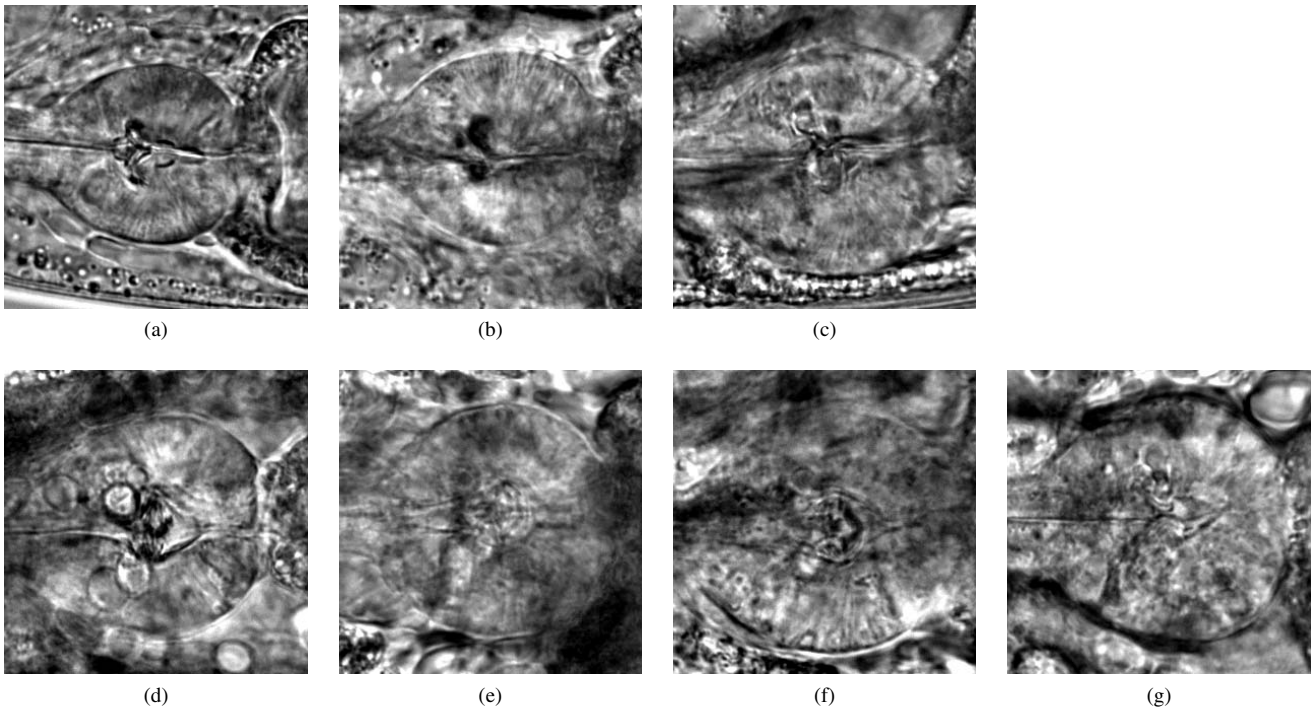

Fig. 1: *C. elegans* terminal bulbs: (a) day 0, (b) day 2, (c) day 4, (d) day 6, (e) day 8, (f) day 10, and (g) day 12.

It has been known that the use of humans for aging study is infeasible because of many difficult encounters of ethical, environmental, and social issues, and most importantly, the natural life span of humans; and therefore animal models are commonly adopted as useful prototypes for aging research [12]. The analysis of the physiological aging on microscopic images of *C. elegans* terminal bulbs is the most challenging problem of the ten aging image datasets studied in [14]. Its classification rate (49%) using 11 features and k-NN was the lowest among the other 10 datasets (HeLa cells, subcellular compartments (CHO), pollen grains, RNAi library of 16 phenotypes of drosophila cells, *C. elegans* muscle age, binucleate and regular cells, malignant lymphoma, mouse liver aging (AL), mouse liver gender (AL), and mouse liver gender (CR)). The light-microscope image dataset of pharynx terminal bulb of *C. elegans* consists of 106 images for day 0, 218 images for day 2, 159 images for day 4, 176 images for day 6, 195 images for day 8, 62 images for day 10, and 54 images for day 12. Each sample is of 300×300 pixels obtained from a differential interference contrast (DIC) microscopy, which is a technique used for contrast enhancement of unstained and transparent images. Figure 1 shows images of the terminal bulb of *C. elegans* of 0, 2, 4, 6, 8, 10, and 12 days of age.

The images are grouped into the stages of the chronological ages of *C. elegans*, where their morphological ages are unknown. It is known that that their physiological and their chronological ages are not fully correlated, given that they are genetically identical and lived in the same environment. The aim is to predict the physiological age of the pharynx terminal bulb of *C. elegans* using nematode-based images. Computerized image analysis methods are necessary because the manual classification of these images is highly difficult with respect to accuracy. Figure 2 shows the reconstructed phylogenetic tree of the physiological ages of *C. elegans* terminal bulbs. The phylogeny was built using the unweighted pair-group method using arithmetic averaging (UPGMA) [15]. Table I is the lower half of the similarity matrix (Euclidean distances) between the pairs of entropy rates of the worm ages, which is used for constructing the phylogeny shown in Figure 2.

TABLE I: Lower half of similarity matrix used for constructing the phylogeny of entropy rates of *C. elegans* ages.

	Day 0	Day 2	Day 4	Day 6	Day 8	Day 10	Day 12
Day 0	0						
Day 2	0.0454	0					
Day 4	0.0798	0.0925	0				
Day 6	0.0833	0.0661	0.0549	0			
Day 8	0.0344	0.0471	0.0379	0.0207	0		
Day 10	0.0095	0.0127	0.0035	0.0137	0.0249	0	
Day 12	0.0092	0.0264	0.0376	0.0172	0.0284	0.0112	0

On a study of age-realted behaviors reported in [16], by using whole-genome expression profiles of 104 individual wild-type *C. elegans* covering the entire nematode lifespan and correlating that profiling with age-related behavior and survival, the study found that a set of genes are associated with the aging process. The use of all gene expression data shows that young worms *C. elegans* with ages from 4 to 8 days are distinctly clustered from all other individual worms. An explanation is that worms actively lay eggs during the period of day 4 to day 8, during which worms are filled with eggs throughout various stages of development. It is

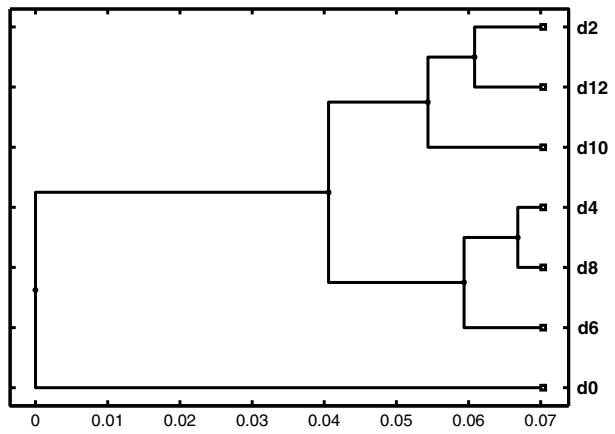

Fig. 2: Phylogenetic tree showing relationships of entropy rates of physiological aging of *C. elegans* terminal bulb on microscopic images, where terminal nodes are days (d0 stands for day 0) and values on the horizontal axis are distance scores. It should be pointed out that the physiological aging of *C. elegans* terminal bulb does not correspond to their chronological aging, and the tree topology appears to match with their gene expression patterns.

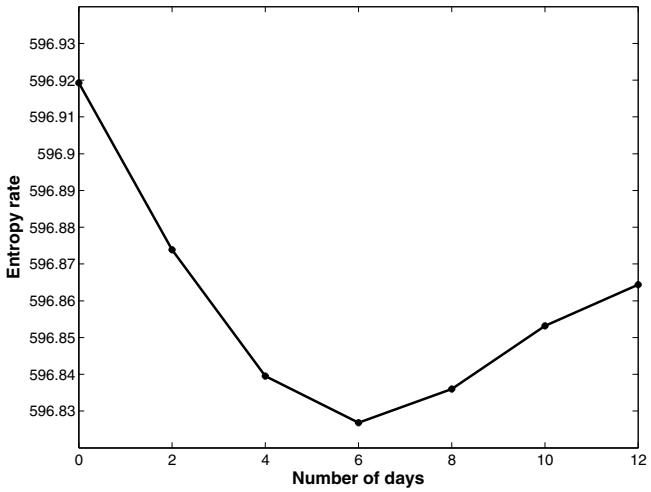

Fig. 3: Mean entropy rates of *C. elegans* terminal bulb aging, which clearly shows a trend, where variances are 2.9375e-04; 9.4624e-04, 0.0025, 0.0021, 0.0019, 0.0017, and 0.0014 for days 0-12, respectively.

therefore the association between gene expression and the development of eggs inside the worms can be discovered. In particular, evidence of substantive changes in gene expression was found in *C. elegans* of 4–8 days old. The association of the physiological aging of *C. elegans* shown in Figure 2 agrees with the association based on gene profiling found in [16], in which days 4, 6, and 8 are grouped in a distinct cluster. Furthermore, day 4 and day 8 belong to the inner group of the clade of days 4, 6, and 8; indicating the symmetrical expression of the transcriptional profiles of the two time points of the

TABLE II: Confusion matrix of classification results using k-NN with k=2.

	Day 0	Day 2	Day 4	Day 6	Day 8	Day 10	Day 12
Day 0	**106**	0	0	0	0	0	0
Day 2	21	**177**	0	0	0	13	7
Day 4	5	42	**91**	0	0	16	5
Day 6	1	20	31	**107**	0	8	9
Day 8	4	38	33	45	**52**	11	12
Day 10	2	0	0	0	0	**60**	0
Day 12	4	0	0	0	0	2	**48**

TABLE III: Classification rates (%) of *C. elegans* terminal-bulb microscopic image dataset.

Method	
Combination of 11 features [14]	49
Image regularity statistic (average of k = 1 to 5)	61

distinct period of nematode reproduction. The worm age of day 0 is the most farther way from other ages as shown in the tree topology, because it was explained in [14] that baby worms are still to grow and therefore very different form adult worms. The physiological conditions of the worms at 0 day, and 4–8 days of age result in locating the worms of days 2, 10, and 12 in the same clade of the tree.

It should be pointed out that the phylogeny shown in Figure 2 is the relationship of the growing rates of the worms according to their physiological appearances. Figure 3, which is the plot of mean entropy rates versus the ages of the worms, bears a striking resemblance to the patterns of the gene expression profiles shown in Figure 4c in [16]. The entropy rates are lowest at days 4-8 and reaches the lowest value at day 6 (making the grouping of day 4 and day 8 together as shown in Figure 2), then decrease with advancing age (entropy rates of day 10 and day 12 are lower than those of day 0 and day 2). The worm growing rate of day 0 of age serves as the base time point in both the phylogenetic tree (Figure 2) and the trend of the entropy rates (Figure 3).

The entropy rates of *C. elegans* terminal-bulb images obtained from the image regularity statistic were also used as image feature values to carry out the classification of physiological ages of the *C. elegans* terminal bulbs using the k-NN (nearest neighbors) method. Two common methods for estimating the rates of misclassification made by a classifier are re-substitution and cross-validation [17]. The re-substitution works by using all data to construct the classifier, and then carries out the testing to each sample. The cross-validation partitions or holds out samples from the dataset, the classifier is then trained with the remaining data to test the held-out samples. For small datasets or sets of limited samples, where estimator variance needs to be considered, the method of re-substitution was reported to be preferred to the method of cross-validation [17], [18]. The re-substitution is adopted here to assess the classification rates of the proposed image regularity statistic. Table II shows the confusion matrix of the k-NN results, where k=2: day 0 = 100%, day 2 = 81%, day 4 = 42% (mostly misclassified as day 2), day 6 = 61% (mostly misclassified as day 4), day 8 = 27% (mostly misclassified as day 2, day 4 and day 6), day 10 = 97%, and day 12 = 89%. In

this case, the entropy rate of day 0 is the most predictive and the next is day 10, whereas day 8 is of the lowest predictive value.

Table III shows the classification results obtained from the work in [14] using a modified k-NN and 11 features (Radon transform, Chebyshev statistics, Gabor filter, multi-scale histogram, first 4 statistical moments, Tamura texture, edge statistics, object statistics, gray-level co-occurrence matrix, and Chebyshev-Fourier transform), and those from the proposed image regularity statistic. The results demonstrate the superior performance of the proposed method for classifying the worm physiological ages on DIC microscopic images, which are known to be very difficult to process using standard image processing techniques.

V. Conclusion

The formulation of a new regularity statistic as a measure of complexity in terms of entropy rate in images has been presented. It is theoretically equivalent to the regularity statistic for quantifying complexity in time series known as ApEn. Based on the experimental results, the proposed image entropy rate is promising as a potential tool for differentiating physiological aging stages. The findings about the relationships of the entropy rates and growth rates of *C. elegans* terminal bulbs and their entropy-rate trend seem to be in agreement with the gene expression profiles reported in literature.

References

[1] V. Chikhman, V. Bondarko, M. Danilova, A. Goluzina, Y. Shelepin, "Complexity of images: experimental and computational estimates compared", *Perception*, vol. 41, pp. 631-647, 2012.

[2] J. Rigau, M. Feixas, M. Sbert, "An information-theoretic framework for image complexity", in *Proc. First Eurographics Conf. Computational Aesthetics in Graphics, Visualization and Imaging*, 2005, pp. 177-184.

[3] TM. Cover, J.A. Thomas, *Elements of Information Theory*, 2nd edition. New Jersey: Wiley, 2006.

[4] J. Perkio, A. Hyvarinen, "Modelling image complexity by independent component analysis, with application to content-based image retrieval", in Artificial Neural Networks – ICANN 2009 (series Lecture Notes in Computer Science) vol. 5769, C. Alippi, M. Polycarpou, C. Panayiotou, and G. Ellinas, Eds. Berlin, Germany: Springer-Verlag, pp. 704-714.

[5] S.M. Pincus, "Approximate entropy (ApEn) as a complexity measure", *Chaos*, vol. 5, pp. 110-117, 1995.

[6] N.N. Cencov, *Statistical Decision Rules and Optimal Inference*. Providence, RI: American Mathematical Society, 1982.

[7] D.H. Johnson, S. Sinanovic, "Symmetrizing the Kullback-Leibler distance", http://www.ece.rice.edu/ dhj/resistor.pdf. Accessed 15 February 2016.

[8] A. Hobson, *Concepts in Statistical Mechanics*. New York: Gordon and Breach, 1971.

[9] K. Jin, "Modern biological theories of aging", *Aging and Disease*, vol. 1, pp. 72-74, 2010.

[10] J. Kenyon, S.L. Gerson, "The role of DNA damage repair in aging of adult stem cells", *Nucleic Acids Research*, vol. 35, pp. 7557-7565, 2007.

[11] J.R. Harris, T.L. Gruenewald, T. Seeman, "An overview of biomarker research from community and population-based studies on aging", in *National Research Council (US) Committee on Advances in Collecting and Utilizing Biological Indicators and Genetic Information in Social Science Surveys*, M. Weinstein, J.W. Vaupel, K.W. Wachter, Eds. Biosocial Surveys. Washington (DC): National Academies Press (US), 2008. Available from: http://www.ncbi.nlm.nih.gov/books/NBK62427/

[12] S.J. Mitchell, M. Scheibye-Knudsen, D.L. Longo, R. de Cabo, "Animal models of aging research: implications for human aging and age-related diseases", *Annu Rev Anim Biosci.*, vol. 3, pp. 283-303, 2015.

[13] A. Burkle, M. Moreno-Villanueva, J. Bernhard, M. Blasco, G. Zondag, J.H.J. Hoeijmakers, O. Toussaint, B. Grubeck-Loebenstein, E. Mocchegiani, S. Collino, E.S. Gonos, E. Sikora, D. Gradinaru, M. Dolle, M. Salmon, P. Kristensen, H.R. Griffiths, C. Libert, T. Grune, N. Breusing, A. Simm, C. Franceschi, M. Capri, D. Talbot, P. Caiafa, B. Friguet, P. Eline Slagboom, A. Hervonen, M. Hurme, R. Aspinall, "MARK-AGE biomarkers of ageing", *Mechanisms of Ageing and Development*, vol. 151, pp. 2-12, 2015.

[14] L. Shamir, N. Orlov, D.M. Eckley, T. Macura, J. Johnston, I.G. Goldberg, "Wndchrm - an open source utility for biological image analysis", *Source Code for Biology and Medicine*, vol. 3:13, 2008. DOI: 10.1186/1751-0473-3-13

[15] C.D. Michener, R.R. Sokal, "A quantitative approach to a problem in classification", *Evolution*, vol. 11, pp. 130-162, 1957.

[16] T.R. Golden, A. Hubbard, C. Dando, M.A. Herren, S. Melov, "Age-related behaviors have distinct transcriptional profiles in C.elegans", *Aging Cell*, vol. 7, pp. 850-865, 2008.

[17] U. Braga-Neto, R. Hashimoto, E.R. Dougherty, D.V. Nguyen, R.J. Carroll, "Is cross-validation better than resubstitution for ranking genes?", *Bioinformatics*, vol. 20, pp. 253-258, 2004.

[18] L. Devroye, L. Gyorfi, G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.