Predicting the Pro-longevity or Anti-longevity Effect of Model Organism Genes With New Hierarchical Feature Selection Methods

Cen Wan, Alex A. Freitas, and João Pedro de Magalhães

Abstract—Ageing is a highly complex biological process that is still poorly understood. With the growing amount of ageing-related data available on the web, in particular concerning the genetics of ageing, it is timely to apply data mining methods to that data, in order to try to discover novel patterns that may assist ageing research. In this work, we introduce new hierarchical feature selection methods for the classification task of data mining and apply them to ageing-related data from four model organisms: *Caenorhabditis elegans* (worm), *Saccharomyces cerevisiae* (yeast), *Drosophila melanogaster* (fly), and *Mus musculus* (mouse). The main novel aspect of the proposed feature selection methods is that they exploit hierarchical relationships in the set of features (Gene Ontology terms) in order to improve the predictive accuracy of the Naïve Bayes and *1*-Nearest Neighbour (1-NN) classifiers, which are used to classify model organisms' genes into pro-longevity or anti-longevity genes. The results show that our hierarchical feature selection methods, when used together with Naïve Bayes and 1-NN classifiers, obtain higher predictive accuracy than the standard (without feature selection) Naïve Bayes and 1-NN classifiers, respectively. We also discuss the biological relevance of a number of Gene Ontology terms very frequently selected by our algorithms in our datasets.

Index Terms—Ageing, Data mining, Classification, Gene Ontology, Feature selection, Naïve Bayes, *1*-Nearest Neighbour.

1 Introduction

The causes and mechanisms of the biological process of ageing are a mystery that has puzzled humans for a long time. Research has, however, revealed some factors possibly involved in ageing. For instance, caloric restriction extends the longevity of many species [1]. Research has identified several pathways regulating ageing in model organisms, such as insulin/insulin-like growth factor-1 (IGF-1) signaling [2]; and mutations in some DNA repair genes lead to accelerated ageing syndromes [3]. In addition, a low degree of protein oxidative damage is associated with longer-lived species [4], and reactive oxygen species (ROS) may play an important role on the process of ageing. Furthermore, some diseases like cancer may be related to ageing, since cell senescence may be a mechanism of tumour suppression [5].

Despite such findings, ageing is a highly complex biological process which is still poorly understood, and much more research is needed in this area. Due to the great difficulty and ethical issues associated with conducting ageing experiments in humans, research on the biology of ageing is usually done by using model organisms. With the growing amount of ageing-related data on model organisms available on the web, in particular related to the genetics of ageing, it is timely to apply data mining methods to that data, in order to try to discover novel patterns that may assist ageing research.

This work addresses the classification task of data mining [6], where each instance (object being classified) consists of a set of features and a class variable. The goal of a classification algorithm is to build, from a set of training instances (the training set), a classification model that predicts the value (or label) of the class variable for an instance, given the values of the features for that instance. The classification model is then used to predict the class values of a different set of testing instances (the testing set). Hence, the testing set is used to measure the predictive performance, or generalization ability, of the model built from the training set.

In the classification task, when the number of features is large (like in this work), it is common to apply feature selection methods, before applying a classification algorithm to the data. Feature selection methods aim to select a subset of the most relevant and non-redundant features [7], out of all input features, in order to try to improve the predictive accuracy of a classification algorithm. Note that feature selection is a hard computational problem, since the number of candidate solutions is 2^m-1, where m is the number of features.

In this work, the instances being classified are genes from four major model organisms, namely: *C. elegans*, *S. cerevisiae*, *D. melanogaster* and *M. musculus*. Each gene has to be classified into one of two classes: pro-longevity or anti-longevity, based on the values of features indicating whether or not the gene is associated with each of a number of Gene Ontology (GO) terms, where each term refers to a type of biological process. Pro-longevity genes

- *C. Wan and A. A. Freitas are with the School of Computing, University of Kent, Canterbury, United Kingdom. E-mail: {cw439; A.A.Freitas}@kent.ac.uk*
- *João Pedro de Magalhães is with the Integrative Genomics of Ageing Group, University of Liverpool, Liverpool, United Kingdom. E-mail: aging@liverpool.ac.uk*

are those whose decreased expression (due to knock-out, mutations or RNA interference) reduces lifespan and/or whose overexpression extends lifespan; accordingly, anti-longevity genes are those whose decreased expression extends lifespan and/or whose overexpression decreases it [8]. We adopt GO terms as features to predict a gene's effect on longevity because of the widespread use of the GO in gene and protein function prediction and the fact that GO terms were explicitly designed to be valid across different types of organisms [9].

GO terms are organised into a hierarchical structure where, for each GO term t, its ancestors in the hierarchy denote more general terms (i.e. more general biological processes) and its descendants denote more specialized terms than t. It is important to consider the hierarchical relationships among GO terms when performing feature selection, because such relationships encode information about redundancy among GO terms. In particular, if a given gene g is associated with a given GO term t, this logically implies that g is also associated with all ancestors of t in the GO hierarchy. This kind of redundancy can have a negative effect on the predictive accuracy of classification algorithms like Naïve Bayes [6].

This work proposes two new feature selection methods that exploit hierarchical relationships among GO terms, in order to minimize the redundancy in the selected GO terms. We use the term "hierarchical feature selection" to refer to feature selection methods that cope with hierarchical relationships among features. The proposed hierarchical feature selection methods work with the Naïve Bayes and 1-NN (nearest neighbour) classifiers in the context of "lazy learning" [10], [11], where a set of features (GO terms) is selected specifically for each testing instance (gene). This is in contrast to the much more common "eager learning" scenario, where the same set of features is selected and used to classify all testing set instances.

This paper is a major extension of our recent paper [12], where we proposed a hierarchical feature selection method that exploits the hierarchical relationships among GO terms in order to improve the predictive accuracy of Naïve Bayes when classifying *C. elegans* genes into pro-longevity or anti-longevity classes. More precisely, this paper extends our previous paper in four main directions. Firstly, in this paper we propose two new hierarchical feature selection methods; which in our experiments obtained higher predictive accuracy than the method proposed in [12]. Secondly, in this paper we report results for genes of four different model organisms, instead of results for just *C. elegans* genes as in our previous paper. Thirdly, in this paper we discuss the biological relevance (for ageing research) of 20 very frequently selected GO terms; whilst in [12] we just briefly mentioned the biological relevance of two GO terms. Fourthly, in this paper we report results for Naïve Bayes and 1-NN classifiers, instead of just for Naïve Bayes in [12].

1.1 Related Work on the Classification of Ageing-Related Genes

Classification methods are widely adopted in bioinformatics, but there are few studies using classification methods for analyzing data on ageing-related genes, as follows. Freitas et al. [13] addressed the classification of DNA repair genes into ageing-related or non-ageing related, and Fang et al. [14] addressed the classification of ageing-related genes into DNA repair or non-DNA repair genes. Both studies used Gene Ontology (GO) terms as features, in addition to other types of features. Li et al. [15] classified *C. elegans* genes into longevity and non-longevity genes. They used a log-odds score to measure the difference in the frequency with which a given GO term occurs in genes of the longevity and non-longevity classes. Huang et al. [16] predicted the effect of a gene's deletion on the longevity (lifespan) of yeast. The three effect classes were: no effect on lifespan, increased or decreased lifespan. For each deleted gene, they removed its downstream lifespan-related genes from the complete lifespan-related gene network and considered the remaining network as the deletion network for that gene. They computed GO enrichment scores (based on the p-value of a hypergeometric test) as functional features of the deletion networks.

It should be noted that all of these studies coped with each GO term individually, without considering the hierarchical relationships between a GO term and its ancestor and descendant terms in the GO hierarchy – unlike this work, where feature selection takes the GO hierarchy into account.

1.2 Related Work on Hierarchical Feature Selection

In the classification task, hierarchical structure can occur in the class labels to be predicted (creating a hierarchical classification problem) or in the features used as predictors (creating a hierarchical feature selection problem). Many hierarchical classification methods have been proposed in the literature [17]. However, our work follows a very different hierarchical feature selection scenario, where the *features* (in our case GO terms), rather than the class labels, are structured into a hierarchy. Hence, we exploit the GO hierarchy's information for conducting feature selection, but still use conventional classifiers to predict the *flat* class labels.

Hierarchical feature selection for the classification task is a very under-explored area, and works in this area are mainly based on linear models for regression (prediction of continuous variables) or classification [18]–[21]. In general, in these works the system has to find the parameters (feature weights) of a linear model that minimizes both the value of a loss function and the value of a regularization term, which penalizes models with large values of feature weights. The need to minimize the regularization term forces the construction of sparse models, where many features with a weight of 0 are eliminated. These methods achieve hierarchical feature

selection by using regularization terms that consider the feature hierarchy. Briefly, a feature can be added into the set of selected features only if its parent feature is also included in that set [18], [20]. This kind of feature selection methods for linear models is very different from the kind proposed here for classification, which is based on identifying redundant information about feature values in the GO hierarchy. Also, the hierarchical feature selection methods proposed here follow the lazy learning approach, unlike the methods proposed in the above studies.

1.3 Organisation of the Remainder of the Paper

Section 2 briefly reviews the background on the GO, Naïve Bayes and 1-NN, lazy learning and feature selection. Section 3 describes the newly proposed hierarchical feature selection methods. Section 4 presents the computational results. Those results are further discussed in Section 5, which also discusses the biological relevance of very frequently selected GO terms in the context of the biology of ageing. Finally, Section 6 presents the conclusion and future research directions.

2 Background

2.1 The Gene Ontology (GO)

The GO consists of a collection of well-defined terms and hierarchical relationships among terms. These hierarchical relationships are mainly "is_a", generalization/specialization relationships, represented by a directed acyclic graph (DAG). For example, in Figure 1, GO:0008150 (biological process) is the root of the DAG; GO:0051704 (multi-organism process) is one of the child nodes of GO:0008150; and GO:0044364 (disruption of cells of other organism) is one of the parents of node GO:0031640 (killing of cells of other organism).

Note that, according to the GO's "is_a" hierarchical relationships, if a gene is associated with GO term t, then the gene is also associated with all GO terms that are ancestors of t in the hierarchy. Conversely, if a gene is not associated with a given GO term t, then the gene is not associated with any of the GO terms that are descendants of t in the hierarchy. Therefore, the GO's data structure contains some redundant information about the GO terms associated with a given gene. For example, in Figure 1, if we know that a gene is annotated with GO:0044364, the information that the gene is annotated with its ancestor terms GO:0035821, GO:0065008, GO:0065007, GO:0051704 and GO:0008150 can be considered redundant, in the sense that those annotations are logically implied by the GO:0044364 annotation. The notion of redundancy refers to the GO terms associated with an individual gene (an instance in our datasets), which suggests that, in order to exploit hierarchical relationships among GO terms when predicting classes, non-redundant GO terms should be selected for each instance separately. This leads to lazy learning, discussed later.

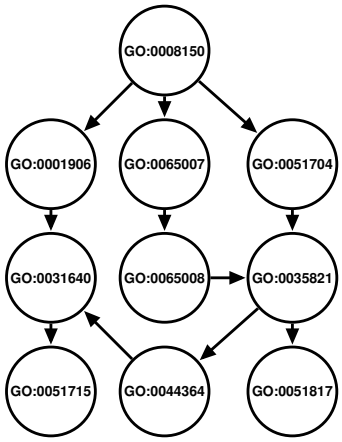

Fig. 1: Example of a small part of Gene Ontology DAG

2.2 Naïve Bayes

Naïve Bayes (NB) is a popular classifier due to its simplicity, relatively powerful predictive ability and its good interpretability. The NB classifier uses the inference formula shown in Equation (1):

$$\mathbf{P}(y|x_1, x_2, ..., x_m) \propto \mathbf{P}(y) \prod_{i=1}^{m} \mathbf{P}(x_i|y) \qquad (1)$$

where m is the number of features and the probability of a class label y given all feature values of an instance is estimated by the product of the prior probability of y times the probability of each feature value x_i given y. Equation (1) is based on the simplifying assumption that features are independent from each other given the class. Clearly, the predictive accuracy of Naïve Bayes is sensitive to the predictive power of individual features; and its accuracy can also be harmed by the use of very redundant features [6]. However, irrelevant or redundant features can be removed by a feature selection method in a preprocessing step, as described later.

2.3 1-Nearest Neighbour (1-NN)

1-NN is a popular "lazy learning" classifier (see Section 2.4). It assigns to a testing instance the class of the training instance which is most similar (or closest) to that testing instance [10], [22], [23]. In our datasets all features are binary, so we use the Jaccard coefficient [24], [25] as the similarity measure in 1-NN, as shown in Equation (2):

$$\mathbf{Jaccard}(i,k) = \frac{m_{11}}{m_{11}+m_{10}+m_{01}} \qquad (2)$$

where m_{11} denotes the number of features with value "1" in both the i_{th} and k_{th} instances simultaneously; m_{10} denotes the number of features with value "1" in the i_{th} instance and value "0" in the k_{th} instance; m_{01} denotes the number of features with value "0" in the i_{th} instance and value "1" in the k_{th} instance. A greater value of the Jaccard coefficient means a higher similarity (closer distance) between the two instances.

2.4 Lazy Learning

In this work we use a "lazy learning" version of Naïve Bayes and the conventional 1-NN algorithm (which naturally performs "lazy learning"). The term "lazy" indicates that the learning process is postponed to the moment when a testing instance is observed and needs to be classified [10], [11]. This is in contrast to the more common "eager learning" approach, where the learning process is performed before any testing instance is observed. Lazy learning builds one specific classifier for each testing instance, whilst eager learning builds one single classifier for all testing instances.

2.5 Feature Selection

Feature selection methods are often used for data preprocessing before classification, in order to filter out redundant and irrelevant features [7]. Feature selection methods can be broadly categorized into the wrapper and the filter approaches. The former uses a classification or regression algorithm to evaluate the performance of candidate features subsets. The latter uses independent feature-evaluation methods, e.g. entropy or chi-square [7], [11], which work regardless of the classification algorithm to be applied to the selected features. The wrapper approach often achieves higher predictive accuracy, since it uses the same algorithm for measuring feature relevance and for classification. However, the wrapper approach is much more computationally expensive than the filter approach. This problem is aggravated in lazy feature selection, since feature selection is performed for each testing instance. Hence, in this work we use the filter approach to select relevant features (GO terms).

Note that the distinction between lazy and eager learning approaches, made earlier, also holds for feature selection methods. Lazy feature selection methods select a set of relevant features for each testing instance separately; whilst eager feature selection methods try to select a single set of features that are relevant to classifying all testing instances.

The main motivation for using the lazy feature selection approach in this work is that it can be used to select a customized set of features (GO terms) for each individual testing instance, which could lead to improved predictive accuracy. Some evidence for this is given in [11], where lazy feature selection has improved Naïve Bayes' predictive accuracy in most experiments. In addition, in [26] the lazy version of a feature elimination approach improved predictive accuracy.

In addition, taking into account that one of the motivations to use feature selection methods is to remove redundant features, in this work a lazy feature selection approach is naturally motivated by the fact that the redundancy associated with the GO's hierarhical relationships refer to individual genes, as discussed earlier. That is, a lazy feature selection method can remove features (GO terms) which are redundant specifically for the classification of a given instance (gene).

2.6 Relevance Measure

As a part of our feature selection method, we use Equation (3) to measure the relevance ($\mathbf{R}$) – or predictive power – of a binary feature X taking value x_1 or x_2.

$$\mathbf{R}(X) = \sum_{i=1}^{n} [\mathbf{P}(y_i|x_1) - \mathbf{P}(y_i|x_2)]^2 \quad (3)$$

where y_i is the *i*-th class and *n* is the number of classes. A general form of Equation (3) was originally used in [27] in the context of nearest neighbour algorithms, and adjusted in [12] to be used as a feature relevance measure for Naïve Bayes. In this work, *n*=2, X is a GO term feature, and Equation (3) is expanded to Equation (4).

$$\begin{aligned}\mathbf{R}(GO) = [\mathbf{P}(Class = Pro \mid GO = Yes) - \mathbf{P}(Class = Pro \mid GO = No)]^2 \\ +[\mathbf{P}(Class = Anti \mid GO = Yes) - \mathbf{P}(Class = Anti \mid GO = No)]^2\end{aligned} \quad (4)$$

This formula calculates the relevance of each GO term as a function of the difference in the conditional probabilities of each class given different values ("Yes" or "No") of a GO term, indicating whether or not a model organism gene is annotated with that GO term.

3 Proposed Hierarchical Feature Selection Methods

We explain the proposed hierarchical feature selection methods in the context of Naïve Bayes, whose predictive accuracy is sensitive to redundant features [28]. However, the proposed methods can be used with any lazy learning classifier, and we will report later results for both Naïve Bayes and 1-NN. As discussed in Section 2.1, the hierarchical relationships among GO terms contain redundancy, but it is not clear which GO terms should be removed to train Naïve Bayes, since feature selection has two goals: minimizing redundancy and selecting features with greater relevance for class prediction. To investigate the relative importance of these two goals, we propose three types of GO hierarchy-based feature selection methods, namely *Select Hierarchical Information-Preserving (HIP) GO Terms*, *Select Most Relevant (MR) GO Terms* and *Select Hierarchical Information-Preserving and Most Relevant (HIP-MR) GO Terms*, as explained below. The *HIP-MR* method was proposed in our recent paper [12], and the other two methods are new. All three methods perform lazy learning, i.e. feature selection is performed separately for each testing instance.

In the description of the feature selection methods in the next subsections, an instance refers to a gene of a model organism, and each instance is described by a set of GO term features. Each feature takes the value "Yes (1)" or "No (0)" for each instance, indicating whether or not (respectively) the gene corresponding to that instance is associated with the corresponding GO term.

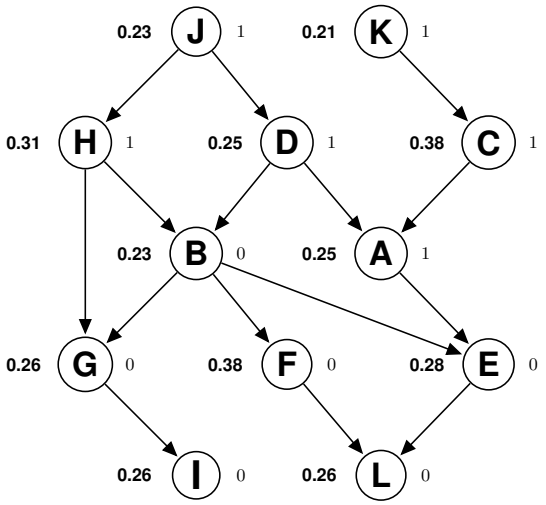

Fig. 2: Example of a small DAG of features

3.1 Select Hierarchical Information-Preserving (HIP) GO Terms

This method focuses only on minimizing the redundancy in the set of selected GO terms, ignoring the relevance values of individual GO terms.

The pseudocode of the HIP method is shown as Algorithm 1, where $Dataset_{<Train>}$ and $Dataset_{<Test>}$ denote the training dataset and testing dataset, and they consist of all GO terms used as features; $Anc(GO_i)$ and $Dec(GO_i)$ denote the set of ancestors and descendants (respectively) of the ith GO term; $Status(GO_i)$ means the selection status ("Selected" or "Removed") of the ith GO term; $Instance_{<n>}$ means the current instance being classified in $Dataset_{<Test>}$; $Value(GO_{i,n})$ denotes the value of GO_i feature ("1" or "0") in that instance; A_{ij} denotes the jth ancestor of the ith GO term; D_{ij} denotes the jth descendant of the ith GO term; $Instance_{<s>}$ means the shorter version of instance n that consists only of GO terms whose status is "Selected".

In the first part of Algorithm 1 (lines: 1-8), it firstly constructs the DAG, finds all ancestors and descendants of each GO term, and initializes the status of each GO term as "Selected". In the second part of Algorithm 1 (lines: 9-24), it performs feature selection for each testing instance in turn, using a lazy learning approach. For each instance, for each GO term GO_i, the algorithm checks its value in that instance. If GO_i has value "1", all its ancestors have their status set to "Removed" – since the value "1" of each ancestor is redundant, being logically implied by the value "1" of GO_i. If GO_i has value "0", all its descendants have their status set to "Removed" – since the value "0" of each descendant is redundant, being logically implied by the value "0" of GO_i.

To show how the second part of Algorithm 1 works, we use as example a hypothetical testing instance with just 12 GO term features, denoted by the letters A-L. Figure 2 shows a small hypothetical DAG specifying the hierarchical relationships among the GO term features of our hypothetical instance. In Figure 2, the relevance and feature value for each GO term is shown on the left (in bold) and on the right (respectively) of the node representing that GO term. Note that the HIP feature selection method uses only information about the GO term feature values and their hierarchical relationships; the GO terms' relevance values are used only by the two other feature selection methods described later.

Algorithm 1 Select Hierarchical Information-Preserving (HIP) GO Terms

```
1: Initialize DAG with all GO terms in Dataset;
2: Initialize Dataset_<Train>;
3: Initialize Dataset_<Test>;
4: for each GO_i in DAG do
5:     Initialize Anc(GO_i) in DAG;
6:     Initialize Dec(GO_i) in DAG;
7:     Initialize Status(GO_i) ← "Selected";
8: end for
9: for each Instance_<n> ∈ Dataset_<Test> do
10:    for each GO_i ∈ DAG do
11:        if Value(GO_{i,n}) = 1 then
12:            for each A_ij ∈ Anc(GO_i) do
13:                Status(A_ij) ← "Removed";
14:            end for
15:        else
16:            for each D_ij ∈ Dec(GO_i) do
17:                Status(D_ij) ← "Removed";
18:            end for
19:        end if
20:    end for
21:    Instance_<s> ← {GO_i : Status(GO_{i,n}) = "Selected"};
22:    NaïveBayes(Dataset_<Train>, Instance_<s>);
23:    Re-assign ∀GO_i : Status(GO_i) ← "Selected";
24: end for
```

With respect to the example DAG in Figure 2, lines 10-20 of Algorithm 1 work as follows. When term A is processed, the selection status of its ancestor terms D, J, C and K will be assigned as "Removed" (lines: 12-14), since the value "1" of A logically implies the value "1" of all of A's ancestors. Analogously, when term B is processed, the selection status of its descendant terms G, I, F, L and E will be assigned as "Removed" (lines: 16-18), since the value of "0" of B logically implies the value of "0" of all of B's descendants.

After processing all terms in the example DAG, the terms selected by the loop in lines 10-20 are A, B and H. Note that these three core GO terms contain the complete hierarchical information associated with all the terms in the DAG of Figure 2, in the sense that the observed values of these three core GO terms logically imply the values of all other GO terms in that DAG.

Next, the current testing instance is reduced to contain only features whose status is "Selected" (line: 21), and that reduced instance is classified by Naïve Bayes (line: 22). Finally, the status of all GO term features is reassigned as "Selected" (line: 23), as a preparation for feature selection for the next testing instance.

3.2 Select Most Relevant (MR) GO Terms

This method performs feature selection considering both the relevance value of individual GO terms and the redundancy among hierarchically-related GO terms. Like the HIP method, for each GO term t in the current instance being classified, MR first identifies the sets of GO terms whose values are implied by the value of t in that instance – i.e. either the ancestors of t, if t has value "1"; or the descendants of t, if t has value "0", for each path from the current node to a root or a leaf node of the GO DAG, depending on whether the current term has value "1" or "0", respectively. Next, MR compares the relevance of t and all terms in the identified GO terms in each path. Among all those terms (including t), MR marks for removal all terms, $except$ the most relevant term. If there are more than one GO terms with the same maximum relevance value in a given path, as a tie-breaking criterion, MR retains the most specific (deepest) term among the set of terms with value "1" or the most generic (shallowest) term among the set of terms with value "0" – since those terms' values logically imply the largest number of other terms' values, among the set of terms being compared.

The pseudocode of the MR method is shown as Algorithm 2, where $\mathbf{R}(GO_i)$ denotes the value of relevance for the ith GO term; $Anc_+(GO_{i,k})$ and $Dec_+(GO_{i,k})$ denote the set of GO terms containing both the ith GO term and its ancestors or descendants (respectively) in the k-th path; *MRT* denotes the most relevant term among the set of GO terms in $Anc_+(GO_{i,k})$ or $Dec_+(GO_{i,k})$; $A_{i,j,k+}$ and $D_{i,j,k+}$ denotes the j^{th} term in $Anc_+(GO_{i,k})$ and $Dec_+(GO_{i,k})$, respectively.

In the first part of Algorithm 2 (i.e. lines 1-9), firstly the DAG will be constructed, then Anc_+ and Dec_+ for each GO term at each path will be initialized, and the relevance ($\mathbf{R}$) value for each GO term will be calculated. In the second part of the algorithm (i.e. lines 10-31), the feature selection process will be conducted for each testing instance using a lazy learning approach.

To show how the second part of Algorithm 2 works, we use again as example the GO DAG shown in Figure 2. When term A (with value "1") is processed (lines: 13 - 18), the GO terms at two paths, i.e. path (a) containing terms J, D and A; and path (b) containing terms K, C and A, are processed. In path (a), the terms having maximum relevance value are D and A; but only term A is selected as the *MRT* (line: 14), since it is deeper than term D in that path. In path (b), only term C is selected as *MRT*, since it has the maximum relevance value. Hence, after processing term A, all terms contained in the two paths have their status set to "Removed", except term C (lines: 15 - 17). Analogously, when term B (with value "0") is processed, the GO terms at three paths, i.e. path (a) containing terms B, G and I; path (b) containing terms B, F and L; and path (c) containing terms B, E and L will be processed. In path (a), both term G and I have maximum relevance value, but G will be selected as the *MRT* (line: 21) since it is shallower than I. In path (b), term F is selected as the *MRT* since it has the maximum relevance value among all terms in that path. In path (c), term E is selected as the *MRT*, since it also has the maximum relevance value. Therefore, after processing term B, the selection status for all terms contained at those three paths will be assigned as "Removed", except terms G, F and E (lines: 22 - 24).

Algorithm 2 Select Most Relevant (MR) GO Terms

```
1: Initialize DAG with all GO terms in Dataset;
2: Initialize Dataset_<Train>;
3: Initialize Dataset_<Test>;
4: for each GO_i in DAG do
5:     Initialize Anc_+(GO_{i,k}) in DAG;
6:     Initialize Dec_+(GO_{i,k}) in DAG;
7:     Initialize Status(GO_i) ← "Selected";
8:     Calculate R(GO_i) in Dataset_<Train>;
9: end for
10: for each Instance_<n> ∈ Dataset_<Test> do
11:     for each GO_i ∈ DAG do
12:         if Value(GO_{i,n}) = 1 then
13:             for each Path_k from GO_i to root in DAG do
14:                 Find MRT in Anc_+(GO_{i,k});
15:                 for each A_{i,j,k+} except MRT do
16:                     Status(A_{i,j,k+}) ← "Removed";
17:                 end for
18:             end for
19:         else
20:             for each Path_k from GO_i to leaf in DAG do
21:                 Find MRT in Dec_+(GO_{i,k});
22:                 for each D_{i,j,k+} except MRT do
23:                     Status(D_{i,j,k+}) ← "Removed";
24:                 end for
25:             end for
26:         end if
27:     end for
28:     Instance_<s>←{GO_i : Status(GO_{i,n})="Selected"};
29:     NaïveBayes(Dataset_<Train>, Instance_<s>);
30:     Re-assign ∀GO_i : Status(GO_i) ← "Selected";
31: end for
```

After processing all GO terms in that example DAG, the selected terms are H, C, G, F and E. Next, the current testing instance is reduced to contain only those five selected features in line 28 of Algorithm 2, and that reduced instance is classified by Naïve Bayes in line 29. Finally, the status of all GO term features is reassigned to "Selected" in line 30, as a preparation for feature selection for the next instance.

Note that, for each set of GO terms being compared when MR decides which terms will have their status set to "Removed", this decision is based both on the relevance values of the GO terms being compared and the redundancy among hierarchically related terms, as explained earlier. Thus, in general the MR method does not select all core GO terms with complete hierarchical information on feature values, as selected by HIP (see Section 3.1). Consider, e.g., the core term B = "0",

which implicitly contains the hierarchical information that terms G, I, F, L and E have value "0". Also, the core term A = "1" implies that terms D, J, C and K have value "1". The GO terms B and A were selected by the HIP method, but neither B nor A is selected by the MR method, because the relevance value of B is smaller than the relevance values of G, F and E; and the relevance value of A is smaller than the relevance value of term C. Hence, we lose the information about the values of nodes B and A, whose values are not implied by the values of terms G, F, E and C (nor implied by any other GO term in the DAG).

On the other hand, the MR method has the advantage that in general it selects GO terms with higher relevance values than the GO terms selected by the HIP method (which ignores GO term relevance values). For instance, in the case of our example DAG in Figure 2, the three GO terms selected by HIP (A, B and H) have on average a relevance value of 0.263, whilst the five GO terms selected by MR (H, C, G, F and E) have on average a relevance value of 0.322.

3.3 Select Hierarchical Information-Preserving and Most Relevant (HIP-MR) GO Terms

Although both HIP and MR select a non-redundant set of GO term features, HIP has the limitation of ignoring the relevance of GO terms, and MR has the limitation that it does not necessarily select all core terms with the complete hierarchical information (terms whose observed values logically imply the values of all other GO terms for the current instance). The HIP-MR method addresses these limitations, by both considering GO term relevance (like MR) and selecting all core terms with the complete hierarchical information (like HIP). The price paid for considering both these criteria is that, unlike HIP and MR, HIP-MR typically selects a large subset of GO term features having some redundancy (although less redundancy than the original full set of features), as will be discussed later.

For each GO term t in the instance being classified, HIP-MR first identifies the GO terms whose values are implied by the value of t in the instance – i.e. the set of terms which are ancestors or descendants of t, depending on whether t has value "1" or "0", respectively. Then, HIP-MR removes GO terms by combining ideas from the HIP and MR methods, as follows. If GO term t has value "1", HIP-MR removes the ancestors of t whose relevance values are not greater than the relevance value of t. If the GO term t has value "0", HIP-MR removes the descendants of t whose relevance values are not greater than the relevance value of t.

Therefore, HIP-MR selects a set of core GO terms where each selected term has the property(ies) of being needed to preserve the complete hierarchical information associated the instance being classified (the kind of GO term selected by HIP) or has a relatively high relevance in the context of its ancestors or descendants (the kind of GO term selected by MR). Hence, the set of GO terms selected by the HIP-MR method tends to include the union of the sets of GO terms selected by the HIP and MR methods separately, making HIP-MR a considerably more "inclusive" feature selection method.

Algorithm 3 Select Hierarchical Information-Preserving and Most Relevant (HIP-MR) GO Terms

1: Initialize DAG with all GO terms in Dataset;
2: Initialize $Dataset_{<Train>}$;
3: Initialize $Dataset_{<Test>}$;
4: **for** each GO_i in DAG **do**
5: Initialize $Anc(GO_i)$ in DAG;
6: Initialize $Dec(GO_i)$ in DAG;
7: Initialize $Status(GO_i) \leftarrow$ "Selected";
8: Calculate $\mathbf{R}(GO_i)$ in $Dataset_{<Train>}$;
9: **end for**
10: **for** each $Instance_{<n>} \in Dataset_{<Test>}$ **do**
11: **for** each $GO_i \in$ DAG **do**
12: **if** $Value(GO_{i,n})$ = 1 **then**
13: **for** each $A_{ij} \in Anc(GO_i)$ **do**
14: **if** $\mathbf{R}(A_{ij}) \leq \mathbf{R}(GO_i)$ **then**
15: $Status(A_{ij}) \leftarrow$ "Removed";
16: **end if**
17: **end for**
18: **else**
19: **for** each $D_{ij} \in Dec(GO_i)$ **do**
20: **if** $\mathbf{R}(D_{ij}) \leq \mathbf{R}(GO_i)$ **then**
21: $Status(D_{ij}) \leftarrow$ "Removed";
22: **end if**
23: **end for**
24: **end if**
25: **end for**
26: Instance$_{<s>}\leftarrow\{GO_i : Status(GO_{i,n})$="Selected"$\}$;
27: NaïveBayes$(Dataset_{<Train>}, Instance_{<s>})$;
28: Re-assign $\forall GO_i : Status(GO_i) \leftarrow$ "Selected";
29: **end for**

The pseudocode is shown as Algorithm 3. In the first part of the algorithm (lines: 1-9), firstly the DAG is constructed, the ancestors and descendants of each GO term are found, and the relevance value of each GO term is calculated by Equation (4). In the second part of the algorithm (lines: 10-29), the feature selection process is carried out by combining ideas of the HIP and MR methods, as explained earlier, for each testing instance.

In the case of our example GO DAG in Figure 2, when GO term A (with value "1") is processed, its relevance value is compared with the relevance values of all its ancestor terms J, D, C and K. Then, terms J, D and K are marked for removal, since their relevance values are not greater than the relevance of A. Next, when term B (with value "0") is processed, none of its descendant terms is marked for removal, since their relevance values are greater than the relevance value of B. This process is repeated for all other GO term features in the instance being classified. At the end of this process, the selected GO terms are: H, C, B, A, G, F and E. Note that in this example HIP-MR selects all GO terms selected by HIP or MR. Actually, as will be shown in Section 4,

HIP-MR tends to select substantially more GO terms than the number of GO terms selected by HIP and MR together. Note that, although HIP-MR selects a GO term subset with less redundancy than the original full GO term set, the terms selected by HIP-MR still have some redundancy, unlike the terms selected by HIP and MR. This is because HIP-MR can select a redundant term t if t has higher relevance than another selected term logically implying t. E.g., in the above example, HIP-MR selects term C, which is redundant with respect to selected term A, since C has higher relevance than A.

4 Computational Experiments

4.1 Data Preparation

We constructed four datasets with data about the effect of genes on an organism's longevity, by integrating data from the Human Ageing Genomic Resources (HAGR) GenAge database (Build 16) [29] and the Gene Ontology (GO) database (version: 2013-08-07) [9]. HAGR provides longevity-related gene data for four model organisms, i.e. *C. elegans*, *S. cerevisiae*, *D. melanogaster* and *M. musculus*. We created one dataset for each of these model organisms. We used the "EntrezID" number for each gene in GenAge to retrieve the list of all GO terms annotated for that gene, by using the "gene2go" file [30], version: 2013-08-06. We used only the biological process GO terms, which can be more naturally interpreted by biologists as predictors of longevity, by comparison with the molecular function and cellular component GO terms. After mapping all genes to their annotated GO terms, we use the "is_a" relationship between GO terms to find all ancestors of each GO term. The final dataset for each model organism consists of one instance for each gene, where each instance consists of a set of binary GO term features (indicating whether or not the gene is annotated with each GO term) and a class value (Pro- or Anti-Longevity).

Additional information about the created datasets is shown in Table 1. The initial number of GO terms is the number of GO terms (features) in the dataset before removing GO terms with frequency of occurence below a user-defined threshold and before running the feature selection methods. The GO term frequency threshold is a user-defined parameter adopted for mitigating the overfitting problem that would happen by constructing models using GO terms with a very low frequency of occurrence (i.e. features whose value "1" occurs in very few instances), since those features do not have a good generalization ability. Hence, it is necessary to investigate what is the most appropriate threshold for the minimum number of occurrences of a GO term. In this work, we experimented with all integer threshold values between 3 and 10. Thus, for each model organism, there are 8 different dataset versions – each version using a different GO term frequency threshold.

TABLE 1: Detailed Information about the Created Datasets

	Caenorhabditis elegans	*Saccharomyces cerevisiae*	*Drosophila melanogaster*	*Mus musculus*
Initial Number of GO Terms	1528	1708	1595	2625
Initial Number of Instances	566	293	121	89
Number (%) of Pro-Longevity Instances	203 (35.9 %)	41 (14.0 %)	81 (66.9 %)	63 (70.8 %)
Number (%) of Anti-Longevity Instances	363 (64.1 %)	252 (86.0 %)	40 (33.1 %)	26 (29.2 %)

4.2 Experimental Methodology

Generally, in our datasets, the distribution of instances belonging to the two classes is imbalanced, as shown in Table 1. Hence, we evaluate the predictive performance of classifiers by using the value of Geometric mean (Gmean), defined as $\mathbf{Gmean} = \sqrt{Sens \times Spec}$, because it takes into account the balance of the classifiers' sensitivity (Sens) and specificity (Spec) [31]. Sensitivity means the proportion of pro-longevity genes that were correctly predicted as pro-longevity, and specificity means the proportion of anti-longevity genes that were correctly predicted as anti-longevity in the testing dataset [32].

For all classifiers evaluated in this paper, the reported values of Sens, Spec and Gmean were computed by a well-known 10-fold cross validation procedure [6].

4.3 Results

We firstly report results comparing the Gmean of four versions of Naïve Bayes (NB), namely standard-NB (without using any feature selection method) and HIP-NB, MR-NB and HIP-MR-NB, which denote NB applied on the set of features selected by the respective hierarchical feature selection method (HIP, MR or HIP-MR). The results are shown in Table 2, where the bold figures denote the highest Gmean value in the corresponding dataset version for each value of the GO term frequency threshold. The figures after "$\pm$" are standard errors.

In terms of average Gmean value among all dataset versions for the four model organisms, MR-NB obtained the highest value, i.e. 61.9%, which is slightly higher than HIP-NB's value, i.e. 61.6%. In terms of performance on individual model organisms, MR-NB obtained the highest Gmean value (averaged over all threshold values) in the *C. elegans* and *S. cerevisiae* datasets; and it obtained the second highest Gmean value in the other two datasets. Conversely, HIP-NB obtained the highest average Gmean value in the *D. melanogaster* and *M. musculus* datasets; and it obtained the second highest Gmean value in the *C. elegans* dataset. In summary, both MR-NB and HIP-NB have been successful feature selection methods, obtaining better results than both the baseline standard Naïve Bayes (without feature selection) and the HIP-MR-NB feature selection method.

The main reasons for the inferior performance of HIP-MR-NB seem to be that it tends to select a much larger

TABLE 2: Sensitivity (%), Specificity (%) & Geometric Mean (%) of Naïve Bayes Classifier

Caenorhabditis elegans												
	Standard-NB			**HIP-NB**			**MR-NB**			**HIP-MR-NB**		
Thre.	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*
T3	52.6 ± 3.0	63.2 ± 4.4	57.7	53.7 ± 3.7	70.8 ± 3.5	61.7	57.9 ± 3.1	68.8 ± 4.7	**63.1**	51.6 ± 3.4	69.4 ± 3.8	59.8
T4	55.8 ± 3.9	61.8 ± 2.5	58.7	54.2 ± 3.9	71.2 ± 2.3	62.1	57.4 ± 4.0	68.8 ± 2.0	**62.8**	51.1 ± 5.0	68.1 ± 2.2	59.0
T5	54.7 ± 2.9	61.5 ± 3.9	58.0	52.1 ± 3.0	72.6 ± 2.4	61.5	57.4 ± 2.0	72.2 ± 3.0	**64.4**	51.6 ± 2.7	69.4 ± 3.8	59.8
T6	60.0 ± 4.6	59.4 ± 4.0	59.7	53.7 ± 4.6	71.9 ± 2.9	62.1	57.9 ± 2.9	69.8 ± 3.5	**63.6**	56.3 ± 4.8	67.4 ± 3.9	61.6
T7	56.8 ± 4.7	61.5 ± 3.1	59.1	55.8 ± 4.7	71.2 ± 2.7	**63.0**	55.3 ± 3.6	71.2 ± 2.6	62.7	56.3 ± 3.5	66.0 ± 2.3	61.0
T8	56.8 ± 4.8	56.8 ± 2.2	56.8	52.1 ± 5.2	70.0 ± 2.5	60.4	53.2 ± 3.4	70.4 ± 3.4	**61.2**	56.8 ± 4.3	62.4 ± 2.9	59.5
T9	57.9 ± 4.0	59.2 ± 4.0	58.5	51.1 ± 5.0	69.0 ± 3.0	**59.4**	48.9 ± 3.3	71.4 ± 3.3	59.1	53.2 ± 3.6	62.4 ± 4.2	57.6
T10	58.4 ± 4.8	57.1 ± 2.4	57.7	48.4 ± 3.6	69.7 ± 2.8	58.1	50.0 ± 3.5	71.4 ± 1.4	**59.7**	52.1 ± 4.1	61.3 ± 2.2	56.5
Ave.	56.6 ± 4.1	60.1 ± 3.3	58.3	52.6 ± 4.2	70.8 ± 2.8	61.0	54.8 ± 3.2	70.5 ± 3.0	**62.1**	53.6 ± 3.9	65.8 ± 3.2	59.4
Saccharomyces cerevisiae												
	Standard-NB			**HIP-NB**			**MR-NB**			**HIP-MR-NB**		
Thre.	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*
T3	45.0 ± 7.3	82.2 ± 2.7	60.8	40.0 ± 8.5	92.3 ± 2.0	60.8	60.0 ± 8.5	80.3 ± 2.4	**69.4**	12.5 ± 7.7	98.1 ± 1.1	35.0
T4	55.0 ± 8.2	83.1 ± 3.4	**67.6**	47.5 ± 10.2	92.3 ± 2.2	66.2	52.5 ± 8.7	80.7 ± 2.9	65.1	20.0 ± 9.0	97.6 ± 1.3	44.2
T5	50.0 ± 6.5	82.0 ± 3.3	64.0	40.0 ± 7.6	90.3 ± 2.3	60.1	52.5 ± 6.9	80.1 ± 3.2	**64.8**	17.5 ± 6.5	97.6 ± 2.0	41.3
T6	45.0 ± 5.0	77.2 ± 2.7	58.9	40.0 ± 4.1	87.4 ± 2.7	59.1	55.0 ± 6.2	79.1 ± 3.5	**66.0**	17.5 ± 3.8	96.1 ± 1.9	41.0
T7	45.0 ± 8.2	78.6 ± 3.2	59.5	37.5 ± 5.6	91.3 ± 1.9	58.5	50.0 ± 6.5	85.0 ± 3.1	**65.2**	20.0 ± 7.3	97.1 ± 1.3	44.1
T8	45.0 ± 7.3	80.8 ± 2.1	60.3	35.0 ± 5.5	91.1 ± 1.7	56.5	45.0 ± 6.2	82.8 ± 3.8	**61.0**	15.0 ± 5.5	96.6 ± 1.2	38.1
T9	47.5 ± 5.8	78.6 ± 2.0	61.1	42.5 ± 9.9	88.1 ± 2.6	**61.2**	42.5 ± 6.5	79.1 ± 2.6	58.0	12.5 ± 5.6	96.0 ± 1.0	34.6
T10	52.5 ± 7.9	78.6 ± 1.9	**64.2**	37.5 ± 6.7	89.1 ± 1.5	57.8	37.5 ± 5.6	80.6 ± 1.2	55.0	20.0 ± 6.2	96.0 ± 1.5	43.8
Ave.	48.1 ± 7.0	80.1 ± 2.7	62.1	40.0 ± 7.3	90.2 ± 2.1	60.0	49.4 ± 6.9	81.0 ± 2.8	**63.1**	16.9 ± 6.5	96.9 ± 1.4	40.3
Drosophila melanogaster												
	Standard-NB			**HIP-NB**			**MR-NB**			**HIP-MR-NB**		
Thre.	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*
T3	73.8 ± 4.4	51.3 ± 10.5	61.5	75.0 ± 3.7	53.8 ± 10.0	**63.5**	70.0 ± 6.2	53.8 ± 10.2	61.4	81.3 ± 4.7	41.0 ± 7.6	57.7
T4	73.8 ± 4.4	48.7 ± 10.8	60.0	75.0 ± 2.6	51.3 ± 10.4	62.0	70.0 ± 3.8	59.0 ± 10.3	**64.3**	85.0 ± 4.5	43.6 ± 8.2	60.9
T5	70.0 ± 5.0	43.6 ± 4.5	55.2	76.3 ± 3.9	48.7 ± 3.2	**61.0**	70.0 ± 6.0	48.7 ± 4.1	58.4	80.0 ± 4.2	38.5 ± 4.0	55.5
T6	72.5 ± 4.5	43.6 ± 9.0	56.2	80.0 ± 3.3	46.2 ± 8.8	**60.8**	76.3 ± 4.4	43.6 ± 8.2	57.7	85.0 ± 3.1	35.9 ± 6.6	55.2
T7	76.3 ± 5.7	43.6 ± 6.3	57.7	80.0 ± 4.2	53.8 ± 8.8	**65.6**	76.3 ± 3.9	53.8 ± 8.0	64.1	83.8 ± 3.3	43.6 ± 6.3	60.4
T8	72.5 ± 4.5	43.6 ± 9.0	56.2	78.8 ± 5.6	48.7 ± 6.7	**61.9**	78.8 ± 5.6	43.6 ± 8.2	58.6	83.8 ± 5.3	38.5 ± 7.6	56.8
T9	75.0 ± 5.6	43.6 ± 10.2	57.2	82.5 ± 6.0	48.7 ± 10.4	**63.4**	75.0 ± 5.3	48.7 ± 10.4	60.4	77.5 ± 5.2	43.6 ± 8.7	58.1
T10	71.3 ± 4.6	41.0 ± 8.5	54.1	81.3 ± 3.8	43.6 ± 7.5	59.5	77.5 ± 4.5	46.2 ± 9.0	**59.8**	75.0 ± 4.6	46.2 ± 8.2	58.9
Ave.	73.2 ± 4.8	44.9 ± 8.6	57.3	78.6 ± 4.1	49.4 ± 8.2	**62.2**	74.2 ± 5.0	49.7 ± 8.6	60.6	81.4 ± 4.4	41.4 ± 7.2	57.9
Mus musculus												
	Standard-NB			**HIP-NB**			**MR-NB**			**HIP-MR-NB**		
Thre.	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*
T3	79.4 ± 5.6	53.8 ± 8.5	65.4	82.5 ± 5.3	57.7 ± 9.7	**69.0**	63.5 ± 5.4	61.5 ± 8.9	62.5	96.8 ± 2.1	30.8 ± 7.4	54.6
T4	73.0 ± 6.4	46.2 ± 8.3	58.1	82.5 ± 5.5	57.7 ± 8.3	**69.0**	66.7 ± 6.2	61.5 ± 5.7	64.0	95.2 ± 3.5	34.6 ± 6.6	57.4
T5	81.0 ± 3.9	42.3 ± 8.3	58.5	82.5 ± 4.3	50.0 ± 9.9	**64.2**	71.4 ± 7.3	57.7 ± 10.9	**64.2**	93.7 ± 3.2	34.6 ± 6.8	56.9
T6	71.4 ± 2.8	46.2 ± 11.0	57.4	81.0 ± 4.2	46.2 ± 9.9	**61.2**	68.3 ± 5.9	53.8 ± 6.6	60.6	95.2 ± 2.4	34.6 ± 10.2	57.4
T7	74.6 ± 5.2	42.3 ± 6.2	56.2	84.1 ± 4.3	46.2 ± 8.2	**62.3**	69.8 ± 6.4	53.8 ± 9.6	61.3	90.5 ± 3.6	34.6 ± 8.4	56.0
T8	69.8 ± 5.7	42.3 ± 12.2	54.3	76.2 ± 6.5	50.0 ± 12.4	**61.7**	74.6 ± 4.5	50.0 ± 11.2	61.1	92.1 ± 2.6	34.6 ± 11.7	56.5
T9	74.6 ± 6.3	50.0 ± 14.2	61.1	81.0 ± 6.2	46.2 ± 11.1	61.2	73.0 ± 7.2	53.8 ± 12.2	**62.7**	92.1 ± 4.4	38.5 ± 12.7	59.5
T10	71.4 ± 5.6	57.7 ± 13.1	**64.2**	76.2 ± 6.0	42.3 ± 12.7	56.8	74.6 ± 5.5	46.2 ± 12.0	58.7	85.7 ± 5.6	30.8 ± 11.3	51.4
Ave.	74.4 ± 5.2	47.6 ± 10.2	59.4	80.8 ± 5.3	49.5 ± 10.3	**63.2**	70.2 ± 6.1	54.8 ± 9.6	61.9	92.7 ± 3.4	34.1 ± 9.4	56.2

TABLE 3: Average Number of Selected GO Terms by Feature Selection Method for the 4 Model Organisms

	Caenorhabditis elegans			*Saccharomyces cerevisiae*			*Drosophila melanogaster*			*Mus musculus*		
Thre.	*HIP-NB*	*MR-NB*	*HIP-MR-NB*	*HIP-NB*	*MR-NB*	*HIP-MR-NB*	*HIP-NB*	*MR-NB*	*HIP-MR-NB*	*HIP-NB*	*MR-NB*	*HIP-MR-NB*
T3	65.3	140.7	265.4	54.3	99.6	218.7	73.3	121.4	228.2	120.6	178.5	330.3
T4	58.6	113.2	223.6	49.4	89.8	185.3	65.2	101.5	190.7	107.4	139.5	264.4
T5	55.7	99.7	201.9	44.5	73.2	151.3	60.4	88.4	164.7	93.1	114.8	215.9
T6	52.4	87.7	182.2	41.5	66.7	134.3	51.9	73.7	139.7	81.8	96.1	188.8
T7	51.1	84.0	170.0	37.3	57.2	117.1	47.2	68.4	122.8	71.8	78.3	160.9
T8	49.4	73.0	152.6	34.2	50.5	106.0	44.4	62.1	108.9	65.7	73.4	145.1
T9	46.7	67.0	142.9	33.2	46.0	98.5	41.2	55.3	97.8	61.0	68.0	133.7
T10	45.5	63.3	135.9	31.7	43.1	85.9	38.8	47.6	87.1	55.5	60.7	117.6
Ave.	53.1	91.1	184.3	40.8	65.8	137.1	52.8	77.3	142.5	82.1	101.2	194.6

number of GO term features, by comparison with HIP and MR (see Section 3.3) and such a larger feature subset contains some redundancy among hierarchically-related features (unlike the non-redundant features selected by HIP and MR), as explained earlier. As evidence for this, Table 3 shows the average number of features selected by each method for each model organism and each dataset version. Each value in the table is the mean number of selected features over the 10 cross-validation iterations. As shown in Table 3, the number of features selected by HIP-MR is always larger (and in most cases substantially larger) than the sum of the number of features selected by HIP and MR. Such larger feature subsets contain many redundant features, reducing the predictive accuracy of Naïve Bayes with the HIP-MR method.

It is also worth observing the effect of different values of the GO term frequency threshold in the Gmean value obtained by the different versions of Naïve Bayes in Table 2. Out of 16 cases (4 versions of NB times 4 model organisms), in 9 cases the threshold value leading to the highest Gmean value was either 3 or 4, which are the most inclusive threshold values - i.e. the values that lead to the largest number of GO term features used as input by the different versions of Naïve Bayes.

As a final note, we also conducted experiments comparing HIP-MR-NB with a simple univariate feature selection method that ranks all features and selects the top-ranked ones. The results of these experiments are reported in [12].

5 Discussion

5.1 Statistical Analysis of the Comparison of Results for the Feature Selection Methods

We chose the combination of Friedman test and Holm post-hoc test as the statistical significance tests applied on the Geometric mean values obtained for the 32 datasets used in our experiments (8 different GO term frequency thresholds times 4 model organisms). The Friedman test is a nonparametric test based on the rankings of each classifier's predictive performance on each dataset, which avoids the problems associated with the assumption of normal distribution made by the t-test and ANOVA [31], [33]. The Holm post-hoc method is used for coping with the multiple-comparison problem when using significance tests, by adjusting the p-values for individual pairwise comparisons. Demsăr [34] argues that in the case of multiple comparisons between one control classifier and other classifiers, the Holm post-hoc test is more powerful than the Nemenyi post-hoc test. We selected MR-NB as the control method, since it obtains the highest average Gmean value (averaged over the 32 dataset versions) among the four methods being compared in Table 2. Comparing the Gmean values of MR-NB as the control method against the values of each of the other methods, at the significance level of 5%, there is no significant difference between the Gmean values of MR-NB and HIP-NB; but MR-NB significantly outperforms both standard-NB and HIP-MR-NB.

Comparing the predictive accuracy of HIP-NB, MR-NB and HIP-MR-NB, it seems that redundancy among the selected GO terms tends to decrease NB's predictive accuracy. As evidence for this, HIP-MR-NB, which selects a set of GO terms with some redundancy, performed considerably worse than MR-NB and HIP-NB, which do not select redundant features. Also, the core GO terms containing the complete hierarchical information in the GO DAG for a given instance seem valuable for prediction, since HIP-NB, which selects such non-redundant core GO terms regardless of relevance, performed about as well as MR-NB.

Finally, we briefly report results of experiments with the 1-NN classifier, in Table 4. HIP-1NN obtained the highest average Gmean value averaged over the 32 dataset versions, viz. 61.4%. It was also the best method in the *S. cerevisiae* and *D. melanogaster* datasets and the second best method in the other two datasets. Then we chose HIP-1NN as the control method to be compared with the other classifiers. The results of the Friedman and Holm post-hoc tests at the 5% significance level show that there is no statistically significant difference between HIP-1NN and MR-1NN, nor between HIP-1NN and standard-1NN; but HIP-1NN significantly outperformed HIP-MR-1NN. This confirms that the non-redundant set of core GO terms containing the complete hierarchical information in the GO DAG is valuable for prediction.

TABLE 4: Sensitivity (%), Specificity (%) & Geometric Mean (%) of 1-Nearest Neighbour Classifier

Caenorhabditis elegans

	Standard-1NN			**HIP-1NN**			**MR-1NN**			**HIP-MR-1NN**		
Thre.	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*
T3	44.2 ± 2.5	71.9 ± 2.2	56.4	38.9 ± 3.5	76.7 ± 2.2	54.6	46.3 ± 4.6	77.8 ± 2.6	**60.0**	40.0 ± 3.9	74.7 ± 3.7	54.7
T4	41.6 ± 3.3	68.8 ± 1.9	53.5	36.3 ± 3.5	76.4 ± 2.8	52.7	42.1 ± 3.7	74.3 ± 1.9	55.9	44.7 ± 3.9	74.0 ± 3.0	**57.5**
T5	43.2 ± 4.1	70.8 ± 2.1	55.3	39.5 ± 4.5	76.0 ± 1.8	54.8	47.4 ± 2.7	77.4 ± 2.3	**60.6**	40.5 ± 4.8	75.0 ± 2.7	55.1
T6	41.6 ± 3.1	68.8 ± 1.7	53.5	42.1 ± 4.3	76.4 ± 1.3	**56.7**	42.1 ± 3.3	75.3 ± 2.9	56.3	37.9 ± 3.7	72.6 ± 2.3	52.5
T7	45.3 ± 3.1	70.1 ± 2.3	56.4	46.3 ± 1.7	75.7 ± 2.5	59.2	46.3 ± 2.7	79.2 ± 2.0	**60.6**	41.6 ± 2.7	71.9 ± 3.0	54.7
T8	41.6 ± 4.3	68.6 ± 3.2	53.4	46.8 ± 4.0	73.9 ± 3.2	**58.8**	44.2 ± 3.7	76.3 ± 3.4	58.1	41.6 ± 4.0	72.8 ± 3.8	55.0
T9	41.1 ± 4.1	68.6 ± 2.4	53.1	48.4 ± 4.2	74.6 ± 2.1	**60.1**	41.6 ± 5.1	75.3 ± 2.0	56.0	40.5 ± 4.6	71.8 ± 1.8	53.9
T10	42.6 ± 4.8	67.9 ± 2.2	53.8	43.7 ± 3.8	72.8 ± 2.3	56.4	46.8 ± 3.2	74.9 ± 2.7	**59.2**	38.9 ± 3.1	72.5 ± 2.2	53.1
Ave.	42.7 ± 3.7	69.4 ± 2.3	54.4	42.8 ± 3.7	75.3 ± 2.3	56.7	44.6 ± 3.6	76.3 ± 2.5	**58.3**	40.7 ± 3.8	73.2 ± 2.8	54.6

Saccharomyces cerevisiae

	Standard-1NN			**HIP-1NN**			**MR-1NN**			**HIP-MR-1NN**		
Thre.	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*
T3	35.0 ± 8.5	88.0 ± 2.7	55.5	40.0 ± 7.6	90.4 ± 1.6	**60.1**	35.0 ± 10.7	89.9 ± 1.8	56.1	30.0 ± 7.3	93.3 ± 1.5	52.9
T4	32.5 ± 6.5	89.4 ± 1.6	53.9	37.5 ± 9.3	91.3 ± 2.6	**58.5**	37.5 ± 10.7	91.3 ± 1.8	**58.5**	27.5 ± 5.8	93.2 ± 2.0	50.6
T5	27.5 ± 7.9	89.8 ± 2.4	49.7	40.0 ± 6.7	89.8 ± 2.3	**59.9**	40.0 ± 7.6	85.9 ± 2.9	58.6	17.5 ± 3.8	91.3 ± 2.2	40.0
T6	35.0 ± 8.5	89.8 ± 2.5	56.1	50.0 ± 8.3	89.8 ± 2.2	**67.0**	42.5 ± 6.5	88.3 ± 3.2	61.3	35.0 ± 6.7	92.7 ± 2.5	57.0
T7	30.0 ± 6.2	90.8 ± 2.8	52.2	42.5 ± 6.5	90.8 ± 1.7	**62.1**	35.0 ± 8.5	88.8 ± 3.1	55.7	30.0 ± 5.0	94.7 ± 1.5	53.3
T8	35.0 ± 9.3	87.2 ± 3.2	55.2	45.0 ± 6.2	90.1 ± 1.7	**63.7**	45.0 ± 9.0	90.1 ± 2.2	**63.7**	35.0 ± 6.7	92.6 ± 1.3	56.9
T9	32.5 ± 6.5	85.6 ± 3.1	52.7	42.5 ± 7.5	84.1 ± 2.3	**59.8**	35.0 ± 6.7	84.1 ± 2.1	54.3	22.5 ± 4.5	89.6 ± 2.8	44.9
T10	32.5 ± 8.4	87.6 ± 2.3	53.4	42.5 ± 7.5	88.6 ± 2.3	**61.4**	37.5 ± 6.7	85.6 ± 2.3	56.7	25.0 ± 6.5	88.6 ± 2.2	47.1
Ave.	32.5 ± 7.7	88.5 ± 2.6	53.6	42.5 ± 7.5	89.4 ± 2.1	**61.6**	38.4 ± 8.3	88.0 ± 2.4	58.1	27.8 ± 5.8	92.0 ± 2.0	50.3

Drosophila melanogaster

	Standard-1NN			**HIP-1NN**			**MR-1NN**			**HIP-MR-1NN**		
Thre.	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*
T3	66.3 ± 7.2	56.4 ± 6.3	61.1	73.8 ± 6.6	56.4 ± 6.9	**64.5**	76.3 ± 3.9	53.8 ± 2.8	64.1	77.5 ± 6.4	51.3 ± 7.6	63.1
T4	68.8 ± 3.4	59.0 ± 8.5	63.7	73.8 ± 6.0	56.4 ± 6.5	64.5	71.3 ± 4.6	66.7 ± 7.5	**69.0**	68.8 ± 4.7	51.3 ± 6.7	59.4
T5	66.3 ± 2.7	53.8 ± 10.0	59.7	75.0 ± 4.6	53.8 ± 9.6	**63.5**	62.5 ± 6.5	61.5 ± 8.2	62.0	70.0 ± 7.0	53.8 ± 10.0	61.4
T6	67.5 ± 8.6	59.0 ± 9.3	**63.1**	73.8 ± 6.6	51.3 ± 10.0	61.5	72.5 ± 5.5	46.2 ± 11.3	57.9	70.0 ± 6.8	53.8 ± 10.3	61.4
T7	67.5 ± 4.2	56.4 ± 11.2	61.7	75.0 ± 3.7	56.4 ± 10.6	**65.0**	72.5 ± 4.5	48.7 ± 9.9	59.4	72.5 ± 4.5	53.8 ± 9.0	62.5
T8	71.3 ± 2.7	51.3 ± 4.9	60.5	71.3 ± 4.2	48.7 ± 10.0	58.9	73.8 ± 2.9	53.8 ± 8.4	**63.0**	70.0 ± 3.8	56.4 ± 8.7	62.8
T9	68.8 ± 4.3	56.4 ± 6.3	62.3	73.8 ± 5.7	53.8 ± 6.0	**63.0**	70.0 ± 5.3	51.3 ± 4.9	59.9	67.5 ± 4.2	53.8 ± 7.6	60.3
T10	68.8 ± 5.4	53.8 ± 7.1	60.8	72.5 ± 4.5	53.8 ± 8.2	**62.5**	65.0 ± 5.2	38.5 ± 10.0	50.0	68.8 ± 5.7	56.4 ± 9.2	62.3
Ave.	68.2 ± 4.8	55.8 ± 8.0	61.6	73.6 ± 5.2	53.8 ± 8.5	**62.9**	70.5 ± 4.8	52.6 ± 7.9	60.7	70.6 ± 5.4	53.8 ± 8.6	61.6

Mus musculus

	Standard-1NN			**HIP-1NN**			**MR-1NN**			**HIP-MR-1NN**		
Thre.	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*	*Sen.*	*Spe.*	*Gmean*
T3	69.8 ± 4.8	57.7 ± 11.1	63.5	82.5 ± 5.1	61.5 ± 5.7	**71.2**	73.0 ± 5.9	50.0 ± 12.2	60.4	85.7 ± 4.2	38.5 ± 10.3	57.4
T4	74.6 ± 5.1	61.5 ± 10.1	**67.7**	73.0 ± 5.6	50.0 ± 12.2	60.4	68.3 ± 7.9	57.7 ± 9.7	62.8	81.0 ± 3.0	46.2 ± 8.4	61.2
T5	74.6 ± 5.6	61.5 ± 7.4	**67.7**	88.9 ± 3.5	50.0 ± 11.2	66.7	66.7 ± 5.5	65.4 ± 10.5	66.0	87.3 ± 5.5	42.3 ± 10.9	60.8
T6	79.4 ± 4.9	57.7 ± 6.6	67.7	84.1 ± 6.0	57.7 ± 9.6	**69.7**	79.4 ± 5.6	57.7 ± 9.9	67.7	85.7 ± 4.6	50.0 ± 8.4	65.5
T7	77.8 ± 4.4	65.4 ± 8.2	**71.3**	77.8 ± 3.7	46.2 ± 6.5	60.0	63.5 ± 7.9	50.0 ± 8.6	56.3	84.1 ± 4.1	50.0 ± 6.8	64.8
T8	73.0 ± 5.1	65.4 ± 8.9	**69.1**	81.0 ± 3.2	53.8 ± 9.6	66.0	76.2 ± 5.7	50.0 ± 11.1	61.7	82.5 ± 5.4	46.2 ± 4.2	61.7
T9	76.2 ± 4.9	65.4 ± 6.9	**70.6**	81.0 ± 3.7	46.2 ± 11.1	61.2	76.2 ± 5.0	50.0 ± 9.8	61.7	79.4 ± 5.7	42.3 ± 10.9	58.0
T10	73.0 ± 5.1	65.4 ± 9.8	**69.1**	74.6 ± 5.5	50.0 ± 6.8	61.1	66.7 ± 7.3	42.3 ± 9.4	53.1	76.2 ± 6.1	50.0 ± 4.3	61.7
Ave.	74.8 ± 5.0	62.5 ± 8.6	**68.3**	80.4 ± 4.5	51.9 ± 9.1	64.5	71.3 ± 6.4	52.9 ± 10.2	61.2	82.7 ± 4.8	45.7 ± 8.0	61.4

TABLE 5: Information About 20 GO Terms Very Frequently Selected by the MR Method

Model Organism	GO Term ID	GO Term Name	Selection Frequency	Rank	P-Value	Relev.	Predicted Class
Caenorhabditis elegans	GO:0006412	translation	478 (100 %)	1	1.15 E-6	0.30	Anti
	GO:0006914	autophagy	478 (100 %)	3	1.57 E-3	0.50	Pro
	GO:0006915	apoptotic process	478 (100 %)	5	4.41 E-3	0.08	Anti
	GO:0006091	generation of precursor metabolites and energy	478 (100 %)	7	1.05 E-2	0.20	Anti
	GO:0032880	regulation of protein localization	478 (100 %)	8	1.82 E-2	0.30	Pro
	GO:0035966	response to topologically incorrect protein	478 (100 %)	9	2.41 E-2	0.23	Pro
	GO:0055085	transmembrane transport	435 (91.0 %)	24	5.26 E-5	0.21	Anti
Saccharomyces cerevisiae	GO:0001302	replicative cell aging	248 (100 %)	1	5.84 E-6	0.35	Pro
	GO:0000183	chromatin silencing at rDNA	248 (100 %)	2	5.67 E-4	0.73	Pro
	GO:0006302	double-strand break repair	248 (100 %)	3.5	7.71 E-3	0.45	Pro
	GO:0016265	death	244 (98.4 %)	6	1.48 E-2	0.53	Pro
	GO:0032200	telomere organization	243 (98.0 %)	7.5	2.95 E-3	0.64	Pro
Drosophila melanogaster	GO:0003006	developmental process involved in reproduction	119 (100 %)	1	3.48 E-3	0.30	Anti
	GO:0007600	sensory perception	119 (100 %)	2.5	1.15 E-2	0.55	Anti
	GO:0006629	lipid metabolic process	119 (100 %)	7	1.89 E-2	0.15	Pro
	GO:0055085	transmembrane transport	119 (100 %)	12	4.26 E-2	0.33	Anti
Mus musculus	GO:0040018	positive regulation of multicellular organism growth	89 (100 %)	2.5	7.28 E-3	0.65	Anti
	GO:0051093	negative regulation of developmental process	89 (100 %)	5	2.24 E-2	0.14	Pro
	GO:0010948	negative regulation of cell cycle process	78 (87.6 %)	19.5	2.24 E-2	0.14	Pro
	GO:0097190	apoptotic signaling pathway	75 (84.3 %)	21	4.04 E-2	0.10	Pro

5.2 On the Statistical and Biological Relevance of a Number of Very Frequently Selected GO Terms

We now discuss the relevance, to the biology of ageing, of 20 GO terms very frequently selected as features by the MR method, among the set of terms whose predictive power was considered statistically significant (p-value < 0.05). Statistical significance was measured using a hypothesis test based on the binomial distribution, similarly to the test used in [13]. This test essentially works as follows. The classification of each gene (instance) based on any given GO term is considered a random trial with the outcome 'success' (if the gene has the class predicted by the GO term) or 'failure' otherwise. For each GO term, the number of trials for the binomial test is the number of genes annotated with that GO term; the number of successes is the number of genes that are annotated with that GO term and belong to the class predicted by that GO term (i.e. the class with the largest number of genes annotated with that GO term in the dataset); and the null hypothesis was represented by a binomial distribution where the probability of occurrence of the class predicted by that GO term is the relative frequency of that class in the dataset. The terms discussed in the remainder of this Subsection are shown in Table 5, whilst the full ranking of all GO terms, for each model organism, is included in a supplementary file. The first three columns of Table 5 are self-explained. The fourth column shows the number (and %) of instances (in the dataset of the corresponding model organism) for which the GO term was selected by MR. The fifth column shows the rank of the GO term (the lower the rank, the better), among the set of GO terms whose p-value was deemed significant. The rank is based on the number of instances for which the GO term is related by MR. The sixth and seventh columns show the p-value and the relevance value (computed by Equation (4)) of the GO term. The eighth column shows the class predicted by each GO term.

Broadly speaking, the top ranking GO terms not only reflect our understanding of biological processes associated with ageing and life-extension in model organisms, but may help identify new putative associations suitable for further studies. As the organism in which single genes were initially associated with ageing, the roundworm *C. elegans* is arguably the best studied model in the context of ageing, with multiple pathways associated with the regulation of longevity [2]. It is the organism

in which more gene manipulations have been shown to extend longevity [8] and unsurprisingly several top ranking GO categories in our results are known to impact on ageing. The top ranking term is "translation" with a strong association with anti-longevity. This is not surprising, since it is well-established that an inhibition of translation extends lifespan in *C. elegans* [2]. Other top categories like "autophagy", "apoptotic process", metabolism ("generation of precursor metabolites and energy") and maintenance of protein homeostasis ("response to topologically incorrect protein") have been linked to ageing [35]. Various top-ranked terms also relate to growth and development, which is not surprising given that developmental pathways in worms can significantly impact on ageing [2], [36]. While all these results fit well with our current understanding of ageing, some categories may point towards novel mechanisms and warrant further investigation like "regulation of protein localization" and "transmembrane transport" associated, respectively, with pro- and anti-longevity.

A similar trend is observed in other model organisms. In yeast, which after worms is the model with most genes associated with ageing [8], top-ranked categories include "chromatin silencing at rDNA", "telomere organization" and "double-strand break repair", all of which have been associated with longevity [35]; in addition to the expected "replicative cell aging" and "death".

In flies, as in worms, some top terms are related to development, including the top category "developmental process involved in reproduction" associated with anti-longevity, and growth including cell division-related categories. Another top category associated with anti-longevity is "sensory perception", which fits well with recent results linking sensory perception, and olfactation in particular, to ageing [37]. Metabolism, with "lipid metabolic process" as the top category associated with pro-longevity, is in line with our understanding of life extension pathways mediated by diet, such as caloric restriction [38]. Intriguingly, "transmembrane transport" is, like in worms, also associated with anti-longevity, which merits further studies.

The top categories from mice partly reflect those found in lower model organisms, such as categories related to development and growth, like "positive regulation of multicellular organism growth" associated with anti-longevity and "negative regulation of developmental process" associated with pro-longevity. These results further emphasize the relationship between developmental processes and ageing, and further strengthen the idea that retarding development and growth can extend lifespan [36]. Also present in mice, as in invertebrates, are terms related to apoptosis ("apoptotic signaling pathway") and cell cycle ("negative regulation of cell cycle process"). Although this likely results from researcher biases, i.e. studying pathways in mice known to be associated with ageing in other model organisms, it highlights the evolutionary conservation of pathways associated with ageing [2].

6 Conclusions

This work proposes two novel hierarchical feature selection methods (HIP and MR) that have been used to select features for the Naïve Bayes and 1-NN classsifiers, in the task of predicting the pro-longevity or anti-longevity effect of genes of the four most widely used biomedical model organisms. These hierarchical feature selection methods were designed to exploit information in hierarchical relationships among Gene Ontology (GO) terms (used as features) in order to reduce redundancy in the set of selected features. The use of the lazy learning approach allowed us to select a subset of GO terms specifically for each testing instance (gene) being classified by Naïve Bayes and 1-NN.

The experimental results showed that both the proposed hierarchical feature selection methods (HIP and MR) improved the predictive accuracy of Naïve Bayes, and the HIP method improved the predictive accuracy of 1-NN; compared with using Naïve Bayes and 1-NN without feature selection. We also discussed the biological relevance of a number of very frequently selected GO terms in the context of the biology of ageing literature.

Concerning future research, the hierarchical feature selection methods proposed here could be applied to other types of hierarchical features; as long as the feature hierarchy is a kind of generalization-specialization hierarchy – where the occurrence of a feature in an instance implies the occurrence of the features' ancestors in that instance – and the classification algorithm follows a lazy learning approach. In addition, our current feature selection methods have two parts: a relevance measure and a feature hierarchy-processing procedure that uses the feature hierarchy to decide which features should be removed. In the future, it would be interesting to design a more complex relevance measure that directly considers the feature hierarchy, which would avoid the need for a separate feature hierarchy-processing procedure. Another research direction would be to use some class imbalancing technique to cope with imbalanced class distributions, which could potentially increase the predictive accuracy in some datasets.

Another research direction consists of integrating the proposed hierarchical feature selection methods with a lazy version of a Bayesian Network-Augmented Naïve Bayes (BAN) classifier, which could increase predictive accuracy in some datasets, but this could lead to overfitting.

Acknowledgments

We acknowledge the support of concurrency researchers at Kent for access to the 'CoSMoS' cluster, funded by EPSRC grants EP/E049419/1 and EP/E053505/1. We thank Daniel Wuttke and Robi Tacutu for valuable discussions about the creation of our datasets. Cen Wan thanks Shuqian Yu for valuable discussions on the biology of ageing. We also thank the anonymous reviewers for their comments. GenAge is supported by a Wellcome Trust grant (ME050495MES) to João Pedro de Magalhães.

REFERENCES

[1] E. J. Masoro, "Overview of caloric restriction and ageing," *Mechanisms of Ageing and Development*, vol. 126, no. 9, pp. 913–922, Sep. 2005.

[2] C. J. Kenyon, "The genetics of ageing," *Nature*, vol. 464, no. 7288, pp. 504–512, Mar. 2010.

[3] A. A. Freitas and J. P. de Magalhães, "A review and appraisal of the DNA damage theory of ageing," *Mutation Research*, vol. 728, no. 1-2, pp. 12–22, Jul./Oct. 2011.

[4] T. B. L. Kirkwood and S. N. Austad, "Why do we age?" *Nature*, vol. 408, no. 6809, pp. 233–238, Nov. 2000.

[5] J. P. de Magalhães, "How ageing processes influence cancer." *Nature Reviews Cancer*, vol. 13, no. 5, pp. 357–365, May 2013.

[6] I. H. Witten, E. Frank, and M. A. Hall, *Data Mining: practical machine learning tools and techniques*. Morgan Kaufmann, 2011.

[7] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*. Norwell, MA: Kluwer Academic Publishers, 1998.

[8] R. Tacutu, T. Craig, A. Budovsky, D. Wuttke, G. Lehmann, D. Taranukha, J. Costa, V. E. Fraifeld, and J. P. de Magalhães, "Human ageing genomic resources: Integrated databases and tools for the biology and genetics of ageing," *Nucleic Acids Research*, vol. 41, no. D1, pp. D1027–D1033, Jan. 2013.

[9] The GO Consortium, "Gene Ontology: tool for the unification of biology," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, May 2000.

[10] D. W. Aha, *Lazy Learning*. Kluwer Academic Publishers, 1997.

[11] R. B. Pereira, A. Plastino, B. Zadrozny, L. H. de C. Merschmann, and A. A. Freitas, "Lazy attribute selection: Choosing attributes at classification time," *Intelligent Data Analysis*, vol. 15, no. 5, pp. 715–732, Aug. 2011.

[12] C. Wan and A. A. Freitas, "Prediction of the pro-longevity or anti-longevity effect of *Caenorhabditis Elegans* genes based on Bayesian classification methods," in *Proc. IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2013)*, Shanghai, China, Dec. 2013, pp. 373–380.

[13] A. A. Freitas, O. Vasieva, and J. P. de Magalhães, "A data mining approach for classifying DNA repair genes into ageing-related or non-ageing-related," *BMC Genomics*, vol. 12, no. 27, pp. 1–11, Jan. 2011.

[14] Y. Fang, X. Wang, E. K. Michaelis, and J. Fang, "Classifying aging genes into DNA repair or non-DNA repair-related categories," in *Lecture Notes in Intelligent Computing Theories and Technology*, D. S. Huang, K. H. Jo, Y. Q. Zhou, and K. Han, Eds., 2013, pp. 20–29.

[15] Y. H. Li, M. Q. Dong, and Z. Guo, "Systematic analysis and prediction of longevity genes in *Caenorhabditis elegans*," *Mechanisms of Ageing and Development*, vol. 131, no. 11-12, pp. 700–709, Nov./Dec. 2010.

[16] T. Huang, J. Zhang, Z. Xu, L. Hu, L. Chen, J. Shao, L. Zhang, X. Kong, Y. Cai, and K. Chou, "Deciphering the effects of gene deletion on yeast longevity using network and machine learning approaches," *Biochimie*, vol. 94, no. 4, pp. 1017–1025, Apr. 2012.

[17] C. N. Silla Jr. and A. A. Freitas, "A survey of hierarchical classification across different application domains," *Data Mining and Knowledge Discovery*, vol. 22, no. 1-2, pp. 31–72, Apr. 2011.

[18] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *The Annual of Statistics*, vol. 37, no. 6, pp. 3468–3497, 2009.

[19] R. Jenatton, J. Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, pp. 2777–2824, 2011.

[20] A. F. T. Martins, N. A. Smith, P. M. Q. Aguiar, and M. A. T. Figueiredo, "Structured sparsity in structured prediction," in *Proc. the 2011 conference on empirical methods in natural language processing (EMNLP 2011)*, Edinburgh, UK, Jul. 2011, pp. 1500–1511.

[21] J. Ye and J. Liu, "Sparse methods for biomedical data," *ACM SIGKDD Explorations Newsletter*, vol. 14, no. 1, pp. 4–15, Jun. 2012.

[22] D. Barber, *Bayesian Reasoning and Machine Learning*. Cambridge University Press, 2012.

[23] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[24] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Prentice Hall, 1988.

[25] A. K. Jain and D. Zongker, "Representation and recognition of handwritten digits using deformable templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 12, pp. 1386–1391, Dec. 1997.

[26] F. Zheng and G. I. Webb, "Efficient lazy elimination for averaged one-dependence estimators," in *Proc. the 23rd international conference on Machine learning*, Pittsburgh, USA, Jun. 2006, pp. 1113–1120.

[27] C. Stanfill and D. Waltz, "Toward memory-based reasoning," *Communications of the ACM*, vol. 29, no. 12, pp. 1213–1228, Dec. 1986.

[28] C. A. Ratanamahatana and D. Gunopulos, "Feature selection for the naive Bayesian classifier using decision trees," *Applied Artificial Intelligence*, vol. 17, no. 5-6, pp. 475–487, Nov. 2003.

[29] J. P. de Magalhães, A. Budovsky, G. Lehmann, J. Costa, Y. Li, V. Fraifeld, and G. M. Church, "The human ageing genomic resources: online databases and tools for biogerontologists," *Aging Cell*, vol. 8, no. 1, pp. 65–72, Feb. 2009.

[30] National Center for Biotechnology Information. (2011) Gene2go. [Online]. Available: ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene2go.gz

[31] N. Japkowicz and M. Shah, *Evaluating learning algorithms: a classification perspective*. Cambridge University Press, 2011.

[32] D. G. Altman and J. M. Bland, "Diagnostic tests. 1: Sensitivity and specificity," *BMJ: British Medical Journal*, vol. 308, no. 6943, p. 1552, Jun. 1994.

[33] J. Derrac, S. Garcia, D. Molina, and F. Herrera, "A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms," *Swarm and Evolutionary Computation*, vol. 1, no. 1, pp. 3–18, Mar. 2011.

[34] J. Demsăr, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine Learning Research*, vol. 7, pp. 1–30, Jan. 2006.

[35] C. López-Otín, M. A. Blasco, L. Partridge, M. Serrano, and G. Kroemer, "The hallmarks of aging," *Cell*, vol. 153, no. 6, pp. 1194–1217, Jun. 2013.

[36] J. P. de Magalhães, "Programmatic features of aging originating in development: aging mechanisms beyond molecular damage?" *The FASEB Journal*, vol. 26, no. 12, pp. 4821–4826, Dec. 2012.

[37] N. J. Linford, T. H. Kuo, T. P. Chan, and S. D. Pletcher, "Sensory perception and aging in model systems: From the outside in," *Cell and Developmental Biology*, vol. 27, pp. 759–785, Nov. 2011.

[38] M. Plank, D. Wuttke, S. van Dam, S. A. Clarke, and J. P. de Magalhães, "A meta-analysis of caloric restriction gene expression profiles to infer common signatures and regulatory mechanisms," *Molecular Biosystems*, vol. 8, no. 4, pp. 1339–1349, Feb. 2012.

Cen Wan obtained his BEng in Computer Science and Technology from Fujian Agriculture and Forestry University, China, in 2009; and his MSc in Computer Science (with distinction) from the University of Liverpool, UK, in 2011. He is currently pursuing the PhD degree in Computer Science in the University of Kent, UK. His research interests are data mining, machine learning, computational biology and bioinformatics. He is a student member of the IEEE.

Prof. Alex A. Freitas obtained his BSc in Computer Science from FATEC-SP, Brazil, 1989; his MSc in Computer Science from UFSCar, Brazil, 1993; his PhD in Computer Science from the University of Essex, UK, 1997; and his MPhil (a research-oriented master's degree) in Biological Sciences from the University of Liverpool, UK, 2011. He is a Professor of Computational Intelligence at the University of Kent, UK. His main research interests are data mining and knowledge discovery (mainly classification) and its application to biology (mainly ageing) and pharmaceutical sciences.

Dr. João Pedro de Magalhães graduated in Microbiology in 1999 from the Escola Superior de Biotecnologia in his hometown of Porto, Portugal, and then obtained his PhD in 2004 from the University of Namur in Belgium, where he worked in the Ageing and Stress Group. Following a postdoc at Harvard Medical School, in 2008 Dr. de Magalhães joined the University of Liverpool where he is now a senior lecturer and leads the Integrative Genomics of Ageing Group (http://pcwww.liv.ac.uk/~aging/). The group's research broadly focuses on understanding the genetic, cellular, and molecular mechanisms of ageing.

APPENDIX - STATISTICAL SIGNIFICANCE OF GO TERMS

This appendix gives more details about how we evaluated the statistical significance of GO terms for the purposes of the analysis of the most frequently selected GO terms reported in Section 5.2. Recall that the proposed hierarchical feature selection methods select a different set of features (GO terms) for each testing instance. Hence, when producing a ranking of GO terms in descending order of their usefulness, it is natural to calculate the ranking based on the number of instances where each GO term is selected to be used as input by Naïve Bayes. MR and HIP are the best feature selection methods in terms of predictive performance in this work, with no significant difference in their performance. However, HIP considers only the redundancy among hierarchically-related GO terms, whilst MR considers both that redundancy and the GO terms' relevance values. Hence, for the purpose of ranking the GO terms in decreasing order of frequency of selection, intuitively the ranking produced when using MR as the selection method is more appropriate, and this ranking criterion is used here.

For each model organism, we produced a ranking of all GO terms occurring in the dataset version with GO term frequency threshold 3 for that organism, since that dataset contains the largest number of GO terms. Note that the ranking criterion based on the frequency of selection when using the MR method does not directly take into account the statistical significance of selected GO terms. Some GO terms may be selected very often by MR due to their high relevance (predictive power), regardless of their statistical significance. Hence, to complement the ranking of GO terms based on their frequency of selection by MR, we also computed, for each GO term, its p-value associated with a statistical significance test, based on the following rationale.

If we had to predict the class of a gene based on a given GO term alone (without using any other feature), we would assign that gene to the class with the largest number of genes (instances) annotated with that GO term. We refer to that class as the class predicted by that GO term. The predictive accuracy associated with the use of that GO term as a predictor is the ratio of the number of instances that are annotated with that GO term and have the class predicted by the GO term divided by the number of instances that are annotated with that GO term.

To evaluate the statistical significance associated with a GO term used as a predictor, we use a significance test based on the binomial distribution, which has two parameters: n, the number of trials, and p, the probability of success in each trial. When applying the significance test, the assignment of the class predicted by the GO term to any given instance annotated with that term is regarded as a random trial with two possible results: success (the class predicted by the GO term equals the true class of that instance) or failure otherwise. The instances classified by the GO term are assumed to be independent from each other, and the number of trials n is the number of instances classified by the GO term – i.e. instances annotated with the GO term. Under the null hypothesis that the value "yes" of the GO term feature is irrelevant for predicting the class of an instance, the probability of observing a successful result is given by the relative frequency of the class predicted by the GO term in the dataset – i.e. the ratio of the number of instances of that class in the dataset divided by the total number of instances (of any class) in the dataset.

Hence, to set up a test of hypothesis for the statistical significance of the predictive power of a given GO term, we consider the observed number of instances that are correctly classified by the GO term, denoted k. That is, k is the number of instances that are annotated with the GO term and belong to the class predicted by the GO term. Let X be a random variable representing the number of successes in a binomial distribution with probability of success p and number of trials n. Under the null hypothesis that the GO term has no predictive power, for each model organism dataset version, the probability of observing exactly k successes, according to the binomial distribution, is given by Equation (5),

$$\mathbf{Pr}(X = k) = \mathbf{C}_k^n \mathbf{p}^k (1 - \mathbf{p})^{n-k}, \tag{5}$$

where $\mathbf{C}_k^n$ is the number of combinations of k elements out of n elements. Finally, for the test of hypothesis, we use Equation (5) to calculate the probability $\mathbf{Pr}(X \geq k)$. If the null hypothesis that the GO term has no predictive power can be rejected at the significant level of 5%, then the GO term's ability to predict its associated class can be considered as statistically significant.