Using prior knowledge in the inference of gene association networks

Isabel A. Nepomuceno-Chamorro[1] · Juan A. Nepomuceno[1] · José Luis Galván-Rojas[1] · Belén Vega-Márquez[1] · Cristina Rubio-Escudero[1]

Abstract

Traditional computational techniques are recently being improved with the use of prior biological knowledge from open-access repositories in the area of gene expression data analysis. In this work, we propose the use of prior knowledge as heuristic in an inference method of gene-gene associations from gene expression profiles. In this paper, we use Gene Ontology, which is an open-access ontology where genes are annotated using their biological functionality, as a source of prior knowledge together with a gene pairwise Gene-Ontology-based measure. The performance of our proposal has been compared to other benchmark methods for the inference of gene networks, outperforming in some cases and obtaining similar and competitive results in others, but with the advantage of providing simple and interpretable models, which is a desired feature for the Artificial Intelligence Health related models as stated by the European Union.

Keywords Gene-gene association networks · Ontology · Semantic similarity measure · Information fusion · Microarray data analysis

1 Introduction

The huge amount of data produced by the biotechnology techniques has grown exponentially in recent years [14]. Nowadays, the invention and application of High-throughput technologies offers scientists from biology and biomedicine the opportunity to gain a better understanding on the behaviour of genes such as the identification of novel gene-gene association, gene expression patterns or gene candidates in disease [19]. The microarray technology has the capacity to monitor changes in RNA[1] abundance for thousands of genes simultaneously, which can be represented as a numerical matrix after preprocessing steps well known as low-level microarray data analysis [28]. In this matrix, the rows correspond to genes, the columns to experimental conditions, and a value in the matrix is the expression value of a gene under a condition. In the field of gene expression data analysis, novel strategies are required to handle the huge amount data and to infer knowledge as gene regulatory networks.

To infer gene regulatory networks, the first step is to extract direct regulatory relationships between genes, i.e., gene-gene associations. The inference of gene-gene associations is based on the concept of guilt-by-association: gene co-expression implies gene co-regulation, i.e., groups of genes that show similar expression profiles also show the same regulatory regime or functionality. Coexpression networks are typically generated using coexpression methods, where each pair of genes is analyzed using correlation statistics as pairwise similarity measures. However, the assumption of guilt-by-association is being reformulated because the co-expression of a group of genes may be the result of an independent activation with respect to the same experimental condition and not due to the same regulatory regime [29]. The RegNet methodology infers gene-gene associations between genes with a similar expression profile under a subset of conditions. RegNet differs from pairwise measures-based methods in that the relationship between one gene and the remaining genes are calculated simultaneously using model trees.

High-throughput technologies have characterised genes by means of multiple functionalities that are stored in pub-

[1] RNA: RiboNucleic acid

✉ Isabel A. Nepomuceno-Chamorro
inepomuceno@us.es

[1] Dpto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Sevilla, Spain

lic databases and open-access data repositories such as the Gene Ontology project (GO) or Kyoto Encyclopedia of Genes and Genomes (KEGG). GO is an ontology which provides a hierarchical structure with three domains: molecular function, biological process and cellular component. A GO annotation is a term in the ontology or vocabulary which is linked to a group of genes. Genes are related to a set of GO annotations with different levels of specificity. GO is freely available for community use in the annotation of genes, gene products and sequences. Many organism model databases and genome annotation groups use the GO and contribute their annotation sets to the GO resource [2]. In the context of gene expression data analysis, GO has been used to validate the quality of computational results from a biological view point [39].

In this paper, we present a novel methodology named PriorRegNet to obtain gene-gene association networks by integrating prior knowledge from GO. PriorRegNet is based on the methodology RegNet [31] to infer gene networks and the fusion of information is made by means of functional annotation files extracted from GO. The biological information is included in RegNet with a GO-based measure that calculates the overlapping among the annotated terms of a pair of genes proposed in [27]. The motivation of our work is to provide a new approach to infer gene-gene associations using prior knowledge, taking the 2-fold advantage of RegNet. First, gene-gene associations are detected favouring more localised similarities over global similarities following the reformulated guilt-by-association assumption. Second, the a priori biological knowledge is used to drive the process and also to generate a simple and explainable model.

The remainder of this paper is organized as follows. In section *Related work*, the benchmark methods and the similarity measures are shown. In Section *Methodology*, a detailed explanation of the proposal is presented. In Section *Experiments*, a blind test is shown as a framework to validate the proposal. Finally, Section *Conclusions* summarizes the most relevant conclusions and future research directions.

2 Related work

The approach used in this work makes use of a well-known method to infer a gene network. The method used in the inference of gene-gene associations is RegNet. It is a model tree-based method to infer gene-gene associations favouring local similarities between genes, i.e., that share the same behaviour under a subset of samples in gene expression data, following the reformulation of the concept of guilt-by-association mentioned before. The idea behind this work is to integrate prior knowledge extracted from GO, i.e., to integrate prior knowledge consists on choosing the pair of genes in the network that are functionally coherent. To this aim, we used Gene pairwise GO semantic similarity measures, that can be defined as a distance among genes according to their information stored in GO. The goal of this type of measures is to report a value for each pair of genes which establishes their functional similarity [30].

In this section, a short state-of-the-art is described in two fields: the benchmark network inference methods and the GO semantic similarity measures.

2.1 Inference of gene-gene associations from microarray data

Inferring gene-gene associations from gene expression data or microarray data is a relevant task since this is the first step to infer gene regulatory networks from several type of sources. There are several methods, see [11] and [45].

2.1.1 Methods based on similarity measures:

These methods are based on the use of statistical measures to extract pairs of genes which have similar expression profiles under the set of experimental conditions. These methods are also known as co-expression networks or dependency networks. Examples of these methods are the proposals presented in [23, 25] and [15]. In these proposals several statistical measures are used as correlation, partial correlation and mutual information measures. The partial correlation methods are also known as Gaussian Graphical Models and it is a full conditional independence model. On the other hand, clustering algorithms represent one of the first methods to infer gene-gene associations networks [13]. Clustering algorithms are distance-based methods.

2.1.2 Bayesian networks:

These methods are probabilistic and the most relevant work is presented by Friedman et al. in [16]. This method is based on the Sparse Candidate algorithm, which selects in an iterative way the set of candidate genes that are related with the target gene maximizing the evaluation function. Other relevant approaches are [40, 43] and [7]. In [40] the authors use an Expectation Maximization algorithm and regression trees to build networks to be maximized using Bayesian punctuation. Finally, [7] uses a classical probabilistic graphical model.

2.1.3 Tree-based methods:

The work presented by Soinov et al. in [41] is a reference method based on the use of decision trees. In this work, the C4.5 algorithm developed by Ross Quinlan in [37] is used to build the decision trees. This algorithm is based on the concept of entropy and how this measure is adjusted between different partitioning of the search space.

The method named REGNET is based on model trees or regression trees, i.e., in the linear similarity between pair of genes under a subspace of the search space. This strongly favours localised similarities over more global similarities, i.e., under a subset of experimental conditions instead of the whole set of experimental conditions. This method was published in [31], the software tool is provided in [33] and it has been used in [32] and [38].

Lastly, GENIE3, presented in [20], uses random-forest regression in combination with transcription factor data to predict the expression of each gene in the dataset. Then, when the expression of genes is predicted, the different models measure the relationship or relevance between each transcription factor and the prediction of each target gene, in order to derive weights that will then be used to establish relationships between genes. This method has been used in [3] and implemented in [1].

2.1.4 Other methods:

A multi-objective evolutionary algorithm for mining quantitative association rules is developed to deal with the problem of network inference in [26]. This work presents the method named GarNet. Finally, Ponzoni et al. present in [36] the method GRNCOP and GRNCOP2, which are combinatorial optimization algorithms.

Furthermore, MFR is a SVM-based method to infer gene networks using prior knowledge [44]. However, the reproducibility is difficult and the comparison with our proposal is not possible because the method used different prior knowledge (COXPRESdb, KEGG and TRRUST databases).

2.2 Semantic similarity measures

Gene Ontology is an ontology where genes are annotated using their biological functionality. This ontology is an open-access repository widely used in the gene expression data analysis and Bioinformatics area. GO is an ontology with a hierarchical structure with three roots or domains[2]:

Cellular Component: the parts of a cell or its extracellular environment.

Biological Process: operations or sets of molecular events with a defined beginning and end. These operations are related to the functioning of integrated living units: cells, tissues, organs, and organisms.

Molecular Function: the elemental activities of a gene product at the molecular level, such as binding or catalysis.

The GO ontology is structured as a directed acyclic graph, and each term has defined relationships to one or more other terms in the same domain, and sometimes to other domains. Each gene is annotated to a set of GO terms with different levels of specificity. And each gene is annotated to a term under an evidence code denoting the type of evidence upon which the annotation is based.

The semantic similarity measures allow obtaining numerical values to show the closeness between a pair of terms in ontology. Every gene is related to a set of GO terms, therefore several gene pairwise GO-based measures have been proposed in the literature. These measures can be classified in: measures based on the node of the graph [10]; measures based on the associations between the terms of the graph [35]; and hybrid measures [35]. Other survey as [17] state a big classification into five categories: methods based on semantic distance, methods based on information content, methods based on properties of terms, methods based on ontology hierarchy, and hybrid methods.

As Pesquita et. al. stated in [34], the hybrid measures obtain good results. Among them, SIMGIC measures obtain better results in general applications. SIMGIC measures can be defined as:

$$simGIC(A,B) = \frac{\sum_{t \in \{GO(A) \cap GO(B))\}} IC(t)}{\sum_{t \in \{GO(A) \cup GO(B))\}} IC(t)} \quad (1)$$

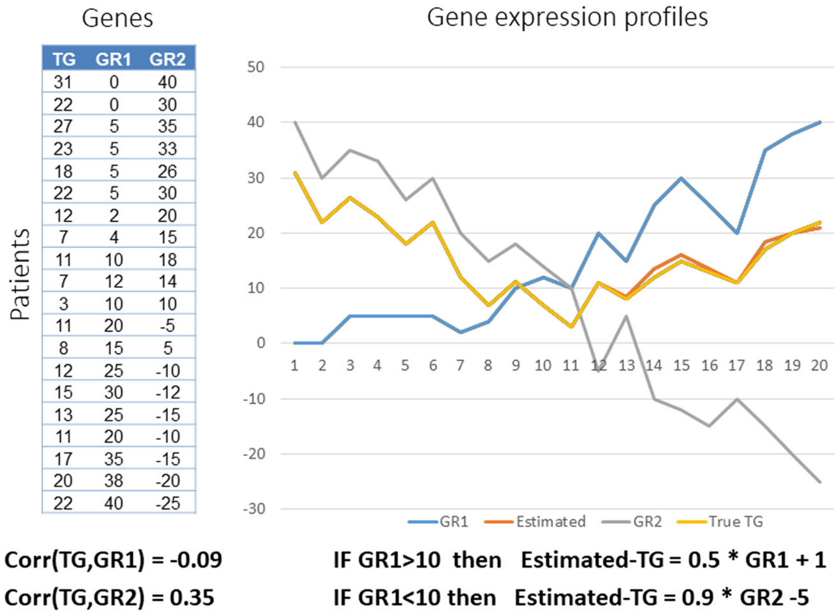
where A and B are the genes under studied, IC is the information content, GO(A) is the GO terms where A is annotated, GO(B) is the GO terms where B is annotated, $t \in \{GO(A) \cap GO(B))\}$ and $t \in \{GO(A) \cup GO(B))\}$ are the terms obtained from the intersection or union between the GO terms of A and B.

The tool GOSSTO [8] (the Gene Ontology Semantic Similarity Tool) implements several gene pairwise GO-based measures. GOSSTO provides a SIMGIC measure together with other five different semantic similarity measures. This can be used as a command line tool or can be integrated with other software packages due to the huge API documentation provided by authors. Furthermore, one of the biggest advantages of GOSSTO is the usability of the tool because other measures can be integrated in an easy manner.

3 Methodology

In this work, we propose the integration of prior knowledge into a method to infer gene-gene association networks. Specifically, we extend the REGNET methodology to include prior knowledge based on the gene pairwise GO-based measure SIMGIC. We integrate the GOSSTO tool into the REGNET methodology.

The method named REGNET is based on model trees or regression trees, i.e., in the linear similarity between pair of



TG	GR1	GR2
31	0	40
22	0	30
27	5	35
23	5	33
18	5	26
22	5	30
12	2	20
7	4	15
11	10	18
7	12	14
3	10	10
11	20	-5
8	15	5
12	25	-10
15	30	-12
13	25	-15
11	20	-10
17	35	-15
20	38	-20
22	40	-25

Corr(TG,GR1) = -0.09

Corr(TG,GR2) = 0.35

IF GR1>10 then Estimated-TG = 0.5 * GR1 + 1

IF GR1<10 then Estimated-TG = 0.9 * GR2 -5

Fig. 1 The method RegNet strongly favours localised similarities over more global similarities. TG can be estimated by GR1 and GR2 using the linear models under a subspace of the search space, i.e., under a subset of conditions of the input dataset [31]

variables (genes) under a subspace of the search space , i.e., under a subspace of experimental conditions. In Fig. 1 we can see an example of regression tree. Let M be a microarray dataset consists of 20 experimental conditions and 3 genes or variables called target gene (TG) and genes R1 and R2. It can be observed that the target gene and the gene GR1 are not correlated, but they have a linear dependency under the subset of experimental conditions where the expression value of GR1 is greater than 10. The linear dependency is $\widehat{TG} = 0.5 * GR1 + 1$ and can be observed in a orange line (estimated line). In a similar way, the TG and GR2 have a linear dependency when the expression value of GR1 is less than 10 and the expression value of GR2 is greater than 10. The regression tree in this case would be a tree rooted on GR1 as the tree on Fig. 2. The method RegNet is based on the existence of linear dependencies between the target gene and several genes under a subspace of the search space. RegNet has three different steps. First, the model trees are built building a forest of model trees, one model tree is build using the M5' algorithm [46] for each gene. Second, the forest of trees is pruned and the linear dependencies are extracted as hypothetical dependencies. Finally, a statistical method described in [5] is applied to reduce the false discovery rate.

The new approach is based on incorporating the semantic similarity heuristic over the tree-based method RegNet. The use of semantic similarity as heuristic for inferring gene association networks allows for the direct search of relationships based on real evidence, which are contrasted in biological databases such as Gene Ontology. The methodology is divided into four phases clearly differentiated, as we can observe in Fig. 3:

Step 1. Building a forest of trees: The first step has as input the microarray dataset and this is an iterative process. In each iteration, a gene is considered as the target gene and the remaining genes are used as input to build a model tree using the M5' algorithm. The M5' algorithm has been proposed by Witten et al. in [46]. The implementation of the M5' algorithm provided by the Weka software package is used. Only the tree with a relative error less than a threshold value θ is taken into account by the method to build the forest of trees.

Step 2. Dependencies extraction: During this phase the dependencies between the target genes and the genes involved in the linear models are extracted from the forest of trees. From the leaves of each tree, the linear models

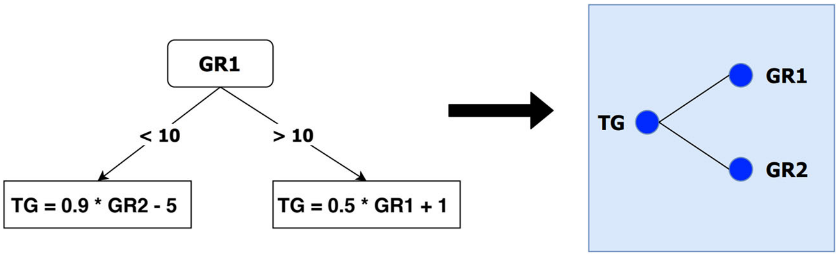


Fig. 2 Regression tree based on data on Fig. 1

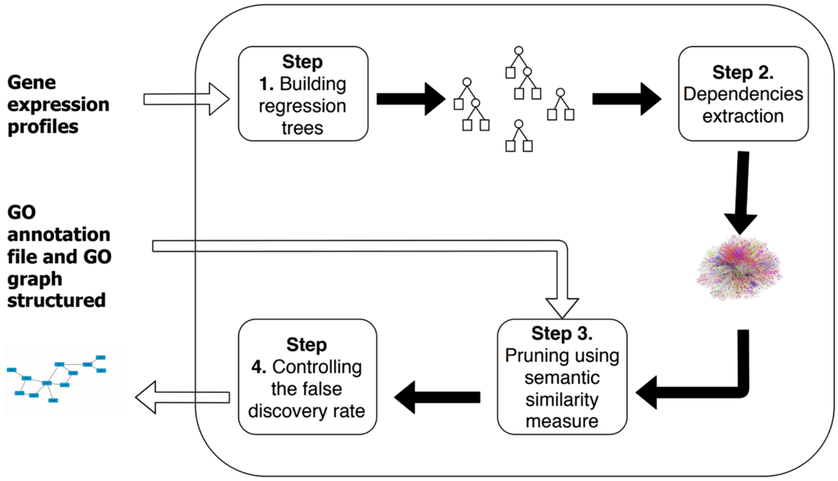

Gene expression profiles
GO annotation file and GO graph structured
Step 1. Building regression trees
Step 2. Dependencies extraction
Step 3. Pruning using semantic similarity measure
Step 4. Controlling the false discovery rate

Fig. 3 Steps in PriorRegNet methodology

are studied and the genes involved in it are extracted and considered as an hypothetical gene dependency with the target gene. Linear dependencies between pair of genes imply that they work in the same regulatory regime [31]. The linear models generated by the M5' algorithm follow the next equation:

$$ML : g_x = \sum_i \lambda_i g_{y_i} \qquad (2)$$

where g_x is the target gene and the set of g_y are a subset of genes from the remaining genes in the input dataset.

Step 3. Semantic similarity measure: In this phase, the fusion of information is made taking as input the set of hypothetical gene-gene dependencies that formed the network and the GO annotation file. The prior knowledge is based on the gene pairwise GO-based measure named SimGIC. In this phase, GOssTo tool is used to generate the semantic similarity value named σ between each gene-gene dependency from the network. If this value is greater than a threshold the dependency is maintained to the hypothetical set of gene-gene dependencies. If this value is less than a threshold the dependency is removed from the hypothetical set of gene-gene associations. To use this semantic similarity measure, the GO graph structure is needed as input to compute the IC together with the gene annotation file.

Step 4. Controlling the false discovery rate: In the last phase, a statistical procedure to control the false discovery rate is applied to the reduced set of hypothetical dependencies. The aim of this step is to control the type I error. The applied procedure is the Benjamini y Yekutieli method [5]. In this method, let $H_0^1, H_0^2, ..., H_0^m$ be the set of null hypothesis and let$p_1, p_2, ..., p_m$ be the p-values from the m null hypothesis. Let $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ be the sorted list of p-values. The Benjamini-Yekutieli procedure defined the K value to reject the hypothesis $H_0^1, H_0^2, ..., H_0^k$. The K-value is calculated as follows:

$$k = max\{i : p_{(i)} \frac{m}{i} \sum_{k=1}^{m} \frac{1}{k} \leq \alpha\} \qquad (3)$$

The hypothesis will not be rejected if there is not an i that satisfies the above equation. A gene-gene dependency will be identified as an edge in the network if and only if there is not any significant monotonic relationship between the two variables, i.e., $H_0 : \rho_{xy} \approx 0$ (where ρ is a correlation measure), taking into account the subspace of the input data identified by the leaf of the linear model in the M5' tree. To test whether a significant monotonic relationship exists, we use the Kendall's Tau as non-parametric measure of association.

4 Experiments

The results of the methodology are compared to other works in the area of gene network inference . The measures for the performance assessment of gene networks, the experimental design and the results are shown in the following subsections.

4.1 Measures for the performance assessment of gene networks

The work [26] is used as a context of comparison together with several measures. These measures are based on a contingency matrix where the gene network provided by the model is compared against a true network or reference network. The TP, FP, TN and FN are defined as follow building the contingency matrix:

- TP: is the number of gene–gene associations obtained by the proposal that also appear in the gene networks used as a true network in the test.

Table 1 Input parameters of GOssTo

GO ontology:	Biological Process
GO evidences:	EXP, IDA, IPI, IMP, IGI, IEP, TAS, IC, ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA, NAS, ND, IEA
GO relations:	is_a, part_of, regulates, positively_regulates, negatively_regulates, has_part

- FP: is the number of gene–gene associations obtained by the proposal that do not appear in the gene networks used as a true network in the test.
- TN: is the number of gene–gene associations not obtained by the proposal that do not appear in the gene networks used as a true network in the test.
- FN: is the number of gene–gene associations not obtained by the method that appear in the gene networks used as a true network in the test.

Based on this contingency matrix, these measures can be defined as follow:

Definition 1 *Network Accuracy:* The accuracy of a network is the proportion of true results (both true positives and true negatives) over the total number of sample cases.

Definition 2 *Network Precision:* is defined as the proportion of the true positives against all the positive results (true positives and false positives).

Definition 3 *Network Sensitivity:* it measures the proportion of true positives which are correctly identified.

Definition 4 *Network Specificity:* The specificity of a gene network measures the proportion of true negative which are correctly identified.

Definition 5 *F1-Score:* F1-Score is the harmonic mean of the precision and sensitivity.

Definition 6 *Number of associations:* it is the number of edges of the graph which model the gene network, that is the number of gene–gene associations detected by the method in the resulted gene network.

4.2 Data set and experimental design

We used the well-known, in the area of gene expression analysis, microarray dataset of Spellman [42] and Cho et al. [9] for the budding yeast (Saccharomyces cerevisiae) cell cycle. The datasets cdc15, cdc28 and alpha-factors were obtained for yeast cell cultures that were synchronized by three different methods and the datasets were defined as statistically independent. Our approach has been trained using a subset of 20 well-described genes which encode important proteins for cell-cycle regulation as Soinov et al. use in their work [41], where the 20 genes are enumerated. This subset of genes are used to compare to several works as established in [26].

We used YeastNet [22], GO [12] and Co-citation [21] as ground truth or true networks as is established in [26]. We use the networks in a blind performance test to compare the output of PriorRegNet against the true networks, to compare PriorRegNet against the approach without the information fusion phase (RegNet method) and to compare also against other benchmark methods as shown in [26]. YeastNet is a network structure report with 102803 potential gene–gene associations among 5483 yeast genes. This network was built mainly from two resources: GO annotation downloaded from the Saccharomyces cerevisiae Genome Database (SGD) and over 29000 Medline abstract that included the word Saccharomyces cerevisiae for perfect matches to each gene pairs in the network. The difference between these networks is that YeastNet is the global network and it can be divided into associations extracted from genes annotated in the Gene Ontology (GO network) or associations that are published in the previous literature Co-citation network. As true network we considerer the subnetwork formed by the 20 well-known genes due to

Table 2 Average values for the gene networks metrics achieved by the proposal

ID	θ	σ	α	Accuracy	Precision	Sensitivity	Specificity
1	[0-100]	[0.10-0.15]	0.05	59.30%	44.24%	5.35%	95.26%
2	[0-100]	[0.16-0.20]	0.05	59.21%	43.08%	5.10%	95.29%
3	[0-100]	[0.21-0.25]	0.05	59.24%	43.06%	5.10%	95.33%
4	[0-100]	[0.26-0.30]	0.05	59.34%	43.90%	4.84%	95.68%

Table 3 Performance of PRIORREGNET and REGNET

		Soinov et al.	BLS	GRNCOP2	GarNet	GENIE3	RegNet	Prior RegNet
YeastNet	Precision	50,00	88,89	93,33	93,75	40,00	100,00	66,67
	Accuracy	48,41	52,09	55,27	55,79	75,00	52,11	61,05
	Sensitivity	3,06	8,19	14,29	15,31	1,00	7,14	5,26
	Specificity	96,74	98,91	98,91	98,91	70,00	100,00	98,25
	F1-Score	5,77	15,00	24,79	26,32	1,95	13,33	9,75
Co-Citation	Precision	50,00	88,89	93,33	93,75	43,00	100,00	100,00
	Accuracy	56,29	60,00	63,16	63,68	76,00	58,42	59,47
	Sensitivity	3,61	9,64	16,87	18,07	1,00	8,13	7,23
	Specificity	97,20	99,07	99,07	99,07	71,00	100,00	100,00
	F1-Score	6,73	17,39	28,57	30,30	1,95	15,04	13,49
GO	Precision	50,00	55,56	73,33	75,00	45,26	71,43	60,00
	Accuracy	54,75	55,24	58,42	58,95	77,00	56,32	55,26
	Sensitivity	3,49	5,81	12,79	13,95	1,00	5,81	3,49
	Specificity	97,12	96,15	96,15	96,16	72,00	98,08	98,08
	F1-Score	6,52	10,52	21,78	23,52	1,96	10,75	6,60

the fact that if we take into account the whole network the number of false negative increases substantially. In the comparison, the metrics described above are measured: accuracy, precision, sensitivity, specificity and F-measure.

In Table 1, the parameter settings used by the semantic similarity measure SIMGIC in the information fusion phase is described. The Obo File and the Goa File are the GO graph structured and gene annotation file, respectively.

In Table 2, the parameter setting used to analyse the performance of our approach is shown. The main parameters are: the threshold value of the relative error from the generated trees, this value varies from 0 to 100 in increments of 10; the threshold value of the semantic similarity measure which varies from 0.1 to 0.3 in increments of 0.1; a level $alpha$=0.05 is fixed for the statistical procedure. Finally, the columns accuracy, precision, specificity and sensibility show the average values of these measures. The SIMGIC measure is fixed from 0.1 to 0.3 after visualizing the arithmetic mean and mode of the SIMGIC measure between all the gene pairs from the dataset. The higher frequency distribution is found under the interval [0.1, 0.3] where more than 80% of pairs of genes have this SIMGIC measure.

4.3 Results

In Table 3, the performance of PRIORREGNET and REGNET can be observed. Furthermore, the proposal is compared against others benchmark methods for the inference of gene networks in a similar way that is established in [26], where the association rule-based method named GarNet, the GRNCOP method [36], the decision tree-based methods [41] and the first order-based method [7] are used as benchmark methods. The contribution in [3] has been added as a more current method. The results have been carried out using YeastNet, Cocitation and GO as reference network in the blind test using the performance measures described on Section 4.1.

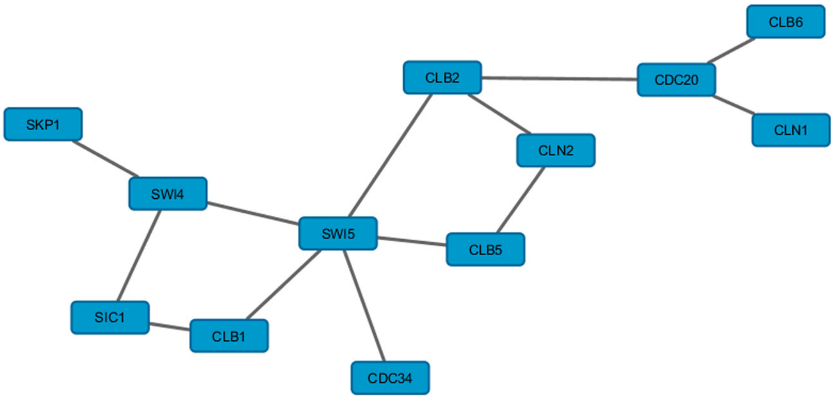


Fig. 4 *Gene-gene association network obtained by* PRIORREGNET

Table 4 GO enrichment analysis to determine whether the subset of genes obtained by PRIORREGNET still maintains common biological behaviour

N	attrib ID	attrib name
7	GO:0000079	regulation of cyclin-dependent protein serine/threonine kinase activity
3	GO:0010696	positive regulation of spindle pole body separation
7	GO:1904029	regulation of cyclin-dependent protein kinase activity
7	GO:0016538	cyclin-dependent protein serine/threonine kinase regulator activity
3	GO:0010695	regulation of spindle pole body separation
5	GO:0000086	G2/M transition of mitotic cell cycle
5	GO:0044839	cell cycle G2/M phase transition
5	GO:0000082	G1/S transition of mitotic cell cycle
5	GO:0044843	cell cycle G1/S phase transition
7	GO:0044770	cell cycle phase transition
7	GO:0044772	mitotic cell cycle phase transition
7	GO:0071900	regulation of protein serine/threonine kinase activity
7	GO:0019887	protein kinase regulator activity
7	GO:0019207	kinase regulator activity
6	GO:0090068	positive regulation of cell cycle process
6	GO:0045787	positive regulation of cell cycle
7	GO:0045859	regulation of protein kinase activity
7	GO:0043549	regulation of kinase activity
8	GO:0051338	regulation of transferase activity
10	GO:1903047	mitotic cell cycle process
7	GO:0001932	regulation of protein phosphorylation
11	GO:0022402	cell cycle process
7	GO:0042325	regulation of phosphorylation
9	GO:0051726	regulation of cell cycle
8	GO:0031399	regulation of protein modification process
7	GO:0019220	regulation of phosphate metabolic process
7	GO:0051174	regulation of phosphorus metabolic process
8	GO:0051301	cell division
9	GO:0032268	regulation of cellular protein metabolic process
8	GO:0030234	enzyme regulator activity
9	GO:0051246	regulation of protein metabolic process
7	GO:0010564	regulation of cell cycle process
8	GO:0098772	molecular function regulator

First column is the number of genes in the network from Fig. 4 with this GO attribute, second and third are the code and name of GO attribute respectively

In Fig. 4, the network of gene-gene associations constructed using the best configuration of PRIORREGNET for the Spellman dataset is shown. Furthermore, the set of genes from this network is used as input of the FuncAssociate tool [6] to report an enrichment analysis as we can see in Table 4. The enrichment analysis determine whether the subset of genes obtained by PRIORREGNET still maintains common biological behaviour and related with cell-cycle regulation. In Table 4, the Gene Ontology attributes report by FuncAssociate is shown. The first column N represents the number of genes in the network from Fig. 4 with this GO attribute, the second and third column are the code and name of the GO attribute.

5 Discussions

To the right of the 6 sub-figures of the Fig. 5, the performance of PRIORREGNET and REGNET can be observed. Furthermore, the performance of PRIORREGNET and REGNET can be observed in the two last columns of Table 3. The methodology with prior knowledge obtains better results in the case of accuracy and having YeastNet and Cocitation as reference networks. In the case of GO as reference network the accuracy is comparable with the results obtained by the methodology without prior knowledge. Furthermore, the proposal is compared against others benchmark methods for the inference of gene networks in a similar way that

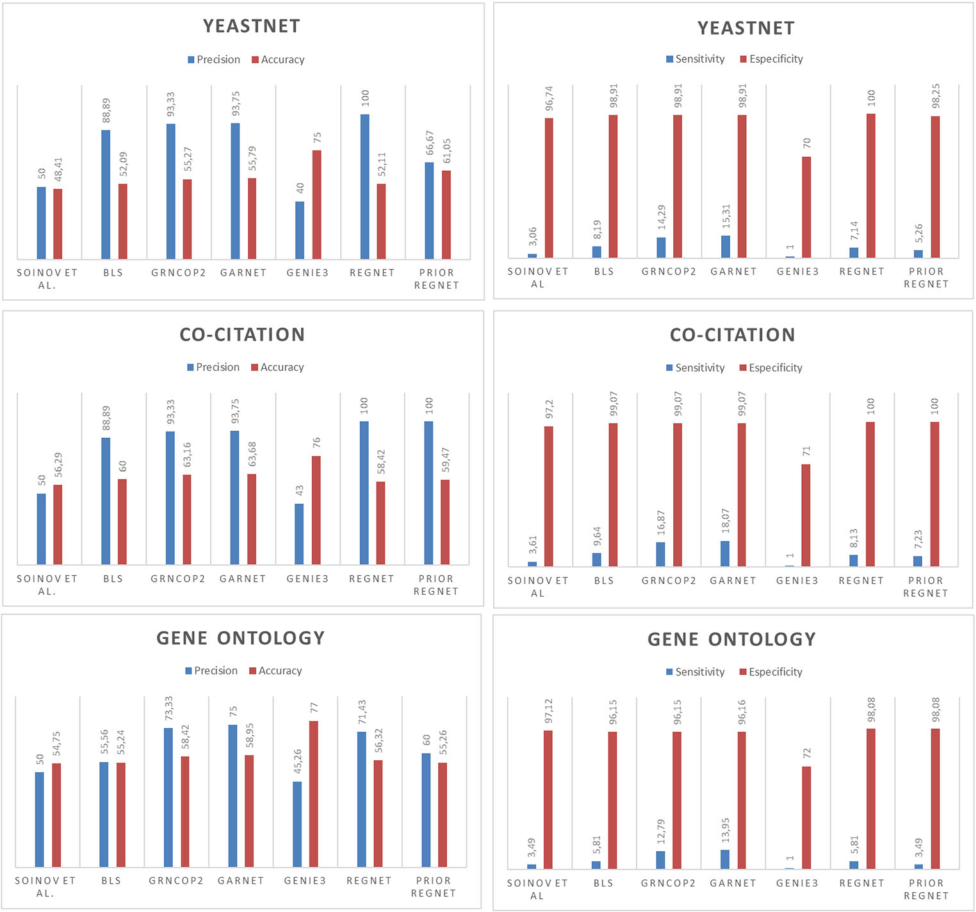


Fig. 5 Experimental results

is established in [26] and the results can be observed from first to second column to seventh column of the Table 3. In the case of YeastNet as true network, the accuracy obtained by the proposal is better than the others except for the GENIE3 algorithm where the results are comparable. In the case of Cocitation and GO the accuracy is not the best but are comparable to the rest of the methods. The precision, sensitivity, specificity and F1-Score measure are shown to remark that the methodology with prior knowledge is better or comparable to the rest of the approaches.

In Fig. 4, the network of gene-gene association constructed using the best configuration of PriorRegNet for the Spellman dataset is shown. The network is a graphical representation of the information comprised in the extracted model trees. Every node in this graph represents a gene and every arc indicates the association between genes. From these associations, CLB1-SIC1, SWI5-CLB2, CLB5-CLN2, CLN2-CLB2 and CDC20-CLN1 are in common with Soinov's network. Finally, the resulting genes are functionally enriched for GO attributes and the great majority of these GO attributes are related, under a *p*-value less than 0.05, with regulator activity of cell cycle and cell division. Thus getting fewer nodes does not means losing biological information. All the enriched-GO terms can be observed in Table 4, where the set of genes from this network is used as input of the FuncAssociate tool [6] to report an enrichment analysis. The enrichment analysis determines whether the subset of genes obtained by PriorRegNet still maintains common biological behaviour related to cell-cycle regulation. In Table 4, the Gene Ontology attributes reported by the FuncAssociate are shown. The first column N represents the number of genes in the network from Fig. 4 with this GO attribute, the second and third column are the code and name of the GO attribute respectively.

6 Conclusions

The integration of prior knowledge into a method to infer gene-gene association networks has been proposed in this work. The proposal is named PriorRegNet . The integration of prior knowledge is based on the use of a semantic similarity measure applied to gene products annotated with Gene Ontology terms named SimGIC .

We used YeastNet, GO and Co-citation as ground truth or true networks as it has already been established. We use the networks in a blind performance test to compare the output networks against the true networks.

The performance of the proposal has been studied against the same method without prior knowledge named RegNet , against a current random forest-based method named GENIE3 and against benchmark methods in the area of the inference networks as is established in [26]. The benchmark methods are a random forest-based method named GENIE3, a multi-objective evolutionary algorithm named Garnet, combinatorial optimization learning method named GRNCOP, a probabilistic graphical model named Bulashevska and a decision-tree-based method named Soinov. It is worth mention that with the expception of Soinov and PriorRegNet and RegNet, the rest of the methods provide blackbox models.

The use of prior knowledge in our proposal PriorRegNet against RegNet improves the accuracy measure of the proposed algorithm regarding the case of YeastNet and Cocitation (see Table 3). In the case of GO as true network, the accuracy is comparable to the results obtained by the methodology without prior knowledge. In the case of comparison against Soinov, PriorRegNet improves the performance of all measures. Finally, in the case of comparison against the rest of the methods, the proposal improves the accuracy for YeastNet and the rest of the measures the results are similar amongst all the approaches.

It is worth to mention that our proposal is based on a deterministic model tree-based method which provides glass-box models, i.e., our proposal provides a simple and explainable model. Our proposal, in comparison to methods that provide simple and explainable model as Soinov or RegNet improve the performance. And in all other cases it provides similar behaviour in performance measures but with the advantage of providing simple models, which is very remarkable as stated in the Joint Research Center's report from 2018 (European Commission) about the interpretability as a requisite of machine learning systems [24].

Future work will be focused on the study of other biological measures to handle other known open-access repository to integrate biological information. The experimentation will be increased with these other measures and others benchmark methods based on the use of prior knowledge. In the first case we could extend using, as for example, Gene Network Coherence from [18] which is a valuable reference to consider as a novel measure to rate the coherence of the dependencies according to different biological databases or IntelliGO from [4]. We take into account it for further research in future work. Finally, we are working on building a Cytoscape App to offer a guided user interface software.

Acknowledgements We would like to thank Spanish Ministry of Science and Innovation for the financial support under project TIN2017-88209-C2-2-R.

References

1. GENIE3 vignette. https://doi.org/10.18129/B9.bioc.GENIE3. https://bioconductor.org/packages/release/bioc/vignettes/GENIE3/inst/doc/GENIE3.html
2. The gene ontology (go) database and informatics resource. Nucleic acids research 32(Database issue), D258–61 (2004). https://doi.org/10.1093/nar/gkh036. https://www.ncbi.nlm.nih.gov/pubmed/14681407
3. SCENIC: Single-cell regulatory network inference and clustering. Nature Methods (2017). https://doi.org/10.1038/nmeth.4463
4. Benabderrahmane S, Smail-Tabbone M, Poch O, Napoli A, Devignes MD (2010) Intelligo: a new vector-based semantic similarity measure including annotation origin. BMC bioinform 11(1):588
5. Benjamini Y, Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency Annals of Statistics. https://doi.org/10.1214/aos/1013699998
6. Berriz GF, King OD, Bryant B, Sander C, Roth FP (2003) Characterizing gene sets with FuncAssociate. Bioinformatics 19(18):2502–2504. https://doi.org/10.1093/bioinformatics/btg363
7. Bulashevska S, Eils R (2005) Inferring genetic regulatory logic from expression data. Bioinformatics (Oxford England) 21(11):2706–13. https://doi.org/10.1093/bioinformatics/bti388
8. Caniza H, Romero AE, Heron S, Yang H, Devoto A, Frasca M, Mesiti M, Valentini G, Paccanaro A (2014) GOssto: A stand-alone application and a web tool for calculating semantic similarities on the Gene Ontology Bioinformatics. https://doi.org/10.1093/bioinformatics/btu144
9. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular Cell 2(1):65–73. https://doi.org/10.1016/S1097-2765(00)80114-8. http://linkinghub.elsevier.com/retrieve/pii/S1097276500801148
10. Couto FM, Silva MJ, Coutinho PM (2005) Semantic similarity over the gene ontology: Family correlation and selecting disjunctive ancestors. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management, CIKM '05. ACM, New York, pp 343–344, https://doi.org/10.1145/1099554.1099658
11. Delgado FM, Gómez-Vela F (2018) Computational methods for gene regulatory networks reconstruction and analysis: A review Artificial intelligence in medicine. https://doi.org/10.1016/j.artmed.2018.10.006
12. Dwight SS, Harris MA, Dolinski K, Ball CA, Binkley G, Christie KR, Fisk DG, Issel-Tarver L, Schroeder M, Sherlock G, Sethuraman A, Weng S, Botstein D, Cherry JM (2002) Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go). Nucl Acids Res 30(1):69–72. https://doi.org/10.1093/nar/30.1.69. http://dblp.uni-trier.de/db/journals/nar/nar30.html#DwightHDBBCFISSSWBC03
13. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95 14863–14868. https://doi.org/10.1073/pnas.95.25.14863
14. EMBL-EBI: Introduction to embl-european bioinformatics institute. https://www.ebi.ac.uk/sites/ebi.ac.uk/files/content.ebi.ac.uk/documents/introduction_to_embl-ebi.pdf
15. Fitch A, Jones M (2009) Shortest path analysis using partial correlations for classifying gene functions from gene expression data. Bioinformatics 25:42–47. https://doi.org/10.1093/bioinformatics/btn574

16. Friedman N (2004) Inferring cellular networks using probabilistic graphical models. Science (New York) 303(5659):799–805. https://doi.org/10.1126/science.1094068. http://www.ncbi.nlm.nih.gov/pubmed/14764868
17. Gan M, Dou X, Jiang R (2013) From ontology to semantic similarity: calculation of ontology-based semantic similarity. Sci World J 2013
18. Gómez-Vela F, Lagares JA, Díaz-Díaz N (2015) Gene network coherence based on prior knowledge using direct and indirect relationships. Comput Biol Chem 56:142–151
19. Gutiérrez-Avilés D, Rubio-Escudero C, Martínez-Álvarez F, Riquelme JC (2014) Trigen: A genetic algorithm to mine triclusters in temporal gene expression data. Neurocomputing 132:42–53. https://doi.org/10.1016/j.neucom.2013.03.061
20. Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring regulatory networks from expression data using tree-based methods PLos ONE. https://doi.org/10.1371/journal.pone.0012776
21. Lee I, Date SV, Adai AT, Marcotte EM (2004) A probabilistic functional network of yeast genes. Science 1555–1558. https://doi.org/10.1126/science.1099511
22. Lee I, LZME (2007) An improved, bias-reduced probabilistic functional gene network of baker's yeast, saccharomyces cerevisiae. PLoS One e988. https://doi.org/10.1371/journal.pone.0000988
23. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics 7 Suppl 1, S7 https://doi.org/10.1186/1471-2105-7-S1-S7. http://www.ncbi.nlm.nih.gov/pubmed/16723010
24. Miron M (2018) Interpretability in AI and its relation to fairness, transparency, reliability and trust. Joint Research Center, EU Commission. https://ec.europa.eu/jrc/communities/en/node/1162/article/interpretability-ai-and-its-relation-fairness-transparency-reliability-and-trust
25. Markowetz F, Spang R Inferring cellular networks–a review. BMC bioinformatics 8 Suppl 6, S5 (2007). https://doi.org/10.1186/1471-2105-8-S6-S5. http://www.ncbi.nlm.nih.gov/pubmed/17903286
26. Martínez B, Isabel A, Nepomuceno C, José C, Riquelme M (2014) Discovering gene association networks by multi-objective evolutionary quantitative association rules. J Comput Syst Sci 80(1):118–136. https://doi.org/10.1016/j.jcss.2013.03.010
27. Mistry M, Pavlidis P (2008) Gene ontology term overlap as a measure of gene functional similarity. BMC Bioinform 9(1):327. https://doi.org/10.1186/1471-2105-9-327. http://www.biomedcentral.com/1471-2105/9/327
28. Nepomuceno JA, Lora AT, Aguilar-Ruiz JS (2011) Biclustering of gene expression data by correlation-based scatter search. BioData Mining 4:3. https://doi.org/10.1186/1756-0381-4-3
29. Nepomuceno JA, Troncoso A, Nepomuceno-Chamorro IA, Aguilar-Ruiz JS (2015) Integrating biological knowledge based on functional annotations for biclustering of gene expression data. Comput Methods Prog Biomed 119(3):163–180. https://doi.org/10.1016/j.cmpb.2015.02.010
30. Nepomuceno JA, Troncoso A, Nepomuceno-Chamorro IA, Aguilar-Ruiz JS (2018) Pairwise gene go-based measures for biclustering of high-dimensional expression data. BioData mining 11(1):4
31. Nepomuceno-Chamorro I, Aguilar-Ruiz J, Riquelme J (2010) Inferring gene regression networks with model trees. BMC Bioinformatics 11(1):517. https://doi.org/10.1186/1471-2105-11-517. http://www.biomedcentral.com/1471-2105/11/517
32. Nepomuceno-Chamorro IA, Jesús S, Aguilar R (2013) Synergies of genes in alzheimer's disease. In: International Work-Conference on Bioinformatics and Biomedical Engineering, IWBBIO 2013, Granada, Spain, March 18-20, 2013. Proceedings. http://iwbbio.ugr.es/papers/iwbbio_008.pdf, pp 51–53
33. Nepomuceno-Chamorro IA, Márquez C, Jesús S, Aguilar-Ruiz AE (2015) Building transcriptional association networks in cytoscape with regnetc. IEEE/ACM Trans Comput Biology Bioinform 12(4):823–824. https://doi.org/10.1109/TCBB.2014.2385702
34. Pesquita C, Faria D, Bastos H, Ferreira A, Falcao A, Couto F (2008) Metrics for go based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 9(Suppl 5):S4. https://doi.org/10.1186/1471-2105-9-S5-S4. http://www.biomedcentral.com/1471-2105/9/S5/S4
35. Pesquita C, Faria D, Falcão AO, Lord P, Couto FM (2009) Semantic similarity in biomedical ontologies. PLoS Comput Biol 5(7):12. https://doi.org/10.1371/journal.pcbi.1000443. http://www.ncbi.nlm.nih.gov/pubmed/19649320
36. Ponzoni I, Azuaje F, Augusto J, Glass D Inferring adaptive regulation thresholds and association rules from gene expression data through combinatorial optimization learning. https://doi.org/10.1109/tcbb.2007.1049. http://www.ncbi.nlm.nih.gov/pubmed/17975273
37. Quinlan JR (1993) C4.5: Programs for machine learning
38. Rodius S, Nazarov P, Nepomuceno-Chamorro I, Jeanty C, Gonzalez-Rosa J, Ibberson M, da Costa RM, Xenarios I, Mercader N, Azuaje F (2014) Transcriptional response to cardiac injury in the zebrafish: systematic identification of genes with highly concordant activity across in vivo models. BMC Genomics 15(1):852. https://doi.org/10.1186/1471-2164-15-852. http://www.biomedcentral.com/1471-2164/15/852
39. Romero-Zaliz RC, Rubio-Escudero C, Cobb JP, Herrera F, Cordón O, Zwir I (2008) A multiobjective evolutionary conceptual clustering methodology for gene annotation within structural databases: a case of study on the gene ontology database. IEEE Trans Evol Comput 12(6):679–701. https://doi.org/10.1109/TEVC.2008.915995
40. Segal E, SMRA, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genet 34:166–176. https://doi.org/10.1038/ng1165
41. Soinov LA, Krestyaninova MA, Brazma A (2003) Towards reconstruction of gene networks from expression data by supervised learning Genome biology. https://doi.org/10.1186/gb-2003-4-1-r6
42. Spellman P, Sherlock G, Zhang M et al (1998) Comprehensive identification of cell cycle–regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9(12):3273–3297. https://doi.org/10.1091/mbc.9.12.3273
43. Steele E, Tucker A, 'T Hoen PAC, Schuemie MJ (2009) Literature-based priors for gene regulatory networks. Bioinformatics (Oxford, England) 25(14):1768–74. https://doi.org/10.1093/bioinformatics/btp277
44. Wang Y, Yang S, Zhao J, Du W, Liang Y, Wang C, Zhou F, Tian Y, Ma Q (2019) Using machine learning to measure relatedness between genes: a multi-features model. Scientific reports 9(1):1–15
45. Wang YR, Huang H (2014) Review on statistical methods for gene network reconstruction using expression data. J Theoret Biol 362:53–61. https://doi.org/10.1016/j.jtbi.2014.03.040
46. Witten IH, Frank E, Trigg L, Hall M, Holmes G, Cunningham SJ (1999) Weka: Practicalmachine learning tools and techniques with java implementations