

An Information-Theoretic Framework for Identifying Age-Related Genes Using Human Dermal Fibroblast Transcriptome Data

Salman Mohamadi

*Computer Science and Electrical Engineering
West Virginia University
Morgantown, WV, USA
Sm0224@mix.wvu.edu*

Donald A. Adjeroh

*Computer Science and Electrical Engineering
West Virginia University
Morgantown, WV, USA
donald.adjeroh@mail.wvu.edu*

Abstract—Investigation of age-related genes is of great importance for multiple purposes, for instance, improving our understanding of the mechanism of ageing, increasing life expectancy, age prediction, and other healthcare applications. In this work, starting with a set of 27,142 genes, we develop an information-theoretic framework for identifying genes that are associated with aging by applying unsupervised and semi-supervised learning techniques on human dermal fibroblast gene expression data. First, we use unsupervised learning and apply information-theoretic measures to identify key features for effective representation of gene expression values in the transcriptome data. Using the identified features, we perform clustering on the data. Finally, we apply semi-supervised learning on the clusters using different distance measures to identify novel genes that are potentially associated with aging. Performance assessment for both unsupervised and semi-supervised methods show the effectiveness of the framework.

I. INTRODUCTION

Aging is a complex process, and is often associated with the process of increasing physiological decline, increased risk of illness, and increased mortality. This process could manifest in terms of various types of decline in physiological function, such as transcriptional decline [1]. Early studies on ageing confirm the contribution of genes to ageing, for instance, in some species such as mouse or dog even a few number of gene mutations resulting from natural selection can lead to a significant extension of life span [2]. From the perspective of cell biology, ageing can be interpreted in terms of variations in gene expression due to both the environment and the genetic makeup. While some studies have showed that the age-associated genes are different among both different tissues and organisms, further analysis on large datasets identified some general common patterns [1]. Accordingly, certain changes in the level of expression, or in the patterns of expression of a subset of genes could imply that those genes are associated with the process of ageing [3]. Furthermore, for a given gene, certain patterns of changes in its expression levels across the human life span could be an indicator of causal or associative relationship between the gene and the ageing process. In parallel with advanced experimental investigations in biological wet-laboratories, detection of the underlying patterns of expression calls for development of novel computational techniques for the huge gene expression datasets.

The process of ageing of an organism (also called senescence [4]), includes nine hallmarks including stem cell reserves depletion, epigenetic alterations, genomic instability, mitochondrial dysfunction and a few other hallmarks [5]. These hallmarks affect and are influenced by transcriptional changes [5]. Studies on the mechanism of these changes could provide some answers on whether these changes are beneficial or detrimental. Accordingly, some studies show that ageing process can be regulated by therapeutic steps for longevity to amplify the protective changes and eliminate the detrimental changes [1], [6].

With an emphasis on the gene expression changes as one of the main elements of ageing process, there are different sets of theories including evolutionary theories and passive theories, trying to explain the hidden mechanisms responsible for gene expression changes associated with senescence [7]. For instance, a study on gene expression data of mice confirmed the ageing theory based on cellular metabolic stability [8]. Although an organism's case-by-case differences in the speed of ageing can be caused by efficiency in cellular and tissue metabolism, the general process of ageing for an organism is associated with changes in its gene expression. From a general perspective, recent studies contrast gene expression changes in terms of global scale versus local scale. This can be interpreted as redirecting efforts toward identifying certain subsets of genes responsible for the ageing process of different cell types (such as brain, pancreatic, kidney, liver, muscle, eye, and cell), tissues, and even species [1]. Recent studies such as Haustead et al [10] and Stegeman et al [1] suggest that less than 2% of genes in human skin are identified with age-related changes in expression levels. See also [10], [11]. Motivated by the foregoing, our goal in this work is to further investigate such subset of genes using advanced machine learning techniques.

In this paper, we develop an information-theoretic computational framework to analyze a dataset of human dermal fibroblasts transcriptome data with gene expression values for 143 individuals. The framework is used to identify hidden patterns of gene expression associated with aging, for some subset of the 27,142 genes. Initially, an unsupervised approach is used to extract important features using information-theoretic measurements, based on which the genes are grouped into clusters with respect to ageing. The clusters are then further analyzed and refined using semi-supervised learning to identify potentially new age-related genes, exploiting prior information from a small subset of genes known to be associated with ageing.

The paper is organized as follows. The next section reviews related work on gene selection for the ageing process. The next section presents our three-step methodology: (1) information theoretic feature measurements, (2) clustering, and (3) semi-supervised learning. Section V provides an evaluation of the proposed methods. Section VI provides a brief discussion and conclusion.

A. Prior Work

Machine learning algorithms have surpassed classical signal processing methods and have gained dramatic attention of computational biologists due to multiple advantages such as accuracy, robustness and good generalization [12], [27], [30], [35]. Specifically, machine learning has showed its capability for genomic data processing. One good example is the study of mechanisms associated with

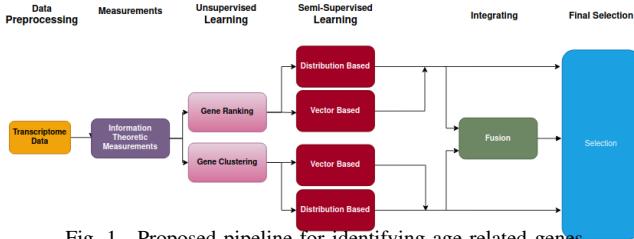


Fig. 1. Proposed pipeline for identifying age-related genes

ageing through gene expression profiling in order to estimate the age [13]. The study of the mechanism of ageing, and in particular, the identification of age-related gene have attracted a lot of attention over the years [14], especially given its implications on the possibility for developing therapeutic approaches to slow down or control the human ageing process [15], [16]. It is hoped that such manipulations can be implemented at different scales, from whole body to organs to small tissues [1]. However, most attention has been directed toward costly experimental studies in laboratories. On the other hand, more recently, computational methods on selecting age-related genes have gained more attention, given their potential for facilitating more cost-effective understanding of mechanism of ageing, for instance, by providing a guide to experimental researchers for more focused investigations. For instance, Srivastava et al [17] employed machine learning tools to predict genes associated with age-related diseases. De MacLhaes et al [18] investigated age-related genes and provided a meta-analysis of work on gene expression data of mice, rats and humans. They considered the noisy nature of gene expression data, and studied how to associate transcriptional changes with the ageing. Their results identified a set of 56 genes as overexpressed over the life span, and another set of 17 as underexpressed. Uddin et al. [19] performed an analysis age-related variations in the gene expression data of porcine tissues in order to identify some reference genes. They identified 9 genes as the reference genes, and used these for data normalization of the gene expression data.

Avelar et al [20] perform integrative computational analysis to select and validate a subset of age-related genes. Based on their comparative study they claim that human tissues' gene over-expression with age can be used to characterize ageing and identifying age-related genes as well as age inhibiting genes.

II. METHODOLOGY

In this section, we discuss our dataset from [13] and introduce our computational framework. Briefly, besides a list of genes identified to be age-related by previous studies [14], we have a dataset of human dermal fibroblast gene expression values for 27,142 genes across 143 individuals with age ranging from less than 1 year to 94 years old [13]. We perform a cascade of coarse and fine-grained gene selection using both unsupervised and semi-supervised learning. In the unsupervised approach, we first represent the expression values for each gene using handcrafted information-theoretic measurements. We then apply a clustering algorithm on the new representation to cluster the genes into two groups. The idea behind the binary clustering is to group the genes based on their potential association with ageing. One rationale here is that, as the vector of expression values for each gene are represented over age (time axis), binary clustering on these raw vectors could expose the discriminating dynamic (behavior) of expression values with respect to aging. Another rationale is that, rather than just using the raw data for the binary clustering, the information-theoretic measurements are performed with respect to aging, thus the binary grouping could help to provide the first level

discrimination between genes that are associated with ageing, and those that are not.

We expect that age-related hidden pattern(s) of changes should manifest in the form of variations of certain measurements such as entropy, Kullback-Leibler divergence and correlation between the expression values and time axis built upon the age range. The credibility of this idea will be assessed in by contrasting experimental results using raw data with those obtained using the information-theoretic measurements. Yet, because noise and other possible unknown factors could be contributing to the variations in the indicated measurements along with the age, we cannot be certain that these variations are only due to hidden patterns associated with the ageing process. To address this issue, we refine the results from the unsupervised approach using semi-supervised learning. Thus, using a small set of previously identified genes known to be associated with ageing, we performing semi-supervised learning for final identification of age-related genes from the previously clustered genes.

A. Datasets

For this work, we use the human dermal fibroblast transcriptome data reported in [13] which contains data on gene expression values. The database of gene expression values consists of a matrix of gene expression values from 27,142 genes across 143 individuals. Each row contains the expression values from the genes across 143 subjects in order of age (from less than 1 to 94 years old). Each column captures the expression values for each of 27,142 different genes for a given individual. We note that, the expression values for any given gene in the data set are measured from different individuals (143 subjects with age ranging from less than 1 to 94 years old), rather than across the lifetime of each individual. For a general study like the one proposed here, this could be important in smoothing out the effects on the analysis that may be due to the impact of one individual.

In the context of unsupervised clustering, we consider each row as a vector representing the expression values for a certain gene across the age, where it can potentially take a binary label with respect to the assumption on its association with ageing. However, for the semi-supervised approach, we have the label for a small number of genes known to be associated with ageing, and we seek to annotate the remaining genes.

We also used a database reported by Tacutu et al [14] which listed 307 genes that are known to be associated with ageing, and 243 genes that are most likely associated with aging.

III. UNSUPERVISED APPROACH

Unsupervised learning as a branch of machine learning that is established based on the fact that data points are generated by natural sources with limited complexity, rather than random sources. It essentially deals with problem domains, such as data clustering, annotation, and dimensionality reduction [24], [25]. One of the well-known unsupervised learning algorithms is k -means clustering, where the goal is to assign the data points to K clusters in an iterative fashion [26], [34].

In this work, before applying the k -means algorithm, we take an information-theoretic approach, and compute a few measurements such as entropy, correlation and K-L divergence, and represent vector of expression values corresponding to each gene with these measurements. In fact, extraction of these measurements represent a limited form of differential gene expression over the age range. Thus, the measurements are expected to guide the k -means algorithm to more directly find any potential associations between the genes and the ageing process. Although we will also represent the result of

applying K-Means on the raw data to ponder upon the effectiveness of using these measurements.

Recent studies suggest that less than 2% of genes in the human skin, retina and blood leucocytes are identified with age-related changes in expression level [1], [9]–[11]. Accordingly, since we have a huge number of genes and we want to select only a small subset of genes, we need to eliminate many of them through clustering. Therefore, the main idea is to eliminate most of the genes that are deemed not to contribute to the ageing process. Along this line, three measurements, namely entropy, correlation, and Kullback-Leibler divergence are employed to discriminate the genes in terms of their association with ageing. In the next few sections, we will discuss these measurements and how we employ them to represent the data, before applying k -means algorithm. After considering considering each measurement separately, we combine the individual results using ensemble modeling.

We expect that genes related to ageing or longevity, will typically influence aging, or be affected by aging, in a manner that is consistent over time, along the age axis. Therefore, we divide the expression values of each gene into two age groups, generally representing young subjects and old subjects respectively, and then examine the identified measurements across all the subjects, within each age group, and between the groups. To find the appropriate age threshold to divide the subjects into binary groups, we performed experiments using different ages, namely, 25, 30, 35, 40, and 45. Then we compare the results to determine the best age threshold.

A. Entropy

We compute the entropy of each gene across all 143 subjects and also across the pair of age groups for each age threshold.

The entropy of a random variable was first defined by Claude Shannon in 1948 as the average amount of information or uncertainty in the variable's possible outcomes [30]. In information theory literature, it is represented as follows:

$$E = \sum -p_i \log p_i \quad (1)$$

where p_i is the probability of the i -th outcome. However, we need to compute the entropy of histogram of expression values for each gene as the entropy of the probability distribution of the values within the age bins of the histogram, in which p_i is defined as the number of values within a given age bin divided by the total number of expression values for that gene [29]. Here we empirically set the number of histogram bins to 7, after testing different values. We select the genes based on their corresponding contribution to aging, as measured by the entropy of expression values for each gene across all the subjects, or within the two separate groups of subjects for a given age threshold, for instance, **Group 1** (1 to 25 years old) and **Group 2** (26 to 94 years old) using an age threshold of 25. In fact, posing the gene selection problem within the context of age groups allows for a dynamic consideration in the selection process. That is, if gene functionality depends on the age of the subject, then, we expect the model to be able to recognize it. Also, by using multiple age thresholds, it becomes possible to see which genes play more significant roles at different stages in life, as captured by different age ranges.

Entropy Across All Subjects. As the first scenario, we compute the entropy of genes individually across all 143 subjects and then select the top T genes with the highest entropy. The intuition behind this is that the genes conveying more information through their expression values across age range of 1 to 94, are more probable to be related to the ageing process.

Entropy Across Subject Groups. As the second scenario, we divide the subjects into two groups for a given age threshold, based on their age. For instance, using an age threshold of 35 we will group the subjects into two age groups: Group 1: ages [1-35], and Group 2: ages [36-94]. Then the entropy of the genes across the subjects in each age group are computed individually and top T genes are selected based on their entropy ranking. Thus, each age group will produce one gene ranking.

Entropy Difference. Given the binary grouping, for a given gene it becomes possible to consider its entropy difference across the subjects in the two groups. Thus, as part of the feature measurements, we also compute the entropy between the expression values for subjects in Group 1 and Group 2. We rank the genes based on this difference, and select the top T genes for further consideration.

For our experiments, we have set $T=1000$.

B. Spearman's Correlation

Correlation is another feasible metric for exploring the underlying pattern of relationship between two variables – in this case, between subject age and expression values of each gene from the individual. In other words, correlation can be used to measure the possible association between expression values of the genes and the ageing process. The outcome of the correlation between two variables could be positive, neutral, or negative. With positive correlation both variables change in the same direction; with neutral correlation there is no relationship in the change patterns of the variables; and with negative correlation the variables change in opposite directions. Since we are seeking for any type of age-association, we use the absolute value of the correlation for further computation and inference.

There are two formulae for computing the correlation, namely Pearson's correlation and Spearman's correlation. The Pearson correlation coefficient can be used to measure the strength of the linear relationship between two variables when both variables follow a Gaussian or Gaussian-like distribution. While Spearman's correlation can be used for two variables with a nonlinear relationship between them, and the variables may have a non-Gaussian distribution.

Since we are unsure of the distribution and possible relationships between these two variables, gene expression values and age, we need to use Spearman's correlation as our metric to select the genes with most contribution. The formulation for Spearman's correlation coefficient is presented as follows:

$$L = \frac{C(R(x), R(y))}{S(x)S(y)}, \quad (2)$$

where x and y are our variables, $R(x)$ and $R(y)$ are rank of variable x and y , S is the standard deviation and C represent the notation for covariance.

Similar to entropy, for each gene we also compute the Spearman's correlation between the expression values and age, for all subjects, within each of the two age groups, and the difference in the correlation for subjects across the two age groups.

C. Kullback-Leibler Divergence

The Kullback-Leibler (KL) divergence (also called relative entropy) measures the difference between two probability distributions [30]. Suppose we have two discrete probability distributions P and Q which are defined on the same probability space, Z , the Kullback-Leibler divergence from Q to P is defined as follows:

$$D_{KL}(P||Q) = \sum_{z \in Z} P(z) \log \frac{P(z)}{Q(z)}, \quad (3)$$

For our purposes, given the age threshold, for each gene we compute the KL divergence for the two groups of expression values (as determined by the given age threshold). The genes are then ranked by their KL-divergence over the two groups, and we select the top T genes with maximum divergence.

Given the information-theoretic measurements, we can combine the individual top T rankings, to obtain one integrated set of genes for more refined analysis, for instance, using semi-supervised machine learning methods. For instance, we can combine the top T genes, by using simple intersection or union for the individual top- T ranked gene sets. An alternative approach is to apply further unsupervised learning techniques on the data, for instance, using k -means clustering, but this time using the information-theoretic measurements as the features. The clustering results will then be passed to the semi-supervised learning algorithm for final refinement of the identified genes. In this work, we show results for the two approaches.

D. Union of Selected Genes Using Measurements

As it is presented in Table I, the measurements provide different performance in gene selection, where the best performance is associated with K-L divergence. This could be due to the fact that each gene relationship with aging would only be captured by only one or a few of measurements, and not all of them. Considering this, one way to select a subset of genes from 27142 genes and pass them to next stage (semi-supervised approach) is to find the union between top genes selected by each measurement as the mentioned subset. The result of this union of top 1000 genes selected by each of the nine measurements in Table I provides us with 6799 genes. We pass it to semi-supervised gene selection stage where the results are presented in Table II.

E. K-means Clustering

A popular clustering method is the k -means algorithm. Though this algorithm is well known, given the significance of clustering in our approach, we briefly describe the algorithm below, for completeness. The k -means clustering algorithm, takes a training set $x^{(1)}, \dots, x^{(m)}$, and the goal is to group the data points into a few predefined number of coherent clusters. To describe a customised version of the algorithm for this work, say we are given feature vectors for each gene (data point) $x^{(i)} \in \mathbb{R}^n$ where n is the dimension of the feature vector. The algorithm initially selects k (here $k=2$) random data points as the centroid of the k clusters, and then according to the distance of the remaining data points to each of these centroids, assigns a cluster to the data point. Next, the mean of each cluster is computed as the new centroid of the cluster, and the update process of assigning clusters to the data points repeats until all data points remain at the assigned centroids and consequently the centroids could not be updated anymore.

More specifically, here rather than 143 measurements, we represent genes individually by nine information-theoretic measurements as described above, and apply k -means algorithm on the gene expression dataset to assign each of the 27,142 genes to one of the binary clusters. As explained earlier, we here we set $k = 2$ since we wish to classify the genes as being associated with age, or otherwise. Thus, the goal is to predict $k = 2$ centroids and a label $c^{(i)}$ for each gene as our data points. The k -means clustering algorithm is as follows:

1. Randomly initialize the cluster centroids μ_1 and μ_2 .

2. Repeat until convergence {

For each $i \in \{1, 2, \dots, 27142\}$, set:

$$c^{(i)} = \operatorname{argmin}_j \|x^{(i)} - \mu_j\|^2$$

For every $j \in \{1, 2\}$, set:

$$\mu_j = \frac{\sum_{i=1}^M 1\{c^{(i)}=j\} x^{(i)}}{\sum_{i=1}^M 1\{c^{(i)}=j\}} \}$$

IV. SEMI-SUPERVISED APPROACH

Unsupervised gene selection, provided us with two clusters of genes, which beside the theoretical support, we experimentally examined that the smaller cluster, with 5669 genes, includes previously known genes, and hence makes a good case for further refinement for gene selection. Thus, we seek to further analyse this cluster to select a small subset of genes using a semi-supervised approach. In order to efficiently and reliably identify new genes associated with ageing, we need to leverage our prior knowledge, i.e., we use 'previously identified subset of age-related genes' in the literature. Here the central idea is to use the information of vector of expression values from each of 307 identified genes to select a new subset of age-related genes. Therefore, we need an appropriate measurement of the similarity between the vector of expression values for each gene in the cluster with 5669 genes and the set of 307 known genes. Mutual information could be one good metric to find the most similar genes to the set of known genes. However, mutual information between two variables requires an exact or at least approximate estimation of joint probability distribution. However, due to the lack of further information such as multiple vectors rather than one vector of expression values for each gene, we can not compute or approximate the joint probability distribution. Instead, we use Jensen-Shannon divergence as a probability-based similarity measurement, which is based on the K-L divergence, except unlike K-L divergence, it always has a finite value and it is symmetric. On the other hand, in contrast to the probability-based similarity measurement, we also use an efficient sample-based similarity measurement to explore the similarity both from probabilistic point of view and sample-based point of view. Cosine similarity as a sample-based similarity measurement, could be a feasible measurement of the similarity in features between two vectors of expression values in this case.

However, one should consider the fact that these 307 known genes, obviously would not all function in the same way. For instance, some of the genes as the cause of ageing, would associate with ageing, while some of them are those genes that are just influenced by ageing [1], and some of them might even indirectly slow down the whole process of ageing. To address this, we initially analyse the probability space representing all these 307 genes, aiming at finding some subspaces each associated with specific type of gene functionality. This is akin to identifying anonymous gene families.

A. Sub-Space Analysis

The goal is to perform sub-space search and clustering of the set of 307 known genes by employing k -means for $K = 2, 3$, and then comparing the hypothesized age-related genes resulting from the unsupervised learning stage against the members of the respective clusters of the known genes for further similarity analysis. As it is formerly presented, the k -means algorithm provides us with k clusters of data points, here clusters of genes. We set number of clusters $k = 2$ and $k = 3$ and refine the gene selection using the Jensen-Shannon divergence (JSD) and cosine similarity. For $k = 2$ as binary clustering, the k -means algorithm provided us with two clusters consisting of 176 and 131 known genes, respectively. For $k = 3$, we had clusters consisting of 51, 117 and 139 known genes, respectively. Latter the results corresponding to each number of clusters will be assessed.

Age-related genes are not expected to all function in the same way. Some may have a causal relationship with ageing, that is,

influencing the aging process, while the function of some others will be influenced by ageing. Even so, some genes with causal relationship with ageing could speed up ageing or slow it down. Moreover, some studies have suggested the possibility of age-related dynamic changes in form of monotonic over-expression, monotonic under-expression, or a mixture of these over the life span. Accordingly, our rationale behind the sub-space search is that these types of different associations, functionalities or monotonic/non-monotonic patterns of expression should manifest in terms of "different" clusters of age-related genes. Later in this work, we will empirically examine the role of this type of subspace analysis in terms of clustering of the known genes [18].

B. Jensen-Shannon Divergence

Jensen-Shannon divergence (JSD) or information radius is an information-theoretic approach to measure the similarity between two probability distributions [30]. This similarity measurement which has been widely employed in bioinformatics and genome comparison, is represented as follows:

$$\begin{aligned} JSD(A||B) &= \frac{1}{2}D(A||M) + \frac{1}{2}D(B||M), \\ M &= \frac{1}{2}(A + B) \end{aligned} \quad (4)$$

where A_i and B_i represent the probability distribution for two given vectors of gene expression data, and $D(\cdot, \cdot)$ is K-L divergence as was introduced earlier in Section III-C.

C. Cosine Similarity

Cosine similarity is a well-known approach for measuring similarity between two vectors in inner product space. With this measure of similarity, we do not need to estimate any probability distribution, we only need to measure the cosine of the angle between the two vectors. Let \mathbf{A} and \mathbf{B} be two non-zero vectors of gene expression values; the cosine similarity between A and B is computed using the dot product and the magnitude as follows:

$$S(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}} \quad (5)$$

where A_i and B_i represent the elements of the corresponding vectors and n is the number of elements of each vector. It is noteworthy to mention that we perform the cosine similarity on both raw vectors of expression data for genes (as presented in Table II and III) as well as a vectors of nine measurements for genes individually. However, we observed that the results for the raw vectors are better. This could be due to the fact that in order to extract the phase information from vectors, we should solely consider the original raw values of the vectors rather than handcrafted features of that vectors.

D. Similarity to the Set of Known Genes

The unsupervised learning stage will produce two initial sets of selected genes, namely, one with 6492 (6799-307) genes using union of top $T = 1000$ ranked genes, and the other one with 5362 (5669-307) genes from k -means. We will then refine these initial selections using semi-supervised methods to select those that are most to the set of known genes, using cosine similarity and JSD.

As noted earlier, age-related genes are not necessarily monolithic – they could still cluster into different groups, depending on the basis of the clustering. Thus, we perform the similarity based refinement for our gene selection for three scenarios based on the clustering on the known genes, namely using one cluster, two clusters, and three clusters, respectively.

First Scenario: One Cluster. To perform gene selection for this scenario, we compute the average JSD as well as average cosine similarity of each gene with all 307 genes and then select the top Q genes with highest similarity values. We do this for both JSD and cosine. Accordingly we will have:

$$S_j = \sum_{i=1}^{307} \frac{|S_{i,j}|}{307}; \quad JSD_j = \sum_{i=1}^{307} \frac{JSD_{i,j}}{307}, \quad (6)$$

where $|S_j|$ is the absolute value of similarity between gene number j (out of 53622) and the whole set of 307 genes known to be age-related. Similarly, JSD_j is the average of the JSD between the j -th gene and each of the 307 known genes. We use absolute value of cosine similarity as we are interested in any type of association, whether positive or negative similarity. To select the most similar genes in the sense of expression values, we rank the genes based on cosine similarity as well as JSD separately, and select first Q genes for each method. For our experiments, we set Q to 500, which is close to the 2% of genes in the human skin presumably with gene expression changes associated with ageing [1], [10].

Second Scenario: Two Clusters.

$$\begin{aligned} S_{jc1} &= \sum_{i=1}^{176} \frac{|S_{i,j}|}{176}, & S_{jc2} &= \sum_{i=1}^{131} \frac{|S_{i,j}|}{131}; \\ S_j &= \max \{S_{jc1}, S_{jc2}\}; \\ JSD_{jc1} &= \sum_{i=1}^{176} \frac{|JSD_{i,j}|}{176}, & JSD_{jc2} &= \sum_{i=1}^{131} \frac{|JSD_{i,j}|}{131}; \\ JSD_j &= \max \{JSD_{jc1}, S_{jc2}\} \end{aligned} \quad (7)$$

where S_{jc1} and S_{jc2} are the similarity between genes number j (out of 5452) and cluster 1 with 176 genes and cluster 2 with 131 genes, respectively. Next, we rank the genes separately, based on S_j , and JSD_j , respectively. Then, for each ranking, we select the top Q genes.

Third Scenario: Three Clusters. Similarly, we perform clustering to divide the 307 genes into 3 clusters, and compute the similarity values:

$$\begin{aligned} S_{jc1} &= \sum_{i=1}^{139} \frac{|S_{i,j}|}{139}, & S_{jc2} &= \sum_{i=1}^{117} \frac{|S_{i,j}|}{117}, & S_{jc3} &= \sum_{i=1}^{51} \frac{|S_{i,j}|}{51}; \\ S_j &= \max \{S_{jc1}, S_{jc2}, S_{jc3}\}. \end{aligned} \quad \text{Similarly, we compute:} \quad (8)$$

$$JSD_j = \max \{JSD_{jc1}, JSD_{jc2}, JSD_{jc3}\}$$

where S_{jc1} , S_{jc2} and S_{jc3} are the similarity between genes number j (out of 5452) and clusters 1, 2 and 3, respectively. We select top Q ranked genes separately, based on S_j and JSD_j respectively.

V. RESULTS AND PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed methods, for both supervised and semi-supervised approaches. Results of the unsupervised approaches are used as the input to the semi-supervised learning approach for final gene selection.

A. Evaluation of Unsupervised Approach

In order to evaluate the unsupervised clustering approach, we contrast the results with the list of the 307 genes [14] known to be associated with aging. The results using each of the measurements are presented in Table I. As the table shows, clearly, KL divergence provided the best performance – recognizing 254 genes out of 307 known age-related genes, within its top $T = 1000$ ranked genes. The best age threshold appears to be at age 40, followed by age 35. Incidentally, these are quite close to the age range 38 to 40, which

TABLE I

PERFORMANCE OF EACH MEASUREMENT IN RANKING THE GENES WITH RESPECT TO ASSOCIATION WITH AGING, AT A GIVEN AGE THRESHOLD (τ_A). ENTRIES SHOW NUMBER OF THE 307 KNOWN AGE-RELATED GENES THAT ARE RECOGNIZED WITHIN THE TOP $T = 1000$ GENES SELECTED BY EACH MEASUREMENT AT THE INDICATED AGE THRESHOLD.

#	Measurement	$\tau_A = 25$	$\tau_A = 30$	$\tau_A = 35$	$\tau_A = 40$	$\tau_A = 45$
1	Entropy (all subjects)	232	232	232	232	232
2	Entropy (Group 1)	112	166	201	207	132
3	Entropy (Group 2)	108	118	196	192	127
4	Entropy difference	128	124	138	162	125
5	Corr. (all subjects)	214	214	214	214	214
6	Corr. (Group 1)	79	125	177	183	155
7	Corr. (Group 2)	112	108	128	163	125
8	Corr. difference	83	103	116	129	132
9	K-L Divergence	201	209	239	254	233

TABLE II

RESULTS USING THE UNION OF THE INDIVIDUAL TOP $T = 1000$ RANKED GENES FROM THE NINE MEASUREMENTS AS INPUT TO THE REFINEMENT STAGE WITH SEMI-SUPERVISED LEARNING. RESULTS SHOW THE NUMBER OF THE 243 KNOWN GENES FOUND IN THE TOP $Q = 500$ SELECTED GENES USING COSINE SIMILARITY, JSD, AND THEIR COMBINATION IN THREE SCENARIOS OF CLUSTERING ON 307 KNOWN GENES.

Metric	1 cluster	2 clusters	3 clusters
Cosine	200 (82.3%)	211 (86.8%)	169 (69.54%)
JSD	221 (90.9%)	218 (89.7%)	178 (73.2%)
Combined	218 (89.7%)	223 (91.7%)	163 (67%)

Belsky et al [32] found to be significant in considering the pace of ageing in young adults. Rahman and Adjeroh [33] also observed that human physical activity and various biomarkers of biological ageing show a turning point around this age range. These show that the age range 38 to 40 could represent a major dividing point for most individuals in terms of age progression.

Besides, the k -means algorithm provides us with two clusters of 5669 and 21,473 genes respectively. Although the second cluster contains even genes with all measurements equal to zero, the first cluster (5669 genes) contains all the genes that are known to be age-related. Therefore, we consider this smaller cluster as the one that most probably would contain other age-related genes that are not yet known. Hence, we pass this cluster to the semi-supervised algorithm to leverage knowledge from known genes to further refine the selection of new genes that are most likely to be age-related.

B. Results of Semi-supervised Learning

To this end, besides the list of 307 known genes used to train the semi-supervised learning algorithm, we use an additional 243 genes also known to be associated with aging to test our proposed methods for identifying novel genes with a high likelihood of association with aging.

Table II shows the results using the combined top T ranked genes from the individual information-theoretic measurements, using unsupervised learning as input to the semi-supervised method. With T set at 1000, taking a union of all the top T ranked genes (at $\tau_A = 40$, see Table V-A) resulted in a total of 6799 unique genes, which were then fed to the semi-supervised learning algorithm. The combined cosine distance and JSD results were obtained by simply computing the average rank from the two methods.

For the semi-supervised scenarios using different number of clusters, the results show that clustering using $k = 2$ clusters tend to

TABLE III

RESULTS USING THE CLUSTER OF SELECTED GENES AS AS INPUT TO THE REFINEMENT STAGE WITH SEMI-SUPERVISED LEARNING. SIMILAR TO TABLE II, RESULTS SHOW THE NUMBER OF THE 243 KNOWN GENES FOUND IN THE TOP $Q = 500$ SELECTED GENES.

Metric	1 cluster	2 clusters	3 clusters
Cosine	228 (93.8%)	238 (97.9%)	199 (81.8%)
JSD	231 (95%)	223 (91.7%)	165 (67.9%)
Combined	234 (96.2%)	243 (100%)	188 (77.3%)

lead to better results for similarity measurements. Tables II and III show that final selection refinement using inputs as the gene clusters based on clustering in the high-dimensional feature space performed better than using the union of the top T ranked genes from individual information theoretic measurements. The results indicate that, at $k = 2$ clusters, the top $Q = 500$ identified genes contained most, if not all of the 243 known genes. Thus, each method identified over 250 other genes that are potentially associated with aging. Further analysis showed that, aside from the 243 known genes, taking the intersection of the results from the two approaches (combined results using $k = 2$ clusters) resulted in a total of 82 unique genes. Thus, these represent novel age-related genes identified by our proposed approach.

VI. DISCUSSION AND CONCLUSION

As presented in Table I, one can easily contrast the result of each of the measurements independently. As it is clear, based on the recognition of known genes among top 1000 selected genes, K-L divergence and entropy across all subjects presented the best selection performance.

On the other hand, in order to assess the effectiveness of representing each gene with nine measurements to k -means rather than utilizing raw vector of gene expression values as the input to k -means algorithm, we perform the k -means on the raw vector of gene expression values in order to cluster the whole set of 27142 genes into two cluster. The result of k -means clustering on the raw data set, presents two clusters with 9479 and 17633 genes respectively. The second cluster consists of genes with no expression values along the age range, while the smaller cluster apparently is more associated with ageing as it contains known genes. As this cluster (9479 genes) represents more genes than the cluster (5669 genes) attained by k -Means on measurements, one might consider it less reliable than the smaller cluster. However, we passed this cluster to the semi-supervised method for further investigation. The top 500 genes picked based on cosine similarity to the set of known genes, contain only 56% of the 243 remaining known genes. This strongly implies the advantage of the k -means on the measurements (as our preprocessing step) over k -means on the raw gene expression data.

In this work, we present and apply a computational approach using an information-theoretic framework on a dataset of human dermal fibroblast gene expression data, in order to identify novel genes associated with human ageing. First we apply information-theoretic measurements on the data to compute important features for our analysis. We then rank the genes based on these measurements, and perform binary clustering on the genes. Next, we leverage prior knowledge about the subset of genes already known to be age-related to further refine the initially identified genes in the clusters. We presented results showing the performance assessment of the proposed framework, indicating the accuracy and effectiveness of the approach. The results could be further improved with availability of

more known age-related genes. Yet, our results could be used to further guide laboratory experiment studies on the mechanisms of ageing and age-related genes.

ACKNOWLEDGMENT

Work supported in part by grants from the US National Science Foundation (Award #1920920, #1816005), and from CITEr, an NSF I/UCRC.

REFERENCES

- [1] Stegeman, R., and V. M. Weake. "Transcriptional signatures of aging." *Journal of Molecular Biology* 429, no. 16 (2017): 2427-2437.
- [2] Miller, Richard A. "Kleemeier award lecture: are there genes for aging?" *Journals of Gerontology Series A: Biomedical Sciences and Medical Sciences* 54, no. 7 (1999): B297-B307.
- [3] Sahin, Ergün, and Ronald A. DePinho. "Linking functional decline of telomeres, mitochondria and stem cells during ageing." *Nature* 464, no. 7288 (2010): 520-528.
- [4] Williams, George C. "Pleiotropy, natural selection, and the evolution of senescence." *Evolution* (1957): 398-411.
- [5] López-Otín, Carlos, María A. Blasco, Linda Partridge, Manuel Serrano, and Guido Kroemer. "The hallmarks of aging." *Cell* 153, no. 6 (2013): 1194-1217.
- [6] Kenyon, Cynthia J. "The genetics of ageing." *Nature* 464, no. 7288 (2010): 504-512.
- [7] Kowald, Axel, and Thomas BL Kirkwood. "Can ageing be programmed? A critical literature review." *Aging Cell* 15, no. 6 (2016): 986-998.
- [8] Brink, Thore C., Lloyd Demetrius, Hans Lehrach, and James Adjaye. "Age-related transcriptional changes in gene expression in different organs of mice support the metabolic stability theory of aging." *Biogerontology* 10, no. 5 (2009): 549-564.
- [9] Harries, Lorna W., Dena Hernandez, William Henley, Andrew R. Wood, Alice C. Holly, Rachel M. Bradley-Smith, Hanieh Yaghootkar et al. "Human aging is characterized by focused changes in gene expression and deregulation of alternative splicing." *Aging Cell* 10, no. 5 (2011): 868-878.
- [10] Haustead, Daniel J., Andrew Stevenson, Vishal Saxena, Fiona Marriage, Martin Firth, Robyn Silla, Lisa Martin et al. "Transcriptome analysis of human ageing in male skin shows mid-life period of variability and central role of NF-κB." *Scientific Reports* 6, no. 1 (2016): 1-12.
- [11] Yoshida, Shigeo, Beverly M. Yashar, Suja Hiriyanna, and Anand Swaroop. "Microarray analysis of gene expression in the aging human retina." *Investigative Ophthalmology & Visual Science* 43, no. 8 (2002): 2554-2560.
- [12] Angermueller, Christof, Tanel Pärnamaa, Leopold Parts, and Oliver Stegle. "Deep learning for computational biology." *Molecular Systems Biology* 12, no. 7 (2016): 878.
- [13] Fleischer, Jason G., Roberta Schulte, Hsiao H. Tsai, Swati Tyagi, Arkaitz Ibarra, Maxim N. Shokhirev, Ling Huang, Martin W. Hetzer, and Saket Navlakha. "Predicting age from the transcriptome of human dermal fibroblasts." *Genome Biology* 19, no. 1 (2018): 1-8.
- [14] Tacutu, Robi, Daniel Thornton, Emily Johnson, Arie Budovsky, Diogo Barardo, Thomas Craig, Eugene Diana et al. "Human ageing genomic resources: new and updated databases." *Nucleic Acids Research* 46, no. D1 (2018): D1083-D1090.
- [15] Hekimi, Siegfried, and Leonard Guarente. "Genetics and the specificity of the aging process." *Science* 299, no. 5611 (2003): 1351-1354.
- [16] Blagosklonny, Mikhail V. "Prospective treatment of age-related diseases by slowing down aging." *The American Journal of Pathology* 181, no. 4 (2012): 1142-1146.
- [17] Srivastava, Isha, Lokesh Kumar Gahlot, Pooja Khurana, and Yasha Hasija. "dbAARD & AGP: A computational pipeline for the prediction of genes associated with age related disorders." *Journal of Biomedical Informatics* 60 (2016): 153-161.
- [18] De Magalhães, João Pedro, João Curado, and George M. Church. "Meta-analysis of age-related gene expression profiles identifies common signatures of aging." *Bioinformatics* 25, no. 7 (2009): 875-881.
- [19] Avelar, Roberto A., Javier Gómez Ortega, Robi Tacutu, Eleanor J. Tyler, Dominic Bennett, Paolo Binetti, Arie Budovsky et al. "A multidimensional systems biology analysis of cellular senescence in aging and disease." *Genome Biology* 21, no. 1 (2020): 1-22.
- [20] Mi, Huaiyu, Anushya Muruganujan, Dustin Ebert, Xiaosong Huang, and Paul D. Thomas. "PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools." *Nucleic Acids Research* 47, no. D1 (2019): D419-D426.
- [21] "The Gene Ontology resource: enriching a GOld mine." *Nucleic Acids Research* 49, no. D1 (2021): D325-D334.
- [22] Ashburner, Michael, Catherine A. Ball, Judith A. Blake, David Botstein, Heather Butler, J. Michael Cherry, Allan P. Davis et al. "Gene ontology: tool for the unification of biology." *Nature Genetics* 25, no. 1 (2000): 25-29.
- [23] Hinton, Geoffrey E., and Terrence Joseph Sejnowski, eds. *Unsupervised learning: foundations of neural computation*. MIT Press, 1999.
- [24] Bengio, Yoshua, Aaron C. Courville, and Pascal Vincent. "Unsupervised feature learning and deep learning: A review and new perspectives." *CoRR*, abs/1206.5538 1 (2012): 2012.
- [25] Hartigan, John A., and Manchek A. Wong. "Algorithm AS 136: A k-means clustering algorithm." *Journal of the Royal Statistical Society. Series c (applied statistics)* 28, no. 1 (1979): 100-108.
- [26] Mohamadi, Salman, Hamidreza Amindavar, and SM Ali Tayaranian Hosseini. "Arima-garch modeling for epileptic seizure prediction." In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 994-998. IEEE, 2017.
- [27] Shannon, Claude Elwood. "A mathematical theory of communication." *The Bell System Technical Journal* 27, no. 3 (1948): 379-423.
- [28] To, Hien, Kuorong Chiang, and Cyrus Shahabi. "Entropy-based histograms for selectivity estimation." In *Proceedings of the 22nd ACM International Conference on Information & Knowledge Management*, pp. 1939-1948. 2013.
- [29] Thomas M. Cover and Joy A. Thomas. "Elements of information theory." Wiley Publishers, 2006, 978-0-471-24195-9
- [30] Hayflick, Leonard, and Paul S. Moorhead. "The serial cultivation of human diploid cell strains." *Experimental Cell Research* 25, no. 3 (1961): 585-621.
- [31] Belsky, Daniel W., Avshalom Caspi, Renate Houts, Harvey J. Cohen, David L. Corcoran, Andrea Danese, Honalee Harrington et al. "Quantification of biological aging in young adults." *Proceedings of the National Academy of Sciences* 112, no. 30 (2015): E4104-E4110.
- [32] Rahman, Syed Ashiqur, and Donald A. Adjeroh. "Deep learning using convolutional LSTM estimates biological age from physical activity." *Scientific Reports* 9, no. 1 (2019): 1-15.
- [33] Sinaga, Kristina P., and Miin-Shen Yang. "Unsupervised K-means clustering algorithm." *IEEE Access* 8 (2020): 80716-80727.
- [34] Mohamadi, Salman, Donald A. Adjeroh, Behnoush Behi, and Hamidreza Amindavar. "A new framework for spatial modeling and synthesis of genomic sequences." In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 2221-2226. IEEE, 2020.