Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: ???
Original Paper

OXFORD

Data and Text Mining

New KEGG pathway-based interpretable features for classifying ageing-related mouse proteins

Fabio Fabris,* and Alex A. Freitas

School of Computing, University of Kent, Canterbury, Kent, CT2 7NF, United Kingdom

*To whom correspondence should be addressed.

Associate Editor: ???

Received on ???; revised on ???; accepted on ???

Abstract

Motivation: The incidence of ageing-related diseases has been constantly increasing in the last decades, raising the need for creating effective methods to analyse ageing-related protein data. These methods should have high predictive accuracy and be easily interpretable by ageing experts. To enable this, one needs interpretable classification models (supervised machine learning) and features with rich biological meaning. In this paper we propose two interpretable feature types based on KEGG pathways and compare them with traditional feature types in hierarchical classification (a more challenging classification task regarding predictive performance) and binary classification (a classification task producing easier to interpret classification models). As far as we know, this work is the first to: (1) explore the potential of the KEGG pathway data in the hierarchical classification setting, (2) use the graph structure of KEGG pathways to create a feature type that quantifies the influence of a current protein on another specific protein within a KEGG pathway graph, and (3) propose a method for interpreting the classification models induced using KEGG features.

Results: We performed tests measuring predictive accuracy considering hierarchical and binary class labels extracted from the Mouse Phenotype Ontology (MPO). One of the KEGG feature types leads to the highest predictive accuracy among five individual feature types across three hierarchical classification algorithms. Additionally, the combination of the two KEGG feature types proposed in this work results in one of the best predictive accuracies when using the binary class version of our datasets, at the same time enabling the extraction of knowledge from ageing-related data using *quantitative influence* information.

Availability: The datasets created in this paper will be freely available after publication.

Contact: ff79@kent.ac.uk

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Ageing-related diseases are affecting an increasing number of people. At the same time, delaying ageing in humans seems to be more and more plausible in the not so distant future. Biologists can already extend the lifespan of several animal species such as the fruit fly and the mouse. The potential economical benefit of investing on this type of research is clear: it is projected that the economical value of adding 2.2 extra healthy years to the human population is $7.1 trillion dollars over 50 years in the United States alone (Goldman *et al.*, 2013).

One of the aims of ageing-research is to treat ageing as a whole, reducing the incidence of many different ageing-related diseases at the same time, instead of focusing on individual diseases. This approach promises to be much more effective than the current approach of treating individual diseases and has the potential of stopping the trend of increasing costs of treating ageing-related diseases (Goldman *et al.*, 2013). One way to study the ageing process holistically is to use data mining algorithms to find connections between genes or proteins that are known to be ageing-related and other genes or proteins that have unknown function using the ever increasing freely accessible biological data.

Two databases of interest for ageing experts are the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2016) and the Mouse Phenotype Informatics (MPI) (Eppig *et al.*, 2015) databases. The MPI database contains, among other data, the definition of an ontology

of ageing-related terms that describe the phenotype of several allele-mutations. The KEGG database also contains several types of information about genes and proteins, including pathway information presented in a graphical way that allows biologists understanding the interactions of proteins in complex biological processes.

We address two types of classification (supervised machine learning) problems: binary classification, where instances (proteins) are annotated with the presence or absence of a class (indicating if the protein is ageing related); and hierarchical classification; where the classes to be predicted (protein functions) are organized into a hierarchy, where more generic functions are ancestors of more specific functions (Silla Jr. and Freitas, 2011a). Note that this is a more complex but more rewarding problem than conventional classification, since the latter ignores hierarchical relationships among classes. We address the hierarchical classification task because the terms in the MPI ontology are hierarchical, and address the binary classification task because it produces classification models easier to interpret, from an ageing-biology perspective. In both tasks, it is crucial to describe each instance (protein) by a set of features (protein properties) that has both good predictive power and rich biological meaning.

The contributions of this work are three-fold: (1) the integration of data from the MPI and KEGG databases to create a new numerical KEGG feature type with rich biological meaning. This new feature type quantifies how an instance (protein) influences other proteins, the idea being that proteins that influence other proteins in a similar way have similar function; (2) the new investigation of the use of the binary KEGG feature type in the context of hierarchical classification; and (3) proposing a method for interpreting classification models generated using KEGG features.

The construction of specially tailored, meaningful features for specific problems is part of the *feature engineering* process (Forman, 2002; Yepes *et al.*, 2015). The objective is to introduce carefully crafted features for the type of problem being addressed. In the bioinformatics field, it has been common to use features that are easily extractable from the protein sequence or from some database containing several protein properties. These features, although valuable, often lack the preciseness an expert needs to reach a meaningful biological conclusion. This works differs from current practice by creating a new KEGG pathway-based feature type that encodes precise and meaningful relations between proteins.

This paper is organised as follows: Section 2 describes how we built our ageing-related datasets, including the proposed KEGG features. Section 3 reports the predictive power of our features across hierarchical and binary classification algorithms and the interpretation of a binary classification model using some of the proposed features. Finally, in Section 4 we discuss the results of our work and draw conclusions.

2 Methods

2.1 Creation of the ageing datasets using the Mouse Genome Informatics dataset

To study the biological aspects of ageing/longevity using hierarchical classification algorithms, we have built 7 datasets containing features extracted from the proteins encoded by the genes in the *Phenotypes and Mutant Alleles* section of the *Mouse Genome Informatics* (MGI) database. The MGI provides the two primary sources of data of our datasets: (1) the definition of the *Mammalian Phenotype Ontology* (MPO), the source of class labels to be predicted, and (2) a list of genotypes annotated with the phenotypes present in the MPO, the source of the features (predictors).

The MPO is organized as a DAG (Directed Acyclic Graph), where each node represents a phenotype (an ontology term) and each edge a "IS-A" relation between phenotypes. Because of the structured organization of the class labels, this is a *hierarchical classification* problem, where the class labels of the instances are organized in a graph, usually a DAG or tree. The nodes of the graph represent class labels and edges are 'IS-A' relationships among class labels. This structural organization means that if an instance is annotated with a given (specific) class label, it is implicitly annotated with all ancestor (more generic) class labels.

The MPO contains 10,907 terms in total, and 113 terms under the term MP:0010768 (ageing/mortality) part of the hierarchy, our research focus. We consider only the 113 ageing-related terms as class labels for our study, and discard the others. Considering all 10,907 terms would generate classification models more focused on predicting non-ageing-related terms, generating models with less interest for the biology of ageing. After further discarding MPO terms with less than 10 instances, we end up with 81 MPO terms, the hierarchical class labels to be predicted.

With the class hierarchy defined, we must create our instances. In the MGI database, 11,532 genotypes are annotated with at least one of the 113 mortality/ageing-related ontology terms. Each genotype is formed by a list of allele-mutations. Each allele-mutation contains (among other information) one or more protein-encoding genes, which in turn are associated with particular mutations. Therefore, using the MPO hierarchy we can associate a protein (instance) with one or more phenotypes (hierarchical classes). Figure 1 shows these relations graphically.

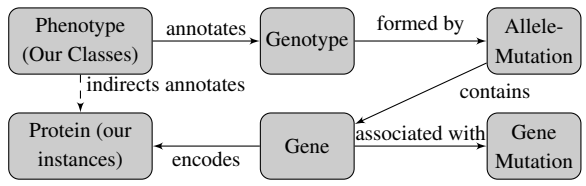

Fig. 1. Relationships among MPI elements and the instances in our datasets. Filled edges represent relationships present in the MGI database. The dashed edge represents the indirect relation that we use for our datasets. Note that we ignore mutation information.

Note that our instances are proteins encoded by standard genes, not gene mutations, because, as discussed later, information about proteins is much richer and precise than information about gene mutations.

However, choosing to use proteins as instances (instead of gene mutations) has the disadvantage of risking annotating the same protein with contradictory MPO terms. This may happen because two different mutations on the same gene may have contradictory effects. E.g., one mutation may over-express the protein encoded by a gene, while some other mutation on the same gene may under-express that protein, possibly leading to opposite MPO terms being associated with the same gene with different mutations. Since this mutation information is not available for the classifier, these apparent contradictions (opposite MPO terms annotating the same instance) may reduce classification performance and interpretability. However, we consider this compromise acceptable since the lack of information about particular gene mutations makes the use of classification algorithms considering gene mutations as instances inviable.

Following this approach, we merged the annotations associated with the same gene, keeping all MPO terms that were associated with the different mutations of that gene. After this step, the 11,532 gene-mutations were reduced to 5,045 genes (without mutation information) keeping all annotations associated with different gene mutations.

The next step is to retrieve the Entrez Id (unique gene identifier) for each one of the 5,045 genes associated with the mortality/ageing phenotypes. Genes without an Entrez Id were discarded, further reducing the number of instances to 4,575. Finally, we retrieved the UniProt Id associated with each Entrez Id., using the UniProt ID Mapping Tool. This gives us information about the protein product associated with each gene. Genes having the same UniProt Id were discarded, leaving us with the final number of 3,886

proteins (instances), each instance linked with one protein and a list of mortality/ageing phenotypes (MPO terms used as class labels).

For the list of 3,886 proteins, we derive five datasets, each with a different feature type: numeric features, protein motifs features, Protein-Protein Interaction (PPI) features, and two types of KEGG pathway features, explained later. We briefly describe next these features.

Numeric dataset - We extracted the following numeric features from the amino acid sequence of the proteins, described by Salama and Freitas (2013); Silla Jr. and Freitas (2011b): "Amino Acid Composition" (21 features, 20 from standard amino acids and one for *Selenocysteine*), "Composition" (3 features), "Transition" (3 features), "Distribution" (15 features), and "Z-Values" (15 features), "Sequence Length", and "Molecular Weight", totalling 59 features.

Protein motif dataset - The binary motif features represent the presence or absence of a motif in the amino acid sequence of the protein. A motif is a template describing 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 motif features studied in Silla Jr. and Freitas (2011b): Interpro, Pfam, Prosite, and PRINTS. We have considered the motifs occurring in at least 1% of proteins (instances) in the dataset, to avoid classifier overfiting, resulting in a total of 95 motif features.

Protein-Protein Interaction (PPI) dataset - 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 proteins). Interacting partners of one protein often give away hints of its function (Sharan *et al.*, 2007). We have used the BioGrid[1] database to extract PPIs and have only considered features representing interacting partners occurring in 1% or more instances in the dataset, to avoid classifier over-fitting. This resulted in a total of 13 PPI features.

KEGG Pertinence (KEGGP) pathway dataset - KEGG pathways are directed-graph representations of interactions between several types of biological products (e.g., genes or proteins). To build our KEGG pathway features we have parsed the KGML representations of the mouse KEGG pathways under the condition that at least 1% of our instances must be present in the pathway in other for the pathway to be considered. This generated a total of 221 KEGGP features.

We have created two pathway feature datasets. The first, similarly to the PPI and motif datasets, contains binary features informing the pertinence of each instance (protein) into several KEGG pathways. We call this dataset KEGGP (KEGG Pertinence) from now on. Pertinence features based on KEGG pathways have already been explored in other works involving data-mining, e.g., (Jungjit *et al.*, 2014; Keerthikumar *et al.*, 2009).

KEGG Influence (KEGGI) pathway dataset - In this dataset, the KEGG pathway features represent the relative influence of an instance (the reference protein) on the other proteins that are downstream in relation to the instance in some KEGG pathway. This feature quantifies an influence that an instance (reference protein) has on the downstream proteins of a KEGG pathway, the idea being that proteins that have a common influence on a set of downstream protein share similar function. Consider that one ageing-related protein affects a set of downstream proteins in a given way. If another protein affects the downstream proteins in a similar way, then it is likely that that protein is also ageing-related.

The use of complex KEGG-based pathway features for data-mining has been proposed in other works: Zhang and Wiemann (2009) proposed a software tool to construct a graph-based model of KEGG pathways. Xia and Wishart (2010) used graph-based KEGG features for *metabolomics analysis*. Chen *et al.* (2010) used characteristics extracted from the KEGG pathway graph to classify the pathways into "biologically meaningful" or not. Breitkreutz *et al.* (2012) correlated the complexity of cancer-related KEGG pathways to patient survivability. Despite being previously used for different goals, as far as we know, this paper is the first work proposing complex KEGG-based features for the classification of protein functions.

The influence score for a given protein p has the minimum value of 0.0 when the reference protein (P_{ref}) does not influence p at all, because p is not "downstream" of (i.e., cannot be reached from) P_{ref}.

Figure 2 shows an example of the calculation of the proposed "influence" score for a hypothetical instance (reference protein) and a set of downstream proteins. Proteins P_1, P_2 and P_6 in have a score of 0.0, since they are not downstream of P_{ref}.

The score of a given protein p that is downstream of P_{ref} has the maximum value of 1.0 if, when P_{ref} is removed from the pathway, the downstream protein p becomes unreachable from the proteins that are not downstream proteins of P_{ref}. The biological meaning that we want to capture is that a knockout on P_{ref} would nullify the standard behaviour of the downstream protein p. Proteins P_3 and P_7, in Figure 2, have a score of 1.0 since if P_{ref} is removed from the pathway, proteins P_3 and P_7 will be disconnected from the KEGG pathway graph defined by the set of proteins that are not downstream proteins.

If the score of a given protein p that is downstream of P_{ref} has a value of 0.5, it means that P_{ref} accounts for half of the influence that the downstream protein p receives. Removal of P_{ref} would not nullify completely the standard behaviour of the downstream protein p, because there would be one more protein (which is not downstream in relation to P_{ref}) that also affects p, therefore the influence of P_{ref} on p is 50%. Protein P_5, in Figure 2, has a score of 0.5 because if one removes protein P_{ref} from the graph, protein P_5 would still be reachable from protein P_2, which is not a downstream protein.

In practice, to calculate the value of the features for each instance, we need to build two sets of proteins: the first, the *downstream proteins*, comprises the proteins that are downstream of the current instance, P_{ref}. The second set, the *non-downstream parent proteins*, contains the proteins that are not downstream of P_{ref} but are the parents of a protein that is downstream of P_{ref} - e.g., proteins P_2 and P_6 in Figure 2. Finally, for each downstream protein, the influence score is equal to $1/(1+p_{effect})$, where p_{effect} is the number of non-downstream parent proteins that have an effect (direct or indirect) on the downstream protein. We consider that a non-downstream protein has an effect on the downstream protein if the non-downstream proteins can reach the downstream protein.

To illustrate these concepts in detail lets us consider protein P_8 (see Figure 2), which is in the set of downstream proteins of P_{ref}. Because both non-downstream parent proteins affect P_8 (both P_2 and P_6 can reach P_8) the value of the influence score for P_8 is $1/(1+2) = 0.3$.

This gives us a set of downstream protein scores for the instance. We repeat this procedure for every available KEGG pathway. If the same downstream protein occurs more once in the same pathway, we keep the highest score. We discard the features (downstream proteins) with value > 0.0 in less than 1% of the instances, totalling 1331 features. We call this KEGG pathway dataset KEGGI (KEGG Influence) from now on.

Combined Datasets - We have created two datasets by combining some feature types. The first dataset was created by joining all 5 feature types into a single dataset. We call it the 'ALL' dataset. The goal of creating this dataset is to investigate if joining feature types from different domains increases the overall predictive performance of the classifiers.

We have also joined the KEGGI and KEGGP datasets to create a new dataset called "KEGGPI". The KEGGPI dataset combines two similar feature types with complementary characteristics: while the KEGGP feature type provides the coarse-grained information about the pertinence

[1] http://thebiogrid.org

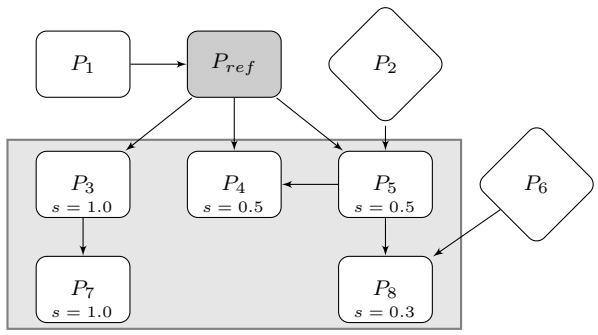

Fig. 2. Example of score values (s) for five downstream proteins (P_3, P_4, P_5, P_7, P_8) in relation to a reference protein P_{ref}. Diamond-shaped nodes represent proteins that are parent of some downstream protein but are not downstream protein themselves.

of a protein in a KEGG pathway, the KEGGPI feature type encodes the fine-grained information of the influence of a protein in a KEGG pathway. We expect that by combining these two feature types, with different strengths, will result in models with superior predictive performance.

2.2 Hierarchical classification algorithms

Interpretability is a desirable characteristic when designing features for classification (Freitas, 2013) as long as predictive power is not sacrificed. In order to check if our newly proposed KEGG features have at least comparable predictive power in relation to the other three types of features (numeric, PPI and motifs) used in (Fabris and Freitas, 2014), we use the following hierarchical classification algorithms.

2.2.1 The Predictive Clustering Tree (PCT) algorithm

The PCT algorithm (Struyf *et al.*, 2005) creates a decision tree finding binary splits that recursively divide the training data in two disjoint clusters until a given threshold that measures the quality of the split is not met. In the testing phase, the algorithm finds which cluster (leaf node of the tree) the instance belongs to using the tree induced in the training phase and then returns a class probability vector that represents the probability of the instance belonging to each one of the hierarchical classes. This class probability vector is calculated by first creating a binary vector for each instance in the cluster. The i-th position of this binary vector has the value '1' if the instance is annotated with the i-th class label and '0' otherwise. The PCT's final class probability vector is the average of the binary vectors of the instances in the cluster. Note that the class probability vectors are guaranteed to be consistent with the class hierarchy (the computed probability of each child class is always smaller than or equal to the probability of its parent classes).

2.2.2 The Hierarchical Dependence Network Using Non-Structural Relationships (HDN-NSR) algorithm

A dependence network is a Probabilistic Graphical Model (PGM) where nodes represent random variables (features or classes) and directed edges represent dependencies among variables (Heckerman *et al.*, 2001).

The Hierarchical Dependence Network (HDN) algorithm (Fabris and Freitas, 2014) is a type of PGM that uses the Gibbs Sampling algorithm to predict the probability of a protein belonging to each one of the hierarchical classes. The HDN algorithm uses the relationship given by the class hierarchy to create the edges of the Dependency Network. The main advantage of the HDN algorithm is that, contrary to most PGMs (e.g. Bayesian Networks), it allows for loops in the graph-representation of the dependencies among the random variables.

This work uses the HDN-NSR variation of the HDN algorithm (Fabris and Freitas, 2015), which uses a more sophisticated procedure to find the relationships between the classes of the hierarchy. It has been shown that HDN-NSR has overall better predictive performance than HDN and greater potential for finding relations that were not initially present in the class hierarchy, possibly useful for biologists studying the ageing process.

The HDN-NSR algorithm uses a standard classification algorithm with probabilistic outputs to estimate the probability of each hierarchical class. We have chosen to use the SVM (Support Vector Machine) classifier due to its high predictive power, and we applied the *F-test* feature selection method (Hall, 1999) to reduce the feature space. To train the classifier for each class c_i we 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.

After we run the HDN-NSR classifier we limit the value of the probability of each class to the minimum probability of its parents, to maintain the classification consistence across the class hierarchy.

2.2.3 The Local Hierarchical Classifier (LHC) algorithm

The LHC algorithm is a collection of flat binary classification algorithms trained to predict independently each one of the classes in the hierarchy. We have again used the SVM algorithm as a base classifier applying the *F-test* feature selection algorithm prior to training the algorithms. We have also used the same strategy to define the positive and negative examples that we used for the HDN-NSR 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 same procedure to guarantee prediction consistence used in the HDN-NSR algorithm, limiting the probability of each class to the probability of its parents.

2.3 Measures estimating the predictive performance of hierarchical classification algorithms

We have used three measures of predictive accuracy ($AU(\overline{PRC})$, $\overline{AUPRC}_w$, and $\overline{AUPRC}$ (Vens *et al.*, 2008)) based on the Area Under the Precision Recall Curve ($AUPRC$). In the flat classification context this measure works by constructing a PR curve (a plot of the classifier's precision as a function of its recall) thresholding the output (class probability) of the classifier. 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 (Boyd *et al.*, 2013). A perfect classifier would have an $AUPRC$ of 1.0. For more detail on how this measure is calculated, see (Vens *et al.*, 2008).

2.4 Interpreting the classification model induced using the KEGG pathway features

Initially we generated a classification model using the KEGG features and all the 81 hierarchical classes in the MPO dataset. However, this led to results that were difficult to interpret, because the ageing-related proteins are much less common (85 out of 3886) than the mortality-related proteins. For this reason, the classification models focused on discriminating the mortality-related classes much more than the ageing-related classes.

The high class imbalance of the original binary class dataset (only 2% of instances belong to the 'ageing' class) is detrimental for classifiers predicting the 'ageing' class, and consequently for interpreting the models. To tackle this problem, in another experiment, we have introduced two

simplifications for generating interpretable models: (1) we have joined all ageing-related classes into a single ageing-related class and all mortality-related classes into a single non-ageing-related class, transforming the hierarchical classification problem into a binary classification problem and (2) we have under-sampled the mortality-related proteins to a 1/1 ratio of instances between the two classes in the training set.

Also, instead of using the PCT algorithm for generating the classification models, we used conventional algorithms for binary classification that generate interpretable models. Namely, we have used the J48 algorithm, which generates a decision tree, the Decision Table (DT) algorithm, which generates a table with a set of conditions that must be satisfied for an instance to be classified as ageing-related, the PART algorithm, that builds several C4.5 decision trees, extracting rules from the "best" leafs, and the JRip algorithm, which builds a rule list by incrementally growing and pruning the model until a given stopping criterion is met. All four binary classification algorithms are available in the Weka data-mining tool (Hall *et al.*, 2009).

3 Results

3.1 Predictive accuracy results for hierarchical classification (with 81 classes)

Table 1. Predictive accuracies of the three hierarchical classification algorithms over the 7 used datasets. Numbers in boldface represent the top result in the row. Boldfaced ranks represent the best (smaller) ranks. Daggers (†) denote t-tests results rejecting the null hypothesis of equivalence between the best feature type (in boldface) and the current feature type, concluding that the models generated using best feature type are statistically superior ($\alpha = 0.025$). Due to lack of space, the HDN-NSR algorithm is referred to simply as 'HDN'.

		Dataset (Feature Type)						
Mea.	Alg.	KEGGP	KEGGI	KEGGPI	PPI	Motifs	Num.	ALL
$AU(\overline{PRC})$	PCT	**0.715**	0.706†	0.714	0.709†	0.710†	0.711	0.714
	LHC	0.716†	0.709†	0.714†	0.709†	0.710†	0.718†	**0.722**
	HDN	0.718†	0.710†	0.715†	0.710†	0.711†	0.718†	**0.721**
Avg. Rank		2.3	6.7	3.3	6.3	5.0	2.7	**1.7**
$\overline{AUPRC}_w$	PCT	**0.556**	0.547	**0.556**	0.544†	0.545†	0.537†	0.545†
	LHC	0.551†	0.536†	0.541†	0.541†	0.539†	0.551†	**0.566**
	HDN	0.546†	0.526†	0.544†	0.529†	0.536†	0.552†	**0.563**
Avg. Rank		2.7	5.7	3.0	5.7	5.3	3.7	**2.0**
$\overline{AUPRC}$	PCT	**0.146**	0.141†	0.144†	0.136†	0.136†	0.128†	0.135†
	LHC	0.141†	0.141†	0.143†	0.130†	0.134†	0.134†	**0.147**
	HDN	0.138†	0.129†	0.139†	0.124†	0.131†	0.134†	**0.147**
Avg. Rank		2.3	4.3	**2.0**	6.3	4.7	5.7	2.7

Table 1 shows the predictive accuracy results of the 3 algorithms we have tested in the 7 hierarchical datasets (5 different feature types and 2 combined feature types). Note that in this work we are interested mainly in comparing *datasets* (feature types), not *algorithms*. So, Table 1 shows, for each accuracy measure, the average rank of each dataset. The average rank is calculated by first assigning a rank varying from 1 (highest predictive accuracy) to 7 (lowest accuracy) to each dataset for each combination of classification algorithm and measure. Next, for each measure, the values displayed in the "Avg. Rank" rows of Table 1 are calculated by averaging the ranks of each dataset across algorithms. The best (smaller) average rank for each accuracy measure is highlighted in boldface.

For each combination of hierarchical algorithm and accuracy measure in Table 1, we have applied the *paired t-test* with the *Hochberg correction* (Demsar, 2006) for multiple comparisons (using the individual results on the 10 folds of the cross-validation process) to check if the predictive accuracy of the model induced using the best dataset in the row is statistically significantly different from the accuracy of the model induced by the same classification algorithm using the other dataset. Statistically significant results are marked with a dagger (†).

Note that the 'ALL' dataset is statistically significantly better than all other 6 datasets when using LHC and HDN-NSR. When using the PCT algorithm the best dataset is KEGGP. Overall, considering all 3 algorithms, the best (smallest) average rank was obtained by the 'ALL' dataset for the $AU(\overline{PRC})$ and $\overline{AUPRC}_w$ measures, while the combined KEGGPI dataset had the best rank for the $\overline{AUPRC}$ measure.

We can also observe that the rank of the KEGGPI feature type was better than the rank of the KEGGP feature type only when using the $\overline{AUPRC}$ measure. We can explain this behavior by analysing the bias of the $\overline{AUPRC}$ measure. This measure weighs all hierarchical classes equally, including those with relatively few proteins. So, classification models that use a wider range of features types (that can better predict more classes) are favored in relation to models which are better at predicting hierarchical classes with more instances using more general feature types.

To find out which *individual* feature representation is the best, we removed the combined datasets and performed a second statistical analysis. In this second study, the KEGGP feature type is always either statistically significantly better than all other feature types or is in the group of statistically equivalent feature types that include the best feature type.

It is also import to note that although the KEGGI feature type carries more complex information than the KEGGP feature type overall, the latter produces more accurate models. In fact, we have observed that PCT models generated using the KEGGP feature type have substantially more splits than the ones generated using KEGGI features. This is due to the smaller number of non-zero feature values present in the KEGGI dataset, which culminates in hierarchical classes with too few instances with non-zero feature values for a good classifier to be induced.

3.2 Results for binary classification

3.2.1 Predictive Accuracy Analysis

We tested 4 well-known algorithms that generate interpretable classification models from binary class datasets: J48, Decision Table (DT), JRip and PART; all available in the Weka data mining tool (Hall *et al.*, 2009). Table 2 shows the Area Under the ROC (Receiver Operating Characteristic) curve (AUROC) measure results obtained by the 4 classification algorithms for the 7 used datasets. The rankings of the feature types are calculated in the same way as described in Section 3.1.

The AUROC measure informs us the quality of the probabilities' ranking given by the classification model. That is, the AUROC measure has the maximum value of 1.0 if, for all ageing-related class instances, this class' probability estimated by the model is higher than the estimated probabilities assigned to the non-ageing-related instances. A random classifier is expected to have a AUROC measure of 0.5.

By analysing Table 2 we can conclude that few feature types have AUROC values statistically significantly worse ($\alpha = 0.05$) than the best performing feature type. This happened in three cases (shown by a dagger (†)), all when using the PART algorithm: when comparing the 'ALL' dataset against the Motif, PPI and KEGGI datasets.

The combined KEGGPI and KEGGP feature types had the best (smallest) joint predictive accuracy rank across the 4 classification algorithms in Table 2. If one is interested only in predictive accuracy, one could use just the KEGGP feature type instead of the combined KEGGPI feature type. However, when model interpretation is important (as it is the case here) using the KEGGPI feature type has the advantage of providing additional, more precise information, while maintaining

Table 2. AUROC measure results for the classification algorithms on the binary class dataset described in Section 2.4. Boldface numbers highlight the best result. Daggers (†) next to a result indicate statistically worse results than the best result for the algorithm in the row, according to a paired t-test using the Hochberg step-up correction (Demsar, 2006) ($\alpha = 0.05$). The rank in boldface indicates the best (smallest) rank.

Alg.	Dataset (Feature Type)						
	KEGGP	KEGGI	KEGGPI	PPI	Motifs	Numeric	ALL
J48	**0.584**	0.505	**0.584**	0.508	0.507	0.495	0.566
DT	**0.605**	0.533	0.593	0.518	0.518	0.484	0.541
JRip	0.556	0.523	**0.563**	0.520	0.505	0.514	0.461
PART	0.581	0.499 †	0.583	0.504 †	0.485 †	0.543	**0.585**
Avg. Rank	**1.8**	4.8	**1.8**	4.5	6.0	5.8	3.5

predictive accuracy. Therefore, the KEGGPI feature type is useful for the binary classification problem we are studying.

3.2.2 Interpreting results for binary classification

We have interpreted the model generated for the KEGGPI feature type induced by the DT (Decision Table) algorithm from the binary class dataset described in Section 2.4. This choice of classification algorithm/dataset was made for 3 reasons: the KEGGPI and KEGGP datasets were tied as the best dataset in Table 2; the KEGGPI dataset comprises the KEGGP and KEGGI feature types, so we can interpret both at the same time, and; the DT algorithm had the best predictive performance across the 7 feature types (winning in 4 out of 7 feature types).

In Table 3 we show the classification *rules* created by the DT algorithm for predicting ageing-related protein functions. A rule is a set of feature values that a protein must have to be classified as either ageing related or non-ageing related. The first rule of Table 3 means: if a protein is not present in the KEGG pathways in columns 2-7; *and* the protein influences protein P11440 (Cyclin-dependent kinase 1), present in pathway mmu04110 (Cell cycle); *then* the protein is likely to be ageing related. The last two columns of this table show the coverage (number of instances classified by the rule) and the accuracy (percentage of correctly classified instances) of each row (rule). Note that the rule containing the 'yes' condition for the KEGGI feature type (first row) had the best accuracy (28%) with good coverage (21 instances). At first glance, an accuracy of 28% seems small, but recall that only 2% of instances are ageing-related, so an accuracy of 28% is actually a 14-fold increase in relation to the prior probability of the 'ageing-related' class.

In Figure 3 we show how we can use the information given in column 1 (from the KEGGI feature type) to interpret the rules created by the DT. This figure shows part of the KEGG pathway mmu04110 and highlights the influence of several proteins on the protein P11440 (CDK1). Our results suggest that if a reference protein P_{ref} has any influence in CDK1 (feature value > 0), then P_{ref} is more likely to be ageing-related.

Note that the highlighted proteins are, at the same time, instances and features for other proteins. For example, the instance representing protein "Chk1,2" (reference protein) influences protein CDK1 (feature) according to our score. Therefore, this instance ("Chk1,2") has a non-zero value in the feature associated with influence on protein CDK1. At the same time, CDK1 is also an instance, having features with a non-zero value in the set of proteins it influences. Note that not all proteins in the KEGG pathway are instances in our ageing-dataset, i.e., not every protein associated with a feature is an instance.

In contrast, the KEGGP feature type provides a different type of information; the conditions involving this feature merely inform the user if a protein (an instance) is present in a KEGG pathway or not. E.g., the seventh row of Table 3 (second most accurate rule) informs us that if a protein *is* in the KEGG pathway mmu04917 (Prolactin signaling pathway), *is not* present in the other selected pathways and has no influence on protein P11440 in pathway mmu04110, it is likely ageing-related. This suggests that proteins in the 'Prolactin signaling pathway' may have some influence on the ageing process, so some other proteins present in the same pathway could be candidates for further investigation.

Table 3. Classification rules that predict the 'ageing' class, generated by the DT algorithm using the KEGGPI dataset. The first column presents a KEGGI feature; its name shows the Uniprot Id of the protein that is being potentially influenced by an instance and (after the '_') the Id of the KEGG pathway where the influence can occur. The next 7 columns show binary KEGGP features, indicating whether or not an instance belongs to the corresponding KEGG pathway. The last two columns show the coverage and the % accuracy of each rule. Due to lack of space, we have suppressed the "mmu0" prefix in the KEGG pathways ids. Each row shows the conditions that must be satisfied for a protein to be predicted as 'ageing-related'. The selected KEGG pathways are: mmu04110 (Cell cycle), mmu04151 (PI3K-Akt signaling pathway), mmu05168 (Herpes simplex infection), mmu04660 (T cell receptor signaling pathway), mmu04380 (Osteoclast differentiation), mmu04350 (TGF-beta signaling pathway), mmu04917 (Prolactin signaling pathway) and mmu03420 (Nucleotide excision repair).

P11440_04110	4151	5168	4660	4380	4350	4917	3420	Cov.	%Ac.
> 0	No	No	No	No	No	No	No	21	28
= 0	Yes	No	No	No	No	No	No	139	6
= 0	No	Yes	No	No	No	No	No	59	12
= 0	No	No	Yes	No	No	No	No	27	15
= 0	No	No	No	Yes	No	No	No	21	10
= 0	No	No	No	No	Yes	No	No	49	10
= 0	No	No	No	No	No	Yes	No	13	23
= 0	No	No	No	No	No	No	Yes	20	20

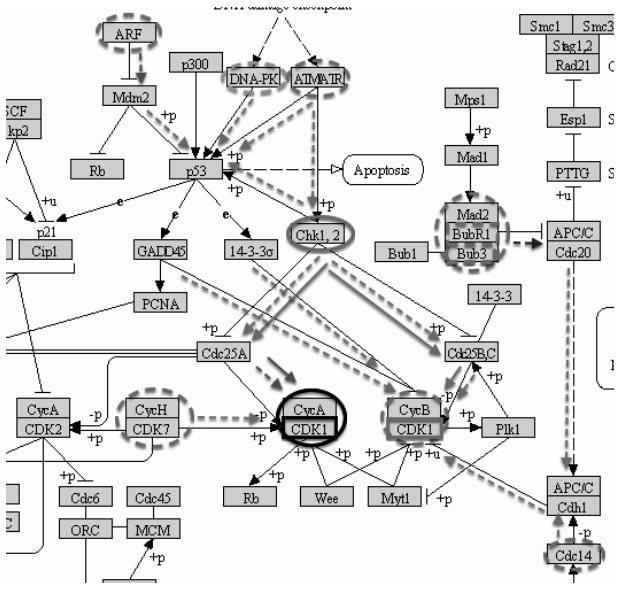

Fig. 3. Graphical representation of the KEGG pathway mmu04110 (Cell cycle) with highlighted interesting proteins and interactions. The protein complex highlighted with a solid black line contains the protein "Cyclin-dependent kinase 1" (CDK1), which occurs in the feature selected by the DT algorithm (see Table 3) The highlighted dashed grey proteins represent ageing-related proteins that influence CDK1. The highlighted solid grey proteins represent non-ageing-related proteins that influence CDK1. Solid grey edges represent all possible influence paths from a solid grey protein to CDK1. Dashed grey edges represent all possible influence paths from the dashed grey protein to CDK1.

4 Discussion

We have presented the construction of two KEGG feature types for the classification of ageing-related protein functions engineered specifically with the goal of predicting protein function using interpretable features.

The advantages of using the KEGG-derived features types are two-fold: (1) the good predictive performance of the models induced using these two features together, and (2) the improved interpretation potential of using richer features to represent the instances. In fact, the KEGG pathway seems to be a very appropriate database to use when interpretation is required, since it is focused on integrating not only biological data from several sources, but also concepts *about* the data (Kanehisa *et al.*, 2011).

The down-side of relying on such rich source of data is that, in order to compute values of the KEGGP and KEGGI features for an instance, the corresponding gene or protein must first be characterised into some KEGG pathway, which involves laborious wet-lab experimentation. So, an uncharacterised protein represented only by its amino acid sequence cannot be classified using KEGGP and KEGGI features (nor using PPI and Motif features). In fact, in this scenario, out of the 5 feature types used in this work, only the 'Numeric' feature type could be used, which is arguably the most difficult to interpret due to its low level of abstraction.

The KEGG Pertinence (KEGGP) feature type, used for the first time for hierarchical classification in this work, had the best performance according to our statistical analysis compared to three other feature types and the KEEGI feature type, a new KEGG feature type proposed here.

The combined KEGGPI dataset (using both KEGGP and KEGGI features) had the best mean rank on the binary class dataset, tied with the KEGGP feature type. Although the KEGGP feature type has a simpler interpretation, if a richer, more precise model interpretation is desired (as it is the case here), the combined KEGGPI feature type is more suitable, as it contains both the easier to interpret KEGGP feature type (at a higher level of abstraction) and the KEGGI feature type (with finer-grain information). To illustrate this point, we have shown how the KEGGPI feature type can be used for generating biological knowledge using the Decision Table algorithm, which generates interpretable classification models.

We have also contrasted the interpretation of the KEGGI feature type with the interpretation of the simpler KEGGP feature type and concluded that the complementary nature of these two feature types provides a good range of biological information: the KEGGI feature type presents more precise information to the user, enabling a richer interpretation of the classification model: it quantifies the influence of a current (reference) protein on another specific protein in a given KEGG pathway. On the other hand, the KEGGP feature type tells the user the higher-level information of which KEGG pathways are important for discriminating between ageing and non-ageing related proteins.

Funding: This work was supported by Capes, a Brazilian research-support agency [grant number 0653/13-6 to F.F.].

Acknowledgements: We thank Dr. J. P. de Magalhães for valuable discussions about the early phase of this work.

References

Boyd, K., Eng, K.H. and Page, C.D. (2013). Area Under the Precision-Recall Curve: Point Estimates and Confidence Intervals. *Mach. Learning and Know. Discovery in Databases*, **8190**, 451–466.

Breitkreutz, D. et al (2012). Molecular signaling network complexity is correlated with cancer patient survivability. *Proc. National Academy of Sciences*, **109**(23), 9209–9212.

Chen, L. et al (2010). Analysis of protein pathway networks using hybrid properties. *Molecules*, **15**(11), 8177–8192.

Demsar, J. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Mach. Learning Res.*, **7**, 1–30.

Eppig, J.T. et al (2015). The mouse genome database (mgd): facilitating mouse as a model for human biology and disease. *Nuc. Acids Res.*, **43**(D1), D726–D736.

Fabris, F. and Freitas, A.A. (2014). Dependency Network Methods for Hierarchical Multi-label Classification of Gene Functions. *Proc. 2014 IEEE Comp. Intel. and Data Mining*, 241–248.

Fabris, F. and Freitas, A.A. (2015). A Novel Extended Hierarchical Dependence Network Method Based on non-Hierarchical Predictive Classes and Applications to Ageing-Related Data. In *Proc.* 27^{th} *IEEE Intl. Conf. on Tools with Artificial Intel.*, 294–301.

Forman, G. (2002). Feature engineering for a gene regulation prediction task. *ACM SIGKDD Explorations Newsletter*, **4**(2), 106–107.

Freitas, A.A. (2013). Comprehensible Classification Models - a position paper. *ACM SIGKDD Explorations Newsletter*, **15**(1), 1–10.

Goldman, D.P. et al (2013). Substantial health and economic returns from delayed aging may warrant a new focus for medical research. *Health Affairs*, **32**(10), 1698–1705.

Hall, M. et al (2009). The weka data mining software: An update. *SIGKDD Explor. Newsl.*, **11**(1), 10–18.

Hall, M.A. (1999). *Correlation-based feature selection for machine learning*. Ph.D. thesis, The University of Waikato, New Zealand.

Heckerman, D. et al (2001). Dependency Networks for Inference, Collaborative Filtering, and Data Visualization. *Journal of Mach. Learning Res.*, **1**, 49–75.

Jungjit, S. et al (2014). Extending multi-label feature selection with KEGG pathway information for microarray data analysis. In *2014 IEEE Conf. on Comp. Intel. in Bioinfo. and Comp. Biology*, 1–8, Hawaii. IEEE.

Kanehisa, M. et al (2011). KEGG for integration and interpretation of large-scale molecular data sets. *Nuc. Acids Res.*, **40**(D1), D109–D114.

Kanehisa, M. et al (2016). KEGG as a reference resource for gene and protein annotation. *Nuc. Acids Res.*, **44**(D1), D457–D462.

Keerthikumar, S. et al (2009). Prediction of candidate primary immunodeficiency disease genes using a support vector machine learning approach. *DNA Res.*, **16**(6), 345–51.

Salama, K.M. and Freitas, A.A. (2013). ACO-Based Bayesian Network Ensembles for the Hierarchical Classification of Ageing-Related Proteins. In *Evoltionary Computation, Mach. Learning and Data Mining in Bioinfo.*, volume 7833 of *LNCS*, 80–91.

Sharan, R., Ulitsky, I. and Shamir, R. (2007). Network-based prediction of protein function. *Molecular Systems Biology*, **3**(1).

Silla Jr., C.N. and Freitas, A.A. (2011a). A Survey of Hierarchical Classification Across Different Application Domains. *Data Mining and Know. Discovery*, **44**(1-2), 31–72.

Silla Jr., C.N. and Freitas, A.A. (2011b). Selecting different protein representations and classification algorithms in hierarchical protein function prediction. *Intelligent Data Analysis*, **15**(6), 979–999.

Struyf, J. et al (2005). Hierarchical Multi-classification with Predictive Clustering Trees in Functional Genomics. In *Progress in Artificial Intel.*, volume 3808 of *LNCS*, 272–283.

Vens, C. et al (2008). Decision Trees for Hierarchical Multi-label Classification. *Mach. Learning*, **73**(2), 185–214.

Xia, J. and Wishart, D.S. (2010). Metpa: a web-based metabolomics tool for pathway analysis and visualization. *Bioinfo.*, **26**(18), 2342–2344.

Yepes, A.J.J. et al (2015). Feature engineering for medline citation categorization with mesh. *BMC Bioinfo.*, **16**(1), 113.

Zhang, J.D. and Wiemann, S. (2009). KEGGraph: a graph approach to KEGG pathway in r and bioconductor. *Bioinfo.*, **25**(11), 1470–1471.