Learning Directed-Acyclic-Graphs from Multiple Genomic Data Sources

Fabio Nikolay, Marius Pesavento
Communication Systems Group, Technische Universität Darmstadt

***Abstract*—In this paper we consider the problem of learning the topology of a directed-acyclic-graph, that describes the interactions among a set of genes, based on noisy double knockout data and genetic-interactions-profile data. We propose a novel linear integer optimization approach to identify the complex biological dependencies among genes and to compute the topology of the directed-acyclic-graph that matches the data best. Finally, we apply a sequential scalability technique for large sets of genes along with our proposed algorithm, in order to provide statistically significant results for experimental data.**

***Index Terms*—Gene networks, discrete optimization, big data, graph learning**

I. Introduction

In genomics research and systems biology, uncovering the interactions among a set of genes with respect to a specified cell function of a biological system, e.g., the fitness of a specific bacteria strain, has recently attracted much attention, [1], since it is fundamental for understanding the biological processes that underlie the specific cell function under study. The interactions among the genes under study can be characterized by an *in-tree*, which is a directed-acyclic-graph (DAG) with a common root node where all edges are orientated towards the root. In the context of systems biology such an in-tree is often simply referred to as a DAG, [2], [3]. The hierarchical relationship between two genes in a DAG describes their hierarchical interaction type [3]. Since DAGs cannot be observed directly, they are hidden quantities and only the specified cell function of the organism under study, referred to as the phenotype, can be monitored. The term phenotype describes the particular manifestation of a biological attribute of an organism that can be observed. For instance, a common biological attribute of bacteria is growth measured in colony size, where a particular size of the bacteria colony is a phenotype of this biological attribute. The role of the studied genes in the cell machinery, the hierarchical interaction types of the genes, as well as the DAG, that describes the latter ones, can be learned by means of knock-out experiments where a gene or a set of genes is functionally disabled and the phenotype is measured, [3], [4], [5]. Based on their single-knockout (SK) and double-knockout (DK) phenotypes, the gene pairs can be classified into one out of five hierarchical relationship classes, according to [3]. From the hierarchical relationship classes the DAG can be inferred [3]. For DAG inference in general, a variety of probabilistic methods have been developed. One prominent example is the Chow-Liu algorithm [6]. However, these methods can only learn the DAGs underlying the data up to Markov equivalence. This means that in many cases only the skeleton of the DAG can be learned but not necessarily the specific orientation of the edges [6]. To reconstruct the DAG with all its edge orientations, a variety of methods based on scoring the measurements or on thresholding the genetic-interaction (GI)-profile data, which is commonly based on Pearson correlation of the SK and DK phenotypes, e.g. [7]-[8] respectively, have been developed. In [5], we have presented the Genetic-Interactions-Detector (GENIE) algorithm which formulates the DAG reconstruction based on the hierarchical relationship classes of [3] as a linear integer program (LIP). However, methods as presented in [5], [7]-[8] have at least one of the following three important disadvantages: D1) poor performance in DAG reconstruction, D2) no easy incorporation of additional side information, and D3) no use of prior knowledge. Although showing the above mentioned disadvantages D1) – D3), methods as presented in [7]-[8] are among most wide-spread algorithms for detecting interactions among a set of genes, i.e., the DAG underlying the different types of data. In this paper, we propose the GI-profile extended GENIE (GI-GENIE) algorithm that is an extension of the GENIE-method of [5] by incorporating GI-profile data into the DAG estimation. Furthermore, the proposed GI-GENIE algorithm is able to easily incorporate prior knowledge. Thus, our proposed method is able to overcome the three main disadvantages summarized in D1) – D3). Finally, we provide statistically stressable statements regarding the topology of the DAG underlying the experimental data of [9] on the yeast microorganism, based on the sequential-scalability (SEQSCA) technique of [4] and the proposed GI-GENIE algorithm.

II. System Model

Given a cell process and a species, the functional dependencies among a set of genes $\mathcal{G} = \{1, ..., G\}$, with $G = |\mathcal{G}|$ elements, can be characterized by a genetic-interaction-map (GI-map), [10], that is essentially a DAG with a common root node called reporter level R. In particular, an arbitrary DAG $\mathcal{D}$ can be described as a graph $\mathcal{D} = (\mathcal{G}_\mathcal{D}, \mathcal{E}_\mathcal{D})$ with a set of nodes $\mathcal{G}_\mathcal{D} = \{\mathcal{G} \cup R\}$ and the set of directed edges $\mathcal{E}_\mathcal{D} = \left\{ \{i, j\}, ..., \{j, l\} \right\}$ [4]. Since genetic interactions can only be observed through the reporter, all edges are always orientated in such a way that each path parting from any arbitrary gene $i \in \mathcal{G}$ always terminates in the root node R and any gene appears on the path at most once, i.e., there exist no cycles in the graph. Hence, DAG $\mathcal{D}$ is always connected via its root node R. The reporter node R is an artificial node, i.e., not

ISBN 978-0-9928626-7-1

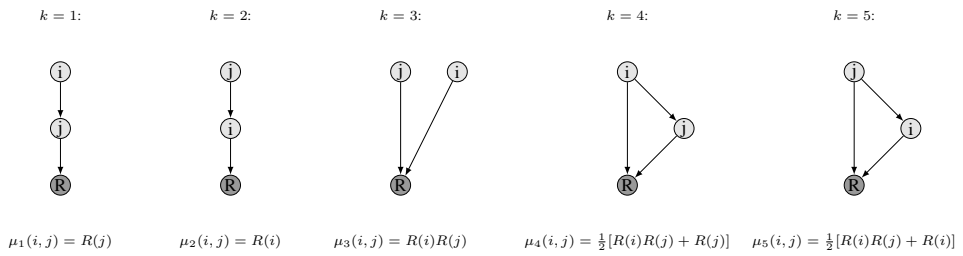

Fig. 2: Possible hierarchical relationship classes between two arbitrary genes i, j of DAG $\mathcal{D}$ according to [3]

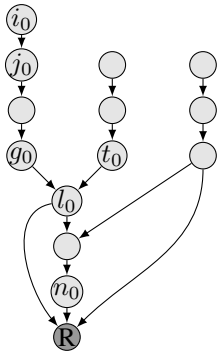

Fig. 1: DAG $\mathcal{D}_0$ of 13 genes and root node R

a gene, in the concept of a DAG representing the measured phenotype of the specific cell process under study. In order to provide a better comprehension of the information encoded in a DAG we give a simple example, similar to that in [4], based on the DAG $\mathcal{D}_0$ displayed in Fig. 1. In $\mathcal{D}_0$ there exists a direct edge from gene i_0 to gene j_0, i.e. $\{i_0, j_0\} \in \mathcal{E}_{\mathcal{D}_0}$, which indicates that the activity of gene i_0 controls the activity of gene j_0. Hence, gene i_0 only affects the phenotype via gene j_0 and not directly. We emphasize that in this model the existence of edge $\{i_0, j_0\}$ in the DAG only describes the hierarchical functional dependency between genes i_0 and j_0 and not the quantitative effect of gene i_0 on gene j_0.

Let us denote $R(i) \in \mathbb{R}$ as the phenotype for a single gene $i \in \mathcal{G}$ functionally disabled. In the same way, we define the phenotype for the DK of genes $i, j \in \mathcal{G}$ as $R(i,j) \in \mathbb{R}$. Let the datasets $\mathcal{R}_i = \{R(i,1), ..., R(i,G)\}$ and $\mathcal{R}_j = \{R(j,1), ..., R(j,G)\}$ contain all DK phenotypes involving genes $i \in \mathcal{G}$ or $j \in \mathcal{G}$. The GI-profile data $\rho(i,j)$ for genes $i, j \in \mathcal{G}$ can be computed as the Pearson correlation between the samples of the datasets $\mathcal{R}_i$ and $\mathcal{R}_j$, respectively. Since the gene pairs i, j and j, i are identical, it is sufficient to consider only gene pairs $i, j \in \mathcal{G} : j > i$. Throughout this paper we mostly omit the specification that j is greater than i for notational convenience. However, we sometimes explicitly state again that only gene pairs $i, j \in \mathcal{G}$ for $j > i$, with their corresponding data types, are considered for clarity of presentation. In genomics research it is a common assumption that given an edge between two genes i, j in DAG $\mathcal{D}$, the GI-profile $\rho(i,j)$ obtains a large value with high probability. Furthermore, according to [3], each pair of genes $i, j \in \mathcal{G} : j > i$ belongs to exactly one out of five hierarchical relationship classes that are characterized in Fig. 2. The hierarchical relationship classes $k \in \mathcal{K} = \{1, ..., 5\}$ of [3] are defined according to the model $\mu_k(i,j)$ in which the single knock-out phenotypes $R(i)$ and $R(j)$ are related with the DK phenotype $R(i,j)$. Given that the gene pair i, j belongs to the hierarchical relationship class k then the observed DK phenotype $R(i,j)$ is described by the model $\mu_k(i,j)$ provided in Fig. 2. We remark that the five hierarchical dependency graphs in Fig. 2 do not reflect the absolute adjacency relations, but the hierarchical relations between genes i, j in DAG $\mathcal{D}$, see [3], [4]. To clarify this further, consider the example DAG $\mathcal{D}_0$ of Fig. 1. All paths from gene i_0 to node R pass through gene j_0, i.e., they are in a linear pathway with gene i_0 upwards of gene j_0. Hence, the pair of genes i_0, j_0 belongs to class $k = 1$. Since all paths from gene i_0 to the reporter level R do not pass through gene t_0 and all paths from gene t_0 to the reporter level do not pass through gene i_0, genes i_0 and t_0 belong to the hierarchical relationship class $k = 3$, i.e., they are independent of each other. With the same line of argument, we can determine the hierarchical relationship class for each pair of genes in DAG $\mathcal{D}_0$. Generally, there are strong dependencies among the hierarchical relationship classes of [3]. If some gene pairs belong to a specific class then this has strong implications for all other pairs, as shown in detail in [4]. Let us consider the case that DAG $\mathcal{D}_0$ was not known and only the hierarchical relationship classes for genes i_0 and j_0, i.e., genes i_0 and j_0 belong to class $k = 1$, as well as the hierarchical relationship class for genes i_0 and g_0, i.e., genes i_0 and g_0 belong to class $k = 1$, were available. By definition of the hierarchical dependency graphs in Fig. 2 and the assumptions, that genes i_0 and j_0 belong to class $k = 1$ as well as that genes i_0 and g_0 belong to class $k = 1$, we conclude that all paths from gene i_0 to R pass through genes j_0 and g_0. Thus, either all paths from gene g_0 to R pass through gene j_0, or vice versa. Consequently, genes j_0 and g_0 either belong to the hierarchical relationship class $k = 1$, or $k = 2$, see [5]. Given the SK/DK phenotypes and the GI-profile data, we can classify the gene pairs i, j to exactly one out of the five hierarchical relationship classes of Fig. 2 [3] and reconstruct the DAG topology jointly. Thus, we can formulate the gene

ISBN 978-0-9928626-7-1 © EURASIP 2017

pair classification and the DAG topology reconstruction jointly as a coupled multi-hypotheses test, which we address in Section 3 by a linear integer programming approach.

III. GI-GENIE-ALGORITHM

In this section, we present the proposed GI-GENIE algorithm which jointly formulates the gene pair classification and the corresponding DAG topology estimation. In order to quantify the mismatch between the measured DK phenotype $R(i,j)$ and the expected phenotype $\mu_k(i,j)$ under the hypothesis that the gene pair i,j belongs to class $k \in \mathcal{K}$ given its respective SK value, we consider a simple quadratic score, [3],

$$s_k(i,j) = \big(R(i,j) - \mu_k(i,j)\big)^2 \; k \in \mathcal{K}, \; \forall i,j \in \mathcal{G} : j > i. \tag{1}$$

Let us define the following class-selection variables

$$\alpha_k(i,j) = \begin{cases} 1 & \text{if } i,j \text{ are in class } k \\ 0 & \text{else} \end{cases} \qquad k \in \mathcal{K}, \quad \forall i,j :\in \mathcal{G} : j > i \tag{2}$$

and edge-selection variables

$$\beta(i,j) = \begin{cases} 1 & \exists \text{ edge between i,j} \\ 0 & \text{no edge} \end{cases} \qquad \forall i,j \in \mathcal{G} : j > i \tag{3}$$

Note that in contrast to the class selection $\alpha_k(i,j) = 1$ for $k \in \mathcal{K}$, the edge selection $\beta(i,j) = 1$ does not capture any directionality, i.e., no hierarchical information about the graph topology. The topology $\mathcal{E}_\mathcal{D}$ of any DAG $\mathcal{D}$ can be represented by the corresponding set of class-selection variables $A^\mathcal{D} = \bigcup_{i,j} \Big\{\alpha_1^\mathcal{D}(i,j), ..., \alpha_5^\mathcal{D}(i,j)\Big\}$ together with the corresponding set of undirected edges $\{\beta(i,j)\}$ for all $i,j :\in \mathcal{G} : j > i$. The GI-GENIE-algorithm yields an estimate $\mathcal{E}_{\text{GI}}$ of the true DAG topology $\mathcal{E}_\mathcal{D}$ by computing sets $\text{A}^{\text{O}_{\text{GI-GENIE}}}$ and $\Big\{\hat{\beta}(i,j)\Big\}$ which are estimates of the true set of class-selection variables and edge-selection variables, $A^\mathcal{D}$, $\{\beta(i,j)\}$, respectively. Based on SK, DK and GI-profile data, the proposed GI-GENIE-algorithm is formulated as the following LIP:

$\text{O}_{\text{GI-GENIE}}$:

$$\begin{aligned} \min_{\{\alpha_k(i,j),\beta(i,j),z_l(i,j)\}} \quad & \lambda_d \sum_{i=1}^{G} \sum_{j=i+1}^{G} \Big(\sum_{k=1}^{|\mathcal{K}|} s_k(i,j)\alpha_k(i,j) \Big) \\ & - \lambda_s \sum_{i=1}^{G} \sum_{j=i+1}^{G} \Big(\sum_{l} z_l(i,j) \Big) \\ & - \lambda_c \sum_{i=1}^{G} \sum_{j=i+1}^{G} \rho(i,j)\beta(i,j) \\ & + \lambda_p \sum_{i=1}^{G} \sum_{j=i+1}^{G} \beta(i,j) \end{aligned} \tag{4a}$$

$$\text{s.t.} \quad \alpha_k(i,j) \in \{0,1\} \; \forall k \in \mathcal{K}, \; \forall i,j \in \mathcal{G} : j > i \tag{4b}$$

$$\sum_{k=1}^{|\mathcal{K}|} \alpha_k(i,j) = 1 \; \forall i,j \in \mathcal{G} : j > i \tag{4c}$$

$$\mathcal{L}, \text{topology constraints of [4]} \tag{4d}$$

$$\beta(i,j) \in \{0,1\} \;\; \forall i,j \in \mathcal{G} : j > i \tag{4e}$$

$$z_l(i,j) \in \{0,1\} \;\; \forall l \in \mathcal{G} \setminus \{i,j\}, \quad \forall i,j \in \mathcal{G} : j > i \tag{4f}$$

$$1 - \alpha_3(i,j) \geq \beta(i,j) \quad \forall i,j \in \mathcal{G} : j > i \tag{4g}$$

$$\mathcal{L}_c \implies \text{additional topology constraints} \tag{4h}$$

$$|\mathcal{G}| - 2 + \beta(i,j) \geq 1 + \sum_{l \in \mathcal{G} \setminus \{i,j\}} z_l(i,j) \quad \forall i,j \in \mathcal{G} : j > i \tag{4i}$$

where the scalars $\lambda_d, \lambda_s, \lambda_c, \lambda_p$ are non-negative weighting constants to balance the impact of the SK, DK measurements and the GI-profile data, respectively, on the estimates. Furthermore, the weighting constants $\lambda_d, \lambda_s, \lambda_c, \lambda_p$ also compensate for the different value domains of the two data types. The functionality of the binary slack variables $z_l(i,j) \, \forall i,j,l \in \mathcal{G} : j > i, l \neq i, l \neq j$ and of the second summand in (4a) will be explained below. Furthermore, the logical implications among the selection variables $\alpha_k(i,j)$ are modeled by the topology constraints $\mathcal{L}$ in Eq. (4d) which directly correspond to the coupling among the hierarchical relationship classes of [3] as mentioned in Section 2. For details, see [4]. Program $\text{O}_{\text{GI-GENIE}}$ can be solved efficiently by branch-and-bound (BB) methods [11]. Note that λ_s is assumed to be a very small non-negative constant, i.e. $0 \leq \lambda_s \ll 1$. The classification-mismatch (CM) term $\lambda_d \sum_{i=1}^{G} \sum_{j=i+1}^{G} \Big(\sum_{k=1}^{|\mathcal{K}|} s_k(i,j)\alpha_k(i,j) \Big)$ in the objective Eq. (4a) seeks to minimize the classification mismatch, while the GI-profile (GIP) term in the objective Eq. (4a), i.e., $-\lambda_c \sum_{i=1}^{G} \sum_{j=i+1}^{G} \rho(i,j)\beta(i,j) + \lambda_p \sum_{i=1}^{G} \sum_{j=i+1}^{G} \beta(i,j)$, strives to detect an edge between genes i,j if the GI-profile $\rho(i,j)$ is larger than the predefined threshold $\frac{\lambda_p}{\lambda_c}$. The joint selection of the set of class-selection variables $\text{A}^{\text{O}_{\text{GI-GENIE}}}$ and the set of edge-selection variables $\Big\{\hat{\beta}(i,j)\Big\}$ is highly coupled. Any pattern of hierarchical relationship classes $\text{A}^{\text{O}_{\text{GI-GENIE}}}$ implies a specific set of edges $\Big\{\hat{\beta}(i,j)\Big\}$ and any set of edges $\Big\{\hat{\beta}(i,j)\Big\}$ implies a specific pattern of hierarchical relationship classes $\text{A}^{\text{O}_{\text{GI-GENIE}}}$. For instance, given that genes i,j are in class three, i.e., $\alpha_3(i,j) = 1$, there cannot be an edge between both genes. Thus, $\beta(i,j) = 0$ as modeled in (4g). To generally account for this coupling, set $\mathcal{L}_c$ contains a plethora of linear integer inequalities that model the mutual logical dependencies between $\text{A}^{\text{O}_{\text{GI-GENIE}}}$ and $\Big\{\hat{\beta}(i,j)\Big\}$. To exemplarily describe this, let us focus on equations (6a) to (6c) as given below:

$$1 - \beta(i,j) \geq \alpha_1(i,j) + \alpha_1(i,l) + \alpha_2(j,l) - 2 \tag{6a}$$

$$1 - z_l(i,j) \geq \alpha_1(i,j) + \alpha_1(i,l) + \alpha_2(j,l) - 2 \tag{6b}$$

ISBN 978-0-9928626-7-1 © EURASIP 2017

$$\frac{1}{2}\left(\alpha_1(i,l)+\alpha_2(j,l)\right) \geq \alpha_1(i,j) - z_l(i,j) \qquad \text{(6c)}$$

where we assume in the following that $i,j,l \in \mathcal{G} : l > j > i$. Given that $\alpha_1(i,j) = 1$ and $\alpha_1(i,l) + \alpha_2(j,l) = 2$ for at least one gene l, there is gene l situated between genes i and j in DAG $\mathcal{D}$. Hence, there cannot be an edge between genes i,j, i.e., $\beta(i,j) = 0$. This is strictly enforced by the constraint in Eq. (6a). Similarly, given that $\alpha_1(i,j) = 1$ again, but now assume that $\alpha_1(i,l) + \alpha_2(j,l) \leq 1$, then there is no gene l between genes i,j in DAG $\mathcal{D}$, so that $\beta(i,j) = 1$ must hold. To see this, the right-hand-side (RHS) of Eq. (6b) is in this case equal to or less than 0, so the slack variables $z_l(i,j)\ \forall l \in \mathcal{G} \setminus i,j$ are set to 1 by the slack term $-\lambda_s \sum_{i=1}^{G}\sum_{j=i+1}^{G}\left(\sum_l z_l(i,j)\right)$ in the objective function in (4a) and finally constraint (4i) enforces that $\beta(i,j) = 1$. Now, assume that $\beta(i,j) = 1$ and $\alpha_1(i,j) = 1$, i.e., there is an edge between genes i,j. Then the RHS of Eq. (6a) and Eq. (6b) must be less than 0. In this case $\alpha_1(i,l)+\alpha_2(j,l) \leq 1$, hence there cannot be a gene l between genes i,j in DAG $\mathcal{D}$. Finally, assume, for example, that $\beta(i,j) = 0$ and $\alpha_1(i,j) = 1$, i.e., there is no edge between genes i,j. Then there must be at least one gene l which is situated between genes i,j in DAG $\mathcal{D}$. Since $\beta(i,j) = 0$, the RHS of Eq. (4i) forces at least one slack variable $z_l(i,j)$ to be 0 meaning that the corresponding gene l is situated between genes i,j. To see that gene l is between genes i,j, i.e., $\alpha_1(i,l) + \alpha_2(j,l) = 2$, draw your attention to Eq.(6c). Since the RHS of Eq.(6c) amounts to 1 in this case, the left-hand-side (LHS) must also be 1, hence $\alpha_1(i,l)+\alpha_2(j,l) = 2$ holds. However, there exist many more logical coupling constraints between the set of hierarchical relationship classes $A^{\mathcal{D}}$ and the set of edge selection variables $\{\beta(i,j)\}_{\forall i,j\in\mathcal{G}:j>i}$ that have been omitted due to space limitations. Furthermore, the cases $j > i > l : l,j,i \in \mathcal{G}$ and $j > l > i : l,j,i \in \mathcal{G}$ have to be considered as well. For a detailed explanation, see [4]. Finally we remark that prior knowledge regarding the interaction between genes i,j, e.g., obtained from gene databases, can be easily incorporated into the GI-GENIE-method by, for instance, setting the corresponding edge variable $\beta(i,j)$ to zero or one. We obtain an estimated set $\mathcal{E}_{\text{GI}}$ of the true topology $\mathcal{E}_{\mathcal{D}}$ of DAG $\mathcal{D}$ based on the computed set of edge selection variables $\left\{\hat{\beta}(i,j)\right\}$ of program $\text{O}_{\text{GI-GENIE}}$ in (4) where we infer the directionality of the edges according to $\text{A}^{\text{O}_{\text{GI-GENIE}}}$.

IV. Real Data Results

Since discovering genetic interaction maps, i.e., DAGs, for specific organisms is an ongoing field of research and the knowledge on genetic interactions is far away from being complete, there is generally no *ground-truth* to directly compare with, even not for yeast which is one of the best understood organisms. Therefore, we base the evaluation of the detection performance of the GI-GENIE method on the biological knowledge that genetic interactions are generally rare and furthermore on the successful detection of known interactions provided by the well known *yeast database* of [15]. We remark that to be able to make statistically significant statements about large sets of genes, we have applied the proposed GI-GENIE algorithm along with the SEQSCA-technique from [4]. The SEQSCA-technique firstly decomposes a large set of genes $\mathcal{G}$ into a sequence of S small subsets $\mathcal{G}_s$, with $|\mathcal{G}_s| = N_{\text{S}}$. The topology $\mathcal{E}_s$ of the DAG underlying each subset $\mathcal{G}_s$ is computed by a DAG-estimation algorithm, e.g., the GI-GENIE algorithm. In each iteration s of the SEQSCA, the adjacency matrix $\boldsymbol{A}_s$ is inferred from $\mathcal{E}_s$ and used to update the reliability matrix $\boldsymbol{M}^{(s)}$. After S iterations the SEQSCA-method yields the averaged reliability matrix $\boldsymbol{M} \in (0,1)^{|\mathcal{G}|\times|\mathcal{G}|}$ where each entry $[\boldsymbol{M}]_{i,j\in\mathcal{G}}$ denotes the empirical probability that genes i,j interact with each other that is computed from the sequence of reliability matrices $\boldsymbol{M}^{(1)},...,\boldsymbol{M}^{(S)}$. The SEQSCA is summarized in Tab. I, for details see [4]. To demonstrate

Initialization: $\boldsymbol{M}^{(0)} = \boldsymbol{0}_{N\times N}$; $\boldsymbol{A}_0 = \boldsymbol{0}_{N_S\times N_S}$; frequency counter $n_{i,j}^{(0)} = 0$

Repeat:

1 : Select subset $\mathcal{G}_s$ of size N_S from $\mathcal{G}$; draw each gene from $\mathcal{G}$ with equal probability without replacement
2 : Frequency update: $n_{i,j}^{(s+1)} = n_{i,j}^{(s)} + 1$ for all $i,j \in \mathcal{G}_s$
3 : Estimate DAG topology $\mathcal{E}_s$ of set $\mathcal{G}_s$; $\Longrightarrow \boldsymbol{A}_s$
4 : Update reliability matrix $\boldsymbol{M}^{(s)}$: $\left[\boldsymbol{M}^{(s+1)}\right]_{i,j} = \left[\boldsymbol{M}^{(s)}\right]_{i,j} + [\boldsymbol{A}_s]_{\kappa_i,\kappa_j}, \forall i,j \in \mathcal{G}_s, \quad \kappa_i \in \{1,...,N_S\}\ \forall i \in \mathcal{G}_s$

Until: $s = S$;
Set $[\boldsymbol{M}]_{i,j} = \left[\boldsymbol{M}^{(S)}\right]_{i,j} / n_{i,j}^{(S)}\ \forall i,j \in \mathcal{G}$

TABLE I: Summary of the proposed SEQSCA-algorithm

the benefit of using multiple data-types instead of only one data type, we compare the reliability matrix results obtained from SEQSCA and GI-GENIE with those obtained from SEQSCA and GENIE [5] which only utilizes SK/DK data. We have applied the above mentioned algorithms to the data set reported in [9] to obtain the reliability matrices for the GENIE based SEQSCA as well as for the GI-GENIE based SEQSCA, $\boldsymbol{M}_{\text{G}}$, $\boldsymbol{M}_{\text{GI}}$, respectively. For computational reasons, we only considered the first 200 genes, i.e., $|\mathcal{G}| = 200$, of the *query genes list* of [9]. Fig. 3 shows $\boldsymbol{M}_{\text{G}}$ obtained by the GENIE-based SEQSCA. In Fig. 4 we have displayed $\boldsymbol{M}_{\text{GI}}$ obtained by the proposed GI-GENIE-based SEQSCA. For both results, we decomposed $\mathcal{G}$ into a sequence of $S = 5e4$ subsets $\mathcal{G}_s$ of equal size $N_s = 10$. In Fig. 3, we observe that 78% of the gene pairs i,j considered by $\boldsymbol{M}_{\text{G}}$ of the GENIE-based SEQSCA of [4] interact with each other with an emperical probability of less than 20%, i.e., $[\boldsymbol{M}_{\text{G}}]_{i,j} \leq 0.2$. Hence, the GENIE-based SEQSCA of [4] yields approximately sparse results. This is a good performance in terms of sparsity, since it is known from biology that genetic interactions are generally very rare. Furthermore, we observe from the reliability matrix $\boldsymbol{M}_{\text{GI}}$ that the proposed GI-GENIE algorithm predicts genetic interactions with a much lower frequency which means a very good performance in terms of sparsity. We have computed the

ISBN 978-0-9928626-7-1 © EURASIP 2017

Method:	Γ
SEQSCA & GENIE	53%
SEQSCA & GI-GENIE	74%

TABLE II: Acceptance ratios; $\epsilon = 0.05$

acceptance ratio

$$\Gamma = \frac{N_{\mathrm{r}}}{N_{\mathrm{t}}} \tag{8}$$

where N_{r} is the number of interactions found with high significance ($[\boldsymbol{M}_{\mathrm{G}}]_{i,j}, [\boldsymbol{M}_{\mathrm{GI}}]_{i,j} \geq 1 - \epsilon$) and which are deposited in the data-base of [15] as well, that is used as a performance measure to assess the detection quality. N_{t} is the total number of highly significant interactions. Given the confirmed interactions at [15] for our set of genes under study, we remark that evaluating the number of confirmed interactions, that we have also found with our proposed method, would not be a reasonable performance metric, since [15] combines knowledge and experimental results of numerous sources. In contrast to that, our results only use the dataset of [9] based on colony size measurements of yeast which may not reflect all existing interactions. As depicted in Tab. II, we have computed Γ in Eq. (8) for both, the GI-GENIE-based SEQSCA and the GENIE-based SEQSCA. It is obvious that the GI-GENIE-based SEQCA outperforms the GENIE-based SEQSCA of [4], since the acceptance ration for the GI-GENIE-based SEQSCA is significantly higher than the one of the GENIE-based SEQSCA.

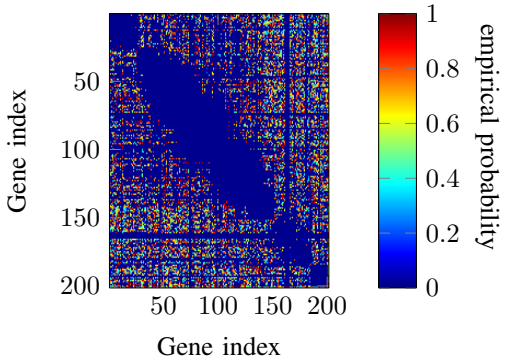

Fig. 3: Reliability matrix $\boldsymbol{M}_{\mathrm{G}}$; $S = 50000$ subsets considered; subset size $N_S = 10$

V. Conclusion

In this paper we have presented the GI-GENIE algorithm which detects the topology of the DAG underlying the observed SK and DK phenotypes, as well as additional side information, i.e., GI-profile data. Due to the use of multiple data types, the proposed GI-GENIE algorithm outperforms comparable algorithms as presented in [5] which can only make use of SK/DK data.

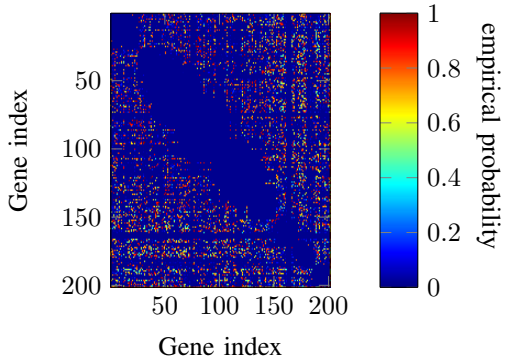

Fig. 4: Reliability matrix $\boldsymbol{M}_{\mathrm{GI}}$; $S = 50000$ subsets considered; subset size $N_S = 10$; $\lambda_d = 1e3$, $\lambda_s = 8e{-}6$, $\lambda_c = 1$, $\lambda_p = .85$

References

[1] A.H.Y. Tong et al. *Systematic genetic analysis with ordered arrays of yeast deletion mutants*, Science 294, 2001

[2] A. Shojaie, G. Michailidis *Discovering graphical Granger causality using the truncating lasso penalty*, Department of Statistics, University of Michigan, ECCB 2010, Vol. 26, 2010

[3] A. Battle, M.C. Jonikas, P. Walter, J.S. Weissman and D. Koller *Automated identification of pathways from quantitative genetic interaction data*, Molecular Systems Biology 6, 2010

[4] F. Nikolay, M. Pesavento, G. Kritikos, N. Typas *Learning Directed-Acyclic-Graphs from Large-Scale Genomics Data*, Communications System Group, TU Darmstadt, arXiv: *http://arxiv.org/abs/1609.02794*

[5] F. Nikolay, M. Pesavento *Learning Directed-Acyclic-Graphs from Large-Scale Double-Knockout Experiments*, C, Communications System Group, TU Darmstadt, EUSIPCO 2016, Budapest, 2016

[6] K.P. Murphy *Machine Learning - A Probabilistic Perspective*, The MIT Press, Camebridge, Massachusetts, ISBN 978-0-262-01802-9, 2012

[7] B. Snijder, P. Liberali, M. Frechin, T. Stoeger and L. Pelkmans, *Predicting functional gene interactions with the hierarchical interaction score*, Nature Methods, Vol. 10, November 2013

[8] S.J. Dixon, M. Constanzo, A. Baryshinkova, B. Andrews, C. Boone *Systematic Mapping of Genetic Interaction Networks*, The Annual Review of Genetics, 2009

[9] M. Costanzo et al. *DRYGIN - Data Repository of Yeast Genetic Interactions*, Terence Donnelly Centre for Cellular and Biochemical Research, University of Toronto, http://drygin.ccbr.utoronto.ca/ costanzo2009/, 2009

[10] A. Jaimovich et al. *Modularity and directionality in genetic interaction maps*, Nature Methods, Vol. 26, 2010

[11] V. Balakrishnan, S. Boyd, S. Balemi *Branch and bound algorithm for computing the minimum stability degree of parameter-dependent linear systems*, International Journal of Robust and Nonlinear Control,1(4):295317, December 1991

[12] R. Diestel *Graphentheorie*, Springer-Verlag, Heidelberg, ISBN 978-3-642-14911-5, 2012

[13] C.H. Papadimitriou, K. Steiglitz *Combinatorial optimization: algorithms and complexity*, Mineola NY, ISBN 0486402584, 1998

[14] M. Babu et al. *Quantitative Genome-Wide Genetic Interaction Screens Reveal Global Epistatic Relationships of Protein Complexes in Escherichia coli*, PLOS Genetics, Vol. 10, February 2014

[15] *SGD - Saccharomyces Genome Database*, http://www.yeastgenome.org

ISBN 978-0-9928626-7-1 © EURASIP 2017