Discovering gene association networks by multi-objective evolutionary quantitative association rules

M. Martínez-Ballesteros, I.A. Nepomuceno-Chamorro, J.C. Riquelme

ABSTRACT

In the last decade, the interest in microarray technology has exponentially increased due to its ability to monitor the expression of thousands of genes simultaneously. The reconstruction of gene association networks from gene expression profiles is a relevant task and several statistical techniques have been proposed to build them. The problem lies in the process to discover which genes are more relevant and to identify the direct regulatory relationships among them. We developed a multi-objective evolutionary algorithm for mining quantitative association rules to deal with this problem. We applied our methodology named GarNet to a well-known microarray data of yeast cell cycle. The performance analysis of GarNet was organized in three steps similarly to the study performed by Gallo et al. GarNet outperformed the benchmark methods in most cases in terms of quality metrics of the networks, such as accuracy and precision, which were measured using YeastNet database as true network. Furthermore, the results were consistent with previous biological knowledge.

Keywords:
Data mining
Multi-objective evolutionary algorithms
Quantitative association rules
Gene networks
Microarray analysis

1.Introduction

Since late 1990s, the interest in microarray technology has exponentially increased due to its ability to monitor the expression of thousands of genes simultaneously. Microarray technology has revolutionized the biological research because it allows to study thousand of genes or even whole genomes [1].

As molecular biology is rapidly evolving into a quantitative science, it increasingly relies on computational algorithms to make sense of high-throughput data. One of the main goals in Microarray analysis is the reconstruction of gene regulatory processes and a key task is the inference of regulatory interactions among genes from gene expression data [2]. Our aim is to infer the relationships between genes from an organism in a particular biological process. This relationships can be modeled in several levels of abstraction, these levels range from the detailed gene regulatory processes (where a chain of intracellular reaction activates a regulatory molecule, transcription factors until a protein is synthesized) to the high models of abstraction named gene association networks. In the reconstruction of gene regulatory processes, building gene association networks has been proven to provide useful insights for such task, the reconstruction of gene regulatory processes. A gene association network can be defined as a graph in which nodes represent genes and edges represent the influence between them. Our goal in this work is the inference of gene association networks from Microarray datasets.

There are several statistical methods to infer gene association networks from Microarray data. A microarray dataset is a bidimensional data structure where conditions are experiments or sources and the columns are gene expression values. In our problem the conditions will be the instances and the gene expression values will be the attributes or features. These methods range from relatively straightforward correlation-based methods to more sophisticated methods based on the concept of conditional independence. In general, these methods based on pairwise similarity measures are very useful to determine whether two genes have a strong global similarity under all conditions in the dataset. However, there could be strong local similarities over a subset of conditions, which could not be detected by them [3]. In this context, the discovery of Association Rules (AR), and particularly of Quantitative Association Rules (QAR), is a popular methodology that allows the discovery of significant and apparently hidden relations among attributes in a subspace of the instances from the dataset. Therefore, we developed a multi-objective evolutionary algorithm for mining QAR to favor the detection of localized similarities over a more global similarity. Furthermore, as it can be observed in the review of the state-of-the-art [4], methods based on the discovery of QAR have not been used to infer gene associations from microarray data. However, qualitative AR have been used to infer gene association networks but this approach needs a discretization step that our proposal avoids.

Our proposal, henceforth named GarNet (Gene–gene associations from Association Rules for inferring gene NETworks), is based on the well-known multi-objective evolutionary approach NSGA-II [5] to discover QAR with adaptive intervals without performing a previous discretization. NSGA-II algorithm has been selected instead of SPEA-II [6] algorithm because it performs better than SPEA-II due to the powerful crowding operator that keeps diversity in the population and generates a more uniform Pareto front. Furthermore, NSGA-II is considered as the paradigm within the MOEA research community [7]. GarNet carries out an inference process based on an iterative rule learning to extract gene–gene associations and builds gene networks by the intersection of the gene–gene associations retrieved from the QAR found in several input microarray datasets. To summarize, our proposal presents mainly two improvements: it favors the detection of localized similarities and avoids the discretization step of other approaches based on the discovery of AR [8].

In this work, we focus on the analysis of a set of genes that encode proteins important for cell-cycle regulation. We applied GarNet to a well-known microarray data of yeast cell cycle and we compared our approach against several benchmark methods focused on the same biological problem. For performance analysis we applied as benchmark methods a decision-tree-based method [9], a regression-tree based method [3], a probabilistic graphical model [10] and combinatorial optimization algorithm [11,12]. The performance analysis was organized in three steps similarly to the study performed by Gallo et al. [12]. GarNet outperformed the benchmark methods in most cases in terms of quality metrics of the networks, such as accuracy and precision, which were measured using YeastNet database as a true network. Furthermore, the results were consistent with previous biological knowledge.

The remainder of the paper is organized as follows. In Section 2, a summary of the benchmark methods to infer gene networks and to extract AR is presented. In Section 3, a detailed explanation of the methodology and the algorithm are presented. Section 4 reports the performance analysis, parameter settings and comparison analysis together with the biological relevance of the experiments. Finally, Section 5 summarizes the most relevant conclusions and future works.

2. Related work

The related work is divided into two parts: the first one describes the methods to infer gene networks from microarray data in the literature and the second one describes data mining techniques to build AR.

2.1. Inferring gene networks: a review

There are several methods to infer gene–gene association networks from gene expression data. These methods range from rather straightforward correlation-based methods to more sophisticated models, such as Bayesian network models.

One of the first approaches to the problem was clustering algorithms [13,14]. These approaches are based on a simple assumption, which is still used in functional genomic, called the guilt-by association heuristic. This assumption suppose that co-expression means co-regulation, i.e. if two genes show similar expression profiles, they are supposed to follow the same regulatory programme.

In order to formalize the idea of similar expression behavior, several statistical measures have been proposed in the literature. In correlation-based methods, gene–gene associations are built using correlation as a pairwise similarity measure between gene expression profiles over all the conditions in the dataset. In standard correlation-based methods, the Pearson or Spellman's coefficient has been used to identify gene–gene associations [15]. In this kind of methods, if the correlation between gene pairs is higher than a threshold value (usually 0.95), then it is assumed that these gene pairs interact directly in a relevant biological process or in a signaling pathway [16,17]. As shown in [18], the results provided by these methods are a framework for assigning biological functions to group of genes. In the literature, gene co-expression networks are also known as gene relevance, gene association or gene interaction networks. Different versions of the standard correlation-based method exist, such as one by Obayashi and Kinoshita in [19] that instead of correlation values uses correlation ranks.

Correlation-based methods are very useful to determine strong global similarity between two genes over all conditions in the dataset. This is a relevant constraint due to there might exist a strong local similarity over a subset of conditions which could not be detected with correlation measure. This constraint is taken into account in the model tree-based method proposed by Nepomuceno-Chamorro et al. [3], where they used regression trees as a way to detect linear dependencies localized over a subset of conditions. Similar to this approach, another rule-based method is presented in [9] in which the authors used decision trees as a way to extract dependencies. Inspired in these two techniques, the model tree-based method and the rule-based method, this work proposed a method to favor the detection of localized similarities over a more global similarity. We proposed a multi-objective evolutionary algorithm to extract AR which describe associations between genes.

Correlation measure is extensively used as an indicator of association measure between two random attributes, genes in our case, but cannot be seen as a causal measure between them. However, correlation is still informative about the underlying structure in the network [20], although correlation cannot provide the regulatory mechanisms. In order to explain the regulatory mechanism we can use full conditional models (also known Markow networks), where the correlation between two genes is explained by means of all other genes. As a drawback of these methods, this kind of relationships between genes can only be calculated if the number of samples is larger than the number of attributes, i.e. if the number of genes exceeded the number of distinct gene expression profiles. Full conditional models become especially simple in a Gaussian setting [21], therefore Gaussian graphical models (GGM) are a popular tool to infer gene association networks. In the GGM instead of full conditional models, first-order dependencies are extracted. In [20], authors use Partial Pearson's correlation to extract associations between pairs of genes when this association cannot be explained by means of a third gene. In [22], authors use conditional mutual information to test for first-order dependencies and they presented the algorithm named ARACNE that was applied to expression profiles of human B cells. Finally, the probabilistic model called Bayesian networks extracted dependencies between genes if no subset of the other genes can explain the correlation. In the literature, one of the most significant works in Bayesian methods is [10].

2.2. *Mining association rules: a review*

In data mining, learning a network structure from data became a main focus of attention in the last two decades [23]. In this context, we present the result of applying a data mining technique, specifically, AR, to infer gene association networks from microarray data. An AR stands for existing relationships among the attributes, genes in our case.

In the field of data mining, the learning of AR is a popular and well-known method for discovering interesting relations among variables in large databases [24–26]. The discovery of AR is a non-supervised learning tool since AR is descriptive, unlike classification. Descriptive mining tasks identify patterns that explain or summarize the data, that means, they are used to explore the properties of the data, instead of predicting the class of new data [27].

The classical example of AR is the well-known *market basket analysis*, where the purchase behavior of customers is analyzed in order to discover regularities among products purchased in a supermarket [28]. AR are widely used in many other fields such as the healthcare environment to identify risk factors in the onset or complications of diseases [29–31]. In addition, AR have been applied in visualization area where interactive visual analysis approaches have been developed to represent them [32].

This form of knowledge extraction is based on statistical techniques such as correlation analysis and variance. One of the most used and cited algorithms is Apriori [24]. When the domain is continuous, the AR are known as QAR. In this context, let $F = \{F_1, \ldots, F_n\}$ be a set of features or attributes, with values in $\mathbb{R}$. Let A and C be two disjoint subsets of F, that is, $A \subset F$, $C \subset F$, and $A \cap C = \emptyset$. A QAR is a rule $X \Rightarrow Y$, in which features in A belong to the antecedent X, and features in B belong to the consequent Y, such that X and Y are formed by a conjunction of multiple boolean expressions of the form $F_i \in [v_1, v_2]$ (with $v_1, v_2 \in \mathbb{R}$). Thus, in a QAR, the features or attributes of the antecedent are related with the features of the consequent, establishing an interval of membership values for each attribute involved in the rule. The consequent Y is usually a single expression.

For instance, a QAR could be numerically expressed as

$$\mathrm{CLB1} \in [-0.68, 0.05] \wedge \mathrm{CLN2} \in [-0.85, 0.1] \Rightarrow \mathrm{CLB5} \in [-1.11, 0]$$

where CLB1 and CLN2 constitute the features appearing in the antecedent and CLB5 the one in the consequent.

It is important to measure the quality of the rules to select the best rules and evaluate the results obtained by the proposed algorithm. There are several probability-based measures proposed in the literature to evaluate the generality and reliability of AR obtained in the mining process [33,34]. However, the real value of a rule, in terms of usefulness and actionability is subjective and depends heavily on the particular domain and objectives of the problem at hand.

A review of the published literature for mining rules reveals that there exist many algorithms to find them, most of them based on classical methods such as AIS [28], Apriori [24] or SETM [35]. However, many of these tools discretize the attributes by using a specific strategy and deal with them as if they were discrete [36], obtaining poor results in real continuous datasets. Afterwards, some researchers were focused on discover AR from continuous datasets as [37]. To deal with continuous datasets, [38] presented a classifier with the aim of extracting QAR over unlabeled data streams. The main novelty of this approach lied on its adaptability to on-line gathered data. In [39], an optimization metaheuristic based on rough particle swarm techniques was presented. In this case, the singularity was that this technique determines the intervals for the AR. Finally, an interval clustering-based method is presented in [40] and it is especially useful to mining complex information.

Many algorithms are based on evolutionary algorithms (EA) [41] which have been extensively used for the optimization and adjustment of models in data mining tasks. EA are search algorithms that generate solutions for optimization problems using techniques inspired by natural evolution. Evolutionary computation is usually used to discover AR in both EA [42,43] and Genetic Programming [44,45] due to they offer a set of advantages for knowledge extraction and specifically for rule induction processes [46].

In the last two decades an increasing interest has been developed in the use of EA for multi-objective optimization [47]. The mining process of AR can be modeled as a multi-objective problem in which the measures used for evaluating a rule are different objectives to be maximized [48]. Nevertheless, the majority of the proposed algorithms for mining AR considered as multi-objective approaches are based on a combination of weighted objectives in a single fitness function. Therefore, some problems might be presented since a weighted scheme could be solved by a single-objective optimization. Hence, methods based on concept of Pareto optimal such as SPEA-II [6] (*Strength Pareto Evolutionary Algorithm*) or NSGA-II [5] (*Non-Dominated Sorting Genetic Algorithm*) are better to optimize the AR extraction process because weights and previous information on the problem at hand are not required. Besides the above advantages, these methods are based on populations of non-dominated solutions.

A multi-objective Pareto-based EA was presented in [49] where the fitness function was formed by four different objectives: support, confidence, comprehensibility of the rule (aimed to be maximized) and the amplitude of the intervals that forms the rule (intended to be minimized). Other multi-objective EA to mining AR is proposed in [50]. In this case, the algorithm doesn't take the minimum support and confidence into account and apply the FP-tree algorithm. The objective of fitness function is maximizing the correlation between support and confidence.

The main motivation of this paper is to extend preliminary works such as the proposed algorithms called QARGA [51, 52] and EQAR [53], to a multi-objective approach based on the NSGA-II algorithm able to deal with biological problems. In particular, a non-dominated multi-objective evolutionary algorithm is proposed in this work which is able to find QAR in databases with continuous attributes avoiding the discretization step. Likewise, the proposed algorithm has also been enhanced to include a mechanism for building networks from AR, henceforth called GarNet (Gene–gene associations from Association Rules for inferring gene NETworks).

3. GarNet: gene–gene associations from Association Rules for inferring gene NETworks

The main features of the proposed algorithm are described in this section. First, the method is presented in Section 3.1 as a multi-objective approach named GarNet that extends previous proposals (QARGA and EQAR) into a multi-objective approach, with the aim to discover QAR from continuous datasets. Afterward, the adaptation to deal with a biological problem and the process to build gene association networks from the rules is explained in detail in Section 3.2.

3.1. Evolutionary process of GarNet

As we mentioned before, we proposed in this work a multi-objective evolutionary algorithm to find QAR in continuous datasets avoiding the discretization step. In continuous domains, rules identify subgroups of samples whose features share certain sets of values. Therefore, it is required to express the samples covered by a rule using the range of values allowed, then adaptive intervals instead of fixed ranges are chosen to represent them. The search for the most appropriate intervals has been carried out by means of the GarNet algorithm. Thus, the intervals are adjusted to find QAR with high interpretability, generality, quality and precision.

The proposed algorithm extends the main features of QARGA [51,52] and EQAR [53] adding new features to improve the AR mining task. The most important improvement achieved in GarNet is related with solving the main drawbacks caused by the weighted objective scheme existing in the fitness function.

A multi-objective approach is the best way to perform the best trade-off among all the measures, hence our proposal GarNet is based on the well-known NSGA-II algorithm [5]. Its main purpose is to evolve the population based on the non-dominated sort of the solutions in fronts of dominance. The first front is composed of the non-dominated solutions of the population (the Pareto front), the second is composed of the solutions dominated by one solution, the third of solutions dominated by two, and so on. The evolutionary scheme of the proposed algorithm is described in Fig. 1.

In the population, each individual constitutes a rule. These rules are then subjected to an evolutionary process, in which the mutation and crossover operators are applied. IRL process (Iterative Rule Learning) [55] is performed to penalize instances already covered by rules found by GarNet, in order to emphasize the covering of instances still not covered. The IRL affects the generation of initial population in each evolutionary process which is described in Section 3.1.2. The evolutionary process ends when the number of generations is reached. Thenceforth, the algorithm returns the rule that belongs in the first Pareto front (PF_1) with higher support value (Eq. (3)). The whole evolutionary process is repeated until the desired number of rules is achieved.

The main parts of the evolutionary process of GarNet are defined in the following subsections.

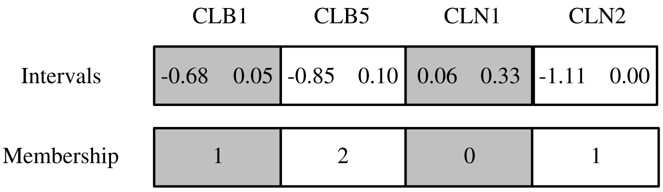
3.1.1. Individuals codification

The lower and upper limits of the intervals of each attribute will be represented by the different genes of an individual. Because the attributes are continuous, individuals are represented by a real coding. An individual consists of a fixed number of attributes n, which represents the number of attribute in the database. The representation of an individual consists in two data structures. The upper structure includes the intervals bounds of all the attributes of the dataset. The bottom structure indicates the membership of an attribute to the rule represented by an individual. The type of each attribute can have three values: 0 when the attribute does not belong to the rule and 1 or 2 if it belongs to the antecedent or the consequent of the

```
Inputs: Maximum number of rules (MaxNumRules), Maximum number of generations (MaxNumGen)
Output: MaxNumRules best rules found

Multi-objective Evolutionary Process(MaxNumRules, MaxNumGen)
Initialize the rule counter $r = 0$.
Repeat // The IRL loop starts here
  1. Initialize the generation counter $t = 0$.
  2. Initialize parent population $P_{t=0}$ based on instances covered by fewer rules.
  3. Evaluate the individuals of $P_{t=0}$ based on the measures selected as objectives.
  4. $P_{t=0}$ is ranked using the Fast Non-dominated Sort [5] that consists in sorting the individuals of a population in different Pareto fronts ($PF$)
     according to their non-dominance.
     Repeat // Generations of EA loop
       (a) An offspring population $Q_t$ of same size as $P_t$ is generated using crossover and mutation operators over the individuals of $P_t$ selected using
           binary Tournament selection-based method [54].
       (b) The individuals of $P_t$ and $Q_t$ are merged into $R_t$ and the Fast Non-dominated Sort is carried out.
       (c) The next population $P_{t+1}$ consists of the $N$ best individuals of $R_t$.
           Initialize the front counter $i = 0$.
           Repeat // Loop to generate the next population $P_{t+1}$
             If the current level of $R_t$ ($PF_i$, $i$th Pareto front) has less than or equal to $N$ individuals, the individuals of $PF_i$ are added to the population $P_{t+1}$.
             In other case,
               if the current level of $R_t$ ($PF_i$, $i$th Pareto front) has more than $N$ individuals, the best individuals are used to fill the population of
               the next generation ($P_{t+1}$), and for that purpose, the Crowding distance assignment [5] is used in order to sort the population of the
               current level and to select the best individuals that represent the best rules.
             Increment the front counter ($i = i + 1$).
           While the next population $P_{t+1}$ is not completed.
       (d) Increment the generation counter ($t = t + 1$).
     While the maximum number of generations is not reached.
  5. Return best individual, thus, the rule in the first Pareto front ($PF_1$) which reaches a higher support value.
  6. Penalize the instances covered by the best rule found.
  7. Increment the rule counter ($r = r + 1$).
While the number of desired rules is not achieved.
Return the best rules found.
```

Fig. 1. General scheme of the algorithm.



	CLB1	CLB5	CLN1	CLN2
Intervals	-0.68 0.05	-0.85 0.10	0.06 0.33	-1.11 0.00
Membership	1	2	0	1

Fig. 2. Example of an individual of the population.

rule, respectively. If an attribute is wanted to be retrieved for a specific rule, it can be done by modifying the value equal to 0 of the type by a value equal to 1 or 2 depending on the antecedent or consequent.

An illustrative example of an individual codification is depicted in Fig. 2. We suppose that the input dataset has 4 attributes. In particular, the rule $\text{CLB1} \in [-0.68, 0.05] \wedge \text{CLN2} \in [-0.85, 0.1] \Rightarrow \text{CLB5} \in [-1.11, 0]$ is represented. Note that attributes CLB1 and CLN2 appear in the antecedent, CLB5 in the consequent and CLN1 is not involved in the rule.

3.1.2. Initial population

The generation of the initial population in the proposed algorithm was carried out at the beginning of each evolutionary process and is performed such that at least one chosen sample or instance of the dataset was covered. The samples of the dataset are selected based on their level of hierarchy. The hierarchy is organized according to the number of rules which cover a sample. Thus, the records are sorted by the number of rules that are covered and the samples covered by a few rules have a higher priority.

A sample is selected according to the inverse of the number of rules which cover such sample. Intuitively, the process is similar to roulette selection method where the parents are selected depending on their fitness.

Thus, the samples covered by a few rules have a greater portion of roulette and, therefore, they will be more likely selected. In the first evolutionary process, all samples have the same probability to be selected. Constraints to generate individuals are given by the number of attributes that belong to rule represented by an individual, the maximum and minimum number of attributes in the antecedents and consequents and the structure of the rule (attributes fixed or not fixed in consequent).

3.1.3. Genetic operators

The genetic operators implemented in our proposal are Crossover and Mutation and they are described in [51]. The Crossover operator is applied over all individuals generated as offspring. After applying the crossover operator, these individuals are mutated with a given probability. In addition, a new directional Mutation operator has been added in order to introduce diversity in the population. This mutation operator works as follows: If the selected attribute belongs to the antecedent of the rule, i.e. its type is 1, then the type of this attribute is changed to belong to the consequent of the rule and now the type is 2. On the other hand, if the selected attribute belongs to the consequent, i.e. its type is 2, then the type of this attribute is changed to the antecedent and now the type is 1.

3.2. Adapting the AR mining process for modeling biological problems

One of the motivations of this work has been to extend previous algorithms such as QARGA and EQAR into a multi-objective approach based on Pareto optimal as discussed before. Nevertheless, the main challenge proposed in this work is how to adapt these evolutionary algorithms to discover QAR to deal with a biological problem: the inference of gene–gene associations from gene expression profiles provided by microarray technology. Besides the evolutionary features described above in Section 3.1, new one has been added in order to improve the performance of the obtained QAR for this specific biological problem. Furthermore, a mechanism to build gene–gene associations from QAR has been included. These new functionalities are detailed in the following subsections.

3.2.1. Building gene networks from gene–gene associations

GarNet is a rule-based method to infer gene networks from microarray datasets. The general scheme is outlined in Fig. 3. The proposed GarNet carry out an inference process to build a gene network from the QAR obtained in each microarray dataset.

The best QAR are obtained by the evolutionary process described in Section 3.1 and the best individual of each iteration is the rule with higher support value from the first Pareto front of the last generation. In addition, as can be seen in Fig. 3 and described in Fig. 1, an IRL process until reaching the maximum number of rules, is applied to lead the search process towards rules satisfying instances which have not yet been covered or are covered by few rules.

In general terms, the outer loop of GarNet depicted in Fig. 3 is based on others rule-based approaches to infer gene network. GarNet applies an inference process from K microarray datasets. This process is performed in [12] to evaluate automatically the gene–gene associations obtained by different datasets, therefore improving the degree of evidence required for the potential associations to be returned.

The gene–gene associations of the gene network are inferred from the obtained QAR by GarNet for each microarray dataset. The set of best QAR obtained by each microarray is decomposed into sets of attribute pairs as follows:

- First, the attributes belonging to the antecedent and attributes belonging to the consequent for each resulting QAR are identified.
- Afterwards, combinations between the attributes of the antecedent and attributes of the consequent of each rule are performed obtaining pairs of attributes.
- For instance, let the following QAR be:

  $$\mathrm{CLB1} \in [-0.68, 0.05] \wedge \mathrm{CLN2} \in [-0.85, 0.1] \Rightarrow \mathrm{CLB5} \in [-1.11, 0]$$

  The resulting attribute pairs (associations) that can be extracted from this rule are:

  $$\mathrm{CLB1} \Rightarrow \mathrm{CLB5}$$

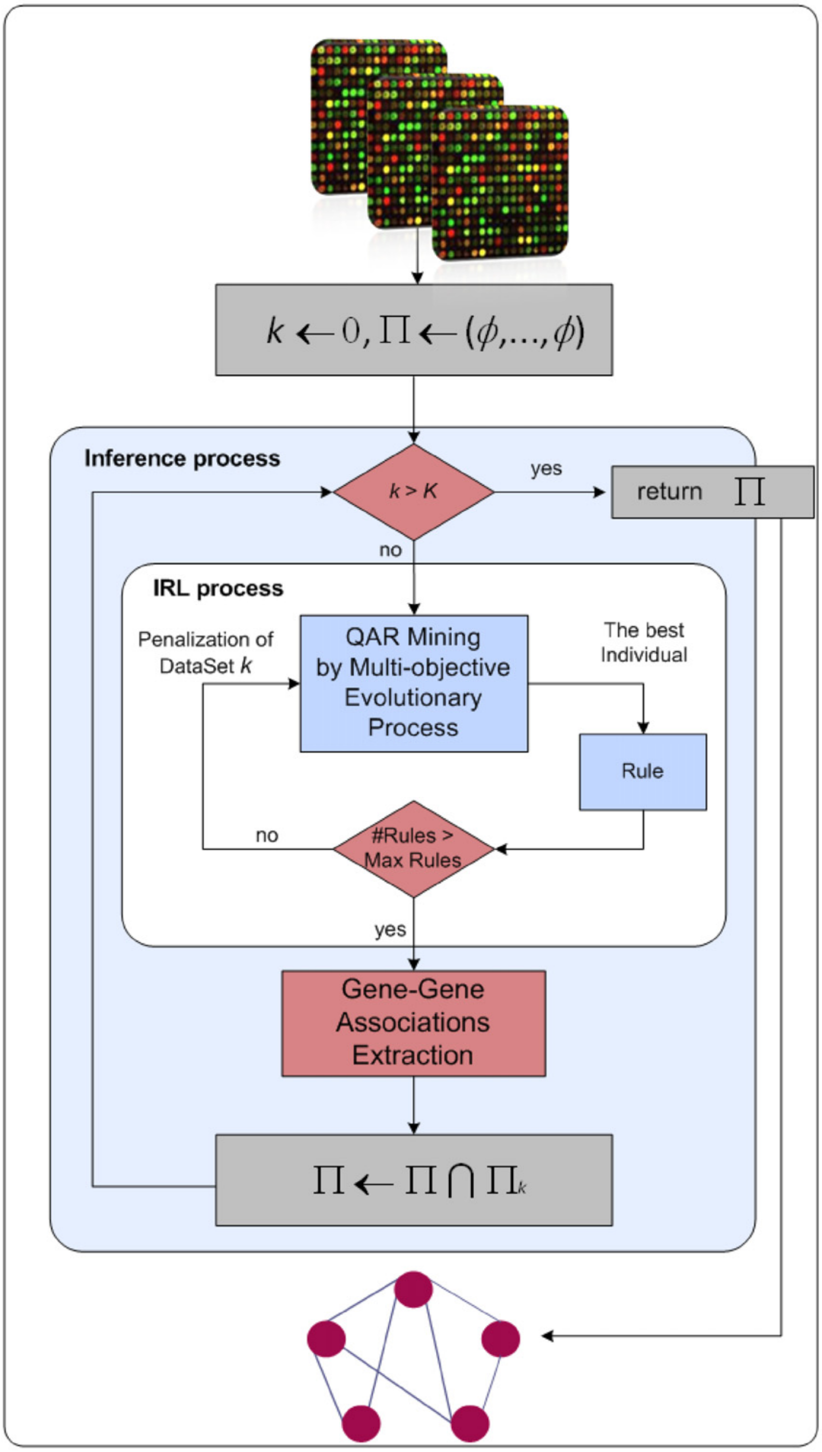
  $$\mathrm{CLN2} \Rightarrow \mathrm{CLB5}$$

After completing the inference process from datasets, the intersection among the attribute pairs found for each input dataset is performed to find the most frequent gene–gene associations, hence, potential and relevant associations. Let K be the number of input datasets used in the inference process. Let Π be the set of attribute pairs obtained from kth-dataset. The output of the inference process is defined as:

$$\Pi = \Pi_1 \cap \Pi_2 \cap \cdots \cap \Pi_K \tag{1}$$

where Π_k, $k = 1..K$, is the set of attribute pairs obtained from the kth-dataset.

The final step involves the inference of the resulting gene network from the intersection of the results obtained for the K datasets. Finally, a gene network can be defined as a graph where nodes are attributes and edges link pairs of attributes extracted from rules. An example of how the decomposition process of a set of QAR into attribute pairs is carried out to build the gene network is defined as follows.



$k \leftarrow 0, \prod \leftarrow (\phi,\ldots,\phi)$
Inference process
$k > K$
yes
return $\prod$
no
IRL process
Penalization of DataSet k
QAR Mining by Multi-objective Evolutionary Process
The best Individual
Rule
no
#Rules > Max Rules
yes
Gene-Gene Associations Extraction
$\prod \leftarrow \prod \bigcap \prod_k$

Fig. 3. General scheme of GarNet algorithm.

Let the following set of QAR obtained for 2 datasets be:

1. **Gene–gene associations extraction** (See Fig. 3.)
   - **Dataset 1:**
     - $\text{CLB1} \in [-0.68, 0.05] \wedge \text{CLN2} \in [-0.85, 0.1] \Rightarrow \text{CLB5} \in [-1.11, 0]$.
     - $\text{CLN1} \in [0.1, 0.45] \Rightarrow \text{CLB5} \in [0.05, 0.5]$.

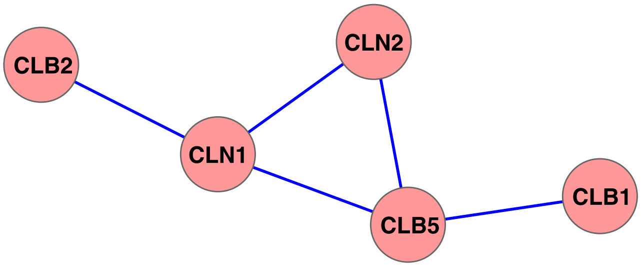


Fig. 4. Example of gene network from the attribute pairs reported by the inference process.

  - $CLB1 \in [-0.68, 0.05] \Rightarrow SWI5 \in [-0.5, -0.01]$.
  - $CLN1 \in [0.02, 0.41] \Rightarrow CLB2 \in [0.1, 0.56] \wedge CLN2 \in [0.03, 0.4]$.

  $\Pi_1 = \{CLB1 - CLB5, CLN2 - CLB5, CLN1 - CLB5, CLB1 - SWI5, CLN1 - CLB2, CLN1 - CLN2\}$.
- **Dataset 2:**
  - $CLB2 \in [0.05, 0.5] \Rightarrow CLB6 \in [0.1, 0.62] \wedge CLN1 \in [0.06, 0.33]$.
  - $CLN1 \in [-0.43, 0.03] \Rightarrow CLN2 \in [-0.85, 0.1]$.
  - $CLN1 \in [0, 0.43] \Rightarrow CLB5 \in [0.02, 0.45]$.
  - $CLB1 \in [0.1, 0.45] \Rightarrow CLB5 \in [0, 0.51]$.
  - $CLB5 \in [-1.11, 0] \Rightarrow CLN2 \in [-0.85, 0.1]$.

  $\Pi_2 = \{CLB2 - CLB6, CLB2 - CLN1, CLN1 - CLN2, CLN1 - CLB5, CLB1 - CLB5, CLB5 - CLN2\}$.

► **Output** $\Pi = \Pi_1 \cap \Pi_2 = \{CLB1 - CLB5, CLB5 - CLN2, CLN1 - CLB5, CLN1 - CLB2, CLN1 - CLN2\}$.

2. **Gene network inference**
   - Graph nodes are: CLB1, CLB5, CLN2, CLB2, CLN1, CLN2.
   - Graph edges from attribute pairs are:
     - CLB1 ⇒ CLB5.
     - CLB5 ⇒ CLN2.
     - CLN1 ⇒ CLB5.
     - CLN1 ⇒ CLB2.
     - CLN1 ⇒ CLN2.
   - The resulting Gene network is shown in Fig. 4.

In an ideal scenario for potential interactions between pairs of attributes, all the combinations between the attributes belonging to the antecedent and the attributes belonging to the consequent are strongly correlated, and therefore, the gene–gene associations derived from the rules present a high statistical dependence. However, in a real scenario all the attributes of the rules might be correlated but when the QAR are split in pair of attributes, they could be present low statistical dependence or be independents. To avoid this drawback, comprehensible and general QAR with high precision and accuracy are desirables, thus, GarNet will try to find QAR with the minimum number of attributes possible, covering a high number of attributes and presenting a high quality. To this aim, GarNet optimizes several measures that are shown in the next section.

3.2.2. QAR quality measures used by GarNet

Probability-based measures [33] have been selected as objectives to be optimized with the aim of selecting the best rules for the biological problem to deal with in this work. Specifically, the support of the rule and accuracy measure are selected to be optimized, respectively, to obtain general and reliable rules. In addition to accuracy and support measures, confidence measure has been considered as a threshold to filter the set of resulting rules.

The description and the mathematical definition of these measures are described as follows:

- *Support*(X): The support of an itemset X is defined as the ratio of instances in the dataset that satisfy X. Usually, the support of X is named as the probability of X.

$$sup(X) = P(X) = \frac{n(X)}{N} \tag{2}$$

  where $n(X)$ is the number of occurrences of the itemset X in the dataset, and N is the number of instances forming such dataset.
- *Support*$(X \Rightarrow Y)$: The support of the rule $X \Rightarrow Y$ is the percentage of instances in the dataset that satisfy X and Y simultaneously.

$$sup(X \Rightarrow Y) = P(Y \cap X) = \frac{n(XY)}{N} \tag{3}$$

where $n(XY)$ is the number of instances that satisfy the conditions for the antecedent X and consequent Y simultaneously.

- *Accuracy*$(X \Rightarrow Y)$: Accuracy measures the degree of veracity of the rules, i.e., the matching degree between the obtained values and the actual data. An accuracy of 100% means that the measured values are exactly the same as the real values. In the AR mining context, the accuracy measures the total percentage of instances in the dataset that satisfy the antecedent and the consequent, and the total percentage of instances in the dataset that do not satisfy neither the antecedent nor the consequent. Accuracy takes values in [0, 1] and values near 1 are expected for a high quality and high veracity rule.

$$acc(X \Rightarrow Y) = sup(X \Rightarrow Y) + sup(\neg X \Rightarrow \neg Y) \tag{4}$$

where $\neg$ means negation, therefore $sup(\neg X \Rightarrow \neg Y)$ is the percentage of instances in the dataset that do not satisfy X and Y simultaneously.

- *Confidence*$(X \Rightarrow Y)$: The confidence is the probability that instances satisfying X, also satisfy Y. In other words, it is the support of the rule divided by the support of the antecedent.

$$conf(X \Rightarrow Y) = P(Y \mid X) = \frac{sup(X \Rightarrow Y)}{sup(X)} \tag{5}$$

3.2.3. *Performance assessment of gene networks*

Several well-known measures, such as accuracy, precision, sensitivity and specificity are usually used to evaluate the quality of the networks obtained by a gene–gene associations-based method. Definitions of these measures are described as follows:

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

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \tag{6}$$

where:
  - *TP* is the number of gene–gene associations obtained by the method that also appear in the gene networks used as test.
  - *TN* is the number of gene–gene associations not obtained by the method that do not appear in the gene networks used as test.
  - *FP* is the number of gene–gene associations obtained by the method that also do not appear in the gene networks used as test.
  - *FN* is the number of gene–gene associations not obtained by the method that appear in the gene networks used as test.

- **Definition 2.** *Network Precision*: The precision of a gene network is defined as the proportion of the true positives against all the positive results (both true positives and false positives)

$$Precision = \frac{TP}{TP + FP} \tag{7}$$

- **Definition 3.** *Network Sensitivity*: The sensitivity of a gene network measures the proportion of true positives which are correctly identified.

$$Sensitivity = \frac{TP}{TP + FN} \tag{8}$$

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

$$Specificity = \frac{TN}{TN + FP} \tag{9}$$

- **Definition 5.** *Network Score*: Different databases for the yeast organism include a score value for each gene–gene association that measures the probability of an interaction representing a true functional linkage between two genes (with stronger associations scoring higher) [56]. In gene networks, the score is measured as the average score values of the rules found by the method assessed.

- **Definition 6.** *Number of associations*: Number of associations of a gene network is the number of edges of the graph which models the gene network, that is the number of gene–gene interactions of the gene network.

Table 1
List of genes analyzed in the dataset.

ORF name	Common name	Description
YMR199W	CLN1	Cyclin, G1/S-specific
YPL256C	CLN2	Cyclin, G1/S-specific
YAL040C	CLN3	Cyclin, G1/S-specific
YGR108W	CLB1	Cyclin, G2/M-specific
YPR119W	CLB2	Cyclin, G2/M-specific
YLR210W	CLB4	Cyclin, G2/M-specific
YPR120C	CLB5	Cyclin, B-type
YGR109C	CLB6	Cyclin, B-type
YMR043W	MCM1	Transcription factor of the MADS box family
YLR079W	SIC1	Inhibitor of Cdc28p-Clb protein kinase complex
YLR182W	SWI6	Transcription factor, subunit of SBF and MBF factors
YBR160W	CDC28	Cyclin-dependent protein kinase
YDL132W	CDC53	Controls G1/S transition, component of SCF-ubiquitin ligase complexes
YDL056W	MBP1	Transcription factor, subunit of the MBF factor
YDR054C	CDC34	E2 ubiquitin-conjugating enzyme
YDR146C	SWI5	Transcription factor
YDR328C	SKP1	Core component of SCF-ubiquitin ligase complexes
YER111C	SWI4	Transcription factor, subunit of SBF factor
YGL116W	CDC20	Cell division control protein
YGL003C	HCT1	Substrate-specific activator of APC-dependent proteolysis

In the next Section 4 the results obtained by GarNet are presented and discussed. The assessment of the resulting gene networks is performed in terms of the quality measures such as accuracy, precision, sensitivity, specificity, score and number of associations previously described.

4. Experimental results

Several studies have been carried out to assess the performance of GarNet in order to realize a fair comparison with other studies published in the literature. The experimental design is described in Section 4.1. In this subsection, we describe the training datasets (Section 4.1.1), the benchmark methods with which GarNet is compared (Section 4.1.2) and the parameter settings (Section 4.1.3). Furthermore, in Section 4.1.2 we report on the database used as a true network for the automatic assessment of the quality of the results obtained by GarNet and the benchmark methods measuring several metrics, such as accuracy and precision.

Results and discussion are presented in Section 4.2. We present the performance of GarNet and the comparison of GarNet against other approaches. The comparative study is organized similar to the study achieved in [12], where it was performed in three different stages due to the availability of the methods and the results reported in the literature. First, we discuss in Section 4.2.1 the performance of GarNet using different parameter thresholds to show the robustness of our approach and identify the most influent metrics in the results. Afterward, we present in Section 4.2.2 the performance of GarNet versus GRNCOP [11] and GRNCOP2 [12]. We compared average values obtaining from modifying several input parameters. Finally, we set the input parameters and we present in Section 4.2.3 the performance of GarNet versus other benchmark methods reported in the literature [3,9–11].

Finally, the biological relevance and a Gene Ontology analysis are presented in Section 4.3.

4.1. Experimental design

4.1.1. Datasets description

We used the microarray dataset of Spellman [13] and Cho et al. [57] 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 [58]. GarNet has been trained using a subset of 20 well-described genes which encode important proteins for cell-cycle regulation. These 20 genes are described in Table 1.

GarNet has been applied to cdc15, cdc28 and alpha-factors datasets. 500 QAR has been generated for each dataset, but only the QAR exceeding a minimum threshold of accuracy (Eq. (4)), confidence (Eq. (5)) and support (Eq. (3)) have been returned. Thereafter, the intersection among the attribute pairs retrieved from the QAR obtained by all datasets is performed to infer the resulting network.

4.1.2. Benchmark methods and description of the true network

The comparative analysis is performed similar to [12] and we compared our approach to other published techniques on the basis of being rule-based method. Consequently, to compare the inference capability of our approach against other methods we selected as benchmark methods: a decision-tree-based method [9], a regression-tree-based method [3] called RegNet, a probabilistic graphical model [10] and combinatorial optimization algorithms named GRNCOP [11] and GRNCOP2 [12].

Furthermore, we report on the database used as a true network for the automatic assessment of the quality of the results obtained by GarNet and the benchmark methods measuring several metrics, such as accuracy and precision.

For the assessment of the quality of the networks obtained by GarNet and the benchmark methods we used YeastNet database [56] as a true network. We used YeastNet as a blind performance test to compare our approach against the benchmark methods measuring several metrics: accuracy, precision, sensitivity and specificity. Regarding YeastNet, this is a network structure report with 102 803 potential gene–gene associations among 5483 yeast genes. This network is reported with a score value for each association. This network was build mainly from two resources. The Gene Ontology (GO) annotation downloaded from the Saccharomyces cerevisiae Genome Database (SGD) [59] and the literature [60], using more that 29 000 Medline abstract that included the word Saccharomyces cerevisiae for perfect matches to the gene pairs.

Furthermore, we analyzed the biological coherence of the results by means of an enrichment analysis using GO and literature mining.

4.1.3. Parameter settings

The values for the main parameters of GarNet are described in this section. It is noteworthy that these values have been used for each analysis carried out to assess the performance of GarNet.

The main parameters of GarNet are: 50 for the size of the population, 50 for the number of generations, 0.1 for the mutation probability p_{mut} of the individuals, 0.2 for the mutation probability p_{mutgen} of each gene in the individual and 50% over the full domain of each attribute for the minimum amplitude of each gene in the individual. The maximum number of attributes which could include both the antecedent and consequent are 10 and 2, respectively. These values are not important since the rules obtained in the experimentation never include more than 5 attributes in total. Note that both the antecedent and consequent have to contain one attribute at least. GarNet has obtained 500 QAR for each dataset described in Section 4.1.1.

Besides these parameters, different ones have been used for each type of analysis. The specific parameters will be described in the corresponding sections.

4.2. Results and discussion

4.2.1. Performance of GarNet using different minimum thresholds of goodness of the QAR

It is well known that one of the shortcomings of the EA is the parameterization of them. The strength of a parametric algorithm depends on a good adaptation to the characteristics of the problem. It is also necessary seeking such adjustments on the parameters according what we are looking for, i.e., a high accuracy or high efficiency.

A parametric sensitivity analysis and a detailed study of some parameters are discussed in this section to ascertain the relative influence of each parameter on the final results obtained by GarNet. The aim of this study is to analyze the applicability and the behavior of GarNet to achieve the most optimal solutions for the problem at hand in this work.

Our goal is to know the minimum quality thresholds of QAR to obtain the best values for the quality metrics in gene networks. Therefore, different configurations of GarNet have been executed by modifying the minimum thresholds of goodness of the QAR obtained in the first phase. Thus, it has been discovered which measures and threshold values for those are needed to be satisfied by a QAR in order to include the relationship between genes defined by such QAR in the network that models the yeast organism. To this aim, we applied GarNet following two phases. In the first phase, we set the parameterization of the algorithm to obtain QAR. Hence, several threshold values have been established for each considered interesting measure in order to obtain the rules that achieve a score greater than the specific value of the measure. Specifically, minimum thresholds for the two objectives (rule accuracy (Eq. (4)) and rule support (Eq. (3))) optimized by GarNet, and also the confidence quality measure (Eq. (5)) of the rule, have been used. Finally, in the second phase, networks are built from the rules and we use three known networks as blind tests to measure several performance metrics. The three networks were YeastNet [56], GO [59] and Co-citation [60] and the performance metrics were precision (Definition 2), sensitivity (Definition 3), specificity (Definition 4), score metrics (Definition 5) and number of associations (Definition 6) defined in Section 3.2.3.

As we mentioned before, the main motivation to carry out this sensitivity study is analyzed such that the behavior of GarNet and these experiments were carried out in a similar way reported in [12], where authors provide the results obtained through the variations of the minimum parameter thresholds. Analogously to that study, the parameters of the first phase were varied. The minimum values for the accuracy and confidence measures to be achieved by the rules vary from 0.6 to 0.9 with increments of 0.05 (7 variations for both), and the minimum values for the support measure vary from 0 to 0.5 with increments of 0.05 (11 variations). That means, GarNet were run 539 ($7 \times 7 \times 11$) times in total. The parametric sensitivity study of the 539 runs performed by GarNet is carried out three times, one for each known network structure used as a blind performance test. In each of these three experiments, the performance metrics of the networks mentioned before are calculated. It is noteworthy that all the metrics have been calculated following the similar idea of Gallo et al. in [12], where the precision, specificity and sensitivity were calculated regarding the reduced search space determined by the 20 genes, which consists of 190 possible gene–gene associations among them. The precision of the 190 possible gene pair-wise combinations was calculated in [12] for each known structure used as test in order to have a reference to validate the results. It is important that the results from each approach should exceed at least these values. The precision of this gene subset reaches the 51.58%, 43.68% and 45.25% values for YeastNet, GO annotations and Co-citation respectively.

Table 2
Average values for the gene networks metrics achieved by GarNet modifying the minimum thresholds of goodness of the QAR and tested in YeastNet.

ID	Accuracy	Confidence	Support	Precision	Sensitivity	Specificity	Score	Number of associations
0	0.6	[0.6–0.9]	[0–0.5]	72.37%	25.36%	83.64%	3.03	39.91
1	0.65	[0.6–0.9]	[0–0.5]	72.84%	25.60%	83.71%	3.05	40.08
2	0.7	[0.6–0.9]	[0–0.5]	75.08%	25.56%	84.57%	3.03	39.25
3	0.75	[0.6–0.9]	[0–0.5]	76.20%	25.59%	85.01%	2.98	38.87
4	0.8	[0.6–0.9]	[0–0.5]	76.09%	24.04%	85.25%	3.08	37.13
5	0.85	[0.6–0.9]	[0–0.5]	76.87%	22.21%	87.03%	3.05	33.70
6	0.9	[0.6–0.9]	[0–0.5]	77.00%	20.73%	87.87%	3.10	31.47
7	[0.6–0.9]	0.6	[0–0.5]	73.28%	27.43%	82.51%	3.03	42.97
8	[0.6–0.9]	0.65	[0–0.5]	73.16%	26.84%	82.39%	3.01	42.50
9	[0.6–0.9]	0.7	[0–0.5]	74.15%	26.03%	83.88%	3.04	40.34
10	[0.6–0.9]	0.75	[0–0.5]	73.45%	25.71%	83.74%	3.03	40.16
11	[0.6–0.9]	0.8	[0–0.5]	76.41%	23.55%	85.87%	3.04	36.08
12	[0.6–0.9]	0.85	[0–0.5]	76.18%	21.98%	87.66%	3.06	32.89
13	[0.6–0.9]	0.9	[0–0.5]	79.65%	17.87%	90.79%	3.12	25.99
14	[0.6–0.9]	[0.6–0.9]	0	58.18%	44.88%	65.02%	2.94	76.16
15	[0.6–0.9]	[0.6–0.9]	0.05	59.18%	45.00%	66.50%	2.91	74.92
16	[0.6–0.9]	[0.6–0.9]	0.1	58.77%	40.69%	69.14%	2.92	68.27
17	[0.6–0.9]	[0.6–0.9]	0.15	61.11%	36.15%	74.80%	2.96	58.61
18	[0.6–0.9]	[0.6–0.9]	0.2	65.95%	29.36%	83.47%	2.94	43.98
19	[0.6–0.9]	[0.6–0.9]	0.25	69.35%	23.80%	88.35%	2.98	34.04
20	[0.6–0.9]	[0.6–0.9]	0.3	82.44%	16.93%	95.63%	2.98	20.61
21	[0.6–0.9]	[0.6–0.9]	0.35	89.03%	11.89%	98.14%	3.05	13.37
22	[0.6–0.9]	[0.6–0.9]	0.4	93.49%	7.75%	99.25%	3.15	8.29
23	[0.6–0.9]	[0.6–0.9]	0.45	98.21%	4.02%	99.91%	3.32	4.02
24	[0.6–0.9]	[0.6–0.9]	0.5	94.28%	2.77%	99.76%	3.41	2.93

Table 3
Average values for the gene networks metrics achieved by GarNet modifying the minimum thresholds of goodness of the QAR and tested in Co-citation.

ID	Accuracy	Confidence	Support	Precision	Sensitivity	Specificity	Score	Number of associations
0	0.6	[0.6–0.9]	[0–0.5]	69.24%	27.73%	84.21%	3.21	39.91
1	0.65	[0.6–0.9]	[0–0.5]	69.24%	27.98%	84.25%	3.23	40.08
2	0.7	[0.6–0.9]	[0–0.5]	72.05%	28.15%	85.16%	3.24	39.25
3	0.75	[0.6–0.9]	[0–0.5]	73.23%	27.95%	85.35%	3.19	38.87
4	0.8	[0.6–0.9]	[0–0.5]	73.18%	26.41%	85.79%	3.28	37.13
5	0.85	[0.6–0.9]	[0–0.5]	73.99%	24.47%	87.49%	3.27	33.70
6	0.9	[0.6–0.9]	[0–0.5]	68.43%	19.59%	85.78%	3.32	31.47
7	[0.6–0.9]	0.6	[0–0.5]	69.37%	29.60%	82.80%	3.23	42.97
8	[0.6–0.9]	0.65	[0–0.5]	69.54%	28.77%	82.60%	3.21	42.50
9	[0.6–0.9]	0.7	[0–0.5]	70.50%	28.09%	84.09%	3.26	40.34
10	[0.6–0.9]	0.75	[0–0.5]	69.48%	27.58%	83.87%	3.23	40.16
11	[0.6–0.9]	0.8	[0–0.5]	72.57%	25.33%	85.93%	3.24	36.08
12	[0.6–0.9]	0.85	[0–0.5]	71.98%	23.83%	87.74%	3.26	32.89
13	[0.6–0.9]	0.9	[0–0.5]	76.19%	19.66%	90.96%	3.31	25.99
14	[0.6–0.9]	[0.6–0.9]	0	51.08%	46.62%	64.98%	3.13	76.16
15	[0.6–0.9]	[0.6–0.9]	0.05	52.14%	46.82%	66.30%	3.10	74.92
16	[0.6–0.9]	[0.6–0.9]	0.1	52.03%	42.66%	69.29%	3.11	68.27
17	[0.6–0.9]	[0.6–0.9]	0.15	56.86%	39.61%	75.95%	3.18	58.61
18	[0.6–0.9]	[0.6–0.9]	0.2	59.68%	31.96%	83.69%	3.16	43.98
19	[0.6–0.9]	[0.6–0.9]	0.25	65.57%	26.56%	88.79%	3.20	34.04
20	[0.6–0.9]	[0.6–0.9]	0.3	79.27%	19.15%	95.59%	3.22	20.61
21	[0.6–0.9]	[0.6–0.9]	0.35	87.38%	13.77%	98.19%	3.28	13.37
22	[0.6–0.9]	[0.6–0.9]	0.4	92.85%	9.07%	99.29%	3.35	8.29
23	[0.6–0.9]	[0.6–0.9]	0.45	97.46%	4.69%	99.88%	3.50	4.02
24	[0.6–0.9]	[0.6–0.9]	0.5	94.28%	3.27%	99.79%	3.55	2.93

The average results of the 539 runs are organized in three parts for each test network which are reported in Tables 2, 3 and 4 respectively. Note that each run is the resulting network after performing the intersection among the attribute pairs derived from the QAR obtained for each input dataset that exceed the minimum thresholds established in each case. The details of each part are described as follows:

- In the first part (rows from 0 to 6), each row represents the average values of the considered performance metrics for the gene networks obtained from executions which have a fixed value in the rules for the minimum accuracy, the

Table 4
Average values for the gene networks metrics achieved by GarNet modifying the minimum thresholds of goodness of the QAR and tested in GO.

ID	Accuracy	Confidence	Support	Precision	Sensitivity	Specificity	Number of associations
0	0.6	[0.6–0.9]	[0–0.5]	57.24%	21.03%	79.02%	39.91
1	0.65	[0.6–0.9]	[0–0.5]	57.59%	21.15%	78.97%	40.08
2	0.7	[0.6–0.9]	[0–0.5]	60.56%	21.31%	80.04%	39.25
3	0.75	[0.6–0.9]	[0–0.5]	60.33%	21.17%	79.78%	39.36
4	0.8	[0.6–0.9]	[0–0.5]	61.22%	19.75%	80.67%	37.13
5	0.85	[0.6–0.9]	[0–0.5]	58.17%	18.31%	82.41%	34.13
6	0.9	[0.6–0.9]	[0–0.5]	61.27%	16.58%	82.97%	31.91
7	[0.6–0.9]	0.6	[0–0.5]	58.90%	22.72%	77.49%	42.97
8	[0.6–0.9]	0.65	[0–0.5]	58.72%	22.73%	77.42%	43.05
9	[0.6–0.9]	0.7	[0–0.5]	58.39%	21.32%	78.86%	40.34
10	[0.6–0.9]	0.75	[0–0.5]	58.76%	21.16%	78.89%	40.16
11	[0.6–0.9]	0.8	[0–0.5]	59.81%	19.40%	80.94%	36.55
12	[0.6–0.9]	0.85	[0–0.5]	59.55%	18.20%	83.19%	33.31
13	[0.6–0.9]	0.9	[0–0.5]	62.07%	14.12%	86.80%	25.99
14	[0.6–0.9]	[0.6–0.9]	0	47.56%	36.82%	56.43%	76.16
15	[0.6–0.9]	[0.6–0.9]	0.05	49.45%	37.67%	58.70%	74.92
16	[0.6–0.9]	[0.6–0.9]	0.1	49.28%	34.40%	62.44%	68.27
17	[0.6–0.9]	[0.6–0.9]	0.15	51.40%	30.67%	68.97%	58.61
18	[0.6–0.9]	[0.6–0.9]	0.2	53.59%	23.95%	77.71%	43.98
19	[0.6–0.9]	[0.6–0.9]	0.25	56.00%	19.30%	83.56%	34.04
20	[0.6–0.9]	[0.6–0.9]	0.3	64.58%	13.35%	91.81%	20.61
21	[0.6–0.9]	[0.6–0.9]	0.35	68.14%	9.10%	95.16%	13.37
22	[0.6–0.9]	[0.6–0.9]	0.4	66.42%	5.55%	96.74%	8.44
23	[0.6–0.9]	[0.6–0.9]	0.45	68.22%	2.77%	98.53%	4.07
24	[0.6–0.9]	[0.6–0.9]	0.5	82.59%	2.39%	99.31%	2.98

minimum confidence varies from 0.6 to 0.9 and the minimum support from 0 to 0.5 summing up 77 runs for each row. Thus, the average results of the row identified by 0 belong to the runs that return the rules that achieve an accuracy value above 0.6, a confidence value above 0.6, 0.65, 0.7, 0.75 and so on until 0.9 and finally, a support value above 0, 0.05, 0.1 and so on until 0.5. The rest of the rows of the first part correspond to each minimum value for the accuracy to be achieved by the rules. The next two blocks follow a similar idea.

- Regarding the second part (rows from 7 to 13), each row indicates the average values of the considered performance metrics for the gene networks obtained from the runs in which the minimum confidence has a fixed value and the minimum accuracy varies from 0.6 to 0.9 and the minimum support varies from 0 to 0.5, adding up 77 runs for each row.
- Finally, the last part (rows from 14 to 24) contains the executions in which the minimum support has a fixed value and the minimum accuracy and the minimum confidence vary from 0.6 to 0.9, adding up 49 runs for each row.

Table 2 shows the average values in terms of precision, sensitivity, specificity, score and number of associations of the gene networks obtained when GarNet are tested using YeastNet network structure [56] as a true network. It can be observed that the average precision of the first part (rows from 0 to 6) reaches values between 72% and 77%. At low values of the accuracy measure (over 60%), GarNet provides at least the 72% of all possible gene–gene combinations which implies that GarNet obtains high quality gene networks although the allowable minimum accuracy in the rules provided by our approach in the first phase is low.

The precision value of the network increases when the minimum accuracy considered in the rules is higher. The average sensitivity takes values from 20% to 26% which are related with the number of gene–gene associations obtained. Contrary to precision, at low values of the accuracy measure, the average sensitivity is higher. This statement makes sense since when the accuracy is higher, the dimensionality of the gene networks increases. The score metric follows a similar behavior to precision. When the minimum accuracy to achieve for the rules is higher, the dimensionality of the network decreases, the precision increases, and therefore, the gene–gene associations are stronger and more relevant.

Similar conclusions can be observed in the second part of Table 2 when runs with a fixed value for the minimum confidence are considered. Nevertheless, we can observe in the first part and the second one respectively that there are no significant variations in the values of the performance metrics regarding to the minimum thresholds for accuracy and confidence. The few changes between the results obtained when low or high values for the minimum of accuracy and confidence are set, imply the high robustness of GarNet in presence of variations of the minimum parameter thresholds.

The third part of Table 2 reports the average result when parameter settings have a fixed minimum support to achieve by the rules and all the possible values for the minimum confidence and minimum accuracy are taken into account. Likewise, when the minimum threshold for the support measure increases, average precision, specificity and score of the gene networks are higher, against the sensitivity and the number of associations obtained whose values are decreasing. Moreover, some other interesting conclusions can be drawn from these results. The precision reaches values above 80% when GarNet

Table 5
Average values for the gene networks metrics achieved by GarNet using the best minimum thresholds of good-ness of the QAR against GRNCOP2 and GRNCOP.

		GarNet Min Sup 0.3	GarNet Min Sup 0.35	GRNCOP2	GRNCOP	RANDOM
YeastNet	Average precision	82.44%	**89.03%**	84.50%	76.69%	51.58%
	Average sensitivity	16.93%	11.89%	16.25%	**28.13%**	–
	Average specificity	**95.63%**	**98.14%**	94.66%	82.43%	–
	Average score	**2.98**	**3.05**	2.79	2.49	1.53033843
Co-citation	Average precision	79.27%	**87.38%**	84.13%	76.48%	43.68%
	Average sensitivity	19.15%	13.77%	19.02%	**30.46%**	–
	Average specificity	**95.59%**	**98.19%**	95.28%	82.76%	**–**
	Average score	**3.22**	**3.28**	2.91	2.50	1.3487118
GO	Average precision	64.58%	68.14%	**70.73%**	52.25%	45.26%
	Average sensitivity	13.35%	9.10%	13.93%	**22.55%**	–
	Average specificity	**91.81%**	**95.16%**	91.48%	76.60%	–
Average number of associations		20.61	13.37	**20.84**	43.73	–

provides the rules with at least 30% of support. Although the sensitivity decreases until 16% due to the reduction of the dimensionality of the gene networks, the average score of the gene–gene associations exceeds the 3 value. It is noteworthy that the support measure is more influential in the results against to the accuracy and confidence measures. Furthermore, most optimal solutions are obtained when the minimum support of the rules increases.

Tables 3 and 4 provide the average values in terms of precision, sensitivity, specificity, score and number of association of the gene networks obtained when GarNet are tested using Co-citation network and GO as a true networks. Note that Table 4 does not contain the average score of the gene networks obtained since GO repository does not provide the score of gene–gene interactions. Besides the results obtained by GarNet, this table reports similar behavior to Table 2 described above. Better results are obtained when the minimum support is higher. Precision values close to 95% are reached for GO network and 85% regarding Co-citation when the minimum support achieves the 50%.

To summarize, Tables 2, 3 and 4 show that results in the third part overcome the results included in the first and second ones. It is due to the support measure has more influence on the results than the confidence and accuracy and also implies getting more optimum results for the yeast organism. It can be concluded from these results that they are more sensitive regarding the support measure instead of accuracy and confidence measures. Furthermore, it is worth mentioned that the most executions performed by our approach obtained high quality results regardless of the minimum values setting for the parameters used, which demonstrates the high robustness of GarNet. Finally, the average results for the precision overcome in all cases the results obtained if the pairs of gene–gene associations were chosen randomly.

In the next subsection, a fair comparative study of GarNet against several benchmark methods is presented in a similar way of the comparative analysis reported in [12].

4.2.2. Performance of GarNet versus GRNCOP2 and GRNCOP

Once the sensitivity study of parameters has been performed, a comparative analysis similar to that presented in [12] is described. To this aim, the best parameter settings have been selected to compare the results obtained by GarNet against other well-known approaches such as the GRNCOP2 [12] which extends the previous version GRNCOP [11]. The selected parameter settings are those identified by rows 22 and 23 in Tables 2, 3 and 4 since the average precision values are higher than 80% without unduly reducing the dimensionality of the gene networks obtained.

As we mentioned before, the comparative study follows the same scheme used in [12] where the improvements of GRNCOP2 were analyzed over GRNCOP. Both algorithms performed several runs varying the accuracy parameter and other specific parameter from 0.6 to 0.9 with increments of 0.05 for the same subset of 20 genes. A total of 56 runs were carried out for each method and the set of associations, i.e. the networks obtained in each case were measured in terms of the precision, sensitivity, specificity and score metrics previously defined.

In this subsection, a fair comparative analysis of GarNet against GRNCOP and GRNCOP2 is performed. To this aim, both best settings from the previous subsection have been selected. Therefore, instead of varying three parameters, the minimum support threshold is fixed at 0.3 and 0.35 and the minimum accuracy and confidence to achieve by the rules is modified similarly to the accuracy parameter and the other specific parameter used by GRNCOP2 and GRNCOP, i.e. from 0.6 to 0.9 with increments of 0.05. Its experiments imply 49 executions.

Table 5 summarizes the average results obtained by our approach compared against GRNCOP and GRNCOP2. The first and second columns show the average results obtained by GarNet in 49 runs when it is applied using a minimum support threshold of 30% and 35%, respectively. The third column describes the average results achieved by GRNCOP2 in 56 runs. The fourth column displays the average results drawn by GRNCOP in also 56 runs. The average results of last column would be the expected values if random sets (uniformly distributed) of gene-pairs were selected (see [12]). Henceforth, GarNet using a minimum support threshold of 0.3 or 0.35 to achieve by the rules is denoted as GarNet$_{0.3}$ and GarNet$_{0.35}$, respectively.

In this table, the average results are presented for each method trained by the cdc15, cdc28 and alpha-factors datasets for the budding yeast organism and tested using YeastNet, GO and Co-citation network according to average precision, sensitivity, specificity, score metrics and average number of associations obtained by each method. The best results for each performance metric have been highlighted but also, when both settings of GarNet overcome GRNCOP and GRNCOP2 are also highlighted although some setting is better than the other in some case.

It can be observed that GarNet presents generally better results than the other approaches in terms of average precision, specificity and score. In the case of GarNet$_{0.3}$, the average number of associations is quite similar to GRNCOP2. The average sensitivity is always higher in GRNCOP but the precision is lower (around 76%), thus, this approach obtains gene networks with higher dimensionality but fewer known associations.

Regarding GarNet$_{0.3}$ tested using YeastNet network structure, it reaches higher values of average sensitivity, specificity and score than GRNCOP2 and GRNCOP. Only the average precision is slightly lower than GRNCOP2, but surpasses GRNCOP. Regarding GarNet$_{0.35}$ also obtains better results for average specificity and average score values against GRNCOP and GRNCOP2. Although the sensitivity is worst compared with GarNet$_{0.3}$, GRNCOP2 and GRNCOP, the average precision overcomes all the other approaches.

The average metric values achieved in the case of Co-citation present the same behavior than YeastNet network structure. Once again, GarNet$_{0.3}$ overcomes GRNCOP in terms of average precision, specificity and score, and also, it reaches better values than GRNCOP2 regarding average sensitivity, specificity and score. In the case of GarNet$_{0.35}$, similar conclusions to YeastNet can be extended for the results obtained in the case of Co-citation.

Respecting GO annotations, both GarNet$_{0.3}$ and GarNet$_{0.35}$ obtain better values according to the average specificity. The average sensitivity is higher for the GRNCOP approach. The average precision presents high values for the GRNCOP2 approach against GarNet$_{0.3}$ and GarNet$_{0.35}$, however, not high differences are detected between them.

It can be observed that when the minimum support is higher, GarNet obtains gene networks more accurate with less the number of associations. GarNet is able to obtain gene networks with size and average precision values similar to GRNCOP2. Moreover, although the average number of associations is similar, the proportion of true negative correctly identified achieved by GarNet is higher and the strength of the associations of the network is also better in contrast to GRNCOP and GRNCOP2. In general terms, the average results obtained by GarNet are successful and also overcome GRNCOP and GRNCOP2 in many cases. Therefore, it can be concluded that GarNet is a valid tool to work on biological problems inferring gene networks, since results are similar and comparable to obtained by other approaches.

With the aim to perform a more detailed comparison analysis, the best gene network of each parameter setting (explained before) has been selected and has been also compared against the best gene network obtained by GRNCOP2 and GRNCOP. Other gene networks obtained by other algorithms dealing with the same biological dataset are also contrasted. The results obtained are discussed in the following Section 4.2.3.

4.2.3. Performance of GarNet versus other benchmark methods

In this section, the inference capability of the best gene networks from each parameter settings is compared against several rule-based methods from the literature. Specifically, a decision-tree-based method [9], a regression-tree-based method [3], a probabilistic graphical model [10] and combinatorial optimization algorithm [12] are contrasted with the best ones inferred by GarNet.

It is noteworthy that the results of every benchmark method were obtained from the original papers and the study proposed in [12]. In this study, only the rules which achieved an accuracy of at least 0.75 on cdc15, cdc28 and alpha-factor were selected. In order to carry out a fair comparative, the same minimum accuracy value has been considered to select the rules obtained by GarNet$_{0.3}$ and GarNet$_{0.35}$. It can be noted that the gene networks obtained by GarNet$_{0.3}$ and GarNet$_{0.35}$, henceforth, are denoted as GarNet$_1$ and GarNet$_2$ respectively. To summarize the parameter setting, both GarNet$_1$ and GarNet$_2$ have been executed on the three datasets with an accuracy threshold of 0.75, a confidence threshold of 0.8, and a support of the rule threshold of 0.3 and 0.35 respectively.

Table 6 presents the comparison of GarNet against the benchmark methods. We can observe the results achieved by the gene networks in terms of precision, sensitivity, specificity, accuracy and score tested with YeastNet network structure, Co-citation and GO annotations. Analogously to Table 5, the best results have been highlighted considering GarNet$_1$ and GarNet$_2$ against the other proposals.

It can be observed that higher number of associations have been obtained in each gene network from GarNet$_1$ and GarNet$_2$ against the other proposals. Although GarNet$_{0.35}$ overcomes GarNet$_{0.3}$ in many cases according the average metrics except in the average sensitivity, GarNet$_1$ presents better values than GarNet$_2$ in most metrics. Only the average score values of GarNet$_1$ are slightly lower than GarNet$_2$ since the size of the gene network of GarNet$_1$ is higher. It could be that both GarNet$_1$ and GarNet$_2$ have the same associations, but because GarNet$_1$ has a higher dimensionality, it might contain other less relevant associations involving a decrease in the average score. Nevertheless, in Table 7 the associations found for each are contrasted to identify similarities and differences among them.

Regarding GarNet$_1$ against the other approaches for the YeastNet structure, GarNet$_1$ presents better results except the average score values due to the dimensionality of gene network as described before. In the case of Co-citation, only the specificity and precision of RegNet are slightly higher than GarNet$_1$ and GarNet$_2$ but minor differences between precision values (93.75% or 95% against 100%) and specificity values (99.07% against 100%) are presented. Furthermore, RegNet achieve a precision of 100% by obtaining a network with a low number of associations (RegNet obtains 7 associations which implies

Table 6
Values for the gene networks metrics achieved by the best settings of GarNet against GRNCOP2, RegNet, Soinov et al. and Bula-shevska and Eils.

		$GarNet_1$	$GarNet_2$	GRNCOP2	RegNet	Soinov et al.	Bulashevska and Eils	RANDOM
YeastNet	Precision	**100**	93.75	93.33	**100**	50.00	88.89	51.58
	Sensitivity	**20.40**	**15.31**	14.29	7.14	3.06	8.16	–
	Specificity	**100**	98.91	98.91	**100**	96.74	98.91	–
	Accuracy	**58.94**	**55.79**	55.27	52.11	48.41	52.09	–
	Score	2.89	2.82	3.04	**3.24**	1.84	2.77	1.53
Co-citation	Precision	95.00	93.75	93.33	**100**	50.00	88.89	43.68
	Sensitivity	**22.89**	**18.07**	16.87	8.13	3.61	9.64	–
	Specificity	99.07	99.07	99.07	**100**	97.20	99.07	–
	Accuracy	**65.79**	**63.68**	63.16	58.42	56.29	60.00	–
	Score	2.92	3.09	3.26	**3.51**	1.85	2.84	1.35
GO	Precision	70.00	**75.00**	73.33	71.43	50.00	55.56	45.26
	Sensitivity	**16.28**	**13.95**	12.79	5.81	3.49	5.81	–
	Specificity	94.23	96.15	96.15	**98.08**	97.12	96.15	–
	Accuracy	**58.96**	**58.95**	58.42	56.32	54.75	55.24	–
Number of associations		**20**	**16**	15	7	6	9	

Table 7
Associations inferred by $GarNet_1$ and $GarNet_2$.

	Id	Gene–gene associations inferred by GarNet	GRNCOP (8/20)	GRNCOP2 (9/15)	RegNet (5/7)	Soinov et al. (3/6)	Bulashevska and Eils (5/9)
Common gene–gene associations	1	CLN3 CLB4					√
	2	CDC34 CDC53					
	3	SWI5 CLB1		√	√		√
	4	SWI5 CLB2		√	√		√
	5	CLB1 CLB6	√	√			√
	6	CLB1 CLN2	√	√			
	7	CLB1 CLB2	√	√	√	√	√
	8	CLB6 CDC53					
	9	CLN1 CLB6	√	√			
	10	CLN1 CLN2	√	√	√	√	
	11	CLN2 CLB5	√	√	√		
	12	CLB2 CLN1	√				
$GarNet_1$ gene–gene associations	20	CLB5 CLB6	√	√		√	
	13	MBP1 CLB2					
	14	CDC53 CLN2					
	15	CDC53 CLB5					
	16	SWI4 CLN1					
	17	CLB6 CLB2					
	18	CLN2 HCT1					
	19	CLB5 CLB1					
$GarNet_2$ gene–gene associations	21	CLB1 CLN1	√				
	22	CLN1 SWI5					
	23	HCT1 CLB2					
	24	CLN1 CLN3					

less than half of the associations of $GarNet_1$). According to GO annotations, $GarNet_1$ and $GarNet_2$ overcome the other approaches in terms of precision, sensitivity and accuracy. Once again RegNet achieve a higher specificity value. The $GarNet_2$ precision values are the best from the table in this case, and the $GarNet_1$ precision values are higher than the *Soinov et al.* and *Bulashevska and Eils* proposals and quite similar to the precision values of GRNCOP2 and RegNet.

It is important to remark that YeastNet was built from several resources such as GO annotations and Co-citation which are association subsets within YeastNet. Therefore, all the associations retrieved in YeastNet could not be present in GO or Co-citation. Hence, if GarNet does not achieve the 100% in Co-citation or GO Annotations is not worrying since it gets this mark in YeastNet structure network.

From the results it can be clearly drawn that GarNet is a tool capable of inferring relevant associations with high precision values that those other methods proposed in the literature. Hence, our goal to adapt an EA to discover QAR dealing with biological problems has been successful achieved.

Furthermore, in Table 7 we can observe the gene–gene associations provided by GarNet. In the first part we can see the associations in common, in the second and third parts we can observe the association obtained in $GarNet_1$ and $GarNet_2$, respectively. Finally, the associations inferred by the benchmark methods are marked. The next Section 4.3 describes the biological relevance of the associations found by GarNet.

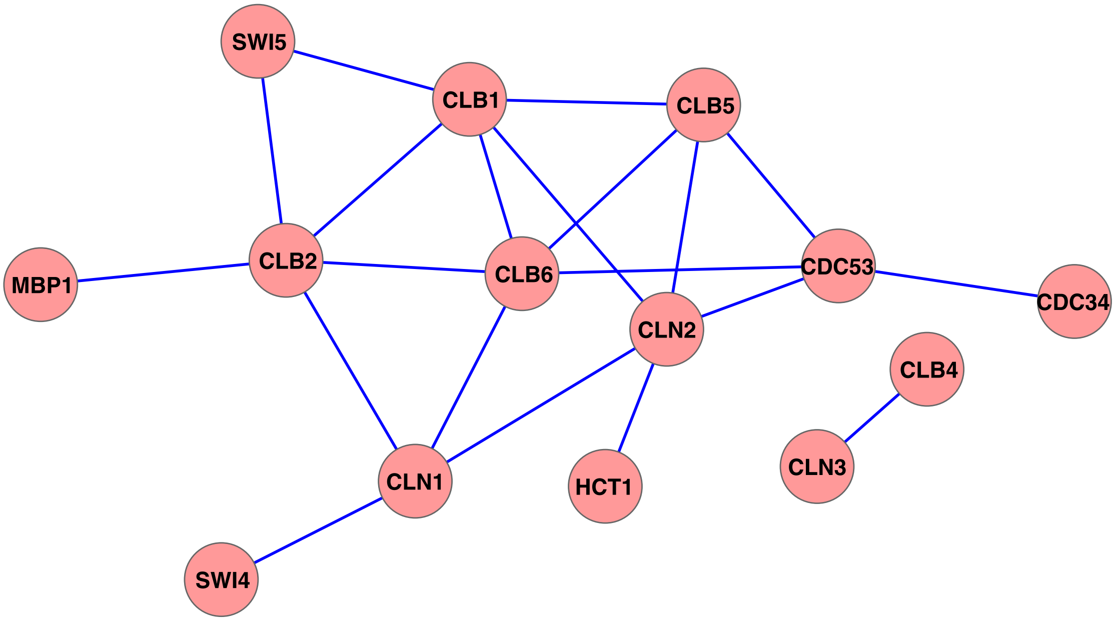


Fig. 5. Gene association network obtained by GarNet$_1$ from the yeast cell cycle datasets.

4.3. *Biological relevance of the results*

In this section, the biological relevance of the gene network inferred by our proposal using the GarNet$_2$ setting is presented. The set of 16 genes from this network and the set of 20 genes from the input microarray data were analyzed in the context of Gene Ontology with the FuncAssociate tool [61]. The idea is to detect whether a loss of information is presented in the results when the number of genes is reduced. Detection of statistically overrepresented GO terms was done with the hypergeometric test, multiple-testing adjustments with the Westfall and Young false discovery rate and a significance level $\alpha = 0.001$. The goal of this analysis is to check that there is no loss of biological information in the genes of the network, because the input dataset is a set of well-known genes in several biological processes (using the specified statistical test and corrections). These 20 genes are involved in 32 GO biological process terms as significantly overrepresented with a p-value less than 0.001 and the network inferred using GarNet$_2$ setting identified by 30 of these 32 GO terms with a p-value less than 0.001. However, there is no loss of information because the 2 different GO terms represent general biological processes: regulation of S phase (GO:0033261) and regulation of cell-cycle process (GO:0010564). Furthermore, the set of genes in the network has specific GO terms involved by the mentioned before, as for example the regulation of S phase of mitotic cell cycle (GO:0007090) and positive regulation of cell0cycle process (GO:0090068).

On the other hand, the biological relevance of the associations inferred by our approach (see Fig. 5) was verified by analyzing whether such associations reflect functional properties relating to the different cell-cycle phases G_1, S, G_2, M and $M \setminus G_1$. These different cell-cycle phases are usually used in the literature.

The association CLB5–CLB6 in GarNet$_1$ is consistent with the knowledge that the maximum of CLB2 transcription is in G2 phase, whereas CLN1, CLN2, CLB5 and CLB6 all have their expression maximum in G1 [62]. The association CLB1–CLB2 is in agreement with CLB2 and CLB1 being expressed simultaneously in G2 [63]. The rules CLB2–SWI5, CLB1–SWI5 are in agreement with the literature: transcription of SWI5 and CLB1 is G_2/M specific and activated in late S phase; the expression pattern of SWI5 is similar to that of CLB1 and CLB2 and the peak of mRNA concentration of SWI5 is in G2 [64] and [65]. The association CLN2–CLN1 is consistent with the knowledge that CLN2 and CLN1 have their transcription maximum in G1 [9] whereas CDC20 is transcribed in late S/G_2 phase. Finally, the associations CLN1–CLN2 and CLN2–CLB5 are not inferred by the benchmark method [9], but they are consistent with observations on the partial functional redundancy existing among CLB5, CLN1 and CLN2, which have been reported in [66] and [67].

The rules which are not supported by the literature are new hypothesis to analyze in the laboratory. Furthermore, these rules reflect the high complexity of the biological process and the regulatory relationships between pairs of genes.

5. Conclusions

In this work, a multi-objective evolutionary algorithm to discover QAR has been proposed to infer gene association networks. The approach named GarNet is based on the well-known NSGA-II and determines the attribute intervals of the rules avoiding the discretization of the attributes as a first step of the process.

GarNet generates new hypothesis of associations among genes, and differs from statistical-based methods as for example correlation-based method such that the dependency is manly detected in a localized region of the space. As a first improvement, our approach strongly favors localized similarities over more global similarity. As a second improvement, our approach avoids the discretization step applied by other approaches for the inference of gene networks based on mining qualitative AR [8].

GarNet mainly performs two steps to deal with the generation of the gene association networks. First, GarNet carries out an inference process based on an IRL to extract gene–gene associations. Then, GarNet builds gene networks by the intersection of the gene–gene associations retrieved from the QAR found in several input microarray datasets.

To evaluate the robustness of our approach, the performance of GarNet was organized in three steps similarly to the study performed by Gallo et al. First, the parameterization of the algorithm was analyzed to show the high robustness of GarNet in terms of the minimum thresholds to be satisfied by the QAR obtained. Second, we applied GarNet to a well-known set of genes from a microarray data of yeast cell cycle and we compared our approach against several benchmark methods. For performance analysis we applied as benchmark methods a decision-tree-based method [9], a regression-tree-based method [3], a probabilistic graphical model [10] and combinatorial optimization algorithm [11,12]. GarNet outperformed the benchmark methods in most cases in terms of quality metrics of the networks (precision, accuracy and others), which were measured using YeastNet database as a true network. The results have shown that the rules obtained have been able to successfully characterize the underlying information, grouping relevant genes for the problem under studied and agreeing with prior biological knowledge. The biological relevance has been analyzed using an enrichment analysis and the literature.

GarNet has proved a valuable tool to infer networks successfully. As a conclusion, an advantage of network reconstruction using GarNet is that the method is able to construct a network correctly, i.e. reproducing the logic of a network consistent with the data. The network reconstructed from cell-cycle yeast datasets is consistent with the knowledge stored in the literature. Furthermore, as future research directions, the method could be improved by adding prior knowledge and different types of molecular data.

Acknowledgments

The financial support from the Spanish Ministry of Science and Technology, project TIN2011-28956-C02-02, and from the Junta de Andalucía, project P11-TIC-7528, is acknowledged.

References

[1] P. Brown, D. Botstein, Exploring the new world of the genome with DNA microarrays, Nat. Genet. 21 (1999) 33–37.

[2] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, D. di Bernardo, How to infer gene networks from expression profiles, Mol. Syst. Biol. 3 (2007).

[3] I.A. Nepomuceno-Chamorro, J.S. Aguilar-Ruiz, J.C. Riquelme, Inferring gene regression networks with model trees, BMC Bioinformatics 11 (2010) 517–525.

[4] M.J. del Jesús, J.A. Gámez, P. González, J.M. Puerta, On the discovery of association rules by means of evolutionary algorithms, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 1 (2011) 397–415.

[5] K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: Nsga-ii, IEEE Trans. Evol. Comput. 6 (2002) 182–197.

[6] E. Zitzler, M. Laumanns, L. Thiele, Spea2: improving the strength Pareto evolutionary algorithm, EUROGEN 3242 (2001) 95–100.

[7] M.J. Gacto, R. Alcalá, F. Herrera, Adaptation and application of multi-objective evolutionary algorithms for rule reduction and parameter tuning of fuzzy rule-based systems, Soft Comput. 13 (2009) 419–436.

[8] F. Du, N. Rao, J. Guo, Z. Yuan, R. Wang, Mining gene network by combined association rules and genetic algorithm, in: W. Liao, Y. Fang, C. Zhu, O.C. Au (Eds.), Proceedings of the Int. Conf. on Communications, Circuits and Systems, IEEE Computer Society, Washington, DC, 2009.

[9] L. Soinov, M. Krestyaninova, A. Brazma, Towards reconstruction of gene networks from expression data by supervised learning, Genome Biol. 4 (2003) R6.

[10] S. Bulashevska, R. Eils, Inferring genetic regulatory logic from expression data, Bioinformatics 21 (2005) 2706–2713.

[11] I. Ponzoni, F.A. Azuaje, D. Juan Glass, Inferring adaptive regulation thresholds and association rules from gene expression data through combinatorial optimization learning, IEEE/ACM Trans. Comput. Biol. Bioinform. 4 (2007) 624–634.

[12] C.A. Gallo, J.A. Carballido, I. Ponzoni, Discovering time-lagged rules from microarray data using gene profile classifiers, BMC Bioinformatics 12 (2011) 123–131.

[13] P. Spellman, G. Sherlock, M. Zhang, V. Iyer, K. Anders, M. Eisen, P. Brown, D. Botstein, B. Futcher, Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization, Mol. Biol. Cell. 9 (1998) 3273–3297.

[14] M.B. Eisen, P.T. Spellman, P.O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, in: Proceedings of the National Academy of Sciences of the United States of America, vol. 95, 1998 pp. 14863–14868.

[15] P. D'Haeseleer, X. Wen, S. Fuhrman, Mining the gene expression matrix: inferring gene relationships from large scale gene expression data, in: Proceedings of the Second International Workshop on Information Processing in Cell and Tissues, 1998, pp. 203–212.

[16] X. Zhou, M. Kao, W. Wong, From the cover: transitive functional annotation by shortest-path analysis of gene expression data, Proc. Natl. Acad. Sci. 99 (2002) 12783–12788.

[17] H.K. Lee, A.K. Hsu, J. Sajdak, J. Qin, P. Pavlidis, Coexpression analysis of human genes across many microarray data sets, Genome Res. 14 (2004) 1085–1094.

[18] C. Wolfe, I. Kohane, A. Butte, Systematic survey reveals general applicability of "guilt-by-association" within gene coexpression networks, BMC Bioinformatics 6 (2005) 227.

[19] T. Obayashi, K. Kinoshita, Rank of correlation coefficient as a comparable measure for biological significance of gene coexpression, DNA Res. 16 (2009) 249–260.

[20] A. de la Fuente, N. Bing, I. Hoeschele, P. Mendes, Discovery of meaningful associations in genomic data using partial correlation coefficients, Bioinformatics 20 (2004) 3565–3574.

[21] F. Markowetz, R. Spang, Inferring cellular networks–a review, BMC Bioinformatics 8 (2007) S5.

[22] A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Favera, A. Califano, Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, BMC Bioinformatics 7 (2006) S7.

[23] C. Borgelt, A conditional independence algorithm for learning undirected graphical models, J. Comput. System Sci. 76 (2010) 21–33.

[24] R. Agrawal, R. Srikant, Fast algorithms for mining association rules in large databases, in: Proceedings of the International Conference on Very Large Databases, 1994, pp. 478–499.

[25] M. Kaya, R. Alhajj, Utilizing genetic algorithms to optimize membership functions for fuzzy weighted association rules mining, Appl. Intell. 24 (2006) 7–152.

[26] B. Alatas, E. Akin, An efficient genetic algorithm for automated mining of both positive and negative quantitative association rules, Soft Comput. 10 (2006) 230–237.

[27] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2006.

[28] R. Agrawal, T. Imielinski, A. Swami, Mining association rules between sets of items in large databases, in: Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, 1993, pp. 207–216.

[29] R.J. Kuo, S.Y. Lin, C.W. Shih, Mining association rules through integration of clustering analysis and ant colony system for health insurance database in Taiwan, Expert Syst. Appl. 33 (2007) 794–808.

[30] V. Mangat, A novel hybrid framework using evolutionary computing and swarm intelligence for rule mining in the medical domain, in: IJCA Proceedings on International Conference on Recent Advances and Future Trends in Information Technology (iRAFIT 2012), iRAFIT 6 (2012) 7–13.

[31] K. Rameshkumar, Extracting association rules from HIV infected patients' treatment dataset, Trends in Bioinformatics 4 (2011) 35–46.

[32] M. Steinbrecher, R. Kruse, Visualizing and fuzzy filtering for discovering temporal trajectories of association rules, J. Comput. System Sci. 76 (2010) 77–87.

[33] L. Geng, H.J. Hamilton, Interestingness measures for data mining: a survey, ACM Comput. Surv. 38 (2006) 9.

[34] G. Piatetsky-Shapiro, Discovery, analysis and presentation of strong rules, in: Knowledge Discovery in Databases, 1991, pp. 229–248.

[35] M. Houtsma, A. Swami, Set-Oriented Mining for Association Rules, in: Proceedings of IEEE Data Engineering Conference, 1995.

[36] M. Vannucci, V. Colla, Meaningful discretization of continuous features for association rules mining by means of a SOM, in: Proceedings of the European Symposium on Artificial Neural Networks, Bruges, Belgium, 2004, pp. 489–494.

[37] T. Fukuda, Y. Morimoto, S. Morishita, T. Tokuyama, Mining optimized association rules for numeric attributes, J. Comput. System Sci. 58 (1999) 1–12.

[38] A. Orriols-Puig, J. Casillas, E. Bernadó-Mansilla, First approach toward on-line evolution of association rules with learning classifier systems, in: Proceedings of the 2008 GECCO Genetic and Evolutionary Computation Conference, 2008, pp. 2031–2038.

[39] B. Alatas, E. Akin, Rough particle swarm optimization and its applications in data mining, Soft Comput. 12 (2008) 1205–1218.

[40] Y. Yin, Z. Zhong, Y. Wang, Mining quantitative association rules by interval clustering, J. Comput. Inf. Syst. 4 (2008) 609–616.

[41] E.D. Goldberg, Genetic Algorithms in Search, Optimization, and Machine Learning, Addison–Wesley Publishing Company, 1989.

[42] V. Pachón Álvarez, J. Mata Vázquez, An evolutionary algorithm to discover quantitative association rules from huge databases without the need for an a priori discretization, Expert Syst. Appl. 39 (2012) 585–593.

[43] X. Yan, C. Zhang, S. Zhang, Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support, Expert Syst. Appl. 36 (2009) 3066–3076.

[44] J.M. Luna, J.R. Romero, S. Ventura, Design and behavior study of a grammar-guided genetic programming algorithm for mining association rules, Knowl. Inf. Syst. 32 (2012) 53–76.

[45] C. Romero, A. Zafra, J.M. Luna, S. Ventura, Association rule mining using genetic programming to provide feedback to instructors from multiple-choice quiz data, Expert Syst. (2012).

[46] J. Alcalá-Fdez, N. Flugy-Pape, A. Bonarini, F. Herrera, Analysis of the effectiveness of the genetic algorithms based on extraction of association rules, Fund. Inform. 98 (2010) 1001–1014.

[47] K. Deb, Multi-Objective Optimization Using Evolutionary Algorithms, John Wiley & Sons, Inc., 2001.

[48] M.J. del Jesús, J.A. Gámez, P. González, J.M. Puerta, On the discovery of association rules by means of evolutionary algorithms, Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 1 (2011) 397–415.

[49] B. Alatas, E. Akin, A. Karci, MODENAR: multi-objective differential evolution algorithm for mining numeric association rules, Appl. Soft Comput. 8 (2008) 646–656.

[50] H.R. Qodmanan, M. Nasiri, B. Minaei-Bidgoli, Multi objective association rule mining with genetic algorithm without specifying minimum support and minimum confidence, Expert Syst. Appl. 38 (2011) 288–298.

[51] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J.C. Riquelme, An evolutionary algorithm to discover quantitative association rules in multidimensional time series, Soft Comput. 15 (2011) 2065–2084.

[52] M. Martínez-Ballesteros, F. Martínez-Álvarez, A. Troncoso, J.C. Riquelme, Mining quantitative association rules based on evolutionary computation and its application to atmospheric pollution, Integr. Comput.-Aided Eng. 17 (2010) 227–242.

[53] M. Martínez-Ballesteros, S. Salcedo-Sanz, J.C. Riquelme, C. Casanova-Mateo, J.L. Camacho, Evolutionary association rules for total ozone content modeling from satellite observations, Chemom. Intell. Lab. Syst. 109 (2011) 217–227.

[54] B.L. Miller, D. Goldberg, Genetic algorithms, tournament selection, and the effects of noise, Complex Systems 9 (1995) 193–212.

[55] G. Venturini, SIA: A Supervised Inductive Algorithm with genetic search for learning attribute based concepts, in: Proceedings of the European Conference on Machine Learning, 1993, pp. 280–296.

[56] I. Lee, Z. Li, E.M. Marcotte, An improved, bias-reduced probabilistic functional gene network of baker's yeast, *Saccharomyces cerevisiae*, PLoS ONE 2 (2007) e988.

[57] R.J. Cho, M.J. Campbell, E.A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T.G. Wolfsberg, A.E. Gabrielian, D. Landsman, D.J. Lockhart, R.W. Davis, A genome-wide transcriptional analysis of the mitotic cell cycle, Mol. Cell. 2 (1998) 65–73.

[58] E. van Someren, L.F. Wessels, M.J. Reinders, Linear modeling of genetic networks from experimental data, in: ISMB'00, 2000, pp. 355–366.

[59] S.S. Dwight, M.A. Harris, K. Dolinski, C.A. Ball, G. Binkley, K.R. Christie, D.G. Fisk, L.F. Issel-Tarver, M. Schroeder, G. Sherlock, A. Sethuraman, S. Weng, D. Botstein, J.M. Cherry, Saccharomyces genome database (SGD) provides secondary gene annotation using the gene ontology (GO), Nucleic Acids Res. 30 (2002) 69–72.

[60] I. Lee, S.V. Date, A.T. Adai, E.M. Marcotte, A probabilistic functional network of yeast genes, Science 306 (2004) 1555–1558.

[61] G. Berriz, O. King, B. Bryant, C. Sander, F. Roth, Characterizing gene sets with FuncAssociate, Bioinformatics 19 (2003) 2502–2504.

[62] K. Chen, A. Csikasz-Nagy, B. Gyorffy, J. Val, B. Novak, J. Tyson, Kinetic analysis of a molecular model of the budding yeast cell cycle, Mol. Biol. Cell. 11 (2000) 369–391.

[63] H. Althoefer, A. Schleiffer, K. Wassmann, A. Nordheim, G. Ammerer, Mcm1 is required to coordinate G2-specific transcription in Saccharomyces cerevisiae, Mol. Cell. Biol. 15 (1995) 5917–5928.

[64] C.J. Loy, D. Lydall, U. Surana, Ndd1, a high-dosage suppressor of cdc28-1N, is essential for expression of a subset of late-S-phase-specific genes in Saccharomyces cerevisiae, Mol. Cell. Biol. 19 (1999) 3312–3327.

[65] J. Toyn, A. Johnson, J. Donovan, W. Toone, L. Johnston, The Swi5 transcription factor of Saccharomyces cerevisiae has a role in exit from mitosis through induction of the cdk-inhibitor Sic1 in telophase, Genetics 145 (1997) 85–96.

[66] C.B. Epstein, F.R. Cross, Clb5: a novel B cyclin from budding yeast with a role in S phase, Genes Dev. 6 (1992) 1695–1706.

[67] K. Levine, K. Huang, F.R. Cross, Saccharomyces cerevisiae G1 cyclins differ in their intrinsic functional specificities, Mol. Cell. Biol. 16 (1996) 6794–6803.