An Extensive Empirical Comparison of Probabilistic Hierarchical Classifiers in Datasets of Ageing-Related Genes

Fabio Fabris, Alex A. Freitas and Jennifer M. A. Tullet

Abstract—This study comprehensively evaluates the performance of 5 types of probabilistic hierarchical classification methods used for predicting Gene Ontology (GO) terms related to ageing. Of those tested, a new hybrid of a Local Hierarchical Classifier (LHC) and the Predictive Clustering Tree algorithm (LHC-PCT) had the best predictive accuracy results. We also tested the impact of two types of variations in most hierarchical classification algorithms, namely: (a) changing the base algorithm (we tested Naive Bayes and Support Vector Machines), and the impact of (b) using or not the Correlation based Feature Selection (CFS) algorithm in a pre-processing step. In total, we evaluated the predictive performance of 17 variations of hierarchical classifiers across 15 datasets of ageing and longevity-related genes. We conclude that the LHC-PCT algorithm ranks better across several tests (7 out of 12). In addition, we interpreted the models generated by the PCT algorithm to show how hierarchical classification algorithms can be used to extract biological insights out of the ageing-related datasets that we compiled.

Index Terms—Ageing, Hierarchical Classification, Protein Classification, Dependence Networks.

1 Introduction

Due to technological advances in medicine and healthcare, human longevity has been significantly increasing for decades [1]. Therefore, diseases associated with the greying population; such as cancers, heart conditions, and neurodegenerative illnesses; are affecting an increasingly large number of people. Therefore, the prospect of slowing down or even reversing the ageing process is attracting researches of several areas. Although a lot of progress has been made in recent years to try to explain why do we age and how to potentially slow-down this apparently relentless process, our understanding of the biology of ageing is in its infancy.

Ageing can be defined as an intrinsic, age-related process of loss of viability and increase in vulnerability [1]. Although the vast majority of animals are impacted by the ageing process, there are some species that have no apparent senescence process, i.e. not only their mortality rate is constant during their adult life, there is no evident age-related physiological functional decline in their organs [2]. Also, studies have found several ageing-related genes in model organisms such as the mouse (*M. musculus*), the fruit fly (*D. melanogaster*), the worm (*C. elegans*), and the yeast (*S. cerevisiae*), which when turned *on* or *off*, considerably affect the lifespan of the organisms. For instance, Friedman and Johnson [3] showed that by manipulating the gene *age-1*, associated with the production of *insulin/insulin-like growth factor 1* (IGF1), it was possible to make the nematode worm *C. elegans* live twice as long. In mice, disruption of the gene *Prop1*, also related to the production of IGF1, may increase their lifespan by 50% [4]. These initial findings were products of laborious "wet-lab" experimentation.

Recently, the cost of extracting genomic and proteomic data from organisms has decreased many-fold. Researchers now have access to vast public datasets of biological data. In general, these datasets contain gene or protein sequence information (e.g. Universal Protein Resource [5]) and possibly a classification of the biological processes the gene or protein is involved in. This classification is often curated in a hierarchical ontology (e.g. Gene Ontology (GO) [6]). The existence of these widely used hierarchical classes justifies the use of hierarchical classification methods to predict hierarchical protein functions and other protein properties such as folding patterns [7].

This work will focus on the hierarchical classification task of data-mining, where the classes to be predicted are organized into a hierarchy. The goal is to build a classification model capable of assigning classes to new instances (with unknown classes) given a dataset containing instances with known classes. Using hierarchically structured classifications of curated datasets such as the GO and the FunCat has a greater potential to generate useful classification results for biologists than the more common approach of performing *flat* classification on the data. Additionally, we are interested in classifiers capable of outputting interpretable models that can be analysed by users to potentially extract new biological knowledge from the data.

Many hierarchical classification algorithms have been proposed [8]; but there has been relatively little comparison of the effectiveness of different types of hierarchical classification algorithms. In this context, this inter-disciplinary

- *F. Fabris and A. A. Freitas are with the University of Kent, School of Computing, Canterbury, Kent, CT2 7NF, United Kingdom. E-mail: ff79,A.A.Freitas@kent.ac.uk*
- *Jennifer M. A. Tullet is with the University of Kent, School of BioSciences, Canterbury, Kent, CT2 7NJ, United Kingdom. E-mail: J.M.A.Tullet@kent.ac.uk*

work offers two types of contribution. First, in terms of contribution to the hierarchical classification literature, we compare the predictive accuracy of five different types of probabilistic hierarchical classification algorithms, most of them with several variations, leading to 17 different variations of probabilistic hierarchical classification algorithms being compared. To the best of our knowledge, this is the first such extensive comparison of probabilistic hierarchical classification algorithms. In addition, we propose a new hybrid hierarchical classification algorithm. Second, in terms of contribution to biology and bioinformatics, this paper describes the creation of the 15 new ageing-related datasets used in our experiments and the analysis of results includes an interpretation of some classification models, discussing the patterns extracted from those models in the context of the biology of ageing literature. The new datasets described in this paper – which will be available after the publication of the paper – involve data from five different model organisms and three different types of predictive features.

The remainder of this paper is organized as follows: Section 2 presents background on hierarchical classification. Section 3 describes the algorithms evaluated in the subsequent sections. Section 4 explains the creation of the ageing datasets used in this work. Section 5 reports the predictive accuracy results and run time of the algorithms we tested. Section 6 offers an interpretation of some of the classification models for an ageing dataset. In Section 7 we conclude our work and give possible lines of future research.

2 Background

Typical classification problems involve a flat set of class labels, i.e., there is no hierarchical relationships among the class labels to be predicted. By contrast, in hierarchical classification problems, the set of class labels is organized into a hierarchy, usually a tree or a DAG (Directed Acyclic Graph), where each node represents a class label and the edges represent generalization-specialization relationships among classes. Hierarchical classification is common in bioinformatics, in particular when predicting gene or protein functions, since such functions are usually specified by a hierarchical scheme like the Gene Ontology [6].

Hierarchical classification algorithms may be divided into two types [8]: global or local. Local Hierarchical Classification (LHC) algorithms build a set of local classification models (base classifiers) by training a traditional (flat) classification algorithm for each (typically small) part of the class hierarchy. By contrast, global hierarchical classification algorithms build a single global classification model predicting classes in the whole class hierarchy.

Considering that each class label is represented as a node in a tree or a DAG, the local base classifiers used by LHC algorithms are usually induced in one of two ways: 1) one local classifier per node, where each classifier is induced to decide if an instance should be annotated or not with a particular class; 2) one local classifier per parent node, where each classifier is induced to decide which child labels (if any) should be assigned to a instance. Next, in the testing phase, the LHC algorithm combines the predictions of the local classification models to predict classes in the whole (global) class hierarchy.

LHC algorithms have the advantage of algorithmic simplicity, since they transform the original hierarchical classification problem into a set of simpler flat classification problems in the training phase, but they produce a large number of local (flat) classifiers, one for each class node or one for each parent node in the class hierarchy, depending on the approach used. Conversely, global hierarchical classification algorithms have the advantage of producing a single coherent global classification model, which tends to be more easily interpreted than a large number of different classification models.

3 The Hierarchical Classification Algorithms Evaluated in our Experiments

In this section we specify the five broad types of probabilistic hierarchical classification algorithms used in our experiments, namely: (1) the standard Predictive Clustering Tree (PCT) algorithm [9]; (2) the Hierarchical Dependence Network (HDN) algorithm [10]; (3) the hybrid HDN-PCT algorithm [10]; (4) a stand-alone Local Hierarchical Classification (LHC) algorithm; and (5) a new hybrid LHC-PCT algorithm, proposed in this work.

PCT is a global hierarchical classification algorithm, but each of the other 4 algorithms is either a local or hybrid global/local hierarchical classification algorithm that needs to use a base local classification algorithm to build different local classification models for different class labels in the hierarchy. Hence, each of the algorithms (2)–(5) has been implemented with two different base local classification algorithms, namely, Naive Bayes (BN) and a Support Vector Machine (SVM); and each algorithm was applied in two scenarios, using all features or only the features selected by the Correlation-based Feature Selection (CFS) method [11] in a preprocessing phase. Hence, we are evaluating in total 17 types of probabilistic hierarchical classification systems: 4 broad types of hierarchical classifiers times 2 base classifiers times 2 feature selection scenarios (using or not the CFS method) plus PCT as a global classifier. To the best of our knowledge, this is the first such extensive evaluation of probabilistic hierarhical classification algorithms.

We now briefly describe each of the above five probabilistic hierarchical classification algorithms. More details can be found in the cited references.

(1) The Predictive Clustering Tree (PCT) algorithm

PCT (Predictive Clustering Tree) is a type of global hierarchical classification algorithms that builds a single decision tree by recursively finding a value for a predictive feature that splits the current set of instances in two clusters, maximizing the similarity of classes within each cluster and the dissimilarity of the classes across the two clusters. The algorithm recurses in each cluster that it forms and eventually stops if the split does not have a good quality (based on some quality measure) or the size of a cluster falls bellow a pre-established threshold [12]. In the prediction phase, to classify an instance $\mathbf{x}$, a PCT algorithm first identifies the cluster associated with that instance and then assigns, to instance $\mathbf{x}$, classes whose probabilities in the class

probability vector of that cluster are greater than a probability threshold. The threshold is varied when computing a Precision-Recall curve, as explained later.

The most well-known version of the PCT algorithm is the Clus-HMC algorithm [9]. There is also an ensemble version of Clus-HMC, called Clus-HMC-Ens [13]. In our experiments we do not use this ensemble version for two reasons: 1) difficulty to interpret the models and 2) the large computational cost.

The PCT algorithm has only one parameter to tune: the s-value that dictates how statistically significantly different two groups of instances in a tree node's split must be in order for the split to be accepted by a F-test. Larger s-values correspond to a more permissive test, and thus a larger decision tree. To tune this parameter we have applied an internal 10-fold cross-validation procedure (using only the training set) in each iteration of the main (external) cross-validation run. We have tested the default set of parameters suggested for the *Clus System* in [9]: $\{0.001, 0.005, 0.010, 0.050, 0.100, 0.125\}$, and chose the one with highest predictive accuracy.

(2) The Hierarchical Dependence Network (HDN) algorithm

Dependence Networks (DNs) are a relatively underexplored type of probabilistic graphical model first described in [14]. Each node of a DN represents a random variable and encodes a probability distribution conditioned on the values of its parents, like in Bayesian Networks (BNs). However, DNs are more flexible than BNs since they allow for cycles in their graphical model. In addition, in a DN the edges coming out of a node n_i connect n_i to the minimal set of other nodes, n_{-i}, that make n_i independent from all other nodes. This set is called the *Markov blanket* of a node. In a conventional classification problem, the Markov blanket of the class node corresponds to the set of predictive features that influences the value of the class variable.

Recently, Guo and Gu [15] and Li et al. [16] proposed DN classification algorithms for multi-label classification, where an instance can be assigned to multiple class labels. This addresses flat classification, since there is no hierarchy among class labels. In this paper we address the more difficult problem of *hierarchical* multi-label classification. Hence, we use a slightly modified version of the hierarchical DN (HDN) algorithm recently proposed in [10], as follows.

The HDN algorithm first creates a graph containing one node for each predictive feature and each class label. Then it builds a local probabilistic classifier for each class node, using the estimated Markov blanket of each class label. In our context of hierarchical classification, the Markov blanket of each class label can include both predictive features and other class labels in the hierarchy. These Markov blankets can be estimated using a feature selection method. For each class label c_i, the feature selection method receives (as input) the set of features and the (pre-selected) subset of class labels containing only the siblings and the parents of the children of c_i in the class hierarchy. The feature selection method then returns the subset of features and the subset of sibling/parents of children of class labels estimated to be relevant for predicting the class label c_i. The reason for pre-selecting the siblings and parents of the children of each class label is that such related labels represent important predictive relationships in the dataset, as encoded in the structure of the class hierarchy (which is defined by human experts). In [10], a simple statistic test of independence, the F-test, was used as a feature selection method; but here we use instead the more sophisticated Correlation-Based Feature Selection (CFS) method [11], comparing its results with not using any feature selection algorithm. Not using a feature selection algorithm means that the Markov blanket of the class labels is not being properly estimated and therefore, strictly speaking, the resulting algorithm is not a DN. However, we consider this a good baseline to check if feature selection is indeed improving the performance of the hierarchical classifiers.

Once the Markov blanket of each class label has been estimated, we use a (flat) classification algorithm – Naive Bayes or a Support Vector Machine (SVM) – to build a classifier that estimates the probability of each class label for each instance being classified. That is, for each class label c_i, a classifier estimates the probability $P(c_i|\mathbf{x}, c_{-i})$, where $\mathbf{x}$ and c_{-i} represent the set of values for the selected features and the set of values for the selected class labels (respectively) in the current instance being classified.

The class imbalance problem is common in the classification of biological data, e.g.: [17], and it is even more common in hierarchical classification problems, where we have usually many class labels with low frequency. Therefore, to train the flat classifiers for class c_i we follow the suggestion of [18] and consider as positive examples the instances annotated with class c_i or any of its descendants, and as negative examples the complementary set of instances. This approach is among the best to deal with class imbalance while training local classifiers for hierarchical classification [18].

Note that the values (presence or absence) of class labels are available during training, but such values are of course unavailable when classifying new instances in the testing phase. Hence, to estimate the most likely class-label distribution for each new instance being classified, we use the Gibbs sampling procedure [14] to query the HDN. The Gibbs sampling algorithm first assigns random values to every node (class label) c_i of the graphical model and then iteratively visits each node, re-sampling the value of each c_i given the values of the nodes it is connected with. That is, it first assigns a random value to every c_i, then samples a value for c_i from $P(c_i|\mathbf{x}, c_{-i})$, updates the value of c_i and proceeds to the next class label. The probability of each class label is estimated by its occurrence frequency after a given number of burn-in iterations, defined by the user.

(3) The hybrid Hierarchical Dependence Network/Predictive Clustering Tree (HDN-PCT) algorithm

From initial experimentation, we have observed that some clusters in the leaf nodes of the tree built by the PCT algorithm had a relatively large number of instances, which could be further used to train a different type of classifier to better exploit the available data. For this reason we apply our HDN algorithm in each cluster produced by the PCT algorithm that has more than min_inst_HDN training instances (a parameter). The resulting hybrid algorithm is named HDN-PCT, and is described in more details in [10].

Note that HDN-PCT is a hybrid global/local hierarchical classification algorithm, since it first produces a global decision tree that potentially predicts all class labels as a whole, and then a set of HDN classifiers (each containing several local base classifiers) for each leaf node of the decision tree with more than min_inst_HDN instances.

(4) The Local Hierarchical Classification (LHC) algorithm

This is a fairly simple and conventional LHC algorithm, producing one local classifier per class node, which can be seen as a strong baseline method, by comparison with the more sophisticated variations of PCT and HDN algorithms described earlier. In our experiments we used, as the local base classification algorithms, NB and SVM, but other standard flat classification algorithms could be used. We have used the same strategy to define the positive and negative examples that we used for the HDN algorithm.

Usually, when using the LHC approach in the testing phase, the top-down strategy is applied: first, the highest-level classes (excluding the root node) are predicted. Then, the algorithm recurses to the children of each positively predicted class, until no positive predictions are made or a leaf node is reached. As we are dealing with probabilistic classifications instead of crisp classifications, we apply the following modification: we recurse to every child of every class but limit the predicted probability of the classes to the probability of its parents, to maintain the classification consistence across the class hierarchy.

(5) The new hybrid Local Hierarchical Classification/Predictive Clustering Tree (LHC-PCT) algorithm

This hybrid hierarchical classification algorithm, introduced in this current paper, can be seen as an extension of the previously described PCT algorithm, which builds a global model where each leaf node is assigned a class probability vector. The PCT model simply assigns, to a new instance reaching a given leaf node, the classes whose probabilities are greater than a certain threshold. Our new hybrid algorithm extends the PCT algorithm in the following way.

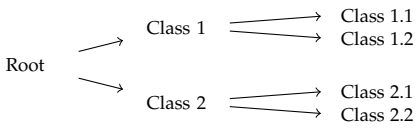
For each leaf node having more than min_inst_LHC (a parameter) instances, the hybrid algorithm builds a local classification model for predicting each class, by running a standard $flat$ classification algorithm from the instances in that leaf node. The idea is that leaves with a large number of instances may be further explored by another classification algorithm, improving predictive performance. Again, we used, as the local (base) classification algorithm in the training phase, NB or a SVM.

The combination of decision trees with other classification algorithms has been recently proposed for (flat) multi-label classification with success [19]; however, as far as we know, it was never tried with the PCT algorithm in the hierarchical classification setting.

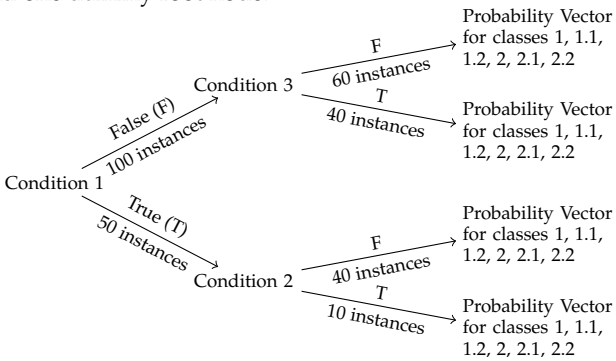
In figures 1(a) to 1(e) we present a graphical representation of the algorithms that we have described so far.

4 Dataset Creation

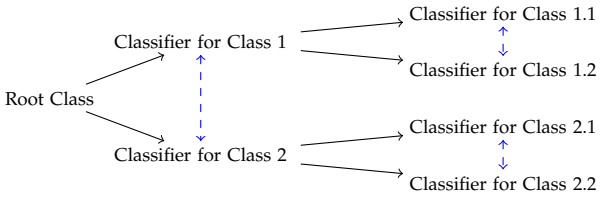
To study the biological aspects of ageing/longevity using our hierarchical classification algorithms, we have built 15



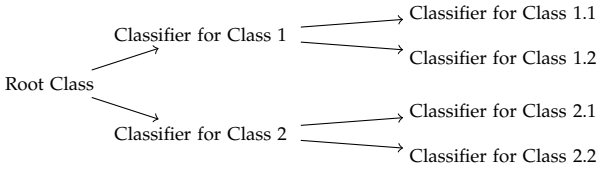
(a) A simple class hierarchy organized as a tree with 6 classes and one dummy root node.



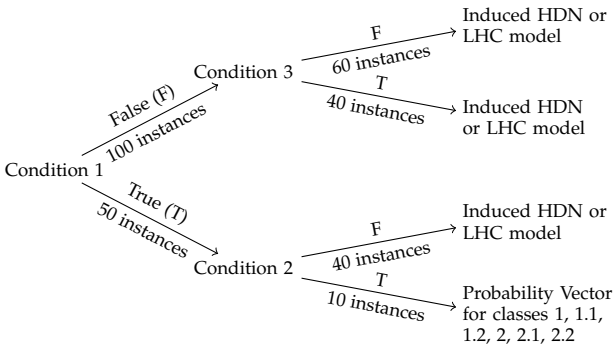
(b) PCT algorithm: the algorithm builds a decision tree by findings splits that maximize the predictive accuracy of the classifier.



(c) HDN algorithm: a flat binary classifier is induced for each class and dependencies (dashed edges) among siblings are considered. Note that the other type of dependency considered by HDN (involving parents of the children of a node) is inexistent in tree hierarchies, as the one in our example.



(d) LHC algorithm: a 'flat' binary classifier is induced for each class and no dependencies between classes are considered.



(e) Hybrid algorithms: for each cluster with more than 30 instances (an example parameter value) in the leaves of the PCT model, a HDN or LHC model (for the HDN-PCT or LHC-PCT algorithm, respectively) is trained. If there are less than or exactly 30 instances, no model is induced. During testing, the PCT model is used to get the induced model and return its predictions. If no model was induced, the PCT's probability class vector is used.

Fig. 1: Graphical representation of the algorithms tested in this work in the context of a simple class hierarchy.

TABLE 1: Number of genes for each species present in the GenAge database.

Species	Number of genes
Saccharomyces cerevisiae (baker's yeast)	825
Caenorhabditis elegans (a type of worm)	741
Homo sapiens (human)	298
Drosophila melanogaster (fruit fly)	140
Mus musculus (house mouse)	112

datasets containing features extracted from the proteins encoded by the genes in the Ageing Gene Database (GenAge) [20]. GenAge is a database of ageing-related genes in a range of species, including human, flies, worms and mice.

The feature extraction procedures used in this paper have also been used in other works dealing with protein classification: [21, 22]. Salama and Freitas [23] have compiled an ageing-related dataset for the hierarchical classification of ageing-related proteins. We build upon their work by updating and expanding the dataset to contain more species and the features used in [21], which focused on the hierarchical classification of generic (not specifically ageing-related) proteins functions. In our datasets, each instance represents an ageing-related gene, and the hierarchical classes to be predicted are Gene Ontology (GO) terms.

All genes were collected from the GenAge database, build 17 (from December 18, 2013). This version contains 298 human ageing-related genes and 1,825 genes from model organisms, related to both ageing and longevity.

The human dataset contains a comprehensive list of genes potentially associated with human ageing. This list contains genes supported by different degrees of confidence, varying from direct evidence linking the gene to human ageing to inconclusive evidence that the gene is related to ageing. The model organism datasets contain genes associated with ageing in non-human organisms. These genes have, in general, a more reliable classification due to easiness of experimenting with these species. Table 1 lists the number of genes for each organism in GenAge, species with less than 5 ageing genes associated with them were discarted.

This leaves us with 5 species. For each species we derive 3 datasets containing numeric alignment independent features, protein motif features and protein-protein interaction features, leaving us with 15 ageing-related datasets. The GenAge database contains the *"Entrez Gene Id"* as an external gene identifier; we use it to retrieve the *"UniprotKB AC ID"* (UniProt Knowledge Database Accession Identifier) protein identifier using the UniProt ID Mapping Tool.

Because more than one protein may be associated with a single gene, 2,855 UniProt identifiers were retrieved from the 2,123 genes. However, from the 2,855 proteins, we discard 1,243 whose functions were not manually reviewed by experts or whose species is one of the 4 that were discarded. After this step, we were left with 1,612 proteins (instances), distributed among organisms as presented in the last column of Table 2.

Finally, we downloaded the amino acid sequence of each protein from the UniProt-SwissProt database, using build 2014_02 of 19 February 2014[1].

1. ftp://ftp.uniprot.org/

The hierarchical classes were created for each model organism by first retrieving the GO terms associated with each protein sequence using the UniProt-SwissProt database. Next, we used the web version of the DAVID tool[2] with default parameters to retrieve the over-expressed GO terms of each model organism, considering only these GO terms in our final dataset. We call these over-expressed GO terms *ageing-related GO terms*, as they occur significantly more often than statistically expected in our datasets of ageing-related proteins.

Therefore, we are dealing with hierarchical ageing-related classes for 2 reasons: 1) the GO terms being predicted are over-expressed according to the DAVID tool in a set of proteins that are ageing-related according to the GenAgen database; 2) the classification algorithms that we tested were designed to deal with hierarchical classes.

Originally, we also compiled datasets based on binary KEGG pathway features, representing the presence or absence of a protein in several KEGG pathways. Although this feature type has arguably a good interpretation potential, we have excluded them from our study because these KEGG pathway features are, in many cases, near synonymous to the GO terms that we want to predict. E.g., the GO term GO:0000718 (nucleotide-excision repair, DNA damage removal), and the KEGG pathway hsa03420 (Nucleotide excision repair) have essentially the same meaning. This closeness unduly inflates the accuracy estimation and harms interpretation, leading to trivial, uninformative rules.

We created 3 types of predictive features, as follows.

(1) Numeric Alignment-Independent Features: We extracted the following numeric features described in [23, 8]: "Amino Acid Composition" (21 features), "Composition" (3 features), "Transition" (3 features), "Distribution" (15 features), and "Z-Values" (15 features). Furthermore, all datasets (Numeric, PPI and Motif) have 2 additional features: "Sequence Length" (the amino acid sequence length), and "Molecular Weight" (the molecular weight of the protein). These 59 features are called alignment-independent, as no alignment procedure, such as "BLAST", is required to be performed on the sequences prior to their calculation.

In addition, following [22], we extended each of the human ageing datasets with the D_n/D_s ratio, which measures the degree of conservation between 2 gene sequences [24]. Using the D_n/D_s ratios from the BioMart tool[3] we extracted 288 D_n/D_s ratios from the human/rhesus genes. 8 genes had no homologs in the *Homologene* dataset and have missing values for this D_n/D_s feature in the datasets.

(2) Protein-Protein Interaction (PPI) Features: This type of binary feature indicates whether or not an ageing-related protein interacts with each of a set of other proteins (which may or may not be ageing-related). Interacting partners of one protein often give away hints of its function [25]. This type of feature was recently used in ageing-related datasets [22]. We have used the BioGrid[4] database to extract PPIs and have only considered features representing interacting partners occurring in 3 or more instances in the dataset, to avoid classifier over-fitting due to rare protein interactions.

2. http://david.abcc.ncifcrf.gov
3. http://www.ensembl.org/biomart/
4. http://thebiogrid.org

TABLE 2: Number of features for each organism (dataset).

Species	Number of features			Number of instances
	Numeric	PPI	Motifs	
Caenorhabditis elegans	59	162	112	263
Drosophila melanogaster	59	105	55	79
Homo sapiens	60	2425	284	301
Mus musculus	59	29	40	107
Saccharomyces cerevisiae	59	4397	296	762

(3) Protein Motif Features: The binary motif features represent the presence or absence of a motif in a protein's amino acid sequence. A motif is a template describing similar sequences of amino acids that occur recurrently in proteins. Motifs serve as a high-level representation of a protein and it is expected that proteins sharing some specific motifs share similar functions. We have used the same 4 motif datasets investigated in [8]: Interpro [26], Pfam [27], Prosite [28], PRINTS [29]. We have only considered motifs occurring in at least 3 proteins (instances) in the dataset, to avoid overfiting as mentioned earlier.

Table 2 shows the number of features of each dataset type and model organism.

5 Computational Results

In this section we present the evaluation of the predictive performance and run time of the 5 hierarchical classification algorithms used in our experiments: the Predictive Clustering Tree (PCT) algorithm, the Hierarchical Dependence Network (HDN) algorithm, the hybrid HDN-PCT algorithm, the Local Hierarchical Classification (LHC) algorithm, and the new hybrid LHC-PCT proposed in this work. Our hybrid algorithms were implemented in the Python programming language. We use the SVM from libSVM [30]. The PCT is implemented in Java and the CFS implementation comes from WEKA.

We have also tested two variations in each of the 4 hierarchical classification algorithms building local classifiers (i.e., all algorithms except the stand-alone PCT): first, we have varied the base local classification algorithm, testing Naive Bayes (NB) and Support Vector Machine (SVM) algorithms. We have chosen these algorithms for their complementary pros and cons: SVM is a complex algorithm having a high predictive performance but producing black-box models, that are difficult to interpret. Additionally SVM can have a significantly lengthy training time. The NB algorithm, on the other hand, is a simple and fast algorithm that produces potentially interpretable models.

Second, we have tested the effect of using the Correlation-based Feature Selection (CFS) algorithm to reduce the number of features available to each local base classification algorithm. CFS takes into account interactions among features and discards redundant features. This is particularly important for the NB algorithm, as over-counting the evidence given by highly correlated features is known to decrease NB's predictive accuracy.

To estimate predictive accuracy we have used 10-fold cross-validation, i.e., we randomly divided each dataset in 10 disjoint folds, train the classification algorithms using 9 folds and test them using the held-out fold. This procedure is repeated 10 times, each time with a different held-out fold, and the results averaged. Tuning of the algorithms' parameters is done using only the training folds.

5.1 Predictive Performance Evaluation

We have used three measures of predictive accuracy: $AU(\overline{PRC})$, $\overline{AUPRC}_w$, and $\overline{AUPRC}$ [9]. These measures are variations of the hierarchical version of the $AUPRC$ (Area Under the Precision Recall Curve) measure, which is used for classifiers with probabilistic outputs. For each class and for each instance, we construct a PR curve (a plot of the classifier's precision as a function of its recall) by thresholding the output (class probability) of the classifier using values in the interval $[0, 1]$. Each threshold is associated with a value of precision and recall, corresponding to a point in the PR space. To obtain a single performance measure from the curve, we calculate the area under the curve using a trapezoidal approximation [31]. A perfect classifier would have an $AUPRC$ of 1.0.

To calculate $AU(\overline{PRC})$, we use the hierarchical versions of precision and recall for a fixed threshold, defined as:

$$hP \equiv \frac{\sum_j |P_j \cap T_j|}{\sum_j |P_j|} \quad \text{and} \quad hR \equiv \frac{\sum_j |P_j \cap T_j|}{\sum_j |T_j|}.$$

Where P_j is the set of predicted classes of the j-th instance and T_j is the set of true classes of the j-th instance.

To calculate $\overline{AUPRC}$ we average all the class-wise $AUPRC$ performances. Similarly, to calculate $\overline{AUPRC}_w$, we calculate the $AUPRC$ of each class and then the average over all classes weighted by the number of instances in each class, that is, $\overline{AUPRC_w} \equiv \frac{\sum_i AUPRC_i \times S_i}{\sum_i S_i}$; where S_i is the number of instances in the i-th class.

Tables 3 to 5 report the results for each of the three evaluation measures. Underlined values represent the best predictive accuracy results in each row, i.e., across several hierarchical classifiers varying the base classifier (SVM or NB) and using or not CFS, for each combination of organism and dataset types. The last row in each table shows the mean rank of a particular combination of hierarchical classifier, using or not CFS and using SVM or NB as base classifier.

5.1.1 Statistical Analysis

We have used the Friedman test, as proposed in [32], to detect if there is any statistically significant difference among the accuracies of the 5 hierarchical classification algorithms for each of the 12 combinations of two base classifiers, 3 performance measures and using or not the CFS feature selection method. Table 6 shows the values of the Iman statistic (used by the Friedman test), the larger the Iman statistic, the larger the difference of predictive accuracies between the classifiers. Numbers in bold denote that the Friedman test has detected some statistically significant difference among the classifiers. For the 11 results with statistically significant differences, we perform the *Hochberg* post-hoc test to check if there are statistical differences between the best performing classifier and the others. The results of the post-hoc test are presented in Figure 2. Asterisks indicate algorithms that are significantly statistically worse than the best performing algorithm (the control).

TABLE 3: Predictive accuracy results with the $AU(\overline{PRC})$ measure (%).

Org.	Feat.	PCT	HDN-PCT				HDN				LHC-PCT				LHC			
			No CFS		CFS		No CFS		CFS		No CFS		CFS		No CFS		CFS	
			NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM
worm	Num.	41.1	34.5	39.9	33.5	39.1	34.4	39.6	33.5	39.2	37.8	44.0	39.5	41.7	37.8	44.0	39.5	41.7
	PPI	41.3	34.9	38.3	35.4	39.5	30.1	38.0	32.9	38.4	39.2	40.6	40.7	40.6	35.8	41.5	40.7	41.1
	Motifs	48.5	44.2	46.9	43.5	47.7	36.0	42.1	36.3	40.6	47.8	48.6	49.3	48.8	41.1	43.9	44.7	43.0
fly	Num.	42.1	40.9	41.9	41.8	41.7	42.3	43.8	43.6	43.6	40.7	42.1	41.8	41.8	42.1	44.0	43.6	43.7
	PPI	49.2	49.2	49.2	49.2	49.2	43.2	44.0	44.3	43.9	49.2	49.2	49.2	49.2	43.0	44.0	44.5	44.2
	Motifs	45.8	45.8	45.8	45.8	45.8	43.1	44.1	43.8	43.8	45.7	45.8	45.8	45.8	43.2	44.0	44.1	44.0
human	Num.	45.5	25.2	40.0	24.4	41.3	23.0	39.8	21.7	41.0	38.9	46.8	43.6	45.7	39.2	47.5	43.8	46.3
	PPI	47.3	42.9	45.7	42.9	43.7	30.1	44.4	32.8	44.0	45.2	45.9	46.0	45.1	41.3	48.4	42.4	46.2
	Motifs	47.2	29.7	48.3	29.1	43.1	25.3	48.7	26.2	42.3	43.4	49.8	47.7	46.9	40.6	50.2	46.9	47.2
mouse	Num.	46.6	40.8	44.9	41.1	44.5	40.2	44.9	40.7	44.5	42.0	47.1	45.7	46.3	42.0	47.1	45.7	46.3
	PPI	45.6	42.6	45.7	42.4	44.9	42.0	46.2	41.3	45.2	45.9	46.6	46.3	46.0	46.1	47.9	46.9	46.6
	Motifs	46.6	38.9	46.8	41.3	44.8	39.0	46.9	41.2	44.7	42.9	47.5	46.5	45.9	42.9	47.5	46.5	45.9
yeast	Num.	42.6	29.8	37.9	29.5	38.4	28.8	33.8	27.6	34.7	36.8	45.2	40.5	43.6	35.0	47.1	36.8	45.6
	PPI	44.9	35.4	42.4	37.7	41.6	33.4	41.8	37.6	43.3	42.0	46.3	44.0	45.0	40.7	46.6	45.4	47.7
	Motifs	42.4	31.8	39.0	31.7	38.5	24.2	39.7	23.8	32.1	41.2	43.5	43.8	43.1	38.6	44.7	43.9	43.8
Avg. Rank		5.3	13.3	7.4	12.7	9.5	15.8	8.9	15.1	11.1	10.2	3.7	5.7	6.1	12.5	3.3	7.0	5.5

TABLE 4: Predictive accuracy results with the $\overline{AUPRC}_w$ measure (%).

Org.	Feat.	PCT	HDN-PCT				HDN				LHC-PCT				LHC			
			No CFS		CFS		No CFS		CFS		No CFS		CFS		No CFS		CFS	
			NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM
worm	Num.	31.9	31.8	31.8	31.9	31.8	31.6	31.6	31.4	30.2	31.8	31.9	31.9	31.8	31.8	32.7	31.9	31.0
	PPI	33.4	31.6	31.6	31.6	31.2	30.8	30.6	30.5	30.0	32.9	32.4	33.5	33.4	32.2	31.6	32.1	32.2
	Motifs	34.6	34.1	34.6	34.1	33.9	32.3	33.1	31.5	31.8	34.9	34.1	35.2	34.7	34.3	33.3	33.7	33.0
fly	Num.	39.8	39.8	39.8	39.8	39.8	38.5	38.4	38.5	38.4	39.8	39.8	39.8	39.8	38.5	38.4	38.6	38.4
	PPI	40.9	40.9	40.9	40.9	40.9	38.5	38.6	38.5	38.5	40.9	40.9	40.9	40.9	38.5	38.5	38.5	38.4
	Motifs	40.4	40.4	40.4	40.4	40.4	38.4	38.4	38.4	38.3	40.4	40.4	40.4	40.4	38.6	38.3	38.5	38.6
human	Num.	36.0	35.9	36.0	35.9	35.9	34.7	34.2	34.1	33.2	36.0	36.0	36.0	36.0	35.1	36.1	35.4	34.3
	PPI	38.2	38.3	38.0	38.3	37.6	39.2	39.0	39.4	37.8	38.2	37.9	38.3	38.0	40.0	39.5	40.1	39.1
	Motifs	38.6	37.8	39.1	37.4	36.9	36.9	39.1	36.7	35.6	39.5	39.1	39.4	39.2	38.8	39.2	38.9	38.6
mouse	Num.	37.6	37.6	37.6	37.6	37.6	36.9	36.8	36.6	36.2	37.6	37.6	37.6	37.6	36.7	36.8	37.1	36.2
	PPI	39.2	38.9	39.1	38.8	38.6	37.5	37.6	37.2	37.1	39.2	38.9	39.7	39.6	38.3	37.3	38.4	38.4
	Motifs	37.8	38.1	38.5	37.5	37.9	36.9	37.3	36.5	36.6	38.7	38.1	38.2	38.2	37.8	37.3	37.6	37.6
yeast	Num.	32.2	32.1	32.3	31.9	31.9	32.3	30.8	31.9	30.0	32.4	32.6	32.4	32.1	33.8	35.6	33.3	33.8
	PPI	36.6	36.2	36.2	36.5	36.0	36.6	36.5	39.5	37.9	36.3	36.3	36.5	36.0	37.4	36.5	40.6	38.3
	Motifs	31.6	30.0	31.3	29.7	30.6	30.2	33.1	30.1	30.2	31.9	31.4	32.3	32.1	33.1	33.6	33.3	32.5
Avg. Rank		6.5	8.7	7.5	9.1	10.2	11.8	11.6	13.6	15.4	5.4	6.9	4.3	6.2	8.7	8.5	7.8	10.6

TABLE 5: Predictive accuracy results with the $\overline{AUPRC}$ measure (%).

Org.	Feat.	PCT	HDN-PCT				HDN				LHC-PCT				LHC			
			No CFS		CFS		No CFS		CFS		No CFS		CFS		No CFS		CFS	
			NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM	NB	SVM
worm	Num.	9.6	9.6	9.6	9.6	9.6	8.9	8.8	8.8	8.5	9.6	9.6	9.6	9.6	8.9	9.2	8.9	8.7
	PPI	9.8	9.5	9.4	9.5	9.5	8.5	8.6	8.5	8.4	9.8	9.6	9.9	9.8	9.0	8.8	9.0	9.0
	Motifs	10.5	10.5	10.5	10.3	10.4	9.0	9.1	8.8	8.9	10.6	10.4	10.7	10.6	9.6	9.1	9.4	9.2
fly	Num.	13.3	13.3	13.3	13.3	13.3	12.8	12.8	12.8	12.8	13.3	13.3	13.3	13.3	12.8	12.8	12.8	12.8
	PPI	13.5	13.5	13.5	13.5	13.5	12.8	12.8	12.8	12.8	13.5	13.5	13.5	13.5	12.8	12.8	12.8	12.8
	Motifs	13.4	13.4	13.4	13.4	13.4	12.8	12.8	12.8	12.8	13.4	13.4	13.4	13.4	12.8	12.8	12.8	12.8
human	Num.	9.4	9.4	9.4	9.4	9.4	8.8	8.7	8.7	8.5	9.4	9.4	9.4	9.4	8.9	9.1	9.0	8.6
	PPI	10.0	10.0	10.0	10.0	9.9	9.9	9.8	9.9	9.6	10.0	10.0	10.0	10.0	10.1	10.0	10.2	9.9
	Motifs	10.3	10.1	10.4	10.0	10.0	9.4	9.9	9.4	9.2	10.5	10.4	10.5	10.5	10.0	9.9	10.0	10.0
mouse	Num.	13.2	13.2	13.2	13.2	13.2	12.7	12.7	12.6	12.6	13.2	13.2	13.2	13.2	12.7	12.7	12.8	12.6
	PPI	13.8	13.7	13.7	13.7	13.7	12.9	12.9	12.9	12.8	13.8	13.7	13.9	13.9	13.1	12.8	13.1	13.1
	Motifs	13.3	13.3	13.4	13.2	13.2	12.8	12.9	12.7	12.7	13.5	13.3	13.4	13.4	13.0	12.8	12.9	12.9
yeast	Num.	8.0	8.0	8.0	8.0	8.0	7.8	7.1	7.8	6.8	8.0	8.0	8.0	8.0	8.2	8.6	8.0	8.1
	PPI	9.6	9.6	9.6	9.6	9.6	9.6	9.4	10.3	9.7	9.6	9.6	9.6	9.6	10.0	9.6	10.7	9.9
	Motifs	7.9	7.6	7.9	7.6	7.8	7.4	8.0	7.1	7.2	8.0	7.9	8.1	8.0	8.7	8.2	8.7	8.4
Avg. Rank		5.5	7.0	6.4	7.4	8.8	13.5	13.8	14.3	15.3	4.3	6.1	3.8	5.9	9.1	11.6	8.7	11.5

TABLE 6: Iman statistic values. Critical value for $\alpha = 0.05$ is 2.54. Values greater than the critical value (in bold face) mean that there is some statistically significant difference in the results. The larger the Iman statistic, the larger our confidence that the difference of predictive accuracies between the classifiers is not just due to chance.

Configuration		Measures		
		$AU(\overline{PRC})$	$\overline{AUPRC}_w$	$\overline{AUPRC}$
No CFS	SVM	**10.83**	2.10	**11.16**
	NB	**20.56**	**5.44**	**15.06**
CFS	SVM	**17.30**	**11.82**	**15.17**
	NB	**18.80**	**10.16**	**17.13**

TABLE 7: Number of times an algorithm in a row was statistically significantly better than the one in a column for each predictive measure.

Measure		PCT	HDN-PCT	HDN	LHC-PCT	LHC
$AU(\overline{PRC})$	PCT		1	1	1	1
	LHC-PCT		1	1		
	LHC		2	2		
$\overline{AUPRC}_w$	LHC-PCT		1	3		
$\overline{AUPRC}$	PCT			1		
	LHC-PCT			3		2

TABLE 8: PPI features with classification coverage > 0.5 in the decision trees built by the PCT algorithm.

Org.	Rank	Feat. Id.	Full Name	Score
Worm	1	LET60	LEThal family member	1.00
	2	wei	Molecular weight	0.94
Fly	1	AMN	Amnesiac	1.00
	2	len	Sequence length	0.92
Human	1	CREBBP	CREB binding prot.	1.00
	2	PTPN11	Tyrosine-prot. phosphatase non-rec. 11	0.83
	3	TP53	Transf.-related Prot. 53	0.76
	4	VTN	Vitronectin	0.52
	5	NOTCH1	Notch homolog 1	0.51
Mouse	1	TP53	Transf. related prot. 53	1.00
	2	POU5F1	POU domain, class 5, transcription factor 1	0.92
Yeast	1	HHT1	Histone H3	1.00
	2	RPS17B	Ribosomal prot. 51	0.85
	3	ATP6	ATP synthase	0.80
	4	PIF1	PIF1 5′-To-3′ DNA Helicase	0.79
	5	FCY2	Purine-cytosine permease	0.77
	6	CYR1	Adenylate cyclase	0.76
	7	HOS2	Hist. deacetylase and subunit of Set3 and Rpd3L complexes	0.73
	8	VPS38	Vacuolar prot. sorting-assoc. prot. 38	0.70
	9	len	Molecular length	0.67
	10	ATG12	Ubiquitin-like prot. ATG12	0.67
	11	IDH2	Isocitrate dehydrogenase (NAD) subunit 2, mitochondrial	0.56
	12	BAS1	Myb-like DNA-binding prot. BAS1	0.52
	13	CLN1	G1/S-specific cyclin CLN1	0.51

TABLE 9: Minimum, maximum and mean of the training time in hours for the measure $AU(\overline{PRC})$ when using NB as a base classifier, across all datasets.

Algorithms		Min.	Max.	Mean
PCT		0.1	3.5	0.7
HDN-PCT (with NB)	No CFS	0.1	6.2	0.9
	CFS	0.1	446.5	34.6
HDN (with NB)	No CFS	0.003	2.1	0.4
	CFS	0.05	576.6	63.0
LHC-PCT (with NB)	No CFS	0.1	6.2	0.9
	CFS	0.1	402.1	33.4
LHC (with NB)	No CFS	0.003	5.0	0.5
	CFS	0.04	576.2	61.6

TABLE 10: Minimum, maximum and mean of the training time in hours for the measure $AU(\overline{PRC})$ when using SVM as a base classifier, across all datasets.

Algorithms		Min.	Max.	Mean
PCT		0.1	3.5	0.7
HDN-PCT (with SVM)	No CFS	0.1	4.4	0.8
	CFS	0.1	443.0	34.6
HDN (with SVM)	No CFS	0.003	1.6	0.2
	CFS	0.1	517.0	62.3
LHC-PCT (with SVM)	No CFS	0.1	4.5	0.8
	CFS	0.1	409.7	32.4
LHC (with SVM)	No CFS	0.002	1.6	0.2
	CFS	0.05	550.9	68.4

Figures 2(a) to 2(d) show the statistical analysis of the results considering the $AU(\overline{PRC})$ measure. The PCT algorithm was significantly better than the other four classifiers when using NB and not using CFS (Figure 2(b)). However, the LHC algorithm had the best rank in two scenarios (Figures 2(a) and 2(c)), and was significantly better than two algorithms in each of these scenarios.

For the $\overline{AUPRC}_w$ performance measure we can see in Figures 2(e), 2(f) and 2(g) that the LHC-PCT algorithm outperformed every other algorithm that we tested.

Finally, considering the measure $\overline{AUPRC}$, the LHC-PCT algorithm outperformed the PCT algorithm in 3 out of 4 occasions, being significantly better than the HDN algorithm in every test.

To summarize the results, we present in table 7 the overall number of times the best performing algorithm was significantly statistically better than each of the others with the measures $AU(\overline{PRC})$, $\overline{AUPRC}_w$ and $\overline{AUPRC}$. In the first row of Table 7, for the $AU(\overline{PRC})$ measure, we can see that both PCT and LHC were statistically significantly better than another algorithm in 4 occasions, while the LHC-PCT algorithm in two occasions. In the second row of Table 7 we can see that LHC-PCT was statistically significantly better than another algorithm in 4 occasions, always in relation to the HDN algorithm. In the last row of Table 7 we can see that, once again, LHC-PCT had better performance than the HDN algorithm in 3 occasions, ranking as the best algorithm in 3 out of 4 times and being statistically significantly better than another algorithm 5 times.

Additionally, we have observed that different predictive accuracy measures tend to favor different types of PCT models (trees). Particularly, models trained at maximizing the $AU(\overline{PRC})$ measure are shallower, and are degenerated (a tree with a single leaf node) more often than the others. Conversely, models trained maximizing the $\overline{AUPRC}$ predictive accuracy measure tend to be larger, with the $\overline{AUPRC}_w$ predictive accuracy measure in between the other two, considering PCT model size. If users of the PCT algorithm desire smaller and higher-level models they should set the parameter s using $AU(\overline{PRC})$, whereas more specific and larger models can be obtained using $\overline{AUPRC}$.

5.2 Training Runtime Analysis

The merit of classification algorithms mainly rests on their predictive performances; however, for some applications, running time is also important. For this reason we present in Table 9 the training times of the algorithms that use NB as a base classifier and in Table 10 the algorithms that use SVM as a base classifier. These tables present, for each combination of hierarchical classification algorithm and whether or not it uses CFS, the minimum, maximum and average training times over all 15 datasets. Each algorithm was run on a Intel Xeon machine with clock speed of 2.27 GHz and 11 GB of RAM memory.

We present the training times in two different tables because the NB and SVM times are not comparable: NB was implemented in the Python programming language, which is generally slower than the C++ language used by the SVM algorithm (from LibSVM library). Also, due to lack of space, we present the training time of the algorithms only for the $AU(\overline{PRC})$ evaluation measure. The training times for the other measures are similar and have the same patterns in terms of identifying faster and slower algorithms.

The tables show that the use of CFS increases training time significantly, as expected. When using the PCT hybrids (HDN-PCT and LHC-PCT) without CFS, the increase in training time in comparison with the training time of the PCT by itself is in general small (recall that in order to use the hybrid algorithms, a PCT model must be trained first). This means that our hybrid algorithms have compatible training times with the PCT algorithm when not using CFS.

6 Classification Model Interpretation

In this section we interpret PCT models to extract potentially relevant ageing-related knowledge from them. It is worth noticing that interpreting the PCT models is valuable for the stand-alone PCT models and the hybrid algorithms that use the same PCT model as a base to induce further classification models (HDN-PCT and HLC-PCT).

6.1 Feature Scores for the PCT Models

One possibly useful information extractable from the PCT models is which features were more "useful" to the decision tree, i.e., features that were used more often to predict ageing-related GO terms. To achieve this, we calculated the "coverage score" for every feature present in the decision tree. Higher scores correspond to more useful features.

The coverage score of a given feature is calculated by building a PCT decision tree (model) using all available

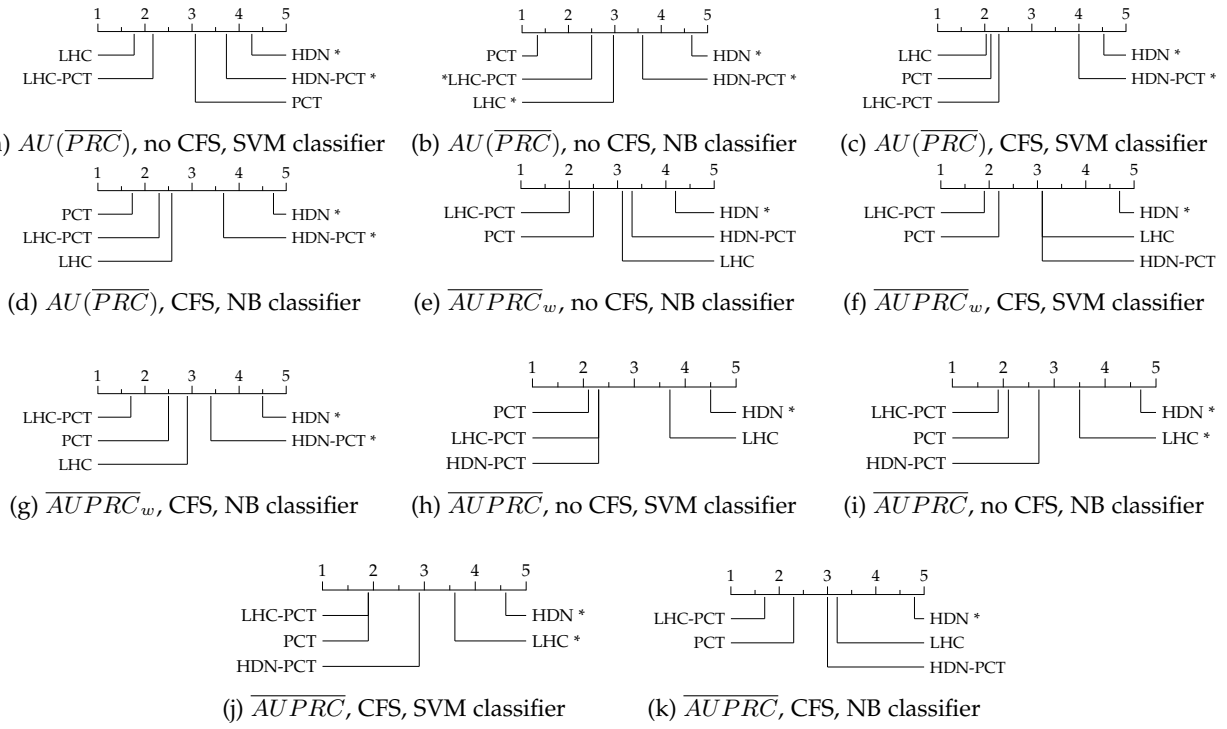


Fig. 2: Statistical analysis of the hierarchical classifiers varying the base classifiers, using or not CFS, and predictive accuracy measure. Numbers in the horizontal axis represent the mean rank of the hierarchical classifier. Algorithms with statistically significantly lower accuracy than the best hierarchical classifier (control algorithm) are marked with an asterisk.

instances in the entire dataset and dividing the number of instances that used that feature for their classification (i.e., the feature occurs in the path from the root to the leaf node where the instance is classified) by the total number of instances that were classified. In particular, a feature in the root node has the maximum coverage of 1, since that feature is used to classify all instances in the dataset; whereas features in deeper nodes have lower coverage, since they are used to classify fewer instances.

Table 8 presents the features with a coverage score greater than 0.5 in the PCT decision trees for the PPI dataset using the s-value that maximizes the $\overline{AUPRC}$ measure. We have chosen this measure because it generates larger decision trees, maximizing the opportunity of finding interesting relations between features and ageing-related GO terms. We focus on the PPI dataset because it has greater interpretability compared to the numeric and motif datasets.

Table 8 shows for each feature: its rank, from most relevant (1) to least relevant; the feature identifier; the full name of the feature; and the feature's coverage score.

Analysing the human dataset, the top-ranked PPI feature is represented by interaction with CREBBP (CREB binding protein), an acetyl transferase involved in the acetylation of histones and other proteins within the cell [33]. CREBBP has diverse functions and is important for development, physiology and disease. One of its targets is the FOXO transcription factor [34], a protein tightly linked to longevity and ageing in several species, including humans [35].

The gene that encodes the protein ranked second in this list is PTPN11 (protein tyrosine phosphatase, non-receptor type 11). Mutations in this gene are known to cause myeloid leukemia, dramatically reducing human life expectancy [36]. In particular, activating mutations in this gene have been shown in cell culture to cause proliferative arrest and senescence. The mechanisms of which could provide insight into the initial onset of PTPN11 related cancers [37].

The third-ranked feature represents interaction with TP53 (transformation-related Protein 53), commonly known as p53 and an important tumor suppressor that has been described as "the guardian of the genome". Its relationship with ageing is complex as decreased expression of the gene leads to tumor formation which increases mortality, but in contrast certain mutations in the gene have also been linked to increased life expectancy in humans [38]. Interestingly, interaction with TP53 was also selected as an important feature in the mouse dataset. This supports the hypothesis that this protein may play an important role in the prediction of ageing-related GO terms.

For the yeast, the top 13 features illustrate a wide variety of molecular interactions making it difficult to comment on individual processes. However, one of the top genes on the list is Ribosomal protein 51. Ribosomal proteins, via their role in translation, are strongly implicated in lifespan regulation in several species and represent a key mode of ageing modulation [39].

6.2 Interpreting Classification Models

In this section we present some classification models (decision trees) generated by the PCT algorithm and the analysis of some over-represented ageing-related GO terms in some leaf nodes of the decision tree. Recall that the PCT algorithm

generates a decision tree that contains, in each node, a test that splits the testing instances into two different paths. When an instance reaches a leaf node (a cluster), a class (GO term) probability vector is assigned to it.

We focus again on the PPI features, which are easier to interpret in comparison with the numeric and motif features, and we discard the human and yeast PCT trees (with 13 and 52 nodes, respectively), due to space limitation.

Figures 3 to 5 show some over-expressed GO terms and the decision tree generated by the PCT algorithm maximizing the $\overline{AUPRC}$ measure for the worm, fly and mouse organisms, respectively. The choice of over-expressed GO terms in the tables in figures 3(b), 4(b) and 5(b) took into account both their small p-value and their relevance to ageing based on current biological knowledge.

In the models, the first number in the pair before the first decision split represents the a-priori mean entropy of the class-labels. Entropy is defined as: $H(\hat{P}) = -\sum_i \hat{P}_i \log(\hat{P}_i)$, where $\hat{P}_i$ is the proportion of instances with class i in the current node of the decision tree. Smaller entropy values represent sets of instances that have more separated (more reliably predicted) classes. The second number in the pair before the first decision split represents the total number of instances classified by the decision tree. In the leaf nodes, the pair of numbers represents the entropy and number of instances in the current leaf node.

In the decision tree shown in Figure 3(a), if a given protein does not interact with "LET60" (LEThal family member 60), then the protein is assigned to a cluster based on its molecular weight. In Figure 3(a) the number of proteins that do not interact with "LET60" is 246, almost all of the 263 instances, and the entropy value did not decrease significantly. Hence, *not* interacting with the "LET60" protein is not a good predictor of ageing-related GO terms. On the other hand, proteins that interact with both "LET60" and "LIN45" (assigned to cluster 3) have a much smaller class entropy compared to other clusters, indicating that these two protein interactions are relevant predictors of ageing-related GO terms.

The table in Figure 3(b), below the decision tree, shows 3 over-expressed ageing-related GO terms in the clusters of the leaf nodes for the worm PPI dataset. This table shows: the cluster ID, the GO term id and name, and the probability that the number of occurrences of the GO term in the current cluster is greater than or equal to the number of observed occurrences, assuming that the occurrence of the GO term in an instance follows a Bernoulli distribution with the GO term's frequency in the training dataset used as the 'success' probability. The table also displays, before the GO terms of each cluster, the feature-based conditions in the decision tree that must be satisfied in order for an instance to be assigned to a particular cluster.

Considering the worm dataset, of the GO terms reaching a significance of 10^{-4} or less, almost half related to developmental processes (full list not shown). This is not surprising as development and, more recently, growth, have been implicated in ageing related processes [40, 41]. The remaining significant GO terms were divided almost equally between involvement in either reproductive processes or signalling pathways. Again, it is difficult to argue against the importance of either of these in ageing, particularly the latter where the use of genetics and molecular biology have allowed key ageing pathways to be dissected in worms [40].

Figure 3(b) suggests that the feature representing interaction with the "LET60" (LEThal family member 60) protein is a good predictor for GO terms related to organism development in worms. This is consistent with the fact that the let-60 gene encodes the *C. elegans* Ras protein, which is central to a variety of different signaling pathways. One of these is the RTK-Ras-ERK pathway, which is well conserved between species [42].

This pathway is critical during development and controls many biological processes during adulthood. Mutations in components of the RAS pathway are also implicated in many human syndromes and diseases, e.g. cancer [43].

In worms, let-60 is expressed in neural, muscle, and hypodermal lineages and its activity is required for proper larval development. It has also been linked to ageing and shown to promote longevity. Two suggested mechanisms for this are: 1) Promoting protein homeostasis via the ubiquitin proteosome system (UPS) [44]; and 2) Activation of the transcription factor SKN-1, which represses insulin-like peptide expression and down regulates the insulin signalling pathway [45]. There are also other, less direct, implications that Ras acts to control ageing and healthspan. Thus, it is interesting that (based on the large coverage score) interaction with LET60 is the most important discriminator for ageing-related GO terms in the worm PPI dataset.

```
(Ent.: 77.50, #Inst.: 263)
LET60 = No (74.92, 246)
|    wei > 96753.69 (88.06, 53)
|         Cluster 0
|    wei <= 96753.69 (68.51, 193)
|         Cluster 1
LET60 = Yes (69.41 17)
|    LIN45 = No   (66.77, 12)
|         Cluster 2
|    LIN45 = Yes (34.77, 5)
|          Cluster 3
```

(a) PCT classification model for the worm PPI dataset.

Clust.	GO term	p-value
	IF LET60 = Yes AND LIN45 = No	
2	GO:0050793 (Regulation of devel. proc.)	10^{-4}
	IF LET60 = Yes AND LIN45 = Yes	
3	GO:0022414 (Reproductive process)	10^{-5}
	GO:0023052 (Signalling)	10^{-5}

(b) Most statistically significant over-expressed ageing-related GO terms in PCT's leaf nodes for the worm PPI dataset.

Fig. 3: PCT model and over-expressed GO terms (classes) in the worm PPI dataset

Figure 4(a) shows the decision tree generated for the fly dataset, where interactions with the AMN (Amnesiac) protein greatly reduce the entropy of the class labels.

In the fly dataset, most over-represented GO terms (full list not shown) were involved in development, food recognition and behaviour, learning and memory. As with the worm, development and growth are not surprising. However, lifespan can be extended dramatically in a wide variety of organisms by reducing caloric intake [46]; thus it

is logical that feeding, and behaviour that influences this, would affect lifespan.

Interestingly, the over-expressed GO terms in cluster 2 are associated with brain development. In fact, AMN has been shown to be required for normal brain development, sleep regulation and adult memory consolidation [47]. Both sleep regulation and memory decline with age in a number of species, and indeed AMN is linked to these processes in an age-dependent fashion in drosophila [48]. Thus, it is interesting that this gene was identified as relevant.

```
(115.88, 79)
AMN = No (112.60, 73)
|     len > 741.0 (127.81, 20)
|         Cluster 0
|     len <= 741.0 (95.58, 53)
|         Cluster 1
AMN = Yes (9.80, 6)
    Cluster 2
```

(a) PCT classification model for the fly PPI dataset.

Clust.	GO term	p-value
	IF AMN = Yes	
2	GO:0007631 (feeding behaviour)	10^{-6}
	GO:0007613 (memory)	10^{-6}
	GO:0048589 (devel. cell growth)	10^{-5}

(b) Most statistically significant over-expressed ageing-related GO terms in PCT's leaf nodes for the fly PPI dataset.

Fig. 4: PCT model and over-expresed GO terms (classes) in the fly PPI dataset

Figure 5(a) shows the decision tree built for the mouse dataset. There are fewer significant GO terms in the mouse dataset and these are enriched for terms implicated in a variety of different regulatory processes (full list not shown), i.e. apoptotic pathways, cell cycle and regulation of gene expression. This contrasts with the worm and fly data where "developmental processes" predominate. However, it does complement the analysis based on feature coverage scores, showing that interaction with p53 (a key signalling molecule) is important for ageing in humans and mice.

7 Conclusion and Future Work

We have compared the predictive performance of 5 types of probabilistic hierarchical classification algorithms in 15 ageing-related datasets. Studies that compare several algorithms and datasets on the task of hierarchical classification of biological data are uncommon, one exception is [49], which focuses on a different type of hierarchical classification algorithms. We have also proposed ways to interpret the generated decision tree models to possibly get biological insights about the ageing process. This kind of interpretation is hardly found in works on hierarchical classification (with a few exceptions such as [50]), presumably due to the large number of class labels; but we have shown how such interpretation can produce comprehensible patterns.

We have concluded that overall, taking into account the results across the 15 datasets used in our experiments, among the classifiers that we tested for this type of problem, the LHC-PCT classifier was the best regarding predictive

```
(190.68 107)
TP53 = No (180.25,98)
|     POU5F1 = No (178.78,92)
|         Cluster 0
|     POU5F1 = Yes (96.07,6)
|         Cluster 1
TP53 = Yes (190.97 9)
    Cluster 2
```

(a) PCT classification model for the mouse PPI dataset.

Clust.	GO term	p-value
	IF TP53 = Yes	
2	GO:051726 (regulation of cell cycle)	10^{-5}
	GO:008630 (intrinsic apoptotic signalling pathway in response to DNA damage)	10^{-4}
	GO:010468 (reg. of gene expression)	10^{-5}

(b) Most statistically significant over-expressed ageing-related GO terms in PCT's leaf nodes for the mouse PPI dataset.

Fig. 5: PCT model and over-expresed GO terms (classes) in the mouse PPI dataset

accuracy in the $\overline{AUPRC}_w$ and $\overline{AUPRC}$ measures, having the best mean rank in 6 out of 7 experiments. Considering the $AU(\overline{PRC})$ measure, two different algorithms were the best performing (LHC and PCT). Only one algorithm (PCT) was statistically significantly better than all the others.

According to our tests, using the CFS algorithm to select features in a pre-processing step improved the predictive accuracy of the hierarchical classification algorithms when using NB as a base classifier. Using CFS in this case resulted in better predictive accuracy results than not using CFS in 8 out of 12 tests. For SVM this result was not observed: using CFS only improved the performance in 3 out of 12 tests.

Regarding our datasets of ageing-related GO terms, our results show that besides being more difficult to interpret, overall the numerical features have inferior predictive accuracy compared to the PPI and protein motif features.

As expected, the predictive accuracies reported in Tables 3 to 5 vary across different species and different protein representations for the data of each species. Hence, for each accuracy measure, we can identify the "best" pair of species and representation as the pair leading to the highest average value of accuracy across the 17 algorithms.

This best pair of species and representation can be interpreted as the pair with the greatest predictive power in general, across all algorithms. According to this criterion, the best pair of species and protein representation in our experiments were the fly dataset using the PPI representation for measures $AU(\overline{PRC})$ and $\overline{AUPRC}_w$, and the mouse dataset using the PPI representation for measure $\overline{AUPRC}$.

Analysing the training time of the algorithms we tested, we conclude that, when not using CFS, the time taken to run the hybrid HDN-PCT and LHC-PCT algorithms (which include the PCT run time and the HDN or LHC run time) is not much greater than the time to run only PCT. When using CFS, the training time of the algorithms is in general greatly incresead, which suggests that if smaller training times are important, one should use the algorithms without CFS.

For future work we plan to develop and test new varia-

tions of the HDN algorithm and apply them to the current and other (novel) ageing-related datasets.

Acknowledgment: The first author is financially supported by CAPES, a Brazilian research-support agency (process number 0653/13-6).

REFERENCES

[1] J. a. P. de Magalhães, "The biology of ageing: a primer," in *An Introduction to Gerontology*, 1st ed., Cambridge, UK, 2011, ch. 2, pp. 22–47.

[2] O. R. Jones *et al.*, "Diversity of ageing across the tree of life," *Nature*, vol. 505, no. 7482, pp. 169–173, Jan. 2014.

[3] D. Friedman and T. Johnson, "A mutation in the age-1 gene in Caenorhabditis elegans lengthens life and reduces hermaphrodite fertility," *Genetics*, vol. 1988, no. 1, pp. 75–86, 1988.

[4] H. M. Brown-Borg, Kurt E. Borg, Charles J. Meliska, and A. Bartke, "Dwarf mice and the ageing process," *Nature*, vol. 387, p. 33, 1996.

[5] The Uniprot Consortium, "The Universal Protein Resource (UniProt) in 2010." *Nucleic Acids Research*, vol. 38, pp. D142–D148, Jan. 2010.

[6] M. a. Harris *et al.*, "The Gene Ontology (GO) database and informatics resource." *Nucleic Acids Research*, vol. 32, no. Database issue, pp. D258–61, Jan. 2004.

[7] C. Lin, Y. Zou, J. Qin, X. Liu, Y. Jiang, C. Ke, and Q. Zou, "Hierarchical classification of protein folds using a novel ensemble classifier," *PloS one*, vol. 8, no. 2, p. e56499, 2013.

[8] C. N. Silla Jr. and A. A. Freitas, "A Survey of Hierarchical Classification Across Different Application Domains," *Data Mining and Knowledge Discovery*, vol. 44, no. 1-2, pp. 31–72, 2011.

[9] C. Vens, J. Struyf, L. Schietgat, S. Dzeroski, and H. Blockeel, "Decision Trees for Hierarchical Multi-label Classification," *Machine Learning*, vol. 73, no. 2, pp. 185–214, Aug. 2008.

[10] F. Fabris and A. A. Freitas, "Dependency Network Methods for Hierarchical Multi-label Classification of Gene Functions," *Proc. of the 2014 IEEE Computational Intelligence and Data Mining*, pp. 241–248, Dec. 2014.

[11] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, The University of Waikato, New Zealand, 1999.

[12] H. Blockeel, M. Bruynooghe, S. Dzeroski, J. Ramon, and J. Struyf, "Hierarchical Multi-Classification," in *Proceedings of the ACM SIGKDD 2002 workshop on multi-relational data mining (MRDM 2002)*, 2002, pp. 21–35.

[13] L. Schietgat, C. Vens, J. Struyf, H. Blockeel, D. Kocev, and S. Dzeroski, "Predicting gene function using hierarchical multi-label decision tree ensembles." *BMC Bioinformatics*, vol. 11, no. 2, pp. 1–14, Jan. 2010.

[14] D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie, "Dependency Networks for Inference, Collaborative Filtering, and Data Visualization," *Journal of Machine Learning Res.*, vol. 1, pp. 49–75, 2001.

[15] Y. Guo and S. Gu, "Multi-Label Classification Using Conditional Dependency Networks," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, vol. 2, 2011, pp. 1300–1305.

[16] L. Li, L. Zhang, G. Li, and H. Wang, "Probabilistic classifier chain inference via gibbs sampling," in *Proc. of the 23rd ACM Inter. Conf. on Infor. and Know. Manag. (CIKM '14)*, New York, NY, USA, 2014, pp. 1855–1858.

[17] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human micrornas by incorporating a high-quality negative set," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, vol. 11, no. 1, pp. 192–201, 2014.

[18] J. Metz, A. A. Freitas, M. C. Monard, and E. A. Cherman, "A study on the selection of local training sets for hierarchical classification tasks," in *Proc. of the Brazilian National Meeting on Artificial Intelligence (ENIA-2011)*, Natal, RN, Brazil, 2011, pp. 572–583.

[19] D. Gjorgjevkikj, G. Madjarov, and S. Dzeroski, "Hybrid Decision Tree Architecture Utilizing Local SVMs for Efficient Multi-Label Learning," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 27, no. 07, p. 1351004, Aug. 2013.

[20] R. Tacutu, T. Craig, A. Budovsky, D. Wuttke, G. Lehmann, D. Taranukha, J. Costa, V. E. Fraifeld, and J. a. P. de Magalhães, "Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing." *Nucleic Acids Research*, vol. 41, pp. D1027–D1033, Jan. 2013.

[21] C. N. Silla Jr. and A. A. Freitas, "Selecting different protein representations and classification algorithms in hierarchical protein function prediction," *Intelligent Data Analysis*, vol. 15, no. 6, pp. 979–999, 2011.

[22] 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. 1, p. 27, Jan. 2011.

[23] K. M. Salama and A. A. Freitas, "ACO-Based Bayesian Network Ensembles for the Hierarchical Classification of Ageing-Related Proteins," in *Evo. Comp., Mach. Learning and Data Mining in Bioinformatics*, ser. Lecture Notes in Computer Science, 2013, vol. 7833, pp. 80–91.

[24] Z. Yang and J. P. Bielawski, "Statistical methods for detecting molecular adaptation," *Trends in Ecology & Evolution*, vol. 15, no. 12, pp. 496–503, Dec. 2000.

[25] R. Sharan, I. Ulitsky, and R. Shamir, "Network-based prediction of protein function," *Molecular Systems Biology*, vol. 3, no. 1, 2007.

[26] S. Hunter *et al.*, "Interpro in 2011: new developments in the family and domain prediction database," *Nucleic Acids Research*, vol. 40, no. D1, pp. D306–D312, 2012.

[27] R. D. Finn *et al.*, "The Pfam protein families database," *Nuc. Ac. Res.*, vol. 36, no. suppl 1, pp. D281–D288, 2008.

[28] C. J. A. Sigrist, E. de Castro, L. Cerutti, B. A. Cuche, N. Hulo, A. Bridge, L. Bougueleret, and I. Xenarios, "New and continuing developments at PROSITE," *Nucl. Acids Res.*, vol. 41, no. D1, pp. D344–D347, 2013.

[29] T. K. Attwood *et al.*, "PRINTS and its automatic supplement, prePRINTS," *Nucleic Acids Research*, vol. 31, no. 1, pp. 400–402, 2003.

[30] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011.

[31] K. Boyd, K. H. Eng, and C. D. Page, "Area Under the Precision-Recall Curve: Point Estimates and Con-

fidence Intervals," vol. 8190, pp. 451–466, 2013.
[32] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
[33] D. C. Bedford and P. K. Brindle, "Is histone acetylation the most important physiological function for cbp and p300?" *Aging (Albany NY)*, vol. 4, no. 4, p. 247, 2012.
[34] H. Daitoku, M. Hatta, H. Matsuzaki, S. Aratani, T. Ohshima, M. Miyagishi, T. Nakajima, and A. Fukamizu, "Silent information regulator 2 potentiates Foxo1-mediated transcription through its deacetylase activity," *Proc. of the Nat. Aca. of Sci. of the U. S. A.*, vol. 101, no. 27, pp. 10 042–10 047, Jul. 2004.
[35] J. Guevara-Aguirre *et al.*, "Growth Hormone Receptor Deficiency Is Associated with a Major Reduction in Pro-Aging Signaling, Cancer, and Diabetes in Humans," *Sci. Trans. Med.*, vol. 3, no. 70, p. 70ra13, Feb. 2011.
[36] H.-A. Hou *et al.*, "Characterization of acute myeloid leukemia with PTPN11 mutation: the mutation is closely associated with NPM1 mutation but inversely related to FLT3//ITD," *Leukemia*, vol. 22, no. 5, pp. 1075–1078, Nov. 2007.
[37] H. Zheng, S. Li, P. Hsu, and C.-K. Qu, "Induction of a tumor-associated activating mutation in protein tyrosine phosphatase ptpn11 (shp2) enhances mitochondrial metabolism, leading to oxidative stress and senescence," *J. of Bio. Chem.*, vol. 288, no. 36, pp. 25 727–25 738, 2013.
[38] D. van Heemst, S. P. Mooijaart, M. Beekman, J. Schreuder, A. J. de Craen, B. W. Brandt, P. Eline Slagboom, and R. G. Westendorp, "Variation in the human tp53 gene affects old age survival and cancer mortality," *Exp. geron.*, vol. 40, no. 1, pp. 11–15, 2005.
[39] N. Charmpilas, I. Daskalaki, M. E. Papandreou, and N. Tavernarakis, "Protein synthesis as an integral quality control mechanism during ageing," *Ageing research reviews*, pp. 75–89, 2014.
[40] C. J. Kenyon, "The genetics of ageing," *Nature*, vol. 464, no. 7288, pp. 504–512, 2010.
[41] D. Gems and Y. de la Guardia, "Alternative perspectives on aging in caenorhabditis elegans: reactive oxygen species or hyperfunction?" *Antioxidants & redox signaling*, vol. 19, no. 3, pp. 321–329, 2013.
[42] C. M. Udell, T. Rajakulendran, F. Sicheri, and M. Therrien, "Mechanistic principles of raf kinase signaling," *Cell. and Mol. Life Sci.*, vol. 68, no. 4, pp. 553–565, 2011.
[43] W. E. Tidyman and K. A. Rauen, "The rasopathies: developmental syndromes of ras/mapk pathway dysregulation," *Current opinion in genetics & development*, vol. 19, no. 3, pp. 230–236, 2009.
[44] G. Liu, J. Rogers, C. T. Murphy, and C. Rongo, "Egf signalling activates the ubiquitin proteasome system to modulate c. elegans lifespan," *The EMBO journal*, vol. 30, no. 15, pp. 2990–3003, 2011.
[45] T. Okuyama, H. Inoue, S. Ookuma, T. Satoh, K. Kano, S. Honjoh, N. Hisamoto, K. Matsumoto, and E. Nishida, "The erk-mapk pathway regulates longevity through skn-1 and insulin-like signaling in caenorhabditis elegans," *Journal of Biological Chemistry*, vol. 285, no. 39, pp. 30 274–30 281, 2010.
[46] W. Mair and A. Dillin, "Aging and survival: the genetics of life span extension by dietary restriction," *Annual Review Biochemistry*, vol. 77, pp. 727–754, 2008.
[47] H. Hashimoto, N. Shintani, and A. Baba, "Higher brain functions of pacap and a homologous drosophila memory gene amnesiac: insights from knockouts and mutants," *Biochemical and biophysical research communications*, vol. 297, no. 3, pp. 427–432, 2002.
[48] T. Tamura, A.-S. Chiang, N. Ito, H.-P. Liu, J. Horiuchi, T. Tully, and M. Saitoe, "Aging specifically impairs amnesiac-dependent memory in drosophila," *Neuron*, vol. 40, no. 5, pp. 1003–1011, 2003.
[49] R. Cerri, G. L. Pappa, A. C. P. Carvalho, and A. A. Freitas, "An Extensive Evaluation of Decision Tree-Based Hierarchical Multilabel Classification Methods and Performance Measures," *Computational Intelligence*, vol. 31, no. 1, pp. 1–46, 2015.
[50] J. Levatic, D. Kocev, and S. Dzeroski, "The use of the label hierarchy in HMC improves performance: A case study in predicting community structure in ecology," *Proc. of the workshop on new frontiers in mining complex patterns at ECML/PKDD2013*, pp. 189–201, 2013.

Fabio Fabris received his bachelor (2010) and masters (2013) degrees in Computer Science from the Federal University of Espírito Santo, Brazil. Before starting his Ph.D. he has worked in research projects aimed at solving real-world problems using data-mining and optimization techniques: finding fraudulent electricity consumers, classifying faults in oil-rig motor pumps and optimizing samples sizes for statistical analysis. He is currently pursuing his Ph.D. in Computer Science in the University of Kent, United Kingdom. His main interest is the application of data-mining and optimization algorithms to solve real-world problems. In particular the classification of ageing-related proteins.

Prof. Alex A. Freitas is a Professor of Computational Intelligence at the University of Kent, UK. He has a PhD in Computer Science from the University of Essex, UK (1997), in the area of data mining; and a research-oriented masters degree (MPhil) in Biological Sciences from the University of Liverpool, UK (2011), doing research on ageing with data mining and bioinformatics methods. At present his main research interests are the development of new classification (supervised learning) methods for data mining and knowledge discovery, as well as the application of such methods to biology (particularly the biology of ageing) and pharmaceutical sciences.

Jennifer Tullet joined the School of Biosciences in Sept. 2014 after conducting postdoctoral research with Prof David Gems (University College London) and Prof Keith Blackwell (Harvard). Prior to that, she obtained her PhD from Imperial College London under the supervision of Prof Malcolm Parker. Jennifer's background covers ageing biology, transcriptional regulation and *C. elegans* genetics. Her research focuses on the molecules and processes that regulate lifespan and influence life-long health.