Causal Discovery from Single-Cell Perturbation Data

Yiqing Sun

Fudan University, School of Mathematical Sciences

October 19, 2023

Table of Contents

1. Introduction
2. Overview of GIES method
3. Two-step approach
4. Numerical experiments
5. Future work

Introduction

Objective : Find plausible causal hypotheses for interactions in gene regulatory network from Perturb-seq data.

- Basic setting: Modeling the gene regulatory network as directed acyclic graphs (DAGs) in causal Bayesian networks.
- Causal inference from purely observational data is considered as a task : do not identify the corresponding DAG completely, but only up to Markov equivalence. Interventions can help to overcome those limitations in identifiability.
- Perturb-seq enables mixtures of genetic perturbations to be assayed in a single reaction, presenting a new opportunity for efficiently inferring GRNs.
- We introduce a two-step approach for causal discovery from single-cell perturbation data.

Table of Contents

1. Introduction
2. Overview of GIES method
3. Two-step approach
4. Numerical experiments
5. Future work

Basic concepts of Causal Structure Learning

Consider a linear structural equation model (SEM) (e.g., Wright 1921) for a p-dimensional random variable $X = (X_1, \cdots, X_p)^t$ under random noise ε:

$$X \leftarrow BX + \varepsilon;$$

where B is a $p \times p$ matrix.

- Graphical model G : each variable X_k is represented by a node k, $k = 1, \cdots, p$
- $B_{j,k} \neq 0 \iff$ there is an edge from node k to node $j \iff X_k$ is a direct cause of X_j, $(k \neq j)$

Structure learning methods: score-based and constraint-based learning, Hybrid methods.

Overview of GIES method

- Score-based :an extended work of the Greedy Equivalence Search (GES) algorithm, greedily maximizes some score function (typically the BIC) for given data in three phases.
- Target graphical object: Interventional Markov equivalence class, a kind of partially directed acyclic graph (PDAG).
- Implementation condition: Known targets in each perturbation; multivariate Gaussian data.

Table of Contents

1. Introduction
2. Overview of GIES method
3. Two-step approach
4. Numerical experiments
5. Future work

Two-step approach

Gene expression information: Count matrix $X = (X_{ik})$;
Perturbation information: binary matrix $Z = (Z_{ik})$.
Assumption: the k-th gene takes up a certain fraction p_k of the total amount n_i of counts in the i-th cell (Townes et al., 2019). Given the expected value $\mu_{ik} = n_i \cdot p_k$, for each gene or protein k, the count data X_{ik} are modelled as Poisson samples without zero-inflation, i.e

$$X_{ik} \sim \text{Poisson}(\mu_{ik}) = \text{Poisson}(n_i \cdot p_k)$$

Step 1:

- In order to address the primary confounding variable, sequencing depth(approximates to n_i), we perform a Poisson regression with log(sequencing depth) as the intercept.
- We use the working residuals from the iteratively reweighted least squares fit to find plausible causal relationships.

Two-step approach

Gene expression information: Count matrix $X = (X_{ik})$; Perturbation information: binary matrix $Z = (Z_{ik})$.
Step 2:

- Some normalization methods, such as log-transformation, variance-stabilizing transformation, and other gene-expression based methods can be applied to the working residuals. Based on the performance, we choose the Freeman-Tukey transformation $\sqrt{x} + \sqrt{x+1}$ (Freeman, Tukey; 1950), It is a modification of the square root transformation and is approximately variance-stabilizing.
- The transformed data is assumed to be multivariate Gaussian-distributed and followed by GIES.

Table of Contents

1. Introduction
2. Overview of GIES method
3. Two-step approach
4. Numerical experiments
5. Future work

Attributes of Current Perturb-seq datasets

Current Perturb-seq datasets are collected and summarized by Green et al. (2022), also at http://projects.sanderlab.org/scperturb/.

- 44 publicly available single-cell perturbation-response datasets with molecular readouts from Homo sapiens and Mus musculus.
- Different perturbations: CRISPR-cas9, CRISPRa, CRISPRi, CRISPR-cas13, cytokines and drugs. 33 datasets in this resource were perturbed using CRISPR and 8 datasets perturbed with drugs.
- 32 datasets measure scRNA-seq exclusively, 5 datasets measure scATAC-seq and 1 measure ATAC and protein simultaneously.

Numerical experiments

We choose PapalexiSatija eccite RNA dataset and ReplogleWeissman2022-K562-essential dataset.

- PapalexiSatija eccite RNA dataset: 18649 genes and 20729 cells, THP-1 cell line, 99 perturbations (actually can be seen as 25), with acute monocytic leukaemia disease.
- ReplogleWeissman2022-K562-essential dataset: 8563 genes and 310385 cells, K562 cell line, 2058 perturbations, with Chronic myeloid leukaemia disease.
- JAK-STAT pathway, PI3K-Akt signaling pathway, MAPK signaling pathway and part of other leukaemia pathways.
- Preprocessing: Filter out cells with few genes (<200) and too-high percent of mitochondria genes ($>20\%$), and filter out genes expressing in less than 5% cells.

Numerical results

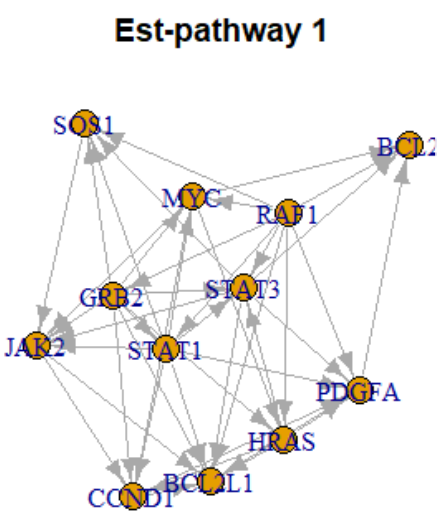
The true MAGs and true DAGs for each of the networks are available in KEGG Pathway Database: https://www.genome.jp/kegg/. The numerical performance of two-step approach is evaluated by TPR, FDR, SHD, SID, MCC as evaluation metric.
Note that a good estimation is implied with small values of FDR, SHD, SID, but large values of TPR and MCC.

	TPR	FDR	MCC	SHD	SID	Running time
pathway1	0.467	0.8409	0.1193	42	111	2.755921 secs
pathway2	0.412	0.8056	0.0488	31	80	0.7783198 secs
pathway3	0.526	0.8701	0.0505	66	134	35.29306 secs
pathway4	0.222	0.9459	-0.166	34	62	15.06681 secs
pathway5	0.714	0.8649	0.0966	24	52	7.91565 secs
pathway6	0.538	0.9176	-0.033	61	128	16.98651 secs

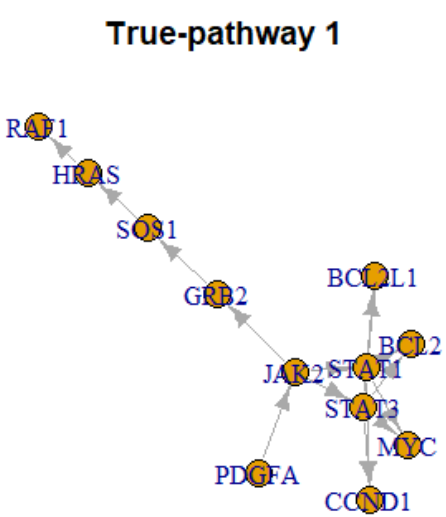
Table 1

Numerical results

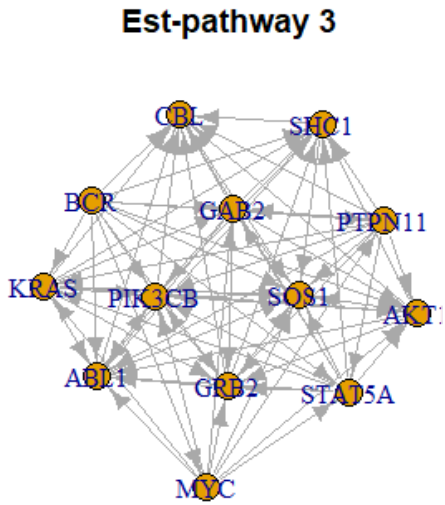
Two-step approach tends to generate a rather dense causal graph and the direction of some edges remains undetected or even reverse.
Example:



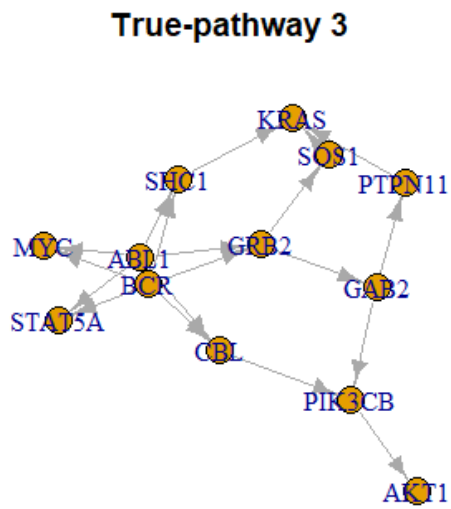
Figure



Figure



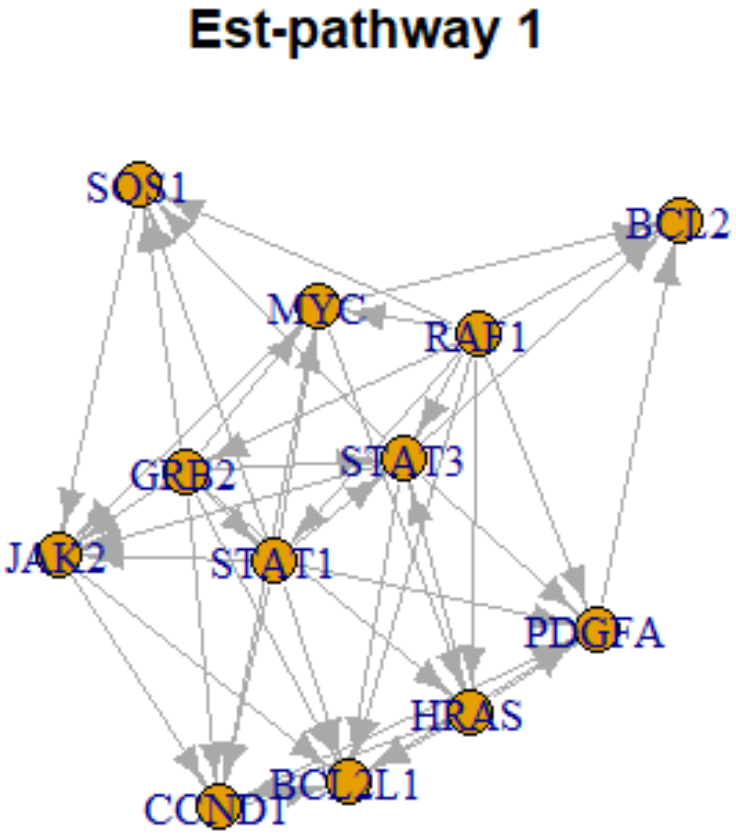
Figure



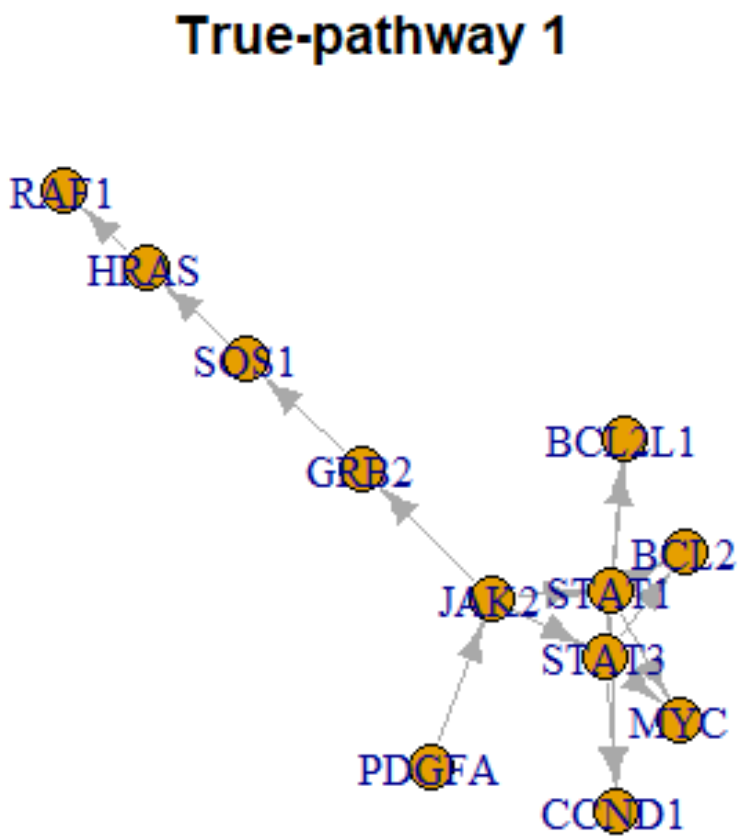
Figure

Plausible reasons

- genes in the pathway may participate in other biochemical reaction and have other causal relationship.



Figure



Figure

- potential presence of confounding variables.

Plausible reasons

- Deviations from the assumed Gaussian distribution.
- Limitations arising from the number of overlapping genes.
- The inherent limitations of the greedy search employed by the GIES method.

Zero inflation

in Poisson regression, we have

$$\text{working residual} = \frac{y_{\text{expected}} - y_{\text{estimates}}}{y_{\text{estimates}}}$$

where y_{expected} is the true value while $y_{\text{estimates}}$ is the estimated value in the model.

Only working residuals of highly-expressed genes perform like Gaussian distributed after transformation:

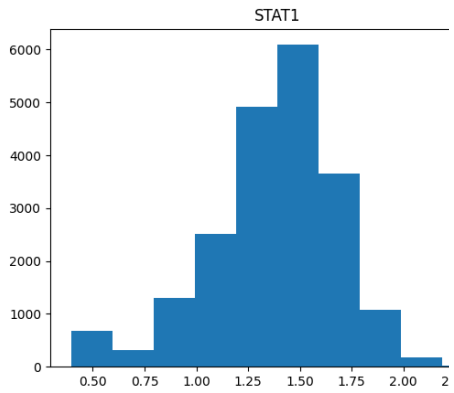


Figure: STAT1

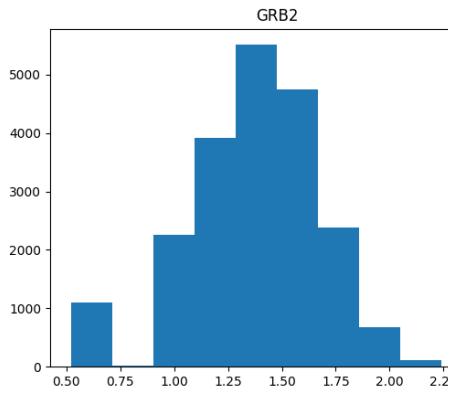


Figure: GRB2

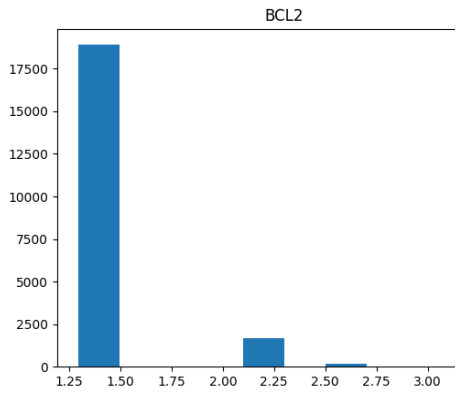


Figure: BCL2

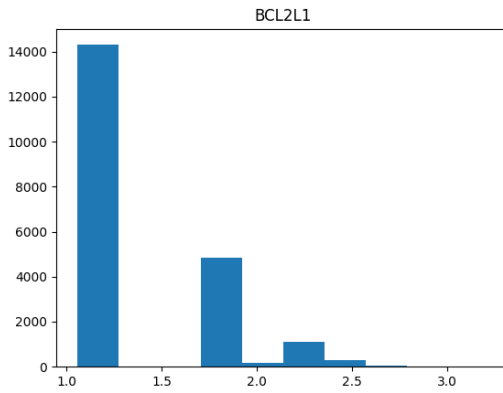
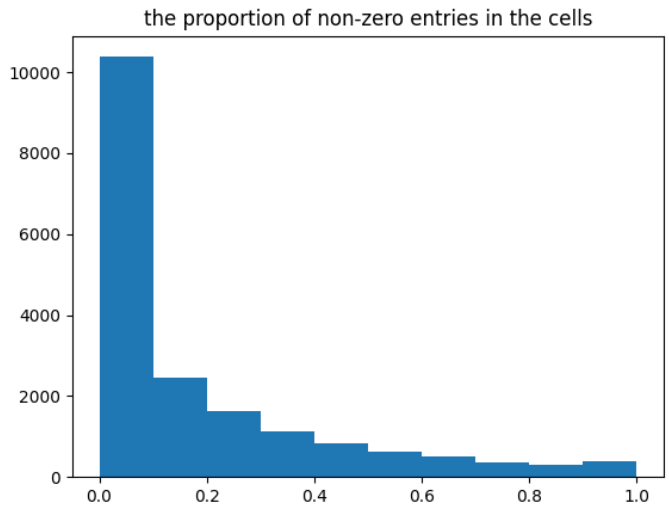


Figure: BCL2L1

Zero inflation



- Applying data imputation techniques for scRNA datasets to recover the underlying signal within the gene regulatory network.
- Revising the original modeling framework. such as hyper-Poisson or negative binomial ,accounting for overdispersion and zero inflation characteristics.

Table of Contents

1. Introduction
2. Overview of GIES method
3. Two-step approach
4. Numerical experiments
5. Future work

Data imputation methods

MAGIC is a graph-based imputation method specifically designed for single-cell data,incorporates local neighborhood structure to impute missing values accurately.

Pathway1						
Imputation?	TPR	FDR	MCC	SHD	SID	Running time
No	0.467	0.841	0.119	42	111	2.755921 secs
MAGIC	0.733	0.814	0.224	49	99	1.655861 secs
Pathway2						
No	0.412	0.806	0.049	31	80	0.7783198 secs
MAGIC	0.235	0.9	-0.15	40	85	0.7909241 secs

Table 2

Zero-inflated model

- Using Hurdle conditional distributions (McDavid et al., 2016), Yu et al. (2020) proposed fully identifiable BNs (ZiDAG) for zero-inflated Gaussian data under a natural and practical assumption.
- Choi et al. (2020) developed zero-inflated Poisson BNs for observational zero-inflated count data. To further accounting for over dispersion in the count data, Choi et al. (2023) propose a new zero-inflated generalized hypergeometric directed acyclic graph (ZiG-DAG) model.

Future work

- Utilise the Hurdle model and Hurdle polynomials in modelling. Develop Zero-inflated GIES method for causal discovery from zero-inflated perturbation data.
- Develop imputation methods for perturbation data, where the perturbation information is considered when constructing a metrics (such as E-statistics). Model the drop-outs (technical zeroes) and actual expression values (biological zeroes) to avoid over-imputing the expression measurements.

Discussion

- Proposed two-step approach for causal discovery (structure learning) from single-cell perturbation data
- Conducted numerical experiments on current Perturb-seq datasets , utilise TPR, FDR, MCC, etc. to test the identifiability and accuracy of our method.
- Implemented data imputation methods to recovering signals.

Although our work is more of an exploratory progress than a completed project, we still hope to provide some information for the modeling, processing and analysing of Perturb-seq data.

Thank you for listening!