Inference on the Structure of Gene Regulatory Networks

Yue Wang

Department of Computational Medicine,
University of California, Los Angeles

Outline

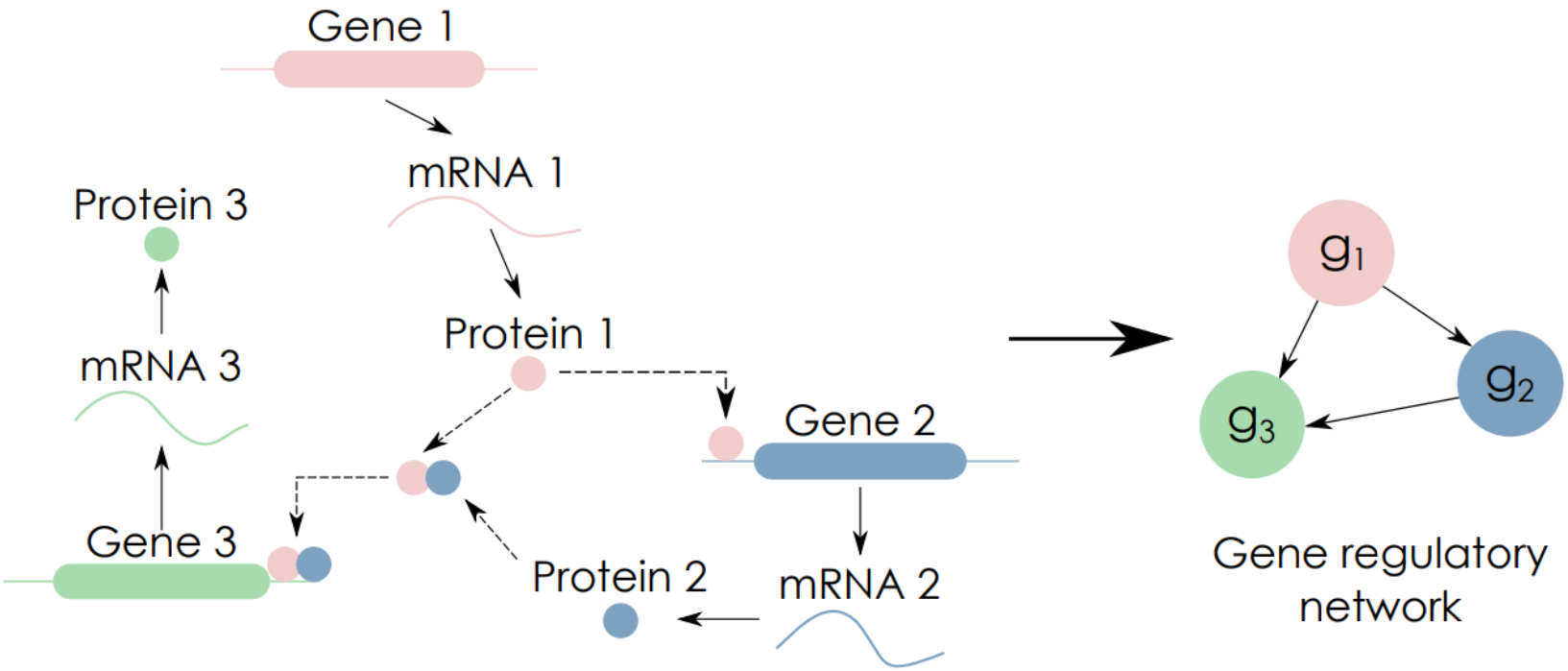
- Introduction to gene regulatory networks (GRN).
- Types of data that can be used to infer GRN structures.
- Mathematical inference methods for GRN structures.

Regulation of gene expression

- Gene expression: genes are transcribed to mRNAs and then translated to proteins.
- Various molecular regulators affect gene expression (change levels of mRNAs and proteins).
- Some regulators are small molecules, such as oxygen, sugars and vitamins. Some regulators are proteins. We focus on regulations between genes.

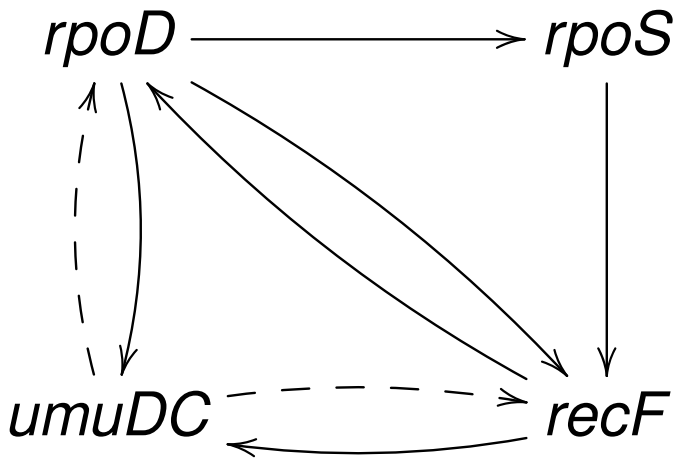
$$\text{gene} \xrightarrow{\text{expression}} \text{protein} \xrightarrow{\text{regulation}} \text{another gene}$$

Regulation of gene expression



Genes and their regulatory relations form a gene regulatory network (GRN).

Regulation of gene expression



- An example of GRN in *E. coli*. Each vertex is a gene. Two types of regulations: solid arrow means activation, and dashed arrow means inhibition.
- We aim at determining the GRN structure.
- For two genes V_i, V_j, does the expression of V_i activates or inhibits the expression of V_j?

Regulation of gene expression

- Genes (DNAs), mRNAs and proteins are generally confined within living cells.
- It is extremely difficult or even impossible to directly determine whether one gene regulates another gene with biochemical methods.
- We have accumulated a large amount of data, e.g., bulk level gene expression data and single-cell level phenotype data. Certain types of data can be used to infer the GRN structure.

Data types: Gene expression vs. Phenotype

- Setup: consider a set of genes $V_1, \ldots, V_n$. Assume this set consists of all genes in a GRN and possibly a few irrelevant genes.
- We can measure the expression levels of these genes, or the level of a phenotype V_0 (e.g., growth rate, drug resistance) which is affected by these genes.
- Determine whether one gene activates/inhibits another gene.

Data types: Single-cell vs Bulk

- Besides “Gene expression” vs. “Phenotype”, there are other dimensions of possible data types.
- The gene expression of a single cell is stochastic. We can measure the levels of $V_1, \ldots, V_n$ for a single cell and repeat many times, so as to obtain a group of random variables $X_1, \ldots, X_n$ that represent the random levels of $V_1, \ldots, V_n$.
- We can also measure these quantities over a large population of cells (bulk level), so that the randomness is averaged out. Then we obtain deterministic results $x_1, \ldots, x_n$.

Data types: Interventional vs. Non-interventional

- We can intervene with certain genes (siRNA, CRISPR, etc.), so that the expression levels of these genes are changed. Then other related genes are also affected.
- We can measure expression levels $x'_1, \ldots, x'_n$ after interfering with certain genes, and compare with corresponding quantities before intervention $x_1, \ldots, x_n$.
- We can also observe without any intervention.

Data types: One-time vs. Time series

- We can measure at a single time point, $X_i(0)$, or measure at multiple time points as a time series, $X_i(0), X_i(1), X_i(2), \ldots$.
- With time series data, we can study the dynamics of gene expression.

Data types: Joint distribution vs. Marginal distribution

- When we measure at single-cell level at multiple time points, we obtain a sequence of random variables $X_i(0), X_i(1), X_i(2), \ldots$.
- Most measurements are destructive, meaning that one cell can be measured only once. If so, we can only obtain the marginal distribution for each time point, $\mathbb{P}[X_i(0) = c_0], \mathbb{P}[X_i(1) = c_1], \mathbb{P}[X_i(2) = c_2]$.
- If the same cell can be measured multiple times, we obtain the joint distribution for multiple time points, $\mathbb{P}[X_i(0) = c_0, X_i(1) = c_1, X_i(2) = c_2]$.
- With the joint distribution, we can obtain more information, such as correlation coefficients.

Data types

- We have four major dimensions: (1) Gene expression or Phenotype; (2) Single-cell or Bulk; (3) Non-interventional or Interventional; (4) One-time or Time series.
- According to these four dimensions, we have $2^4 = 16$ different data types (scenarios).
- In four scenarios (Single-cell + Time series), there is an extra dimension of Joint distribution or Marginal distribution, meaning a total of 20 scenarios.

Data types

		One-Time		Time Series	
		Non-Intervention	Intervention	Non-Intervention	Intervention
Gene Expression	Single-Cell	Scenario 1	Scenario 2	Scenario 3a Joint Scenario 3b Marginal	Scenario 4a Joint Scenario 4b Marginal
	Bulk	Scenario 5	Scenario 6	Scenario 7	Scenario 8
Phenotype	Single-Cell	Scenario 9	Scenario 10	Scenario 11a Joint Scenario 11b Marginal	Scenario 12a Joint Scenario 12b Marginal
	Bulk	Scenario 13	Scenario 14	Scenario 15	Scenario 16

All 20 scenarios, classified by data types.

Question?

Structure inference

- Different scenarios require different mathematical inference methods.
- In order to infer the GRN structure with limited experimental data, we need some assumptions about GRN and data.
- Under these assumptions, the underlying GRN is simple enough, or the experimental data are regular enough, so that they follow certain mathematical models.
- For instance, we can assume the GRN has no cycle, or the gene expression levels satisfy a linear ODE system.

Structure inference

- For each scenario, we discuss what structures can be inferred, and what assumptions are required.
- Scenarios 1/3/8 have been extensively studied. For other scenarios, we invent new mathematical methods, or prove that the GRN structure cannot be inferred.

Structure inference

		One-Time		Time Series	
		Non-Intervention	Intervention	Non-Intervention	Intervention
Gene Expression	Single-Cell	Scenario 1 MF+DAG: partial	Scenario 2 PB: full DAG: partial MF+DAG: full	Scenario 3 a/b 3a Joint: UC: full 3b Marginal: MF+DAG: partial	Scenario 4 a/b 4a Joint: UC: full 4b Marginal: LS: full PB: full DAG: partial MF+DAG: full
	Bulk	Scenario 5 No	Scenario 6 PB: full DAG: partial	Scenario 7 No	Scenario 8 LS: full PB: full DAG: partial

Inference results for different scenarios (part I). MF, DAG, PB, LS: mathematical assumptions required by corresponding inference methods. UC: no assumption required. Full/partial/no means all/some/no GRN structures can be inferred.

Structure inference

		One-Time		Time Series	
		Non-Intervention	Intervention	Non-Intervention	Intervention
Phenotype	Single-Cell	Scenario 9 No	Scenario 10 PB: partial	Scenario 11 a/b No	Scenario 12 a/b PB: partial LS+DAG: partial* PB+LS+DAG: partial*
	Bulk	Scenario 13 No	Scenario 14 PB: partial	Scenario 15 No	Scenario 16 PB: partial LS+DAG: partial* PB+LS+DAG: partial*

Inference results for different scenarios (part II). DAG, PB, LS: mathematical assumptions required by corresponding inference methods. Partial/no means some/no GRN structures can be inferred. Asterisk means activation/inhibition cannot be determined.

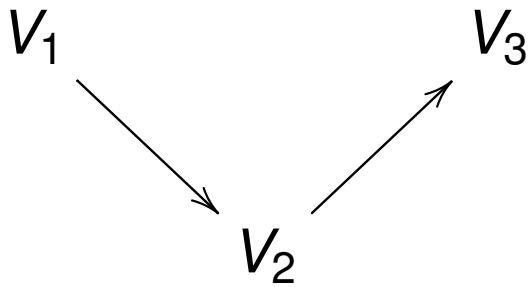
Structure inference

- Gene expression data are more informative than phenotype data.
- Interventional data are more informative than non-interventional data.
- Scenario 4 (gene expression, single-cell, interventional, time series) is the most informative case.
- Nevertheless, for more informative data types, generally the experiments are more difficult, more expensive, and less accurate.

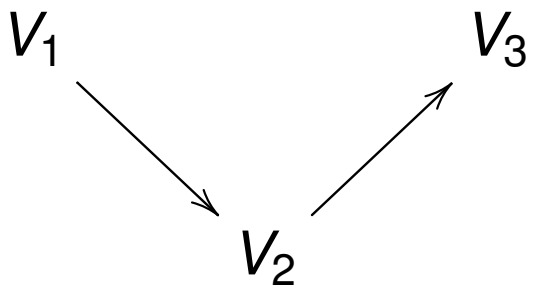
- Question?

Examples of inference methods

- In Scenario 6 (gene expression, bulk, interventional, one-time), we can partially infer the GRN structure under the DAG assumption.
- DAG: directed acyclic graph, meaning that the GRN has no directed cycle.
- GRN is represented by a DAG. Each vertex is a gene, and each directed edge is a regulatory relation.

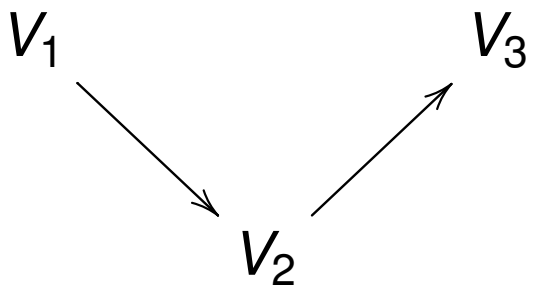


Examples of inference methods

- In a DAG, if there is a directed path from V_i to V_j, then V_i is an ancestor of V_j, and V_j is a descendant of V_i.
- 

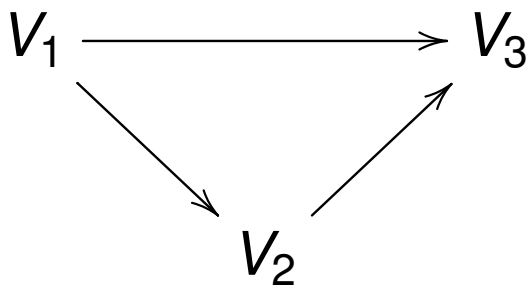
  V_1 has descendants V_2, V_3; V_2 has descendant V_3; V_3 has no descendant.
- If we add intervention on gene V_i, then the descendants of V_i are also affected.

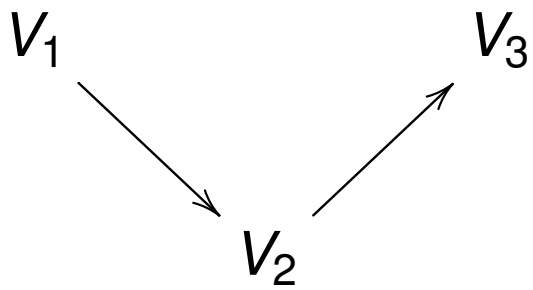
Examples of inference methods

- After adding intervention on gene V_i, if gene V_j is also affected, then in the DAG, V_j is a descendant of V_i.
- With such intervention experiments, we can determine the ancestor-descendant relations between genes.
- Now we have a mathematical problem: given the ancestor-descendant relations of a DAG, how to infer its structure?
- 

V_1 has descendants V_2, V_3; V_2 has descendant V_3; V_3 has no descendant.

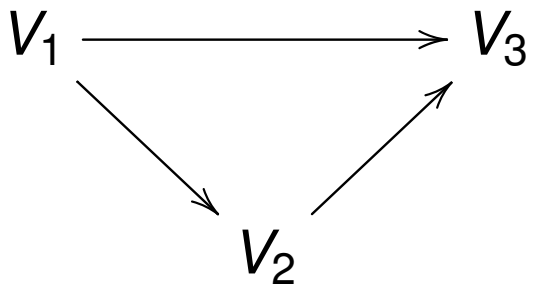
Examples of inference methods

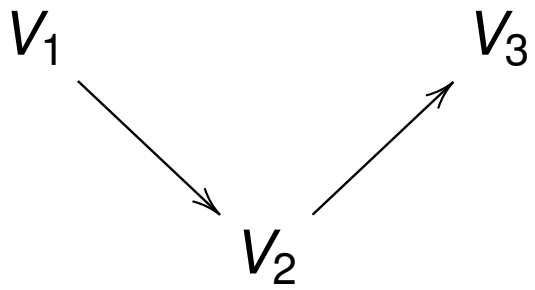
- 



- Two DAGs with the same ancestor-descendant relations are called “AD equivalent”.
- All DAGs that are AD equivalent form an equivalent class.

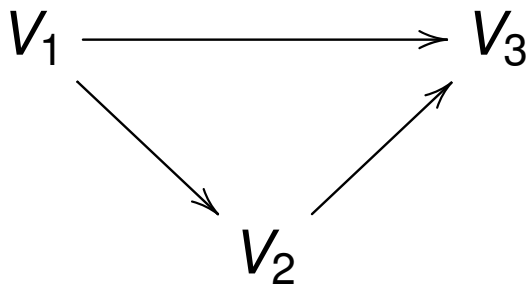
Examples of inference methods

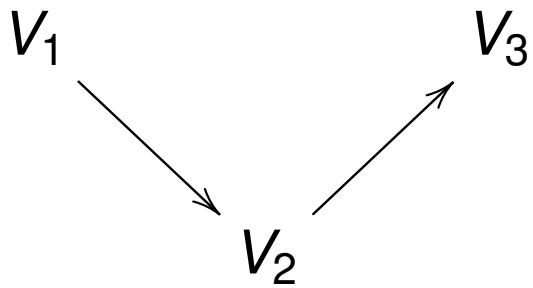
- 



- Using the ancestor-descendant relations, if an edge $V_i \to V_j$ appears in all of these AD equivalent DAGs, we can determine the edge $V_i \to V_j$ exists in the GRN.
- We can determine that the GRN has edges $V_1 \to V_2$ and $V_2 \to V_3$.

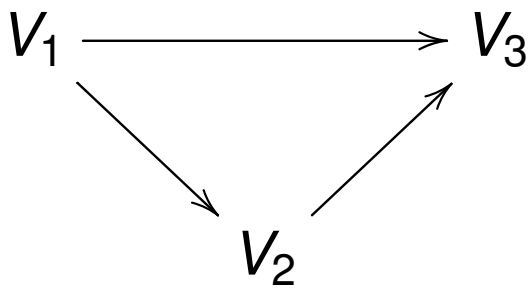
Examples of inference methods

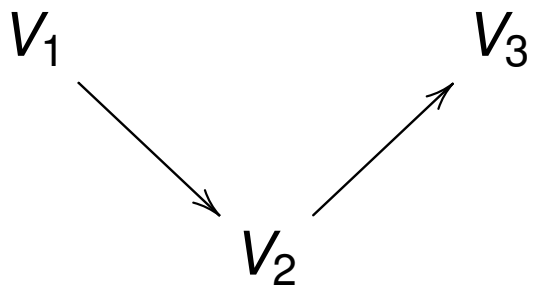
- 



- If an edge $V_i \to V_j$ appears in none of these AD equivalent DAGs, we can determine the edge $V_i \to V_j$ does not exist in the GRN.
- We can determine that the GRN does not have edges $V_3 \to V_2$, $V_3 \to V_1$, and $V_2 \to V_1$.

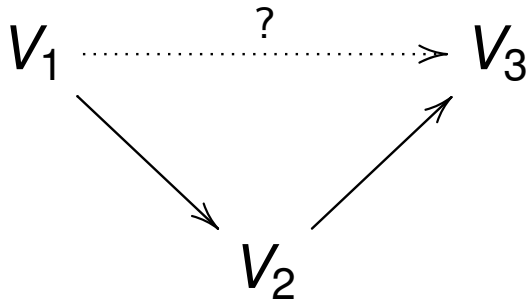
Examples of inference methods

- 



- If an edge $V_i \to V_j$ appears in some but not all of these AD equivalent DAGs, we cannot determine whether the edge $V_i \to V_j$ exists in the GRN.
- We cannot determine whether the GRN has edge $V_1 \to V_3$.

Examples of inference methods

- 

  We can identify two edges in the GRN. One edge is unknown.
- In sum, the GRN structure can be partially inferred.

Examples of inference methods

Given a DAG, we can find out what edges can be determined by ancestor-descendant relations.

Theorem

The following procedure describes how to determine certain edges with ancestor-descendant relations.
(1) If V_j is not a descendant of V_i, then we can determine that the edge $V_i \to V_j$ does not exist.
(2) If V_j is a descendant of V_i, and V_i has another descendant V_k, which is an ancestor of V_j, then we cannot determine the existence of the edge $V_i \to V_j$.
(3) If V_j is a descendant of V_i, and V_i does not have another descendant V_k, which is an ancestor of V_j, then we can determine that the edge $V_i \to V_j$ exists.

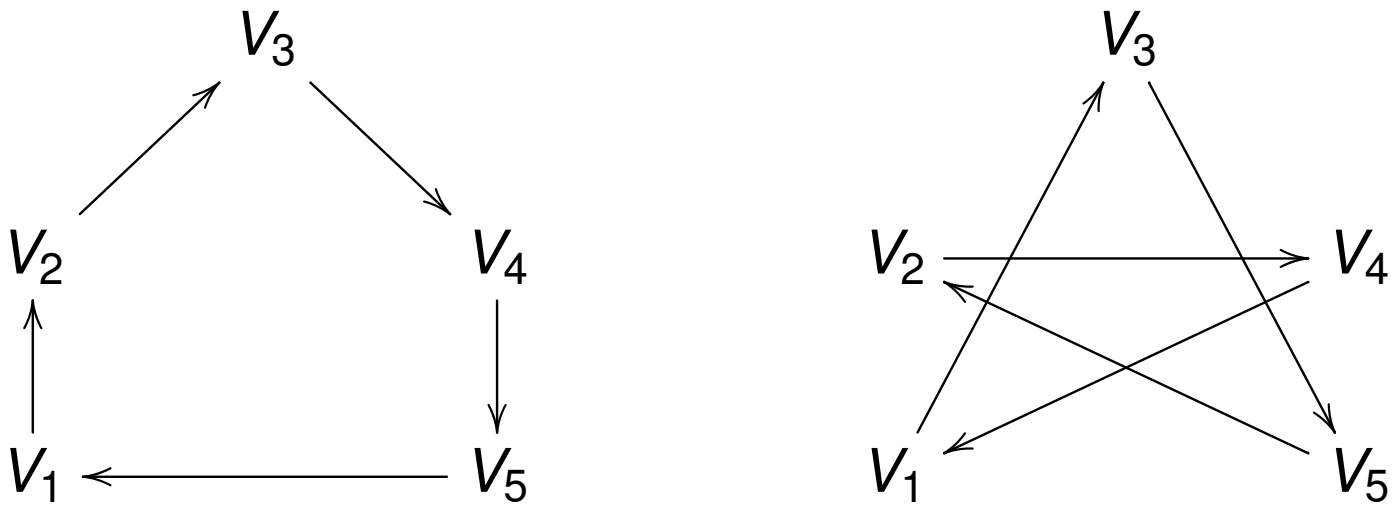
Examples of inference methods

Although not all edges can be inferred, we have a lower bound for edges that can be inferred.

Theorem

If the GRN is a connected DAG with n vertices, then we can use ancestor-descendant relations to identify at least $n - 1$ edges.

Examples of inference methods

- If the GRN has cycles, we might infer no edge.
- 

- These two GRNs share the same ancestor-descendant relations, but they have no common edges. Thus we cannot determine the existence of any edges.

Summary

- Introduce the GRN structure inference problem.
- Classify the inference problem into 20 scenarios.
- Previous studies are unified under a few scenarios. Invent mathematical methods for scenarios that have not been extensively studied.
- This work provides a unified framework to discuss the GRN structure inference problem.

- Questions?

References

- Wang, Y., & Wang, Z. (2022). Inference on the structure of gene regulatory networks. Journal of Theoretical Biology, 539, 111055.
- Wang, Y., & He, S. (2022). Inference on autoregulation in gene expression. arXiv preprint arXiv:2201.03164.