Inference on the Structure of Gene Regulatory Networks

Yue Wang

Irving Institute for Cancer Dynamics and
Department of Statistics, Columbia University

yw4241@columbia.edu

Jan. 18, 2024

Outline

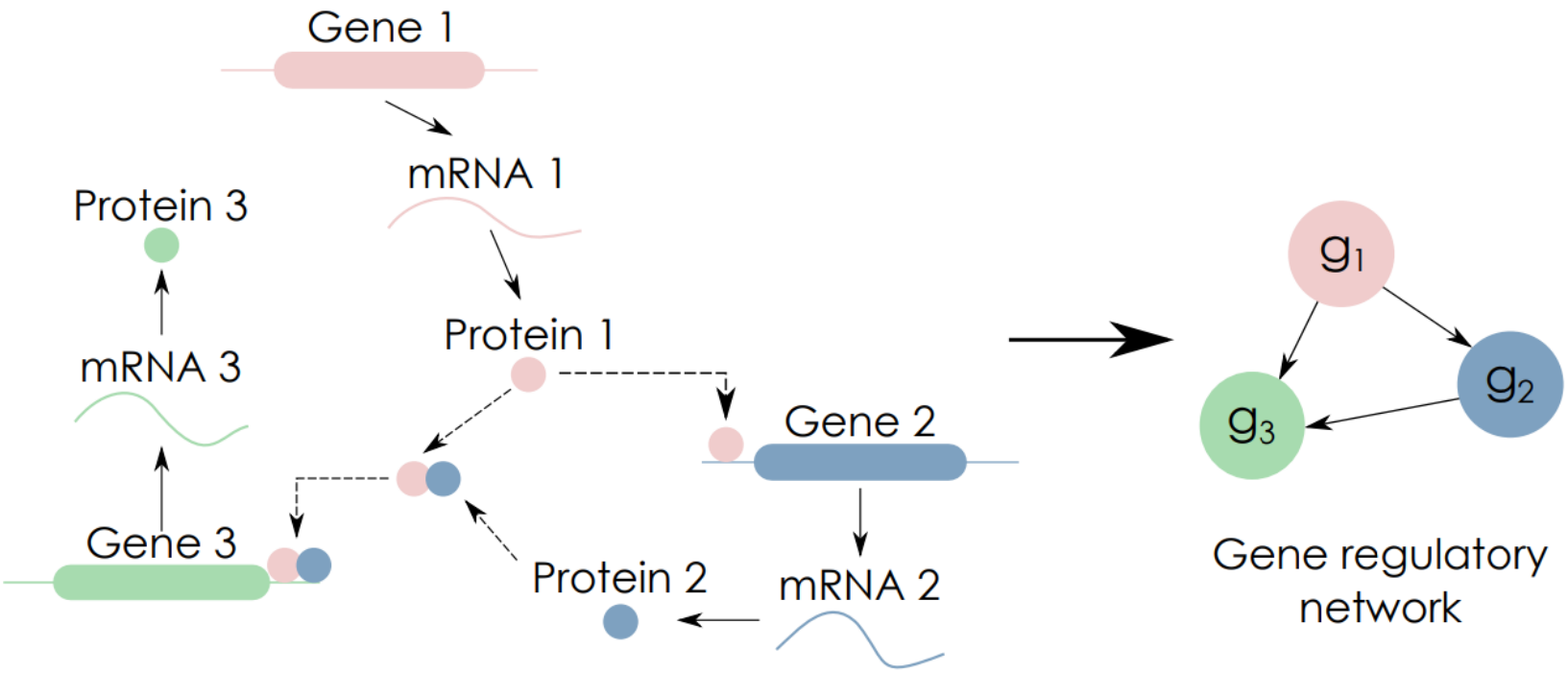
- Introduction to gene regulatory network (GRN).
- Types of data that can be used to infer GRN structures.
- Mathematical inference methods for GRN structures.
- Inference on autoregulation.

Regulation of gene expression

- Genes are sequences of nucleotides in DNA, generally in the cell nucleus.
- 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.

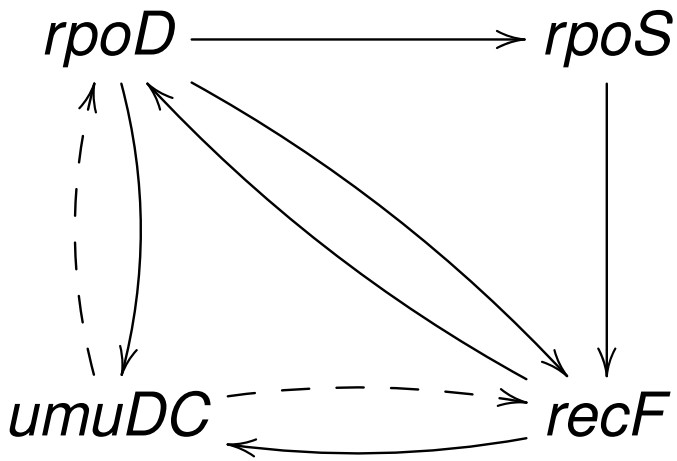
$$\text{gene} \xrightarrow{\text{expression}} \text{protein} \xrightarrow{\text{regulation}} \text{another gene}$$

Regulation of gene expression



Genes and their regulatory relations form a gene regulatory network (GRN). It is a directed graph: vertices are genes, and edges are regulatory relations.

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.
- GRN determines cell function. If the GRN structure is known, we know how the cell fate is determined, and how we can change the cell fate.

Regulation of gene expression

- A central question in biology is to determine the GRN structure.
- For two genes V_i, V_j, does the expression of V_i regulate (activate or inhibit) the expression of V_j?
- 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 gene expression data. Certain types of gene expression data can be used to infer the GRN structure.

Data types: Single-cell vs Bulk

- 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.
- Traditionally, measuring gene expression level (mRNA count or protein count) is not very sensitive. Generally, people measure the overall expression of many cells (bulk level).
- Single-cell measurements (e.g., single-cell RNA sequencing) became popular in last 10 years.
- The gene expression of a single cell is stochastic. When we measure different cells at single-cell level, the results can be different. Measurements on bulk level should produce the same results.

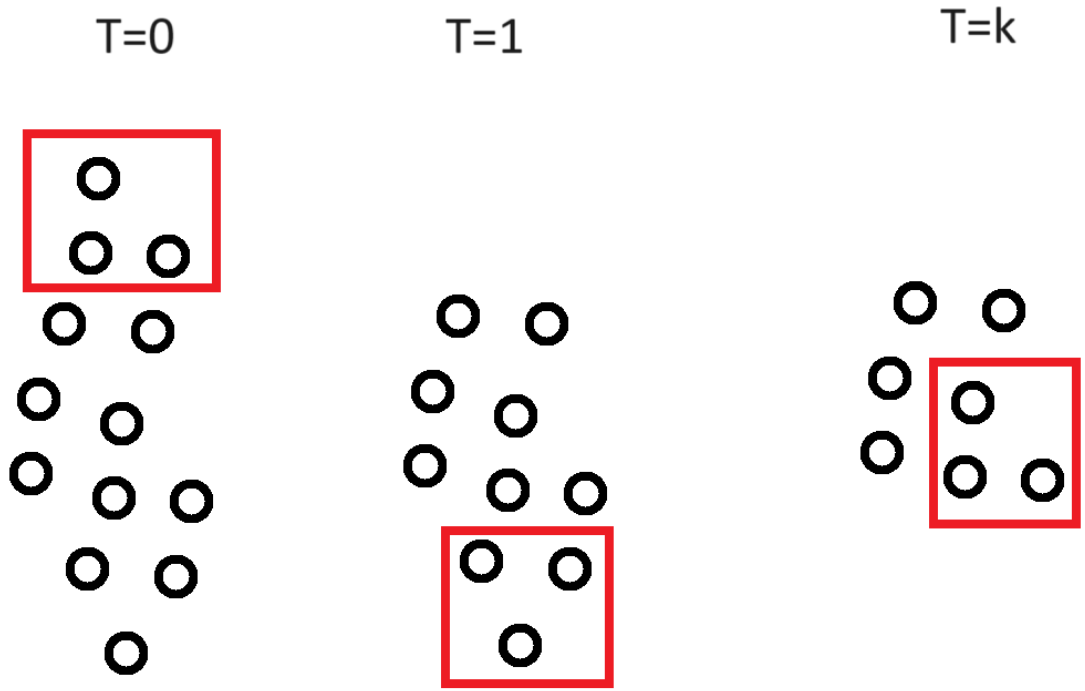
Data types: Single-cell vs Bulk

- We can measure the levels of $V_1, \ldots, V_n$ (mRNA count or protein count) 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 $x_1, \ldots, x_n$.

	Gene V_1	Gene V_2	$\ldots$	Gene V_n
Cell 1	a_{11}	a_{21}		a_{n1}
$\ldots$				
Cell m	a_{1m}	a_{2m}		a_{nm}
Average	x_1	x_2		x_n

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, $X_i(0), X_i(1), X_i(2), \ldots$.

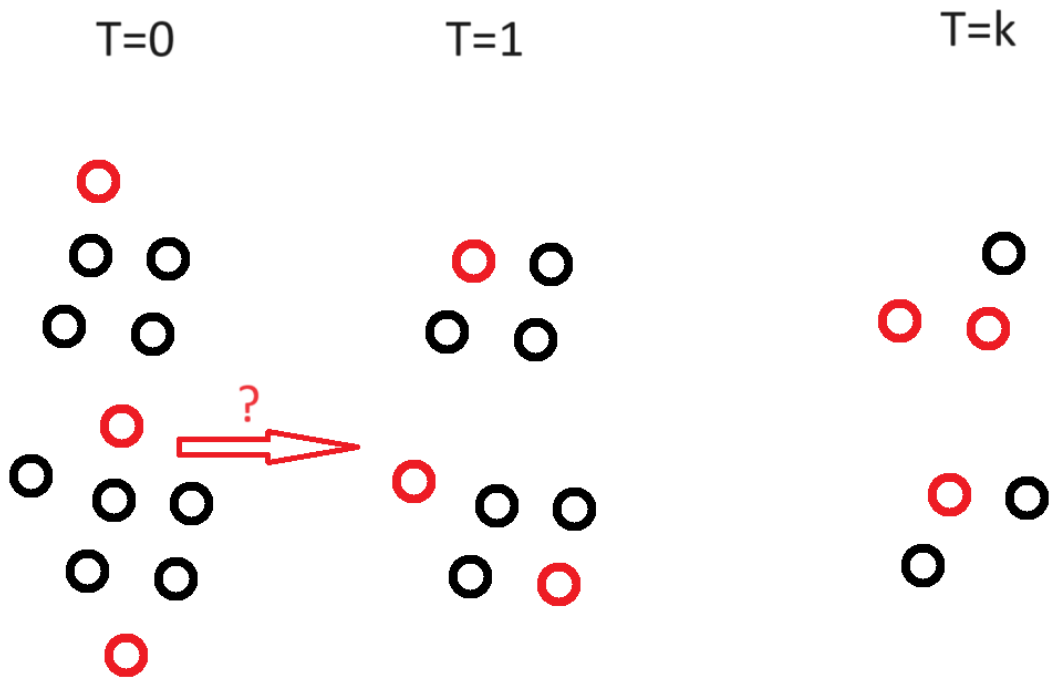


Data types: Joint distribution vs. Marginal distribution

- If we only measure the expression level of a single gene, it is possible to measure the same cell multiple times and obtain time series data (with fluorescent proteins).
- Most measurements are destructive, meaning that one cell can be measured only once. To obtain time series data, we cultivate many cells under the same condition, and sample different cells to measure at each time point.
- For bulk level data, since the results are deterministic, whether we can measure the same cell multiple times does not make a difference. For single-cell level data, things are different.

Data types: Joint distribution vs. Marginal distribution

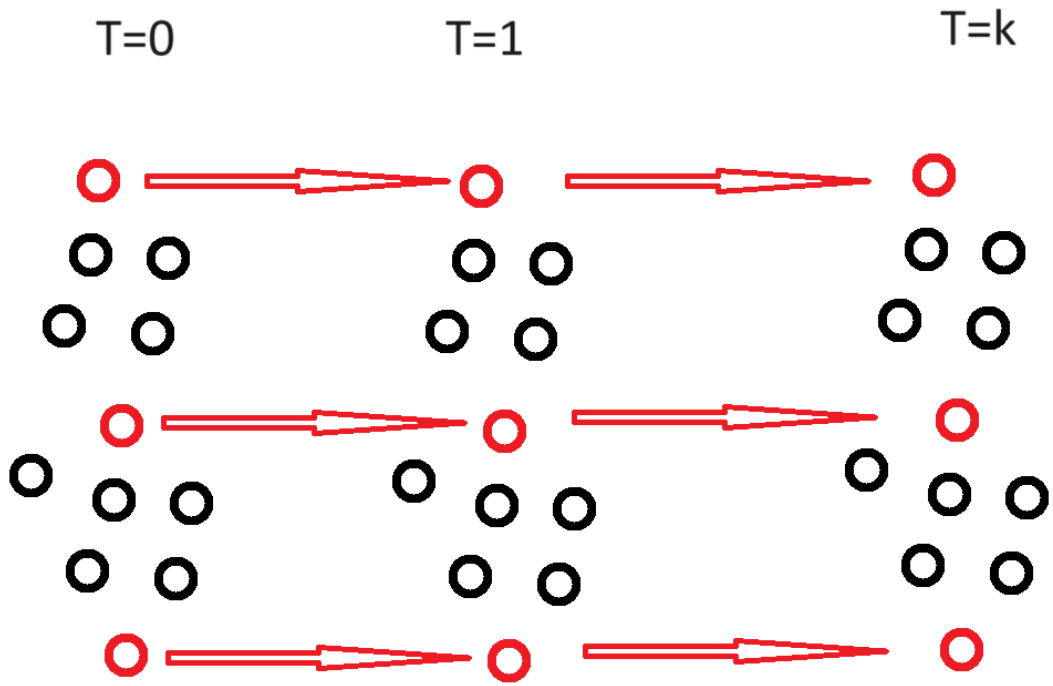
- At single-cell level, if the same cell can be measured only once, we can only obtain the marginal distribution for each time point, $\mathbb{P}[X_i(0) = c_0], \mathbb{P}[X_i(1) = c_1], \mathbb{P}[X_i(2) = c_2], \ldots$.



- With only marginal distributions, we cannot build connections between different time points.

Data types: Joint distribution vs. Marginal distribution

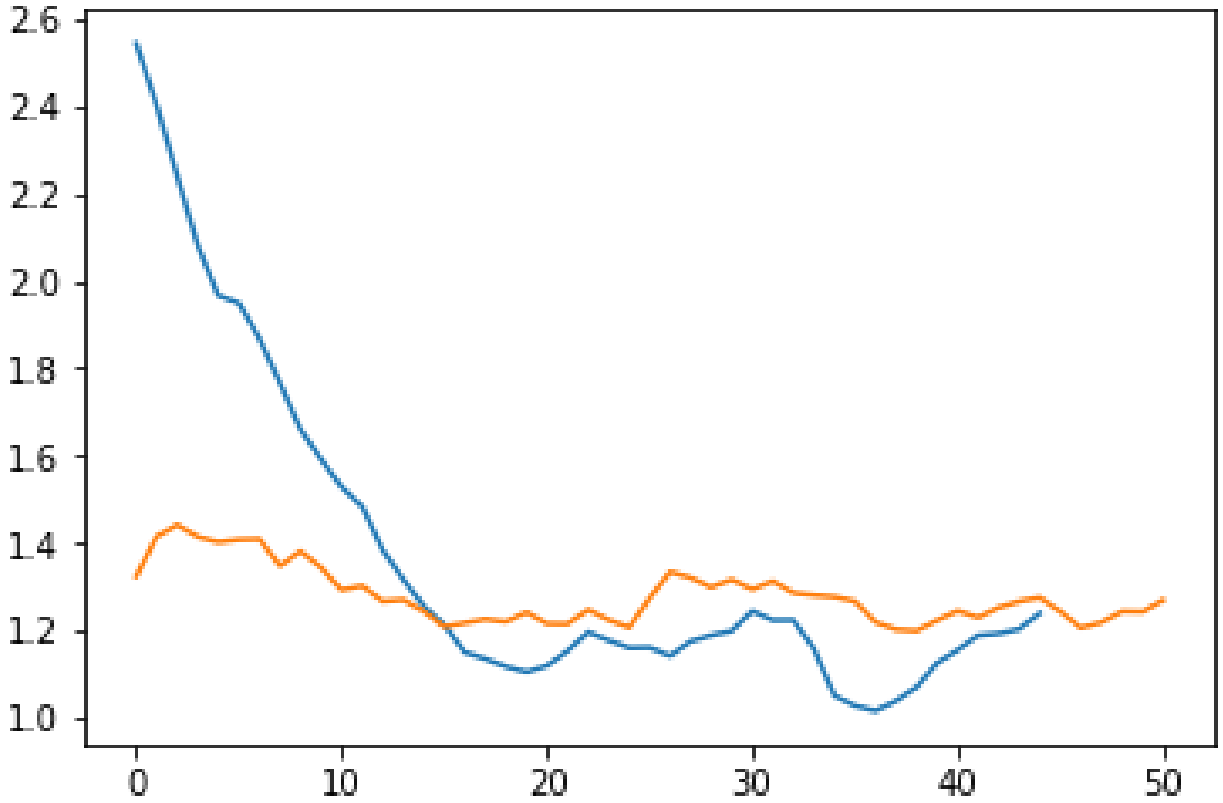
- If the same cell can be measured multiple times, we obtain the joint distribution for multiple time points, $\mathbb{P}[X_i(0) = c_0, X_i(1) = c_1, X_i(2) = c_2, \ldots]$.



- With the joint distribution, we can obtain more information, such as correlation coefficients between different time points.

Data types: Stationary vs. Interventions

- We can measure the expression levels for genes at stationary.
- We can add interventions to drive genes away from stationary, which will converge to the new stationary.



Data types: Stationary vs. Interventions

- For general interventions (drugs, etc.), genes that will be affected are fixed.
- We can intervene with any specific 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 intervention, and compare with corresponding quantities $x_1, \ldots, x_n$ before intervention.
- Experiments of adding specific interventions are time-consuming and expensive. Such data are not common for now.

Data types

- We have three major dimensions: (1) Single-cell or Bulk; (2) One-time or Time series; (3) Stationary, with general intervention, or with specific intervention.
- For Single-cell + Time series data, there is an extra dimension of Joint distribution or Marginal distribution.
- According to these dimensions, we have 15 different data types.

Data types

	One-time		Time series		
	Bulk	Single-cell	Bulk	Single-cell	
				Marginal distribution	Joint distribution
Stationary					
General intervention					
Specific intervention					

All 15 data types.

Questions?

Structure inference

- The goal is to infer GRN structure with gene expression data.
- Different data types 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 directed cycle, or the gene expression levels satisfy a linear differential equation system.

Structure inference

- Some data types are more informative than other data types.
- Bulk $<$ Single-cell
- One-time $<$ Time series
- Marginal distribution $<$ Joint distribution
- Stationary $<$ General intervention $<$ Specific intervention
- Given a more informative data type, we can transform it into a less informative data type. If a GRN inference method works for a less informative data type, it automatically works for a more informative data type.
- Nevertheless, for more informative data types, generally the experiments are more difficult, more expensive, and less accurate.

Structure inference

- For each data type, we discuss whether the GRN structure can be inferred.
- Some data types have specific inference methods. Some other data types do not have specific inference methods. They need to be transformed into less informative types, and the corresponding methods may apply.

Structure inference

	One-time		Time series		
	Bulk	Single-cell	Bulk	Single-cell	
				Marginal distribution	Joint distribution
Stationary	1: No	2: Yes	3: No	4: Ditto	5: Yes
General intervention	6: No	7: Ditto	8: Yes	9: Ditto?	10: Ditto
Specific intervention	11: Yes	12: Ditto	13: Ditto	14: Ditto	15: Ditto

No means the GRN structure cannot be inferred. Ditto means using the same method for a less informative data type. Yes means existing specific inference methods.

Structure inference

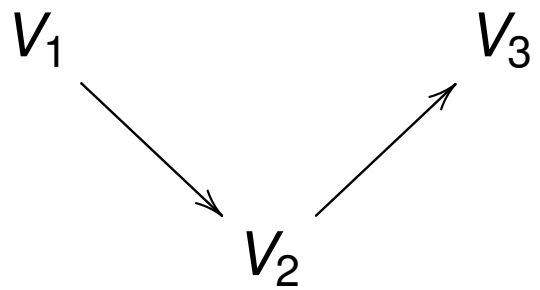
- Data types 2, 5, 8 are commonly studied.
- For data type 11, the GRN structure can be partially inferred.
- Data type 9 might have specific inference methods.
- In practice, data types 2, 8, 9 are more common.

- Questions?

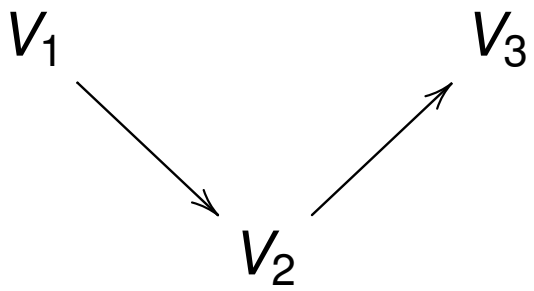
- Next: examples of inference methods for data types 2, 5, 8, 11.
- Such methods need different mathematics: combinatorics, probability, statistics, differential equations, etc. Besides, we need the help of computational mathematics and computer science to efficiently implement such methods.

Data type 11

- For data type 11, bulk level one-time gene expression data under specific interventions, we know that after interfering with any one gene, what other genes are also affected. 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.

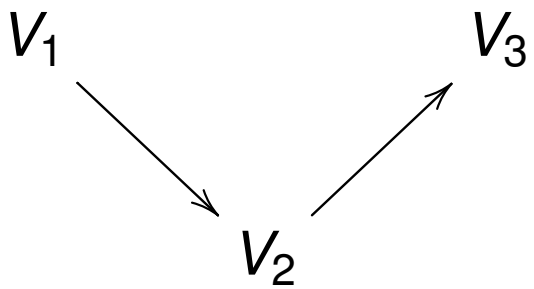


Data type 11

- 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.
- 

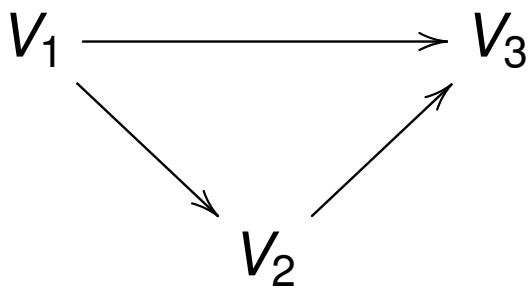
V_1 has descendants V_2, V_3; V_2 has descendant V_3; V_3 has no descendant.
- If we add intervention on gene V_i, then the descendants of V_i are also affected.

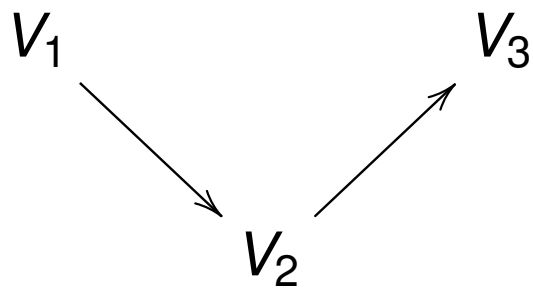
Data type 11

- After adding intervention on gene V_i, if gene V_j is also affected (compared to the situation before intervention), 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.

Data type 11

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



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

Data type 11

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

- 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 $V_2 \to V_3$.

Data type 11

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

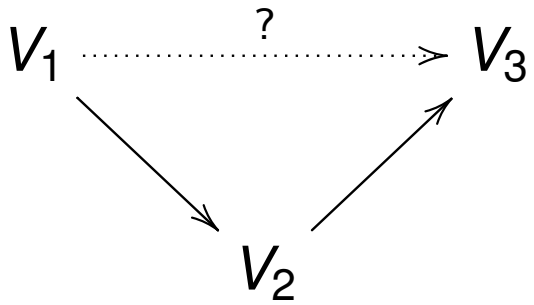
- If an edge $V_i \to V_j$ appears in none of these AD equivalent DAGs, we can determine the edge $V_i \to V_j$ does not exist in the GRN.
- We can determine that the GRN does not have edges $V_3 \to V_2$, $V_3 \to V_1$, and $V_2 \to V_1$.

Data type 11

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

- 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 $V_1 \to V_3$.

Data type 11

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

  We can identify two edges in the GRN. One edge is unknown.
- In sum, the GRN structure can be partially inferred.
- In practice, for a DAG with n vertices, there might be exponentially many AD equivalent DAGs. It is not feasible to find all AD equivalent DAGs to determine which edges can be inferred.

Data type 11

We have a quick algorithm for determining edges 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.

The proof uses combinatorics.

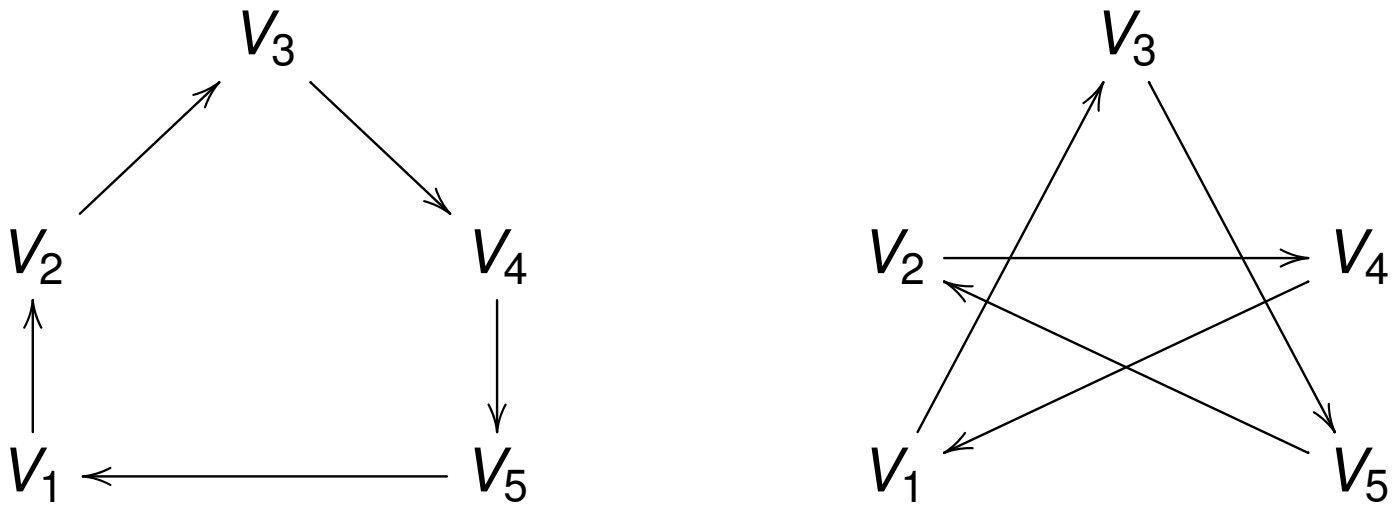
Data type 11

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.

Data type 11

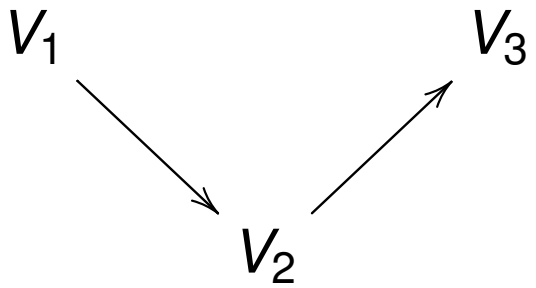
- If the GRN has cycles, we might infer no edge.
- 

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

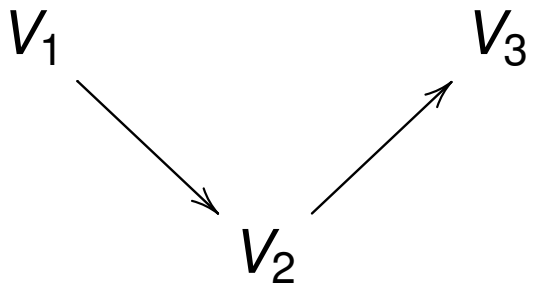
- Questions?

Data type 2

- For data type 2, single-cell level one-time gene expression data at stationary, we have random variables $X_1, X_2, \ldots, X_n$ as the expression levels of genes $V_1, V_2, \ldots, V_n$. All edges can be determined, but the direction of certain edges cannot be determined.
- This is a very common data type in reality, and there have been many methods developed for this data type: LocalBN, TIGRESS, ARACNe, PCA-CMI, GENIE3, GENIX, CausalCell, GRNUlar, etc.
- The basic idea: if gene V_i regulates V_j, then the levels of V_i and V_j are correlated. This means we can use the level of V_i to predict the level of V_j (and vice versa).



Data type 2

- 

  One idea is to calculate the Pearson correlation coefficient ρ for each gene pair V_i, V_j. There is an edge between V_i, V_j if and only if ρ is significantly different from 0.
- If the regulation relationship is $V_i \to V_k \to V_j$, then V_i does not directly regulate V_j, but V_i and V_j might be correlated.
- One can calculate the partial correlation of V_i, V_j conditioned on V_k. This excludes indirect regulations.
- (Partial) correlation coefficients can be replaced by other information theory quantities.

Data type 2

- Another approach is to use the levels of $V_2, \ldots, V_n$ to predict the level of V_1, and check which gene has a higher prediction ability for V_1.
- The prediction method can be regression (TIGRESS), random forest (GENIE3), and others (CausalCell).
- This provides a quantitative measurement for the regulation strength between two genes.
- Add a regularization term to obtain sparse results, since we do not want the GRN to be too dense.

Data type 2

- Correlation coefficient is symmetric with variables: $\rho(X, Y) = \rho(Y, X)$.
- For prediction methods, if X has a high prediction ability for Y, then in general Y also has a high prediction ability for X.
- This is an essential problem for data type 2: how to determine the direction of a regulation relation.
- One solution is to add specific interventions. For a regulation between V_1, V_2 with unknown direction, if adding intervention on V_1 can also affect V_2, then the direction should be $V_1 \rightarrow V_2$.

Data type 5

- Another solution of determining direction is to use time.
- For data type 5, single-cell level time series gene expression data at stationary (joint distribution of different time points is known), we have the joint distribution $\mathbb{P}[X_i(0) = c_0, X_i(1) = c_1, X_i(2) = c_2, \ldots]$ for $i = 1, 2, \ldots, n$. All edges can be determined, including the directions.
- Although obtaining data type 5 is technically difficult (almost impossible), there have been some inference methods: Granger causality, transfer entropy, dynGENIE3, BiXGBoost, TCDF, etc.

Data type 5

- Similar to data type 2, we can use the levels of $V_2, \ldots, V_n$ at time t to predict the level of V_1 at time $t+1$.
- A causal relation can only travel forward along time.
- With the time dimension, some fancy deep learning methods can be applied.

Data type 8

- For data type 8, bulk level time series gene expression data with general intervention, we have deterministic expression levels $x_i(0), x_i(1), x_i(2), \ldots$ for $i = 1, 2, \ldots, n$. All edges can be determined, including the directions.
- There have been some inference methods: DBN, NonlinearODEs, etc.
- Use the $x_i(t)$ data to infer the ODE system:

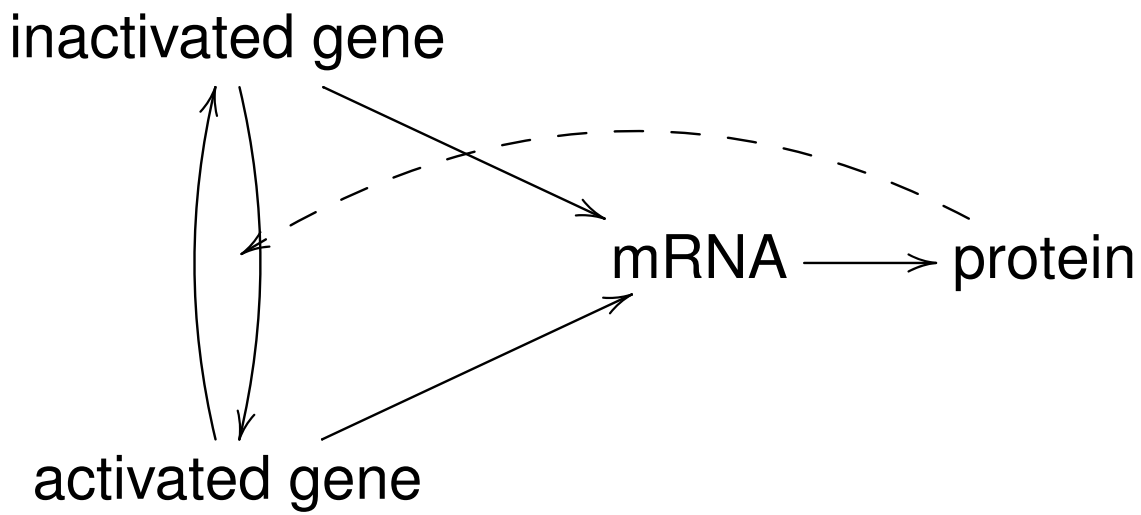
$$\frac{\mathrm{d}x_j(t)}{\mathrm{d}t} = f(x_1, x_2, \ldots, x_n) - cx_j.$$

- Whether f is linear or nonlinear, add a regularization term to obtain a sparse expression of f.

- Questions?

Autoregulation

- So far, we only consider the regulation between two different genes (mutual regulation). Some genes can regulate their own expression, which is called autoregulation.
- Autoregulation includes auto-activation and auto-repression.
- General mechanism of autoregulation:



Autoregulation

- If gene V_1 does not have autoregulation, then its level should satisfy

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = f(x_2, \ldots, x_n) - cx_1.$$

  Here the synthesis rate $f(x_2, \ldots, x_n)$ does not depend on x_1, and the degradation rate cx_1 is linear with x_1.
- Autoregulation means the synthesis rate of one gene depends on itself, and/or the degradation rate has a more complicated form.

Autoregulation

- Determining autoregulation is much more difficult than determining mutual regulation, whether by biochemical methods or by inference methods.
- There are some (not many) methods to infer autoregulation from gene expression data.
- Data types 2, 5, 8 have specific inference methods for autoregulation.

Autoregulation

- For data type 8, bulk level time series gene expression data with general intervention, we have deterministic expression levels $x_i(0), x_i(1), x_i(2), \ldots$ for $i = 1, 2, \ldots, n$. Autoregulation can be fully determined.
- One just needs to fit the expression data $x_i(t)$ to an ODE system, and check if it has the form

$$\frac{\mathrm{d}x_1(t)}{\mathrm{d}t} = f(x_2, \ldots, x_n) - cx_1.$$

Autoregulation

- For data type 5, single-cell level time series gene expression data at stationary (joint distribution of different time points is known), we have the joint distribution $\mathbb{P}[X_i(0) = c_0, X_i(1) = c_1, X_i(2) = c_2, \ldots]$ for $i = 1, 2, \ldots, n$. Autoregulation can be fully determined.
- In this situation, the expression level can be modeled by a birth-death process. Given the joint distribution of expression at different time points, we can directly calculate the birth rate and death rate. Then we just need to check how these rates depend on the expression level.

Autoregulation

- For data type 2, single-cell level one-time gene expression data at stationary, we have random variables $X_1, X_2, \ldots, X_n$ as the expression levels of genes $V_1, V_2, \ldots, V_n$. Autoregulation can be partially determined.
- We assume the GRN has no directed cycle. Otherwise, we cannot distinguish feedback loop from autoregulation.
- Build a continuous-time Markov chain model: consider a Markov chain Y with transition rate q_{ij}. Y represents all other genes that regulate gene V_1.
- The expression level of V_1 is a linear birth-death process X that depends on Y: When $Y = i$, the transition rate from $X = n$ to $X = n + 1$ is F_i, and the transition rate from $X = n$ to $X = n - 1$ is nG_i. Here F_i and G_i only depend on $Y = i$, but not $X = n$, meaning that V_1 does not have autoregulation.

Autoregulation

- The master equation of $[X(t), Y(t)]$ is

$$\begin{aligned}
&\frac{\mathrm{d}\mathbb{P}[X(t)=n, Y(t)=i]}{\mathrm{d}t} \\
=&\mathbb{P}[X(t)=n-1, Y(t)=i]F_i \\
&+\mathbb{P}[X(t)=n+1, Y(t)=i]G_i(n+1) \\
&+\sum_{j\neq i}\mathbb{P}[X(t)=n, Y(t)=j]q_{ji} \\
&-\mathbb{P}[X(t)=n, Y(t)=i](F_i+G_i n+\sum_{j\neq i}q_{ij}).
\end{aligned}$$

Autoregulation

- We consider a quantity: variance-to-mean ratio (VMR):

$$\mathrm{VMR}(X) = \frac{\mathrm{Var}(X)}{\mathbb{E}X} = \frac{\mathbb{E}(X^2) - (\mathbb{E}X)^2}{\mathbb{E}X}.$$

Theorem

In the above model without autoregulation, at stationary, we have VMR$(X) \geq 1$*.*

- Therefore, if we find $\mathrm{VMR}(X) < 1$, then there is autoregulation.

- Questions?

Notes

- Most single-cell data are not very accurate. For instance, single-cell RNA sequencing might only catch 10% genes that are expressing in each cell. A challenge is to develop methods that can deal with missing values.
- Various data pre-processing (e.g., normalization) should be applied before inference.
- A common problem is cell heterogeneity: the sampled cells might not be of the same type. Assume genes V_1, V_2 are independent in type A cells, and they are also independent in type B cells. When we sample from a mixture of type A cells and type B cells, V_1, V_2 might look dependent.

Notes

- There are not many experimental data sets with known GRN. We lack a gold standard to evaluate GRN inference methods.
- In practice, it is common to test inference methods on synthetic data. How can we guarantee that the synthetic data are generated by a mechanism that fits with reality?
- Most inference methods only work on tens of genes. Given a data set of thousands of genes, we need to find a small subset to apply inference methods. (For a gene of interest, select other genes that are highly correlated with it.)
- With the fast development of experimental methods, there will be new data types, and new inference methods are required. Various mathematical techniques might be useful.

Summary

- Introduce the GRN structure inference problem.
- Classify the inference problem into 15 subproblems by data types.
- For different data types, present corresponding inference methods.
- Introduce the autoregulation inference problem and corresponding methods.
- This work provides a unified framework to discuss the GRN structure (including autoregulation) inference problem.

References

- Wang, Y., & Wang, Z. (2022). Inference on the structure of gene regulatory networks. Journal of Theoretical Biology, 539, 111055.
- Wang, Y., & He, S. (2023). Inference on autoregulation in gene expression with variance-to-mean ratio. Journal of Mathematical Biology, 86(5), 87.