PLOS ONE

RESEARCH ARTICLE

Gene Function Prediction from Functional Association Networks Using Kernel Partial Least Squares Regression

Sonja Lehtinen[1,2¤], Jon Lees[2], Jürg Bähler[3], John Shawe-Taylor[4], Christine Orengo[2]*

1 CoMPLEX, University College London, London, United Kingdom, **2** Institute of Structural and Molecular Biology, University College London, London, United Kingdom, **3** Department of Genetics, Evolution and Environment, University College London, London, United Kingdom, **4** Department of Computer Science, University College London, London, United Kingdom

¤ Current address: Department of Infectious Disease Epidemiology, Imperial College London, London, United Kingdom
* c.orengo@ucl.ac.uk

Citation: Lehtinen S, Lees J, Bähler J, Shawe-Taylor J, Orengo C (2015) Gene Function Prediction from Functional Association Networks Using Kernel Partial Least Squares Regression. PLoS ONE 10(8): e0134668. doi:10.1371/journal.pone.0134668

Editor: Peter Csermely, Semmelweis University, HUNGARY

Received: April 27, 2015

Accepted: July 13, 2015

Published: August 19, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors have no support or funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

With the growing availability of large-scale biological datasets, automated methods of extracting functionally meaningful information from this data are becoming increasingly important. Data relating to functional association between genes or proteins, such as co-expression or functional association, is often represented in terms of gene or protein networks. Several methods of predicting gene function from these networks have been proposed. However, evaluating the relative performance of these algorithms may not be trivial: concerns have been raised over biases in different benchmarking methods and datasets, particularly relating to non-independence of functional association data and test data. In this paper we propose a new network-based gene function prediction algorithm using a **com**mute-time kernel and **pa**rtial least **s**quare**s** regression (Compass). We compare Compass to GeneMANIA, a leading network-based prediction algorithm, using a number of different benchmarks, and find that Compass outperforms GeneMANIA on these benchmarks. We also explicitly explore problems associated with the non-independence of functional association data and test data. We find that a benchmark based on the Gene Ontology database, which, directly or indirectly, incorporates information from other databases, may considerably overestimate the performance of algorithms exploiting functional association data for prediction.

Introduction

Network Approaches for Protein Function Prediction

The rapidly increasing volume of genomic and proteomic data has led to a surge of interest in the automatic extraction of functionally meaningful information from these datasets. One key approach is the *in silico* prediction of *gene and protein function*, a broad concept with meanings ranging from a protein's biochemical role to its impact on phenotype. Owing to the scope of

the problem, a variety of data sources and computational approaches have been exploited in gene function prediction. In general terms, prediction methods fall into two broad categories: *de novo* methods seeking to predict function based on intrinsic properties of a gene and *guilt-by-association* (GBA) approaches, which predict new functional labels based on a gene's similarity to already functionally characterised genes.

A number of established GBA-type prediction methods base their predictions on sequence or structural similarity. Recently however, in response to the increasing prevalence of functional association data, there has been considerable interest in developing GBA methods exploiting functional association networks. The premise of these methods is that the functional similarity of two genes depends on 1) how close the genes are in the functional association network (local proximity) and 2) how many paths connect the two (global topology) [1]. There are two main classes of methods exploiting both global topology and local proximity: probabilistic network models and kernel methods.

Probabilistic network models are formalisms for representing dependencies between random variables. In the context of gene networks, these models capture how a gene's function depends on that of its network neighbours. A number of approaches have modelled the problem in terms of belief propagation in these networks [2–5]. GeneMANIA [6], one of the most successful prediction algorithms to date [7, 8], makes use of this approach, implementing Gaussian label propagation. To our knowledge, no prediction algorithm has consistently outperformed GeneMANIA. We therefore benchmark our methods against this algorithm.

The other major class of methods makes use of kernels. Kernel approaches transform functional association networks into functional similarity scores between genes, based on the topology of the network. More specifically, these similarity scores represent inner products between gene vectors in some feature space, where distances between genes reflect their proximity in the network. A number of statistical learning approaches (such as various forms of regression for example) can be expressed in a form which operates on the kernel instead of the original feature space. Thus, kernel representations allow the use of statistical learning approaches on network data. Existing methods have most commonly used diffusion kernels, paired with support vector machines [9] or logistic regression [10]. A related method, FunctionalFlow [11], makes use of a diffusion kernel-like process.

While most existing methods have focused on diffusion kernels, recently, work by Heriche et al compared different kernel functions (i.e. different ways of generating similarity scores between genes from the network) [12]. In this work, the *commute time kernel* [13] was found to perform most robustly: when tested on a number of different benchmarks, this kernel was consistently among the top performers, while other kernels' performance fluctuated significantly.

Heriche et al's work made predictions by treating the kernel as a matrix of functional similarity scores between genes, but did not explore more complex prediction algorithms. Combining a commute time kernel with statistical learning methods therefore seems like a promising approach.

Benchmarking

Accurate evaluation of the performance of prediction methods is essential for meaningful comparison of different algorithms. At a minimum, evaluation requires sets of true positives: genes which are known to be involved in the same function. These true positives are commonly derived from the Gene Ontology (GO) [14], with genes labelled with the same term considered a 'set' sharing the same function. A typical approach to benchmarking is to then use cross-validation: a subset of known labels are hidden, and the performance of the method is assessed by how well the hidden labels are recovered.

However, in the context of protein function prediction, cross-validation can be problematic. There is evidence to suggest that information is transferred between functional association databases (such as BioGrid [15], STRING [16] or KEGG [17] for example) and the GO: Gillis and Pavlidis looked at the source of GO annotations shared by proteins involved in a protein-protein interaction. 13% of these annotations were found to be derived from the publication which reported the interaction between the proteins [18]. Furthermore, the authors found a low ($r = 0.2$) but significant correlation between how well guilt-by-association methods perform for a particular term (as assessed by cross-validation) and the extent of this overlap between network and gene annotation data for this term. Thus, cross-validating a functional association network-based method may not actually reflect the algorithm's ability to predict function for new genes, but rather the extent to which information has been dissipated across databases.

Interestingly, similar problems have also been reported for sequence similarity based prediction algorithms. The GO derives some of its annotations from sequence similarity (for example 'IEA' (inferred from electronic annotation), 'ISS' (inferred from sequence similarity)). Again, this raises the concern that the dataset used for evaluation is not independent from the dataset used for prediction, potentially leading to a biased estimation of predictive performance. Indeed, Rogers and Ben-Hur [19] showed that including these evidence codes when benchmarking a prediction algorithm tends to over-estimate how well sequence similarity based methods perform.

There have been significant efforts to compare prediction algorithms using a more realistic benchmark. Competitions such as CAFA (Critical Assessment of Function Annotation) [20] and MouseFunc [8] evaluate prediction methods based on novel true positives uncovered after the predictions have been made. Thus, unlike cross-validation, this benchmark directly assesses an algorithm's ability to predict novel annotations.

These frameworks are essential for providing fair comparative assessment of prediction methods. However, CAFA-style competitions have also attracted criticism, particularly because of their reliance on GO annotations. There is evidence to suggest the process of label acquisition may be affected by existing annotations, which would extend the problems with cross-validation to benchmarks based on new labels as well. For example, existing 'IEA' annotations for a particular term are highly predictive of which genes will acquire an annotation with an experimental evidence code for the term [21], suggesting 'IEA' annotations may be guiding GO curation and/or target selection for experiments. This effect is strong: Gillis and Pavlidis predicted new labels based on existing 'IEA' annotations and reported performance comparable to the best CAFA entries in the 2011 competition [21]. This suggests that (sequence-based) computational methods may simply be re-creating the 'IEA' annotation and therefore seem to perform well, not because of actual predictive power, but because they mimic the process of annotations becoming incorporated in the GO.

We hypothesise that similar concerns may also be relevant for network-based prediction: the addition of new annotations into the GO may be affected by current functional association network data, either through temporal delays in information transfer into the GO or because choices of which putative gene/function pair to investigate may be partially driven by knowledge of functional associations.

Our contribution

In this work, we develop a prediction algorithm based on a **com**mute-time kernel combined with a **pa**rtial least **s**quare**s** regression (Compass). PLS regression is a dimensionality reduction approach similar to principle component (PC) regression in that it projects the data into a

subspace of the feature space. Instead of a space maximizing the variance of the inputs, however, PLS selects directions that maximize the covariance between inputs and the target. Originally developed for regression problems where features outnumber observations and exhibit multicollinearity PLS has been successfully applied to categorization problems [22], including ones involving genomic data [23, 24]. Indeed, given that high dimensionality and low sample size are common problems in the study of genomic and proteomic data [25], PLS is a promising approach for gene function prediction.

In addition to applying a novel approach to GBA prediction, we construct a simulation of a CAFA-style competition through a *rollback benchmark* [26]. We use functional association networks and GO assignments dating prior to a specific cut-off time to make predictions and evaluate these predictions on annotations acquired after the cut-off date. We use this benchmark to compare the Compass and GeneMANIA algorithms.

We also use the rollback benchmark to explicitly explore potential biases relating to transfer of information between databases. Furthermore, in light of the problems we identify, we develop two additional benchmarks ('RNAi' and 'ageing'), which are not affected by information transfer. In these benchmarks, functionally related gene sets are derived from genes giving rise to a particular phenotype in a genome-wide knock-out experiment. The networks used in prediction pre-date the screens, ensuring information transfer between the test data and network is not possible. We use these benchmarks to further compare the performance of Compass and GeneMANIA.

Methods

Prediction

We implement a prediction algorithm (Compass) which first computes the **com**mute-time kernel of a functional association network and then performs a kernelized form of **pa**rtial least **s**quare**s** regression in the feature space represented by the kernel.

1. Networks are combined by summing the individual adjacency matrices: $A(i,j) = \sum_k A_k(i,j)$, where $A_k(i,j)$ is equal to the weight of the edge between nodes i and j in network k.

2. The commute-time kernel K_{CT} [13] of network with n nodes and adjacency matrix A is computed by: $K_{CT} = L^+$, where L^+ is the Moore-Penrose pseudoinverse of the graph laplacian L, defined by $L = D - A$, where D is the diagonal degree matrix, with entries $D(i,i) = \sum_{j=1}^{n} A(i,j)$. The commute-time kernel assumes the network has one connected component. In this work, if functional networks had more than one connected component, only the largest component was considered, as this resulted in the elimination of a small minority of the nodes. For networks with larger or more numerous smaller components, each component can be treated separately.

3. The kernel matrix is normalized:

$$K_{CT}^{norm}(i,j) = K_{CT}(i,j) / \sqrt{K_{CT}(i,i) * K_{CT}(j,j)}$$

   and centred $K^{centred} = K - 1_N K - K1_N + 1_N K1_N$, where K is the normalized kernel and 1_N is a n-by-n matrix, where all elements are equal to 1/n.

4. For a gene set of interest ('seed set') for which we seek to predict new members, we generate a label vector y. If n is the set of all genes and n_+ is the gene set of interest, $\hat{y}$ is constructed by assigning $y(i) = 1$, if $i \in n_+$, else $y(i) = 0$ and then subtracting the mean, giving $y(i) = 1 - |n_+|/|n|$, if $i \in n_+$, else $y(i) = -|n_+|/|n|$, where $|n|$ is the total number of genes and $|n_+|$ the number of genes in the seed set. This approach treats all genes which are not part of the seed

set as negative examples. Although more sophisticated methods of selecting negative examples exist, these tend to be GO specific. This did not suit our purpose of developing a method not restricted to GO label prediction. Note also that as PLS is multivariate, the approach could be extended to simultaneous prediction for multiple seed sets. This approach is not explored in this paper.

5. We perform a PLS regression in the feature space represented by the commute-time kernel using the kernelized implementation by Shawe-Taylor and Cristianini [27] The dependent variable predicted by the regression model, $\hat{y}$, gives the scores used to rank the genes for membership in the seed set.

The number of components to use in the PLS regression was determined by two-fold cross-validation on the seed set in the GO benchmark (i.e. based on labels discovered prior to the cut-off date). The optimal number of directions was 1 (see S1 Fig). This parametrization was used for all benchmarks.

Network Construction

Functional association networks were downloaded from STRING database (version 8.1, released in June 2009) [16]. For each organism, this gave 7 individual networks, each corresponding to a different indicator of functional association (conserved genome neighborhood, gene fusion, phylogenic co-occurrence, co-expression, database imports, large-scale experiments and literature co-occurrence). STRING weights edges in the networks based on how well these interactions correspond to shared membership in KEGG pathways [17].

Benchmarking

GO Rollback Benchmark. A GO rollback benchmark was constructed using data from yeast (*Saccharomyces cerevisiae*), using a 2009 cut-off date. Evaluation sets were created from the Biological Process (BP) branch of the GO tree. For each GO term, proteins for which the annotation was associated with a date prior to 2010 were taken as the seed set and those associated with a date from 2010 onwards as the test set. GO annotations were filtered by evidence code in order to 1) ensure high quality seed and test sets and 2) avoid predicted annotations, thus minimizing overlap between network data and test set. Specifically, only annotations derived from the evidence codes IC, IMP, TAS, IDA and NAS were used. Proteins not present in any of the functional association networks were ignored and categories with no proteins in the seed or novel set were excluded. This resulted 760 evaluation sets (i.e. GO terms).

RNAi Phenotypic Benchmark. For a complementary interpretation of function, we constructed a rollback phenotypic benchmark from genome-wide knock-out data by considering genes which, when knocked out, give rise to the same phenotype as a set of functionally related genes. Phenotypic data was downloaded from the GenomeRNAi database [28], a repository for RNAi screens. To ensure independence from the network data, only screens performed from 2010 onwards were considered. This benchmark was implemented in human and fly. Five fold cross validation was used to estimate predictive performance on this benchmark. Because of the independence of the network and test data, cross-validation on this benchmark is not subject to the concerns associated with cross-validation based on GO benchmarks.

Ageing Benchmark. Phenotypic benchmarking was also performed on an experimentally derived set of fission yeast (*Schizosaccharomyces pombe*) long-lived mutants from a longevity screen by Sideri et al [29] (see S1 Text for gene list). Predictions were seeded using long-lived mutant *clg1*, *pef1* [30], *pma1* [31], *sck2* and *pka1* [32], which were known prior to Sideri et al's screen.

Comparison to Genemania

GeneMANIA's predictions were generated using the command line tool for the GeneMANIA cytoscape plug-in [33]. The plugin was given the same functional association networks and seed sets as used by our algorithm.

Comparison to Multifunctionality and Degree-Based Prediction

Some authors have expressed concern that to some (potentially considerable) extent, the performance of guilt-by-association methods does not capture genuine *function-specific* insight, but instead reflects a general ranking of gene multifunctionality and/or degree [34]. Indeed, simply ranking genes based on their multifunctionality outperforms GeneMANIA on a disease gene prioritization task [34]. To investigate whether Compass outperforms this type of generalized ranking, we included a comparison against a degree-based and a multifunctionality-based ranking.

For the degree-based prediction, genes are ranked in order of their weighted node-degree in the combined String network. For the multifunctionality-based prediction, genes are ranked according to a multifunctionality score, defined, for gene a, as in the original publication [34] as:

$$s_a = \sum_{i|a \in T_i} \frac{1}{|T_i|(n - |T_i|)}$$

where T_i is GO term i, $|T_i|$ is the number of genes in GO term i and n is the total number of genes. This score is the number of GO terms a gene is labelled with, weighted by the contribution the gene makes to the group.

Results and Discussion

GO Rollback Benchmark

Relative performance of Compass and GeneMANIA. The relative performance of Compass and GeneMANIA at predicting novel GO annotations was assessed using a rollback benchmark (Table 1 and Fig 1). Compass outperforms GeneMANIA when performance is measured in terms of AUC ($p = 2.5 \times 10^{-4}$, two-tailed Wilcoxon signed-rank test), while precision-based measures (mean average precision, P_{mean}, and precision at recall 0.1, $P_{r=0.1}$) show no significant difference between the two algorithms (although both measures are higher for GeneMANIA). As shown in Fig 1, the average precision recall curves cross: on average, GeneMANIA performs better at low recall values while Compass performs better at high recall. This suggests Compass outperforming GeneMANIA on the AUC measure is associated with improved performance for gene-annotation pairings which are more difficult to predict.

To further understand the relative performance of Compass and GeneMANIA, we sought to identify factors affecting the performance of the algorithms. First, we looked at the size of the seed set (i.e. number of training examples) and the specificity of the GO term (in terms of GO level). There was no significant correlation between the number of seed genes and performance for either method. For both methods, performance correlated with GO specificity, with higher performance at greater specificity (Spearman correlation coefficient 0.1740 and 0.1459 for Compass and GeneMANIA respectively $p < 10^{-4}$). However, as this effect is similar for both methods, it is unlikely to explain the difference in performance.

Second, we looked at how a gene's degree affects how successful the algorithms are at predicting annotations for it. As shown in Fig 2, both methods are, as expected, more successful at making predictions for high degree genes. However, Compass outperforms GeneMANIA for

Table 1. Predictive performance on the GO rollback benchmark.

Measure	Compass	GeneMANIA	p-value
AUC	0.8286 (sd 0.1861)	0.8026 (sd 0.2301)	2.5×10^{-4}
P_{mean}	0.0717 (sd 0.1838)	0.0718 (sd 0.1834)	0.0606
$P_{r=0.1}$	0.1000 (sd 0.2396)	0.1020 (sd 0.2462)	0.4725

Predictive performance of Compass and GeneMANIA, as measured by the area under the receiver operating characteristic (ROC) curve (AUC), mean average precision (P_{mean}) and precision at recall 0.1 ($P_{r=0.1}$). The benchmark consists of 760 GO terms. Standard deviations (sd) are reported in parentheses.

doi:10.1371/journal.pone.0134668.t001

genes with very low degree. This suggests that Compass' improved performance on difficult to predict gene-annotation pairings arises from its improved performance on low degree genes. Compass outperforms GeneMANIA for low-degree genes in three out of four of our benchmark sets (Fig 2). It is thus unclear whether this is a general property of the method or a particularity of the benchmark sets used.

Comparison to Degree and Multifunctionality-Based Rankings. Network-based predictors are known to be biased toward high degree genes (Fig 2). This has led to a concern that instead of capturing genuine functional insight, network-based methods may simply be producing a *generic* ranking based on node-degree and/or gene multifunctionality [34]. Indeed, simply ranking genes by multifunctionality (based on the number of GO terms each gene is labelled with) has been reported to outperform GeneMANIA on a disease gene prioritization task [34]. We therefore tested how degree and multifunctionality-based rankings perform on this benchmark. Both generic rankings are clearly outperformed by Compass and GeneMANIA (AUC 0.5940 and 0.5719 for degree and multifunctionality-based rankings respectively). Thus, while a generic ranking gives above random performance (reflecting the tendency of

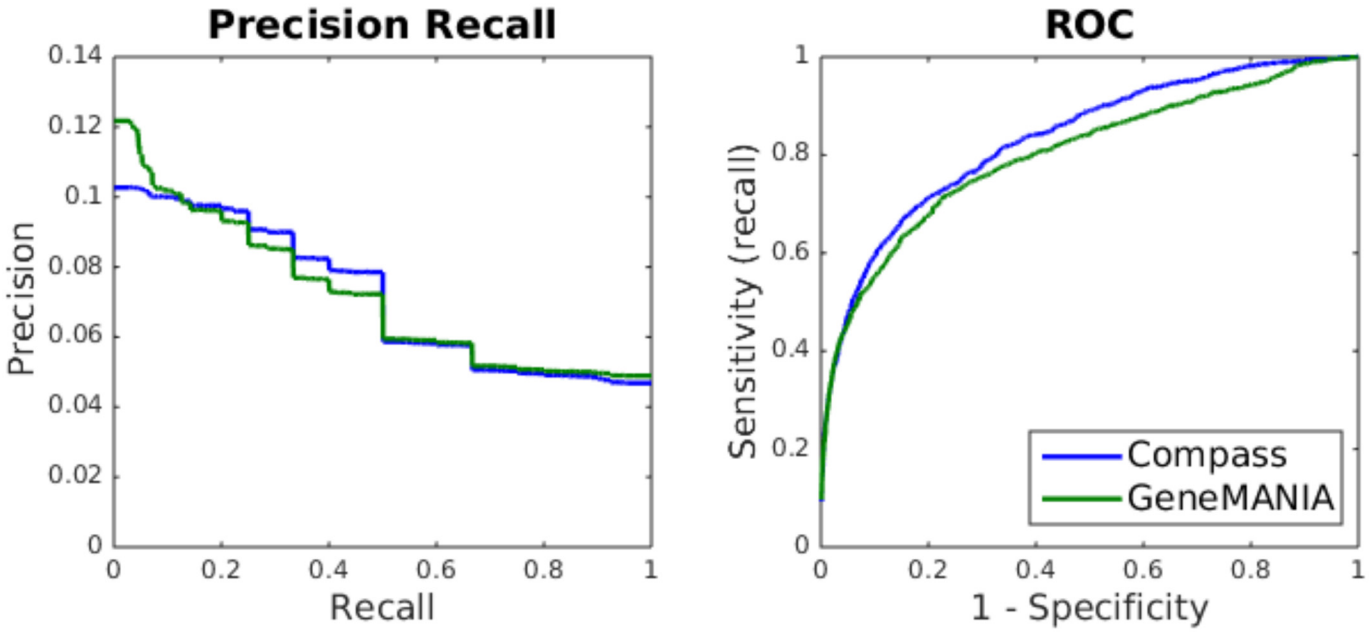

Fig 1. Average precision-recall and ROC curves for the GO rollback benchmark. The curves represent the average curve for the 760 GO terms.

doi:10.1371/journal.pone.0134668.g001

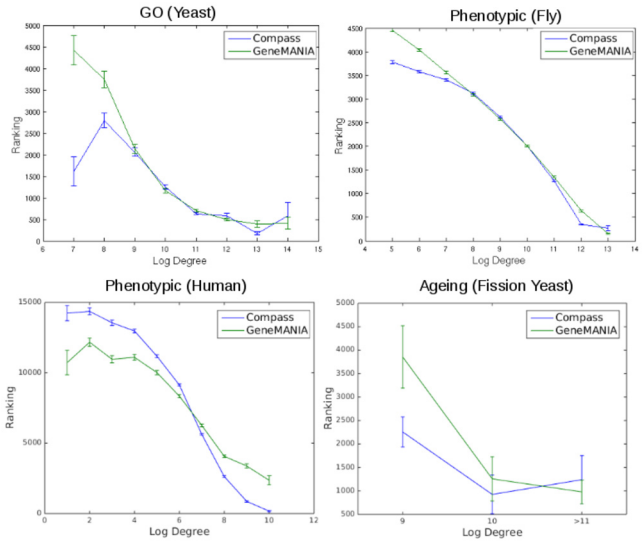

Fig 2. Association between a gene's degree and how well the algorithms predict new annotations for the gene. The figure shows average ranking (top ranking = rank of 1.) for genes grouped according to weighted degree in the String network for all four benchmarks (GO Benchmark, RNAi benchmark in fly and human and fission yeast ageing benchmark). Error bars represent standard error of the mean.

doi:10.1371/journal.pone.0134668.g002

central and well annotated genes to acquire new rankings), the guilt-by-association methods do provide further, function-specific insight on this benchmark.

GeneMANIA Weighting Scheme. The default option for the GeneMANIA algorithm is to integrate multiple networks using query-specific weights, which reflect how well a network captures functional similarity between the query genes. For Compass, on the other hand, networks were combined without query-specific weighting. To ensure that the observed difference in performance was not simply due to GeneMANIA's query-specific weighting of the networks, GeneMANIA was also run without the seed-specific weighting step. The relative performance of the two algorithms was not affected: removal of seed-specific weighting decreased GeneMANIA's performance to AUC 0.800 (compared to 0.803 with the default setting and 0.8286 for Compass).

Cross-Validation vs Rollback with GO Benchmark. Constructing the rollback benchmark gave us the opportunity to explicitly explore potential biases in how GO-based benchmarks assess predictive performance. We started by comparing the rollback to cross-validation. As discussed in the introduction, there are concerns that cross-validation using known labels does not adequately capture how well GBA methods predict *new* annotations. To assess this, we compared how Compass performed on the rollback benchmark with its performance as evaluated by cross-validation using the labels acquired prior to the cut-off date. As expected, performance was higher using cross-validation (see Fig 3), suggesting information transfer between functional association networks and the known labels makes the known labels 'too easy' to predict during cross-validation.

Furthermore, the correlation between AUC (area under receiver operating characteristic curve) as evaluated by cross-validation and using the rollback benchmark was relatively low (Pearson's correlation coefficient of 0.260). This indicates that cross-validation on known protein sets is not a particularly good indicator of performance at predicting novel labels.

Effect of Discovery Date on Label Predictability. As discussed in the introduction, a rollback benchmark does not necessarily guarantee independence between the network and the

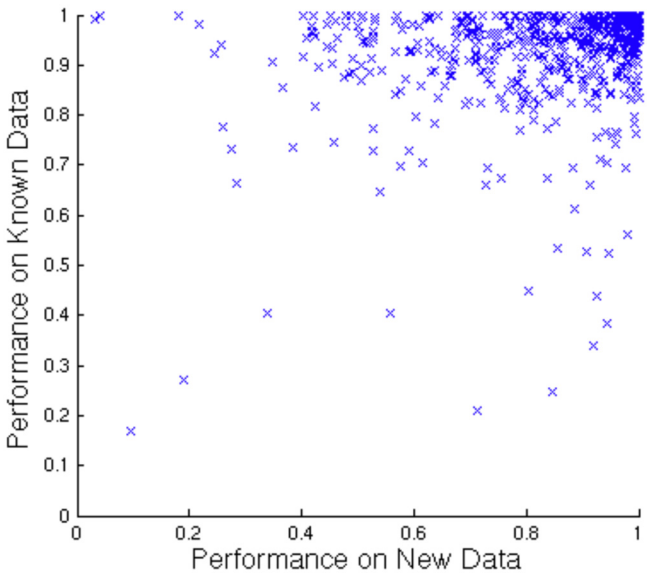

Fig 3. Performance on known vs new labels. Comparison between Compass performance on new data and known data (as measured by two-fold cross validation) on the GO benchmark. Each data point represents performance on one GO term as measured by AUC.

doi:10.1371/journal.pone.0134668.g003

test data. If currently known functional associations do indeed drive label acquisition, we would expect the date a new label was acquired to affect how easily it is retrieved: labels acquired close to the cut-off date would be easier to predict than those acquired later. We therefore looked at the correlation between how highly true positives were ranked and the date the annotation was made (see Fig 4).

There is a significant positive correlation between the ranking and the annotation date for both Compass and GeneMANIA (Spearman Correlation Coefficient (SCC) 0.197 and 0.163 respectively, $p < 10^{-15}$), indicating early new labels are indeed easier to predict (top ranking = rank of 1, see Fig 4).

Temporal Effects by Information Source. The STRING network is composed of functional associations from a number of sources: genome neighbourhood, gene fusion, genome co-occurrence, co-expression, experiments (i.e. high-throughput screens for physical interaction), databases (i.e. curated small-scale interaction screens and annotated pathways) and textmining. Some of these evidence types, such as textmining, may be more susceptible to the temporal effects described above. We therefore looked at the correlation between label discovery date and predictability in each source network individually (Table 2).

After correction for multiple testing using Sidak correction (the corrected significance level $\alpha = 0.0037$, given by $\alpha = 1 - (1 - \bar{\alpha})^{1/k}$, where k is the number of comparisons (14) and $\bar{\alpha}$ is the original significance level (0.05)), the experiment, database and textmining networks all showed significant positive correlations between ranking and discovery date (i.e. earlier labels were easier to predict).

It is not surprising that the effect is seen in these three networks, because these networks represent known protein interactions, whereas genome neighbourhood, gene fusion and co-occurrence networks are classed as *de novo* interaction prediction methods [16]. Indeed, if the temporal effect is due to a lag in information transfer between databases, we would not expect to observe a correlation between discovery date and ease of prediction for the *de novo* networks: GO annotations inferred from these functional associations would be classed as non-

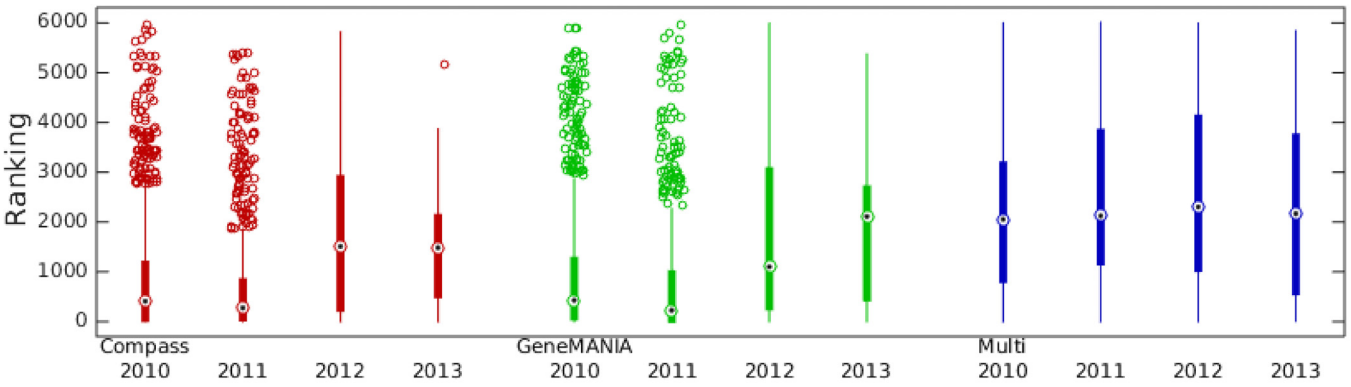

Fig 4. Relationship between how easy an annotation is to predict and the year the annotation was made in the GO. The ease of prediction is measured as the ranking of the gene in a prioritized list (i.e. a rank of 1 indicates the highest prioritized gene). The relationship is shown for Compass (red), GeneMANIA (green) and a multifunctionality-based predictor (blue).

doi:10.1371/journal.pone.0134668.g004

experimental and would therefore be excluded from our rollback benchmark which only included experimental annotations. If, on the other hand, the temporal effect is due to the publicly available functional association data guiding choices of targets for experiments, we would also expect the effect to be clearest in the experimental, database and textmining networks as these information sources are more widely used than the other four.

Phenotype-Based Independent Benchmarks

The results from the GO rollback benchmark are relevant for the design of CAFA or MouseFunc style competitions because they suggest the time period between prediction and assessment (i.e. the time window allowed for new annotations to accumulate) could affect how well a method appears to perform. This issue is particularly noteworthy because genes not labelled with a particular term at the time of evaluation are considered negatives although they could actually represent hidden true positives [35]. This could lead to penalisation of methods ranking the 'more difficult' and not yet discovered labels higher than the more obvious ones. This leads to the concern that competition style benchmarks may encourage the building of tools to mimic experimental discovery as opposed to guiding it. Thus, the re-evaluation of algorithms

Table 2. Correlation between label predictability and date of discovery in different network types.

	C SCC	C p	G SCC	G p
Neighourbood	0.03	5.2×10^{-1}	-0.11	4.5×10^{-2}
Gene Fusion	-0.19	2.2×10^{-2}	-0.20	1.8×10^{-2}
Cooccurrence	-0.26	6.5×10^{-3}	-0.13	1.6×10^{-1}
Coexpression	-0.07	2.3×10^{-2}	0.06	6.1×10^{-2}
Experiments	**0.15**	1.3×10^{-15}	**0.15**	2.0×10^{-15}
Databases	**0.14**	1.2×10^{-5}	**0.20**	3.8×10^{-10}
Textmining	**0.13**	1.2×10^{-9}	**0.08**	3.0×10^{-4}

Spearman correlation (SCC) between the ranking of true positive annotations and the date the annotation was made in the GO rollback benchmarks for Compass (C) and GeneMANIA (G). A positive correlation indicates early labels are easy to predict because a low numerical rank (example rank = 1) indicates an easily predictable gene. Significant correlations ($p < 0.0037$, derived using a Sidak multiple comparison correction $\alpha = 1 - (1 - \bar{\alpha})^{1/k}$, where k is the number of comparisons (14) and α the original significance level (0.05)) have been highlighted (bold).

doi:10.1371/journal.pone.0134668.t002

after a longer wait period could provide valuable insight into their performance. Indeed, CAFA has been designed to allow reassessment of algorithm performance at a later date [20].

As an alternative way of addressing the concerns discussed above, we designed two benchmarks where the network data and the test data were definitely independent. As before, we used the 2009 STRING functional association networks. The gene sets used for testing were

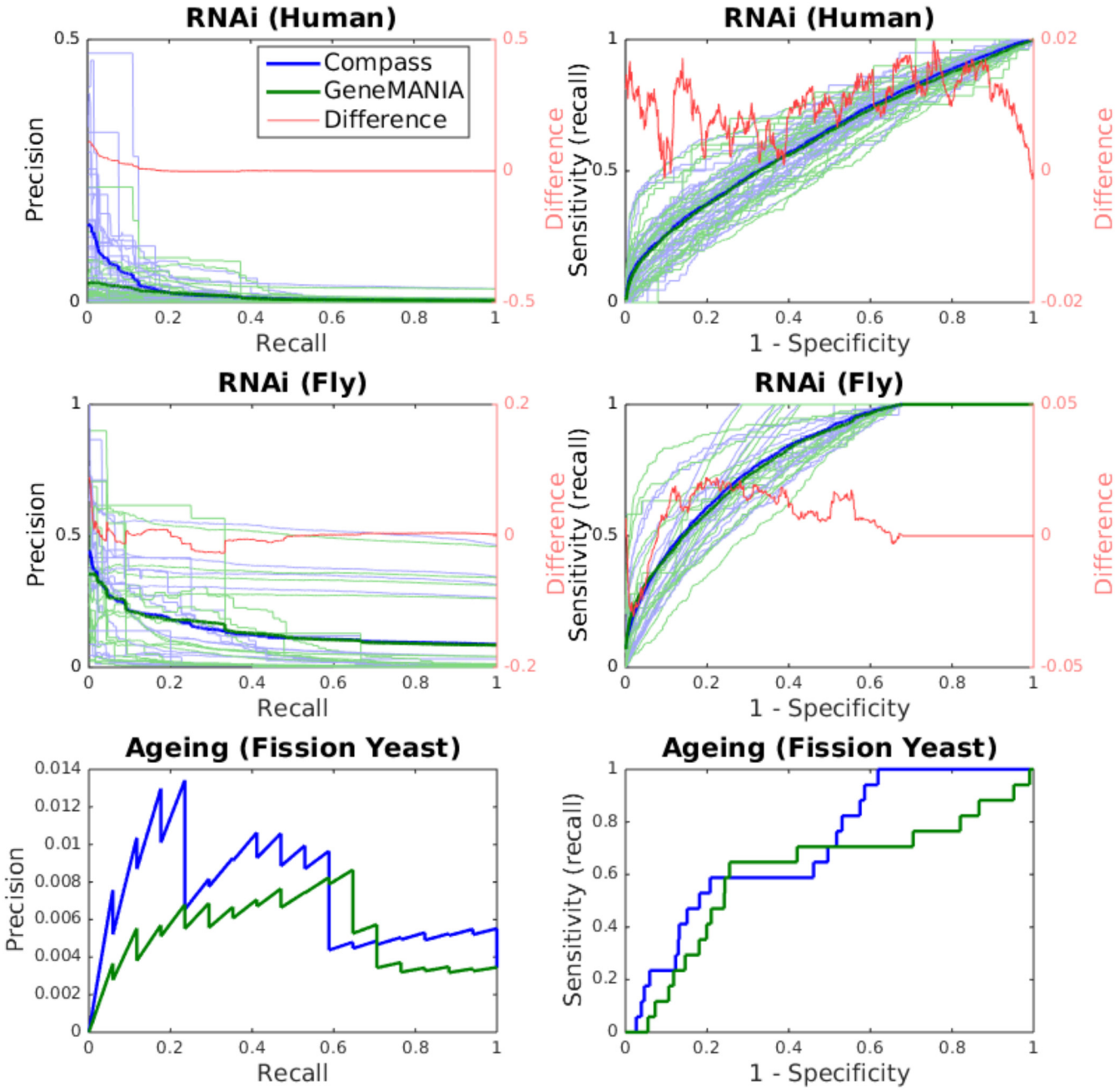

Fig 5. Precision-recall and ROC curves for the phenotypic benchmarks (RNAi and Ageing) for Compass and GeneMANIA. Precision-recall (right) and ROC (left) are shown for all gene sets (27 for human and 18 for fly) in the RNAi benchmark (top two rows). Average curves are also shown, as well as the average difference between the two methods (Compass minus GeneMANIA). The bottom row shows precision-recall and ROC curves for the fission yeast ageing benchmark (one gene set). The precision-recall curves for the plots depicting multiple gene sets (top two rows) have been interpolated for clarity.

doi:10.1371/journal.pone.0134668.g005

Table 3. Summary of Compass and GeneMANIA performance on the phenotype-based benchmarks.

Benchmark/Measure	COMPASS	GeneMANIA	p-value
AUC RNAi (Human and Fly)	0.6542 (sd 0.0773)	0.6442 (sd 0.0927)	0.0175
P_{mean}RNAi (Human and Fly)	0.0628 (sd 0.1101)	0.0593 (sd 0.1103)	0.0051
$P_{R = 0.1}$ RNAi (Human and Fly)	0.1127 (sd 0.1725)	0.0965 (sd 0.1568)	0.0019
AUC Ageing (Fission Yeast)	0.713	0.613	N/A
P_{mean}Ageing (Fission Yeast)	0.0082	0.0057	N/A
$P_{R = 0.1}$ Ageing (Fission Yeast)	0.0104	0.0055	N/A

Predictive performance of Compass and GeneMANIA, as measured by the area under the receiver operating characteristic (ROC) curve (AUC), mean average precision (P_{mean}) and precision at recall 0.1 ($P_{r = 0.1}$), on the phenotype-based benchmarks. The RNAi benchmark consists of 45 gene sets (27 for human and 18 for fly). Standard deviations (sd) are reported in parenthesis. The ageing benchmark consists of 17 genes associated with a long lived phenotype in a knock-out screen. For the RNAi benchmark, p-values are derived from a two-tailed Wilcoxon ranked sum test. For the ageing benchmark (which consists of only one gene set), the statistical significance of the difference in performance was evaluated by comparing how highly each long lived mutant was ranked by Compass and Genemania, giving a p-value of 0.0168 (two-tailed Wilcoxon sing-rank test).

doi:10.1371/journal.pone.0134668.t003

derived from genome-wide knock-out screens performed *after this date*. Test sets consisted of genes giving rise to a particular phenotype when knocked-out. This principle was used to construct two benchmarks: one based on RNAi screens in fly and human and one based on a screen for ageing related genes in fission yeast.

RNAi Benchmark. This benchmark was based on a database of phenotypes observed in various RNAi knock-out screens. As shown in Fig 5 and summarised in Table 3, Compass significantly outperforms GeneMANIA on this benchmark.

Ageing Benchmark. This benchmark was based on long lived mutants (n = 17) identified in a genome wide screen [29]. Prediction was seeded with long lived mutants known prior to the screen (see Methods). On this benchmark, COMPASS outperforms GeneMANIA (Fig 5 and Table 3). The statistical significance of this result was evaluated by comparing how highly each long lived mutant was ranked by Compass and Genemania, giving a p-value of 0.0168 (two-tailed Wilcoxon sing-rank test).

Conclusion

We have proposed a novel guilt-by-association prediction algorithm (Compass) for gene function prediction which outperforms GeneMANIA, a leading network-based prediction algorithm, on a CAFA-style GO-based benchmark and two phenotype-based benchmarks.

Additionally, we have shown that information transfer between databases may affect not only cross-validation, but also benchmarks based on the accumulation of new labels, such as CAFA-style competitions. Our results suggest the length of the wait period between prediction and evaluation affects how well algorithms appear to perform. Thus, re-evaluation of prediction methods after a longer wait period may provide further insight into the relative performance of algorithms. We have also proposed an alternative, phenotype-based, benchmark which may, in context where knock-out phenotype is a relevant indicator of protein function, serve as a complementary assessment method.

Supporting Information

S1 Fig. Compass parametrization. Compass performance on the GO benchmark set, using different number of dimensions for the PLS regression. Performance is measured by area under ROC curve (AUC). Performance is shown estimated from cross-validation on the seed set

('seed') and prediction of new labels ('new'). Error bars represent standard error of the mean. The number of optimal dimensions between seed set and novel set is the same: performance is maximized using a single dimension. This is in line with previous work recommending the use of K-1 dimensions for PLS discriminant analysis, where K is the number of classes [24].
(TIF)

S1 Text. Gene list for ageing benchmark. Experimentally derived set of fission yeast (*Schizosaccharomyces pombe*) long-lived mutants from a longevity screen by Sideri et al [29].
(TXT)

Author Contributions

Conceived and designed the experiments: CO JB JST JL SL. Performed the experiments: SL JL. Analyzed the data: SL. Contributed reagents/materials/analysis tools: JL JST JB CO. Wrote the paper: SL CO.

References

1. Rao VS, Srinivas K, Sujini GN, Kumar SN. Protein-protein interaction detection: methods and analysis. International journal of proteomics. 2014; 2014. doi: 10.1155/2014/147648 PMID: 24693427
2. Deng M, Zhang K, Mehta S, Chen T, Sun F. Prediction of protein function using protein-protein interaction data. Journal of computational biology: a journal of computational molecular cell biology. 2003; 10 (6):947–960. doi: 10.1089/106652703322756168
3. Letovsky S, Kasif S. Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics (Oxford, England). 2003 Jul; 19 Suppl 1(suppl 1):i197–i204. doi: 10.1093/bioinformatics/btg1026
4. Kourmpetis YAI, van Dijk ADJ, Bink MCAM, van Ham RCHJ, ter Braak CJF. Bayesian Markov Random Field Analysis for Protein Function Prediction Based on Network Data. PLoS ONE. 2010 Feb; 5(2): e9293+. doi: 10.1371/journal.pone.0009293 PMID: 20195360
5. Tsuda K, Shin H, Schölkopf B. Fast protein classification with multiple networks. Bioinformatics. 2005 Jan; 21(suppl 2):ii59–ii65. doi: 10.1093/bioinformatics/bti1110 PMID: 16204126
6. Warde-Farley D, Donaldson SL, Comes O, Zuberi K, Badrawi R, Chao P, et al. The GeneMANIA prediction server: biological network integration for gene prioritization and predicting gene function. Nucleic acids research. 2010 Jul; 38(Web Server issue):W214–W220. doi: 10.1093/nar/gkq537 PMID: 20576703
7. Mostafavi S, Ray D, Warde-Farley D, Grouios C, Morris Q. GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. Genome biology. 2008; 9 Suppl 1(Suppl 1):S4+. doi: 10.1186/gb-2008-9-s1-s4 PMID: 18613948
8. Peña Castillo L, Tasan M, Myers CL, Lee H, Joshi T, Zhang C, et al. A critical assessment of Mus musculus gene function prediction using integrated genomic evidence. Genome biology. 2008; 9 Suppl 1 (Suppl 1):S2+. doi: 10.1186/gb-2008-9-s1-s2 PMID: 18613946
9. Lanckriet GR, Deng M, Cristianini N, Jordan MI, Noble WS. Kernel-based data fusion and its application to protein function prediction in yeast. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing. 2004;p. 300–311.
10. Lee H, Tu Z, Deng M, Sun F, Chen T. Diffusion kernel-based logistic regression models for protein function prediction. Omics: a journal of integrative biology. 2006; 10(1):40–55. doi: 10.1089/omi.2006.10.40 PMID: 16584317
11. Nabieva E, Jim K, Agarwal A, Chazelle B, Singh M. Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. Bioinformatics. 2005 Jun; 21(suppl 1):i302–i310. doi: 10.1093/bioinformatics/bti1054 PMID: 15961472
12. Hériché JK, Lees JG, Morilla I, Walter T, Petrova B, Roberti MJ, et al. Integration of biological data by kernels on graph nodes allows prediction of new genes involved in mitotic chromosome condensation. Molecular Biology of the Cell. 2014 Jun; 25(16):mbc.E13–04–0221–2536. doi: 10.1091/mbc.E13-04-0221 PMID: 24943848
13. Yen L, Fouss F, Decaestecker C, Francq P, Saerens M. Graph Nodes Clustering Based on the Commute-Time Kernel. In: Zhou ZH, Li H, Yang Q, editors. Advances in Knowledge Discovery and Data Mining. vol. 4426 of Lecture Notes in Computer Science Springer Berlin Heidelberg; 2007. p. 1037–1045.

14. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics. 2000 May; 25(1):25–29. doi: 10.1038/75556 PMID: 10802651

15. Stark C, Breitkreutz BJJ, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic acids research. 2006 Jan; 34(Database issue):D535–D539. doi: 10.1093/nar/gkj109 PMID: 16381927

16. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, et al. STRING 8–a global view on proteins and their functional interactions in 630 organisms. Nucleic acids research. 2009 Jan; 37(Database issue):D412–D416. doi: 10.1093/nar/gkn760 PMID: 18940858

17. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research. 2000 Jan; 28(1):27–30. doi: 10.1093/nar/28.1.27 PMID: 10592173

18. Gillis J, Pavlidis P. Assessing identity, redundancy and confounds in Gene Ontology annotations over time. Bioinformatics. 2013 Feb; 29(4):476–482. doi: 10.1093/bioinformatics/bts727 PMID: 23297035

19. Rogers MF, Ben-Hur A. The use of gene ontology evidence codes in preventing classifier assessment bias. Bioinformatics. 2009 May; 25(9):1173–1177. doi: 10.1093/bioinformatics/btp122 PMID: 19254922

20. Radivojac P, Clark WT, Oron TRR, Schnoes AM, Wittkop T, Sokolov A, et al. A large-scale evaluation of computational protein function prediction. Nature methods. 2013 Mar; 10(3):221–227. doi: 10.1038/nmeth.2340 PMID: 23353650

21. Gillis J, Pavlidis P. Characterizing the state of the art in the computational assignment of gene function: lessons from the first critical assessment of functional annotation (CAFA). BMC Bioinformatics. 2013; 14(Suppl 3):S15+. doi: 10.1186/1471-2105-14-S3-S15 PMID: 23630983

22. Barker M, Rayens W. Partial least squares for discrimination. J Chemometrics. 2003 Mar; 17(3):166–173. doi: 10.1002/cem.785

23. Pérez-Enciso M, Tenenhaus M. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Human genetics. 2003 May; 112(5–6):581–592. PMID: 12607117

24. Le Cao KA, Boitard S, Besse P. Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics. 2011; 12(1):253+. doi: 10.1186/1471-2105-12-253 PMID: 21693065

25. Bastien P, Bertrand F, Meyer N, Maumy-Bertrand M. Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. Bioinformatics. 2015 Feb; 31(3):397–404. doi: 10.1093/bioinformatics/btu660 PMID: 25286920

26. Lees JG, Heriche JK, Morilla I, Ranea JA, Orengo CA. Systematic computational prediction of protein interaction networks. Physical biology. 2011 Jun; 8(3):035008+. doi: 10.1088/1478-3975/8/3/035008 PMID: 21572181

27. Shawe-Taylor J. Kernel methods for pattern analysis. Cambridge university press; 2004.

28. Schmidt EE, Pelz O, Buhlmann S, Kerr G, Horn T, Boutros M. GenomeRNAi: a database for cell-based and in vivo RNAi phenotypes, 2013 update. Nucleic Acids Research. 2013 Jan; 41(D1):D1021–D1026. doi: 10.1093/nar/gks1170 PMID: 23193271

29. Sideri T, Rallis C, Bitton DA, Lages BM, Suo F, Rodríguez-López M, et al. Parallel Profiling of Fission Yeast Deletion Mutants for Proliferation and for Lifespan During Long-Term Quiescence. G3 (Bethesda, Md). 2014 Dec.

30. Chen BR, Li Y, Eisenstatt JR, Runge KW. Identification of a Lifespan Extending Mutation in the Schizosaccharomyces pombe Cyclin Gene clg1+ by Direct Selection of Long-Lived Mutants. PLoS ONE. 2013 Jul; 8(7):e69084+. doi: 10.1371/journal.pone.0069084 PMID: 23874875

31. Ito H, Oshiro T, Fujita Y, Kubota S, Naito C, Ohtsuka H, et al. Pma1, a P-type Proton ATPase, Is a Determinant of Chronological Life Span in Fission Yeast. Journal of Biological Chemistry. 2010 Nov; 285(45):34616–34620. doi: 10.1074/jbc.M110.175562 PMID: 20829365

32. Roux AE, Quissac A, Chartrand P, Ferbeyre G, Rokeach LA. Regulation of chronological aging in Schizosaccharomyces pombe by the protein kinases Pka1 and Sck2. Aging cell. 2006 Aug; 5(4):345–357. doi: 10.1111/j.1474-9726.2006.00225.x PMID: 16822282

33. Montojo J, Zuberi K, Rodriguez H, Kazi F, Wright G, Donaldson SL, et al. GeneMANIA Cytoscape plugin: fast gene function predictions on the desktop. Bioinformatics. 2010 Nov; 26(22):2927–2928. doi: 10.1093/bioinformatics/btq562 PMID: 20926419

34. Gillis J, Pavlidis P. The Impact of Multifunctional Genes on "Guilt by Association" Analysis. PLoS ONE. 2011 Feb; 6(2):e17258+. doi: 10.1371/journal.pone.0017258 PMID: 21364756

35. Dessimoz C, Škunca N, Thomas PD. CAFA and the open world of protein function predictions. Trends in genetics: TIG. 2013 Nov; 29(11):609–610. doi: 10.1016/j.tig.2013.09.005 PMID: 24138813