Causal Discovery on Flow Cytometry Data

Julian Schmocker

18 12 2019

1 Introduction

In this project, various methods for causal inference were applied on a dataset from the paper *Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data* (1), where they were able to almost reconstruct a known causal signaling pathway using Bayesian networks.

Initially, the aim of this project was to reproduce the analysis from the paper *Predicting Causal Relationships from Biological Data: Applying Automated Cytometry Data of Human Immune Cells* (2), where they tried to discover novel causal relationships from a large collection of public mass cytometry data of immune cells perturbed with a variety of compounds. Unfortunately, it was not possible to download the data from (2) because the Cytobank servers, on which the data was stored, were down. This is why the methods for causal inference were tested on the Sachs, et al. data.

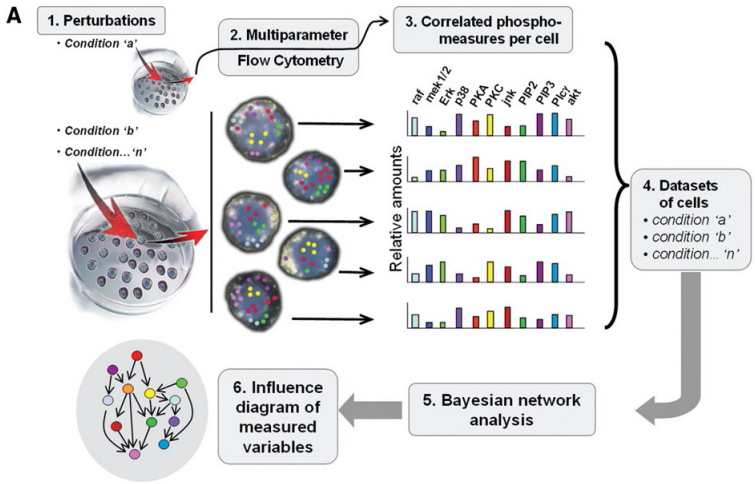

Figure 1: This figure demonstrates how the approach from (1) works. There were 9 different perturbation conditions applied to a set of individual cells. Then the expression levels of different proteins were recorded with a multiparameter flow cytometer. With a Bayesian network analysis they tried to detect the structure of the signaling pathway. Figure from Sachs, et al. (1).

1.1 Biological Background

Signaling pathways represent a cascade of information flow and they play an important part in cellular communication. The cascade is triggered by a signal, which modifies a signaling molecule's structure (chemically or physically) or changes the location of the molecule. This reaction then has an effect other molecules in the cascade. Such a reaction can be seen as a causal effect from one molecule to other molecules.

Often, scientists just study individual pathways. In those situations only one protein was perturbed and then the effects of this perturbation were studied. This approach has the disadvantage that interpathway cross-talk and other properties of networks can not be examined. With the network based causal inference methods, which were applied in this project, it is possible to look simultaneously at several pathways.

Figure 2 illustrates which were the conventionally accepted connenctions of the signaling molecules (via arcs) before publication of the paper. It also shows which protein is influenced by the stimulatory cues (green) or the inhibitory interventions (red).

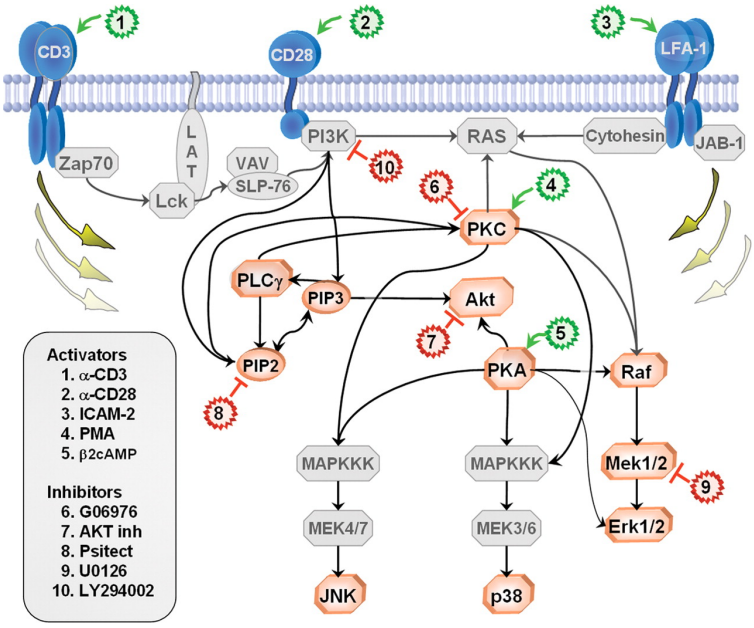

Figure 2: The nodes in color were measured directly. There are no measurements in the data for the gray nodes. Figure from Sachs, et al. (1).

Table 1 gives an overview of the reagents that were used. The effect of the reagent is shortened and simplified.

Table 1: Overview stimulations

Reagent	No.in.Fig2	Reagent.class	Effect.of.Reagent
Anti-CD2/CD28	1,2	General	Activates T cells and induces proliferation and cytokine production
ICAM-2	3	General	Induces LFA-1 signaling and contributes to CD3/CD28 signaling
β2camp	5	Specific	Activates PKA
AKT inhibitor	7	Specific	Inhibits AKT
U0126	9	Specific	Inhibits MEK1/2
PMA	4	Specific	Activates PKC and initiates some aspects of T cell activation
G0076	6	Specific	Inhibits phosphoinositide hydrolysis and PIP2 production
Psitectorigenin	8	Specific	Inhibits PIP2 production
LY294002	10	Specific	PI3K (not measured) inhibitor

For each method that will be applied, we will get a predicted causal graph. The result will be compared with the network that was inferred in Sachs et al. (2005). In their resulting network 15 of the 17 predicted arcs are well-established in the literature. Two of the predicted arcs are not well known, but there exist publications that mentioned those connections. They were able to verify those two connections with experiments.

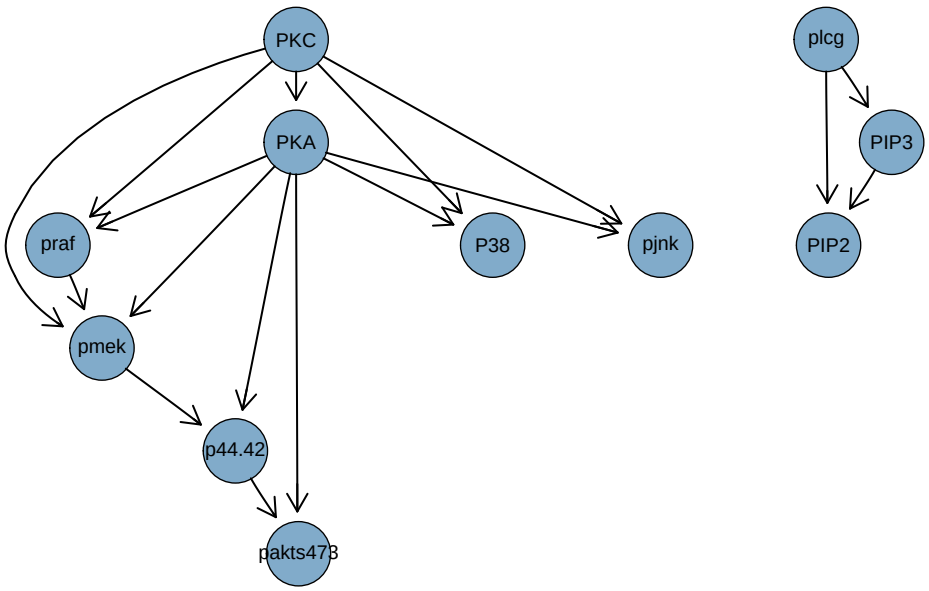

Figure 3: Validated network from Sachs et al. (2005). Protein p44.42 is also known as Erk1/2.

To get an overview how similar the predicted network are with the validated network, the following summary was calculated.

- Arcs that agree with the validated structure (also correct direction)
- Arcs that agree with the validated structure, but point in the wrong direction
- Arcs that don't agree with the validated structure

For each of the classes, we have the maximal number of arcs:

Table 2: Maximal Number of Arcs in each Class

	Correct.Arcs	Reversed.Arcs	Other.Arcs
Possible Arcs	17	17	87

1.2 Data

1.2.1 Sachs (2005)

There were 9 different perturbation conditions employed on the signaling network. For each condition the content of 11 phosphorylated protein and phospholipid components were simultaneously measured from thousands of individual primary immune system cells. The levels were measured with a multiparameter flow cytometer. The 11 phosphorylated protein and phospholipid components correspond to the nodes in figure 3.

The perturbations can be grouped in two different classes. First, there are the general stimulatory cues. Here, the protein signaling paths are active but there are no specific activators or inhibitors for one of the measured 11 phosporylated proteins or phospholipids used. Then, there are the perturbations with specific stimulatory / inhibitory cues. In that case, the perturbation affects only one of the protein. This also has an impact on all the descendants of the perturbed protein.

The point of intervention of the stimulation can be seen in figure 2. The data from this stimulation (7; CD3, CD28, LY294002) will be considered as observational data. The reason for this is that the interventional site of action PI3K was not measured.

Here is an overview of the 9 perturbation conditions used in this analysis:

Table 3: Overview stimulations

No.	Stimulations	Intervention.site.of.action
1	CD3, CD28	-
2	CD3, CD28, ICAM-2	-
3	CD3, CD28, akt-inhibitor	Akt
4	CD3, CD28, G0076	PKC
5	CD3, CD28, Psitectorigenin	PIP2
6	CD3, CD28, U0126	MEK1/MEK2
7	CD3, CD28, LY294002	-
8	PMA	PKC
9	β2camp	PKA

1.2.2 Data preparation

The data was downloaded from the webpage of the article. There was an excel sheet available for each of perturbation conditions that are mentioned in table 2. Initially, a column with the intervention type was added to each dataset (according to table 2) and then all the dataframes were merged into one large dataframe. Thereafter, all the data points that fell more than three standard deviations from the mean were discarded (According to the supplement of the article from Sachs, et al.).

Table 4: Observation per intervention - after removing outliers

	Intervention.Site.of.Action	Number.of.Obervations
0	-	2539
2	pmek	662
4	PIP2	801
7	pakts473	883
8	PKA	680
9	PKC	1149

Additionally, a discretized dataset was created with the *hartmink*-method in which the protein measurements were reduced to 3 levels. This dataset was then used for Bayesian network approach with tabu search. From the article and the supplement, it is not entirely clear how the data was pre-processed. It is only mentioned that their large dataset contains 5400 data points. The dataset that was used in this project contains 6714 data points.

2 Results

In this project 4 different methods were used to infer causal pathways. Initially, they were tested on the original Sachs data (after pre-processing). Furthermore, I created a simulated dataset and examined if it is possible to reconstruct the causal structure.

2.1 Exploratory Data Analysis

The violin plots in figure 4 demonstrate some effects of 5 different perturbations. The first column of plots (dataset 1) shows the distribution of the proteins in an observational setting. In dataset 3 AKT was inhibited.

The distributions look similar to the ones of the observational data because none of the selected proteins is a descendant of AKT.

In dataset 4 PKC was activated and it seems this also activates the other 3 proteins that are on this plot. In dataset 6 the reagent U0126 was employed, which inhibits MEK1 and MEK2. This does not really correspond to the plots. I double checked if the numbering of the datasets are correct, but I did not find any mistake. In dataset 9 PKA was activated.

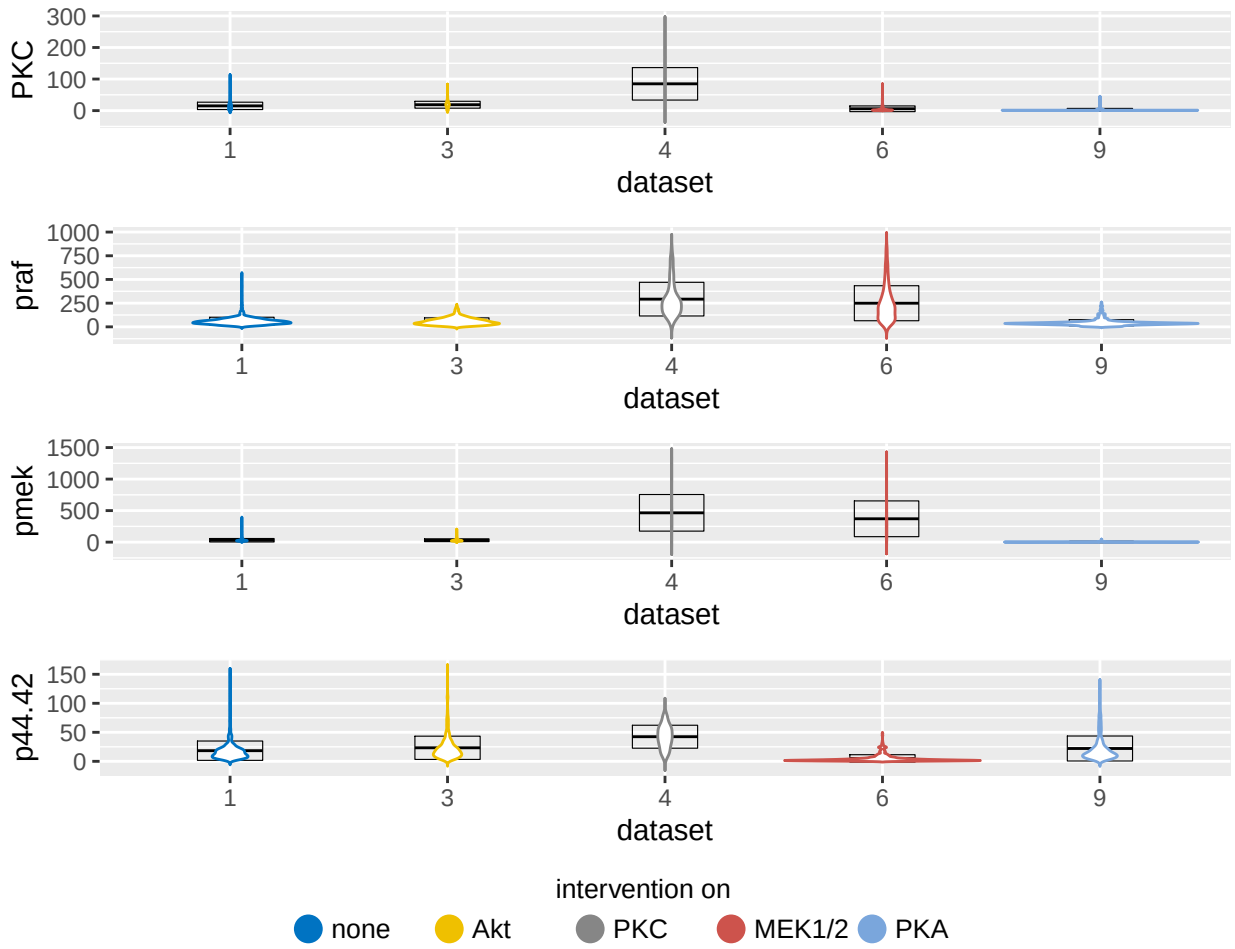

Figure 4: Violin plots for 4 selected phosphorylated protein / phospholipid for 6 different datasets. The crossbars indicate the mean and mean +/- standard deviation.

The biplot indicates that the first two principal components explain around 54% of the total variability. The interventional data for which *pmek* was the site of action seems to form a cluster. The data of the group *PKC* also has a different structure than the rest of the data. From this plot alone, it is not possible to distinguish any other clusters.

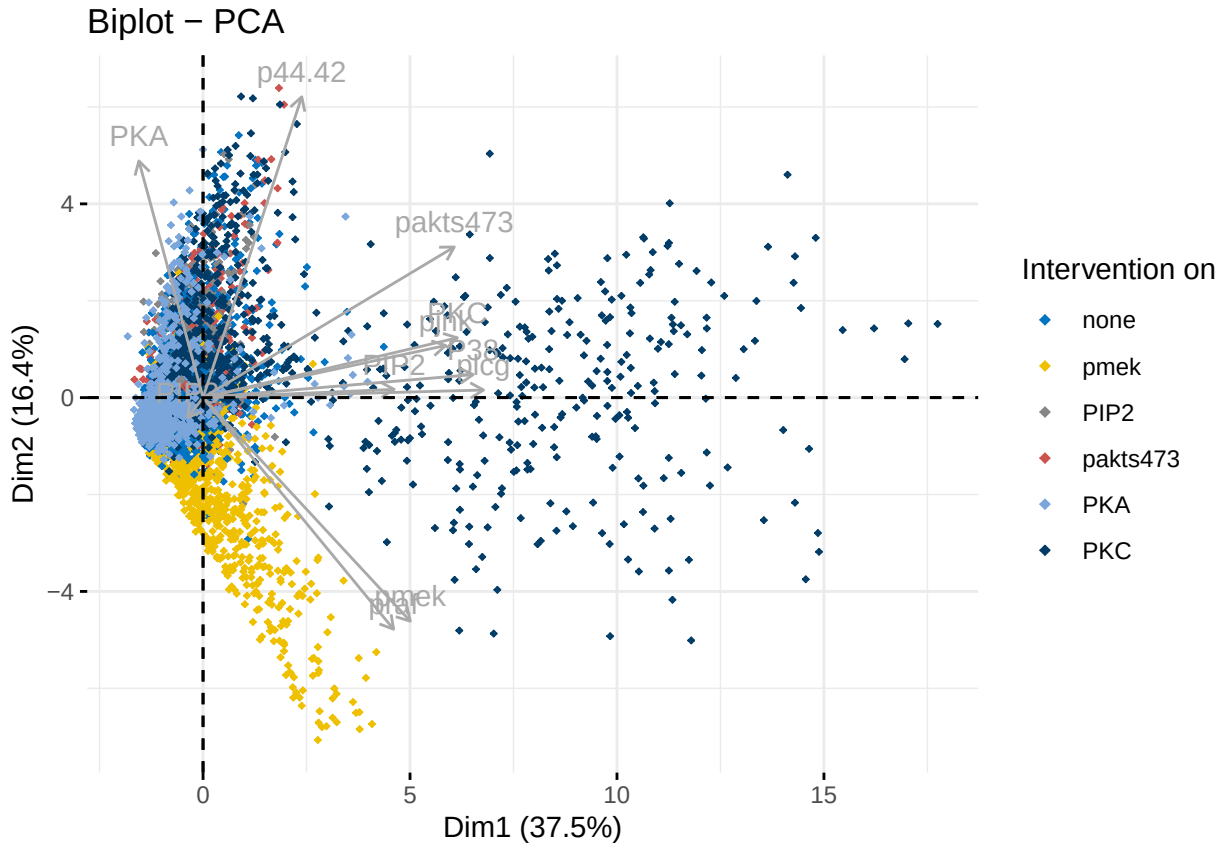

Figure 5: Sachs (2005) data - Biplot

2.1.1 PC-algorithm

The PC-algorithm mainly uses conditional independence relationships to estimate the causal structure. It is essential for this method that the data is generated in an observational setting. Therefore, only the observational data was included in the analysis.

With this method it is possible to predict a small part of the validated network.

Table 5: Classification of the predicted arcs

	correct	reversed	other
number of arcs	4	7	1

The algorithm left some of the predicted arcs undirected. In the summary such arcs are classified at the same time as correct and reversed arcs. The network inferred by this method has the following form. The bold arcs are the ones that are also contained in the validated network.

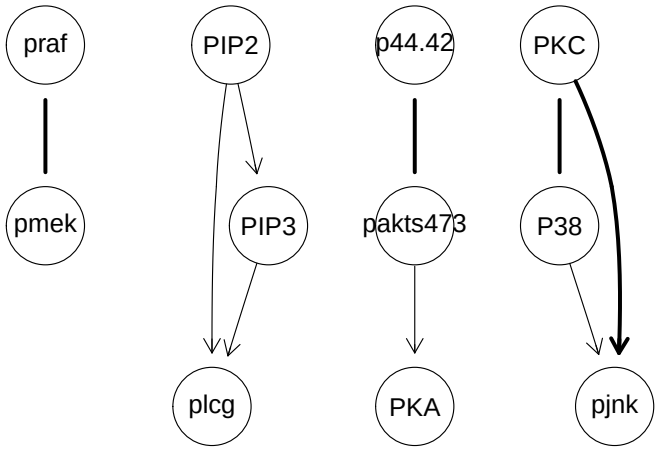

Figure 6: PC-algorithm

2.1.2 Tabu search with modified BDe score (bnlearn)

This approach is the one that is the most similar to the one they used in Sachs, et al. The results are better but there are still 11 arcs in validated network that this method did not predict.

Table 6: Classification of the predicted arcs

	correct	reversed	other
number of arcs	6	7	4

Here is the network that is predicted by this method.

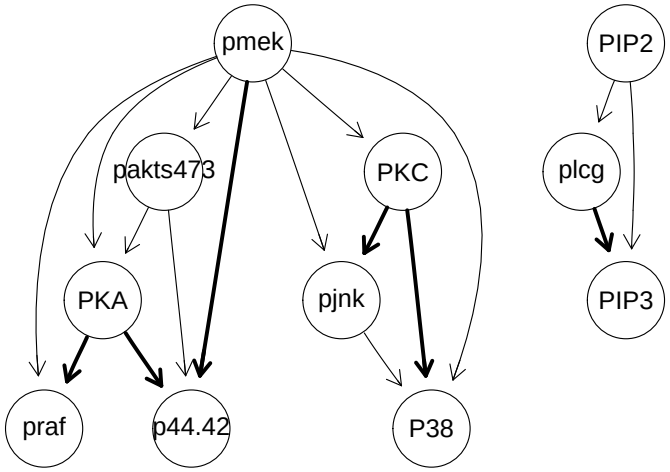

Figure 7: Tabu search with modified BDe score - Result

2.1.3 Greedy interventional equivalence search (GIES) algorithm

The GIES method is a score based method and tries to find the structure that best fits the data.

Table 7: Classification of the predicted arcs

	correct	reversed	other
number of arcs	4	4	5

With this method, we find 4 of the arcs of the validated network. It is interesting that they coincide with the ones from the previous approach.

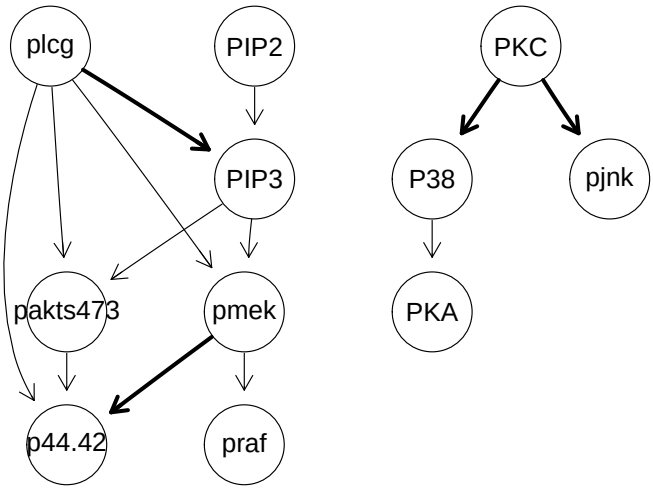

Figure 8: Greedy interventional equivalence search (GIES) algorithm

2.1.4 Backshift algorithm

The last algorithm, that was applied, was the Backshift algorithm. This approach differs significantly from the previous algorithms.

Table 8: Classification of the predicted arcs

	correct	reversed	other
number of arcs	3	4	0

This is probably also the reason that the network from this method is different compared to the previous results. It is interesting that all the predicted arcs are either correct or reversed. 68% percent of the runs converged. This seems to indicate that we don't deal with shift interventions.

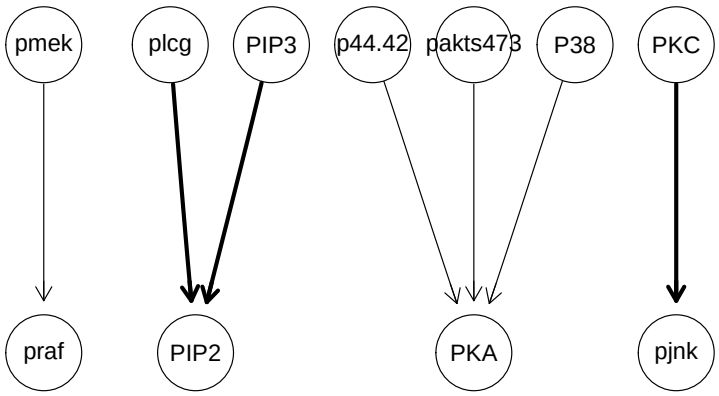

Figure 9: Backshift algorithm

2.2 Simulated data

Since the structure of the real data seems to be complicated and noisy, the causal inference analysis was also conducted on a simulated dataset. It was tested whether it is possible to detect the causal structure of simulated data.

2.2.1 Generating simulated data

The causal structure of the simulated data is the same as the one of the validated network from figure 3. All the causal effects are linear and there were only Gaussian noises used. For each of the 9 perturbation condition 600 data points were created. The observational data was simulated in the following way.

$$X^0_{PKC} = \epsilon_9, \quad X^0_{plcg} = \epsilon_3, \quad X^0_{PIP3} = X^0_{plcg} + \epsilon_5,$$
$$X^0_{PIP2} = X^0_{PIP3} + X^0_{plcg} + \epsilon_4, \quad X^0_{PKA} = X^0_{PKA} + \epsilon_8,$$
$$X^0_{praf} = X^0_{PKA} + X^0_{PKC} + \epsilon_1, \quad X^0_{pmek} = X^0_{PKA} + X^0_{PKC} + X^0_{praf} + \epsilon_2,$$
$$X^0_{p44.42} = X^0_{pmek} + X^0_{PKA} + \epsilon_6, \quad X^0_{pakts473} = X^0_{p44.42} + X^0_{PKA} + \epsilon_7,$$
$$X^0_{P38} = X^0_{PKA} + X^0_{PKC} + \epsilon_{10}, \quad X^0_{pjnk} = X^0_{PKA} + X^0_{PKC} + \epsilon_{11},$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(2,1)$ with $1 \leq i \leq 11$.

I tried to simulate the interventional data in a similar way as in the real experiment. For example in dataset 6 a reagent was applied that inhibited MEK1/2. Here, I used the approach

$$X^6_{pmek} = \mathcal{N}(0, 0.1),$$
$$X^6_{p44.42} = X^6_{pmek} + X^0_{PKA} + \epsilon_6, \quad X^6_{pakts473} = X^6_{p44.42} + X^0_{PKA} + \epsilon_7,$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(2,1)$.

All the descendants of X_{pmek} are influenced by this intervention. The rest of the proteins have the same distribution as in the observational data.

In dataset 9 PKA was activated. This was modeled by

$$X^9_{PKA} = \mathcal{N}(6, 1),$$

$$X^9_{praf} = X^9_{PKA} + X^0_{PKC} + \epsilon_1, \quad X^9_{pmek} = X^9_{PKA} + X^0_{PKC} + X^9_{praf} + \epsilon_2,$$
$$X^9_{p44.42} = X^9_{pmek} + X^9_{PKA} + \epsilon_6, \quad X^9_{pakts473} = X^9_{p44.42} + X^9_{PKA} + \epsilon_7,$$
$$X^9_{P38} = X^9_{PKA} + X^0_{PKC} + \epsilon_{10}, \quad X^9_{pjnk} = X^9_{PKA} + X^0_{PKC} + \epsilon_{11},$$

where $\epsilon_i \overset{iid}{\sim} \mathcal{N}(2, 1)$.

The other interventions were simulated accordingly.

2.2.2 Results on simulated data

The GIES algorithm performs really well with this kind of simulated data. The result of the PC algorithm is also good, considering that it only uses the 1800 observational data points. One has to keep in mind that some of the arcs are not directed and therefore count at the same time as correct and reversed arcs. The Tabu search, which performed best in the previous section, had more issues with the simulated data.

In this setting the BACKSHIFT algorithm is not able to predict a network. The diagonalization did not succeed. The reason for this is probably that the interventions are not shift interventions. The activators are simulated as shift interventions, but not the inhibitory perturbations.

Table 9: Simulated Data - Results

	correct	reversed	other
PC algorithm	11	7	0
Tabu search	7	7	6
GIES	12	1	0
Backshift	0	0	0

3 Methods

Let $G = (V, A)$ be a network structure, a Directed Acyclic Graph (DAG), where $v_i \in V$ are the nodes and $a_{ij} \in A$ represent the arcs in the graph. Each node v_i corresponds to a random variable X_i. Then a multivariate joint probability distribution $P(X)$ can be factorized into smaller local probability distribution according to the arcs of G.

$P(X_1, \cdots, X_n) = \prod_{i=1}^n P(X_i|Pa_i)$, where Pa_i denotes the parents (in the graph G) of each random variable X_i.

If a distribution factorizes in such a way then it is called a Bayesian network. This means that graph structure expresses the conditional independence relationships among the variables.

3.1 PC-algorithm of Peter Spirtes and Clark Glymour

This algorithm belongs to the class of constraint-based structure learning. The idea is that if all the conditional independence relationships in the observational distribution are given, then part of the structure of the graph of G can be inferred.

Unfortunately, it is not possible to identify the true graph G from observational data. But Markov equivalence classes (MEC) of G can be identified. The definition of Markov equivalence class can be found in the book *Elements of Causal Inference* (5). Each MEC can be uniquely represented by a Completed Partially Directed Acyclic Graph (CPDAG).

The PC-algorithm first determines the skeleton of the graph. Secondly, the v-structures are detected by checking for conditional dependence. Eventually, as many of the remaining edges are directed.

Definition: The nodes x, y, z are a *v-structure* if

- The arcs are directed $x \to y \leftarrow z$.
- There is no direct arc between x and z.

This model does not take into account that some of the data was created in an interventional setting. To get a stable result, I applied the model 100 times on a random subset of 1000 observations and then only considered the arcs that appeared in more than 85% of the predicted graphs.

R package: *bnlearn*

3.2 Tabu search with modified BDe score

I tried to use a similar model as the one they used in Sachs, et al, which included an adaptation of the Bayesian scoring metric (3) and simulated annealing. In the paper and in the supplement there are some details missing and it seemed to be challenging to use their approach. During research regarding Bayesian networks, I noticed that in the book *Bayesian Networks in R. with Applications in Systems Biology* there is a section, in which the authors also tried to reproduce the results from Sachs (2005). Therefore, I could just adapt their Bayesian network which uses a tabu search and modified BDe score.

The tabu search is a score based method. This method tests different graph structures and tries to find the structure that best fits the data. The idea is that to each graph G that represents the causal structure a score is assigned. The score measures how well G fits the data. The algorithm starts with the empty graph and searches over the space of DAGs to find the graph with the highest score. Typically the BIC is used for the score. Here, they chose an adaptation of the Bayesian scoring metric (3), the modified BDe score, as the score function.

The predicted network in the book is closer to the one from Sachs (2015). 12 of the their predicted arcs correspond to the one of the validated network, 2 arcs point in the wrong direction and 7 of the predicted arcs are not in the validated network.

They seem to pre-process the data in a different way. Each of their data set contains exactly 600 observations. I did not work with their data because they only provided a version with discrete values. The BACKSHIFT algorithm does not work on data with just 3 levels.

R package: *bnlearn*

3.3 Greedy interventional equivalence search (GIES) algorithm

The Greedy interventional equivalence search (GIES) algorithm is also a score based method and is a generalization of the Greedy Equivalence Search (GES) algorithm of Chickering (2002). The GIES has the advantage that it can be used for interventional data.

The number of DAGs with p nodes grows super-exponentially and therefore it is not possible to compute the score for all the DAGs. Therefore, a greedy search technique with 3 phases is used.

1. Forward phase. Edges are added until a local maximum is reached.
2. Backward phase. Remove edges until a local maximum is reached.
3. Turning phase. Single arrows are reversed in the space of DAGs until the score cannot be increased.

To compare this model with the other approaches, I also applied it 100 times on a random subset of 2000 observations and then only considered the arcs that appeared in more than 85% of the predicted graphs. Because this method uses the full dataset, 2000 instead of 1000 data points were chosen.

R package: *pcalg*

3.4 BACKSHIFT algorithm

The BACKSHIFT method uses different experimental conditions (or environments) with unknown targets to uncover the causal structure of the measured variables. The model assumes a linear causal model. This model can also include cycles and confounders. In this model the experimental conditions are modeled as so-called "shift interventions".

Assume that we have linear causal structure $X \to Y \to Z$.

$$X = \epsilon_x, \qquad Y = b_{xy}X + \epsilon_y, \qquad Z = b_{yz}Y + \epsilon_z.$$

This is the data-generation process under no intervention (environment 0). Here, b_{xy} expresses the strength of the linear causal effect from X to Y and ϵ_x is the noise of the random variable X. A different experimental condition (environment j, with $j \neq 0$) are assumed to be of the form

$$X = \epsilon_x + c_x^j, \qquad Y = b_{xy}X + \epsilon_y + c_y^j, \qquad Z = b_{yz}Y + \epsilon_z + c_z^j.$$

The variable c_x^j describes how much the variable X is shifted in environment j and is assumed to be a realization of a random variable. It is possible that some of the c_i^j's are zero.

A general linear causal structure can be written as

$$X = BX + \epsilon,$$

where $x \in \mathbb{R}^p$ is a random vector and $B \in \mathbb{R}^{p \times p}$. In that case, the data in environment j are observations of the model

$$X^{(j)} = BX^{(j)} + c^{(j)} + \epsilon^{(j)},$$

which can also be written as

$$(I - B)X^{(j)} = c^{(j)} + \epsilon^{(j)}.$$

The non-zero coefficients of the connectivity matrix B correspond to the direct causal relationships. The matrix is not affected by the interventions. A simple joint matrix diagonalization is used to estimate the matrix B and the estimation is based on the covariance of X in environment j, the covariance of c in environment j and the covariance of the noise.

Probably, this method does not perform so well on the simulated data, because the interventional data was not created using shift interventions. Instead, the noise of the observations of the inhibitory interventions has a smaller variance than in the observational setting. Additionally, we exactly know which variables are the target of the intervention. The BACKSHIFT method makes no use of this information.

R package: *backShift*

4 Conclusions and Discussion

It is impressive that they were able to almost reconstruct the conventionally accepted signaling molecule interactions and even predicted some new connections. But the results of this project show that one needs to know the methods, its advantages and disadvantages, in detail to apply them properly.

The pre-processing of the data has a noticeable impact on the resulting network. It matters which outliers are removed and which technique is used to discretize the data. Even small changes in the process can lead to different results. If the "true" network is known, or at least some parts of it, then it looks as if it is possible to apply the causal inference methods successfully. But without any additional knowledge concerning the causal structure it is difficult to detect the true causal structure.

The BACKSHIFT algorithm was not suitable for this data. The other algorithms had in general better results. In this data, there were no cycles and also no hidden variables in the network structure. We also knew exactly which are the site of the interventions. If this were not the case then it would be reasonable to use the BACKSHIFT algorithm. If there are hidden variables present then one could also use the FCI (Fast Causal Inference) algorithm.

Those methods cannot be applied per default on every dataset. The structure of the data is important and each situation needs a suitable approach.

5 References

1. K. Sachs, O. Perez, D. Pe'er, D. Lauffenburger, G. Nolan, *Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data*, Science **308** (2005).
2. S. Triantafillou, V. Lagani, C. Heinze-Deml, A. Schmidt, J. Tegner, I. Tsamardinos, *Predicting Causal Relationships from Biological Data: Applying Automated Causal Discovery on Mass Cytometry Data of Human Immune Cells*, Nature **308** (2005).
3. D. Pe'er, A. Regev, G. Elidan, N. Friedman, *Bioinformatics* 17 (suppl. 1), 2001.
4. R. Nagarajan, M. Scutari, S. Lbre, Bayesian Networks in R. with Applications in Systems Biology, 2013.
5. J. Peters, D. Janzing, B. Schölkopf, Elements of Causal Inference, 2018.