Learning Directed-Acyclic-Graphs from Large-Scale Double-Knockout Experiments

Fabio Nikolay, Marius Pesavento
Communication Systems Group, Technische Universität Darmstadt, Germany
Email: \{nikolay, pesavento\}@nt.tu-darmstadt.de

Abstract—In this paper we consider the problem of learning the genetic-interaction-map, i.e., the topology of a directed acyclic graph (DAG) of genetic interactions from noisy double knockout (DK) data. Based on a set of well established biological interaction models we detect and classify the interactions between genes. Furthermore, we propose a novel linear integer optimization framework called Genetic-Interactions-Detector (GENIE) to identify the complex biological dependencies between genes and to compute the DAG topology that matches the DK measurements best where we make use of the well known branch-and-bound (BB) principle. Finally, we show via numeric simulations that the GENIE framework clearly outperforms the conventional techniques.

Index Terms—Genetic interactions analysis, large scale gene networks, discrete optimization, big data

I. INTRODUCTION

Genetic interaction analysis aims at uncovering the interactions among a set of genes with respect to a specified cell function of a biological system, e.g., bacteria. The interactions among the genes under study are well described by a directed-acyclic-graph (DAG) where the hierarchical relationship between two genes of a DAG describe their hierarchical interaction type. However DAGs cannot be observed directly but only the specified cell function under study which yields observable phenotypes. The role of the studied genes in the cell machinery, the hierarchical interaction types of the genes, as well as the DAG, which describes the latter ones, can only be learned by means of knock-out experiments where a gene or a set of genes is functionally switched off and the phenotype is observed. Traditionally, only single-knockout (SK) experiments have been conducted but those mainly allow for a statement about the importance of a single gene for the investigated cell process and do not convey much information about the interaction among the genes under study. Recently with the technological advances in micro arrays and the development of the synthetic-genetic-array technologies [7] new approaches have been taken that are based on large scale knock-out experiments of pairs of genes. Such double knock-out (DK) experiments are much more powerful for exploring genetic interactions since a DK phenotype of an arbitrary pair of genes generally differs considerably from the superposition of the corresponding SK phenotypes of this pair of genes. According to [6], the gene pairs can be classified to one out of five hierarchical relationship classes based on their SK and DK phenotypes. Further, based on the hierarchical relationship classes the DAG underlying the observed SK and DK phenotypes can be computed which directly reflects the genetic interactions among the genes. In order to detect the DAG underlying the SK and DK phenotypes a variety of statistical methods based on scoring the measurements or on Pearson correlation of the genetic-interaction (GI)-profile, e.g. [1]-[5] respectively, have been developed. Although showing poor performance, methods as those presented in [1]-[5] are the most commonly used methods to detect the DAG underlying the measured DK data. In contrast to that, we propose an approach that is based on the biological system model of [6]. Since the hierarchical relationship classes are mutually dependent classifying each pair of genes to a specific hierarchical relationship class corresponds to a multi-hypotheses. Thus, we formulate this multi-hypotheses test as a linear integer optimization program in order to find the set of hierarchical relationship classes, best matching the observed SK and DK phenotypes. Based on the detected set of hierarchical relationship classes, the set of edges of the DAG which reflects the interactions among the genes can be computed.

This paper is organized as follows. We first summarize the biological system model of [6], present in Section 3 the Genetic-Interactions-Detector (GENIE)-framework of detecting the set of hierarchical relationship classes, that represents a valid DAG and matches the DK measurements best, and continue with presenting numerical results which demonstrate the performance of our algorithm. Finally, we summarize the key parts of this paper and give a brief outlook on future work.

II. SYSTEM MODEL

In this section we first provide a mathematical description of a DAG as well as its biological implications. Furthermore, we introduce the common biological terms and follow with the model of [6] where we provide an intuitive interpretation of how to read and interpret a DAG of a genetic interaction map.

The functional dependencies among a set of genes \( G = \{g_1, ..., g_G\} \), with \( G = |G| \) elements, for a given cell process and specie can be well described by a genetic-interaction-map (GI-map,[9],[10]) which is essentially a DAG with a common root node, i.e., the reporter level \( R \). In particular, an arbitrary DAG \( D \) can be described as a graph \( D = (G_D, \mathcal{E}_D) \) with the set of nodes \( G_D = \{G \cup R\} \) and the set of directed edges \( \mathcal{E}_D = \{\{g_i, g_j\}, ..., \{g_j, g_l\}\} \) where edge \( \{g_i, g_j\} \) denotes an edge from gene \( g_i \) to gene \( g_j \). As the genetic interactions can
 describes the functional dependency between genes in this model. The existence of edge \( i \rightarrow j \) implies that the phenotype via gene \( j \). Let us consider the case that the gene pair \( i, j \in G \), \( j > i \) belongs to the hierarchical relationship class \( k \) then the observed DK phenotype \( R(i, j) \) is described by the model \( \mu_k(i, j) \) of 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 \( D \). To clarify this further, let us consider the example DAG \( D_0 \) 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 \( j_0 \). Thus the pair of genes \( i_0, j_0 \) belongs to class \( k = 1 \). Note that with the same line of argument, we conclude that also genes \( i_0 \) and \( j_0 \) belong to relationship class \( k = 1 \). Since all paths from gene \( i_0 \) to the reporter level \( R \) do not pass through gene \( t_0 \) and vice versa, genes \( i_0 \) and \( t_0 \) belong to the hierarchical relationship class \( k = 3 \) as given in Fig. 2, which states that genes \( i_0 \) and \( t_0 \) are independent of each other and the DK phenotype amounts to \( R(i_0, t_0) = \mu_3(i_0, t_0) \). Finally, let us inspect the structural relation between genes \( t_0 \) and \( n_0 \) in DAG \( D_0 \). It is obvious that gene \( t_0 \) has (at least) one path to node \( R \) which does not pass through gene \( n_0 \), i.e., genes only having paths to \( R \) that do not pass through gene \( n_0 \) do not affect the activity of gene \( n_0 \). Since there is (at least) one other path from gene \( t_0 \) to \( R \) passing through gene \( n_0 \), we can conclude that genes \( t_0 \) and \( n_0 \) belong to class \( k = 4 \). Generally, there are strong implications among the hierarchical relationship classes of [6], i.e., if some pairs belong to a specific class then this has strong implications for all other pairs. Let us consider the case that DAG \( D_0 \) 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, 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 \( k = 2 \). As we have emphasized by the example above, generally if the hierarchical relationship class is known for two arbitrary genes \( i, j \in G : j > i \) as well as for another pair \( i, l \in G : l > i \), then this has strong logical implications on the hierarchical relationship classes genes \( j, l \in G : l > j \) can belong to. Since we can interpret the classification of the pairs of genes \( i, j \in G : j > i \), based on their observed SK and DK phenotypes \( R(i), R(j) \) and \( R(i, j) \), respectively, to exactly one out of the five hierarchical relationship classes as a coupled multi-hypotheses test, we address this problem in the next section by a linear integer optimization program. The proposed linear integer optimization program identifies the set of hierarchical relationship classes, that represents a valid DAG and matches the DK measurements best with respect to the logical coupling between the classes.

III. GENIE-FRAMEWORK

In this section, we formulate the problem of classifying the gene pairs \( i, j \in G : j > i \) into the classes of hierarchical relationships based on the observed SK and DK phenotype values as a linear integer optimization program. Furthermore, we translate the logical implications among the hierarchical relationship classes into constraints that ensure that the detected set of hierarchical relationship classes represents a valid graph. That is the detected set of hierarchical relationship classes represents a graph which is a DAG as defined in Section 2. In order to quantify the mismatch between the measured DK phenotypes \( R(i, j) \) and the expected phenotype \( \mu_k(j, i) \) under the hypothesis that the gene pairs \( i, j \in G : j > i \) belong to class \( k \in K \) given their respective SK values, we propose a simple quadratic score \([6]\) as given in Eq. (1):

\[
s_k(i,j) = \left( R(i,j) - \mu_k(i,j) \right)^2 \quad k \in K, \quad \forall i,j \in G : j > i
\]  

Let us define the following selection variables

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

Then the GENIE problem of classifying the gene pairs \( i, j \in G : j > i \) into the set of hierarchical relationship classes, that represents a valid DAG and matches the DK measurements best can be formulated as

\[
O_{\text{GENIE}} : \quad \min \{\alpha_k(i,j)\} \quad \sum_{i=1}^{G} \sum_{j=i+1}^{G} \sum_{k=1}^{[K]} s_k(i,j)\alpha_k(i,j) \] (3a)

s. t. \( \alpha_k(i,j) \in \{0,1\} \quad \forall k \in K, \forall i,j \in G : j > i \) (3b)

\[
\sum_{k=1}^{[K]} \alpha_k(i,j) = 1, \quad \forall i,j \in G : j > i
\] (3c)

\[L \implies \text{additional topology constraints} \] (3d)

where \( A^{O_{\text{GENIE}}} \) \( \forall i,j \in G : j > i \) denotes the solution to program \( O_{\text{GENIE}} \) and the set of best matching selection variables \( A^{O_{\text{GENIE}}} \) corresponds to the most consistent pattern of hierarchical relationship classes. Program \( O_{\text{GENIE}} \) is a linear integer program which can be solved efficiently by BB-methods [11]. The objective of program \( O_{\text{GENIE}} \) is to minimize the overall mismatch in classifying each gene pair \( i, j \in G : j > i \) to one out of five hierarchical relationship classes. The constraints in (3b) reflect the binary nature of the selection variables, while (3c) represents a multiple choice constraint that enforces the gene pairs \( i, j \) to select only one hierarchical relationship class.

The set \( L \) in (3d) contains additional constraints to ensure that the detected set of selection variables \( A^{O_{\text{GENIE}}} \) always represents a valid graph, i.e., a DAG. In the following we exemplarily derive topology constraints in set \( L \). In order to identify the numerous logical implications among the selection variables \( \alpha_k(i,j) \), \( k \in K \) for all \( i,j \in G : j > i \) we proceed in the following way. We first fix the assumption that genes \( i,j \in G : j > i \) belong to class \( k = 1 \), i.e., \( \alpha_1(i,j) = 1 \). Further we assume that genes \( i,l \in G : l > i \) belong to class \( k' \), i.e., \( \alpha_{k'}(i,l) = 1 \). Then we derive the set of classes \( K'' \) that genes \( j,l \in G : l > j \) can belong to under the assumptions made. In the following, we have formulated the logical implications among the selection variables for \( \alpha_1(i,j) = 1 \) as linear integer inequalities defined in constraints (4a)-(4e) and summarize them as set \( L_1 \)

\[
L_1 = \begin{cases} 
\alpha_1(j,l) + \alpha_2(j,l) \geq \alpha_1(i,j) + \alpha_1(i,l) - 1 & (4a) \\
\alpha_2(j,l) \geq \alpha_1(i,j) + \alpha_2(i,l) - 1 & (4b) \\
\alpha_2(j,l) + \alpha_3(j,l) + \alpha_5(j,l) \geq \alpha_1(i,j) + \alpha_3(i,l) - 1 & (4c) \\
\alpha_2(j,l) + \alpha_4(j,l) \geq \alpha_1(i,j) + \alpha_4(i,l) - 1 & (4d) \\
\alpha_5(j,l) + \alpha_2(j,l) \geq \alpha_1(i,j) + \alpha_5(i,l) - 1 & (4e) \\
\end{cases} \quad \forall i,j,l \in G : l > j > i
\]  

where constraints (4a)-(4e) are convex after the continuous relaxation of the selection variables \( \alpha_k(i,j) \), \( \forall i,j \in G : j > i \). To explain the origin and the functionality of the constraints in \( L_1 \), let us further define a sub-genetic-interactions-map
(SMAP) $S$ as given in Fig. 3 according to the following definition where we adopt the graph notation of [8]:

**Definition:**
Given a non-empty set of edges $E_{\text{in}}$ and a non-empty set of edges $E_{\text{out}}$. Graph $S = (G_S, E_S)$, with set of nodes $G_S$ and set of edges $E_S$, is a SMAP if the following conditions are fulfilled: i) the graph $S$ is acyclic and directed ii) there exists an edge $e_{\text{in}} \in E_{\text{in}}, e_{\text{out}} \in E_{\text{out}}$ such that each path $P$ through graph $S$ incides $S$ via edge $e_{\text{in}}$ and leaves graph $S$ via edge $e_{\text{out}}$.

The DAG $D_1$, as displayed on the left-hand-side (LHS) Fig. 4, consists of genes $i, j$ and SMAPS $S_1$ and $S_2$. It is obvious that genes $i, j$ belong to class $k = 1$, i.e., $\alpha_1(i, j) = 1$. Furthermore, all genes $l \in G_{D_1} \setminus \{R\} : l > j > i$ for which $\alpha_1(i, l) = 1$ must be either located in SMAP $S_1$ or $S_2$. Hence it follows from DAG $D_1$ in Fig. 4 that the gene pair $j, l$ is either in hierarchical relationship class $k = 1$ or $k = 2$, i.e., $\alpha_1(j, l) = 1$ or $\alpha_2(j, l) = 1$.

This logical implication is directly reflected by constraint (4a). Given $\alpha_1(i, j) = 1$ and $\alpha_1(i, l) = 1$ the right-hand-side (RHS) of (4a) amounts to 1. In this case also the (LHS) of (4a) becomes 1 to fulfill the inequality (4a). Thus either $\alpha_1(j, l) = 1$ or $\alpha_2(j, l) = 1$. Reversely, assume that $\alpha_1(i, j) = 1$ and $\alpha_1(i, l) = 1$ does not hold, then the RHS of (4a) is less than 1, i.e., 0 or -1, while the LHS of (4a) is always greater than 0. Hence, constraint (4a) is fulfilled irrespectively of the choice of $\alpha_k(j, l)$, i.e., constraint (4a) enforces no logical implications.

Similarly for DAG $D_2$ of the RHS of Fig. 4, it is obvious that genes $i, j$ belong to the hierarchical relationship class $k = 1$, i.e., $\alpha_1(i, j) = 1$. All genes $l \in G_{D_2} \setminus \{R\} : l > j > i$ which are in a linear pathway upstream of gene $i$, i.e., $\alpha_2(i, l) = 1$, must be located in SMAP $S_3$. Hence it directly follows from DAG $D_2$ that also gene $l$ must be in a linear pathway upstream of gene $j$, i.e., $\alpha_2(j, l) = 1$. This logical implication is compactly represented in constraint (4b). Under the assumption that $\alpha_1(i, j) = 1$ and $\alpha_2(i, l) = 1$, the RHS of (4b) amounts to 1 enforcing $\alpha_2(j, l) = 1$, so that the LHS of (4b) equals the RHS and the inequality in (4b) is fulfilled. Reversely, assume that $\alpha_2(i, l) = 0$, then the RHS of (4b) is less than 1 and hence the LHS of (4b) is always bigger than or equal to the RHS irrespectively of the choice of $\alpha_k(j, l)$, i.e., constraint (4a) enforces no logical implications. We can proceed in the same fashion to explain constraints (4c)-(4e). Furthermore, with the same line of argument we can derive the sets $L_k$ for $k \in K \setminus 1$ which reflect the logical implications among the selection variables under the assumptions that $\alpha_k(i, j) = 1$ for $k \in K \setminus 1$. However, due to space limitations we omit the derivation of the full set of logical implications and refer the interested reader to a future journal version of this paper. The full set of topology constraints $L$ in (3d) can be computed as

$$L = \bigcup_{k=1}^{|K|} \{L_k\}.$$  

The totality of the computed selection variables, i.e., the set $A^{O_{\text{GENIE}}}$, represents DAG $D$ underlying the observed SK and DK phenotypes in a different domain. It can be theoretically proven that this representation is not unique, i.e., the computed selection variable pattern $A^{O_{\text{GENIE}}}$ represents not only the true DAG $D$, but also a set of similar DAGs which are in line with the pattern $A^{O_{\text{GENIE}}}$, but have a minorly different set of edges as the true DAG $D$. Based on the detected set of selection variables, an estimate $\hat{E}_D$ of the true set of edges $E_D$ of DAG $D$ can be computed.

**IV. Simulation Results**

We have generated the ideal SK phenotypes $R(i) \in \mathbb{R}$ for all $i \in G$ as well as the ideal DK phenotypes $R(i, j) \in \mathbb{R}$ for all $i, j \in G : j > i$ according to the model of [6]. We compare our method with the well known GI-profile similarity approach ([12],[6]), where the Pearson correlation between the GI-profiles of genes $i$ and $j$ is computed and an edge in the DAG is detected if the Pearson correlation is above a pre-defined threshold $\ell_{\text{corr}}$, where the directionality is inferred from the selection variable $\alpha_k(i, j)$ corresponding to the least mismatch scoring hierarchical relationship class based...
on the SK and DK phenotypes $R(i)$ and $R(i,j)$ respectively. For the simulations, we consider a total of 10 genes in order to limit the Monte-Carlo simulation time. In Fig. 5 we display the percentage of erroneously detected edges $D_{ed}$ in the detected DAG normalized to the true number of edges $|\mathcal{E}_D|$ as defined in Eq. (6) versus the SNR.

$$D_{ed} = \frac{|\mathcal{E}_D \cup \hat{\mathcal{E}}_D| \setminus \mathcal{E}_D}{|\mathcal{E}_D|}$$

(6)

In Fig. 5 we observe that in the low SNR regime, the Pearson correlation based method performs best in terms of erroneously detected edges, however it fails to improve performance with increasing SNR, because for correct directionality information of the edges this approach relies on the hierarchical relationship classes detected by method $O_{GENIE}$ without considering $\mathcal{L}$. In the high SNR regime, the proposed GENIE-approach clearly outperforms the Pearson correlation based method as well as the method which simply detects the highest scoring class regardless of their coupling with other classes. In Fig. 6 we display the percentage of missing edges $D_{mis}$ in the detected DAG normalized to the true number of edges $|\mathcal{E}_D|$ as defined in Eq. (7) versus the SNR, i.e.,

$$D_{mis} = \frac{|\mathcal{E}_D \cup \hat{\mathcal{E}}_D| \setminus \hat{\mathcal{E}}_D}{|\mathcal{E}_D|}$$

(7)

In Fig. 6 we observe that our proposed GENIE-approach clearly outperforms the competing methods over the entire SNR region. Particularly, the Pearson correlation based method shows poor performance.

![Figure 5: $D_{ed}$ versus SNR; $t_{corr} = .8$; 200 Monte Carlo runs](image1.png)

![Figure 6: $D_{mis}$ versus SNR; $t_{corr} = .8$; 200 Monte Carlo runs](image2.png)

V. CONCLUSION AND OUTLOOK

In this paper, we have demonstrated the benefits of the GENIE-framework over conventional techniques for detecting the topology of the DAG which is underlying the SK and DK data measured. Especially in the high SNR-regime, our proposed algorithm outperforms the conventional techniques on simulated data. Future work may consider a robust modification of the presented GENIE-approach with respect to noise as well as an incorporation of correlation data into the GENIE-framework in order to enhance the reliability of the detection results even further.

ACKNOWLEDGMENT

The authors would like to thank George Kritikos and Nassos Typas from the European Molecular Biology Laboratory for their help and support.

REFERENCES


