Gene Regulatory Network Inference with Covariance Dynamics

Yue Wang

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

yw4241@columbia.edu

May 01, 2024

Outline

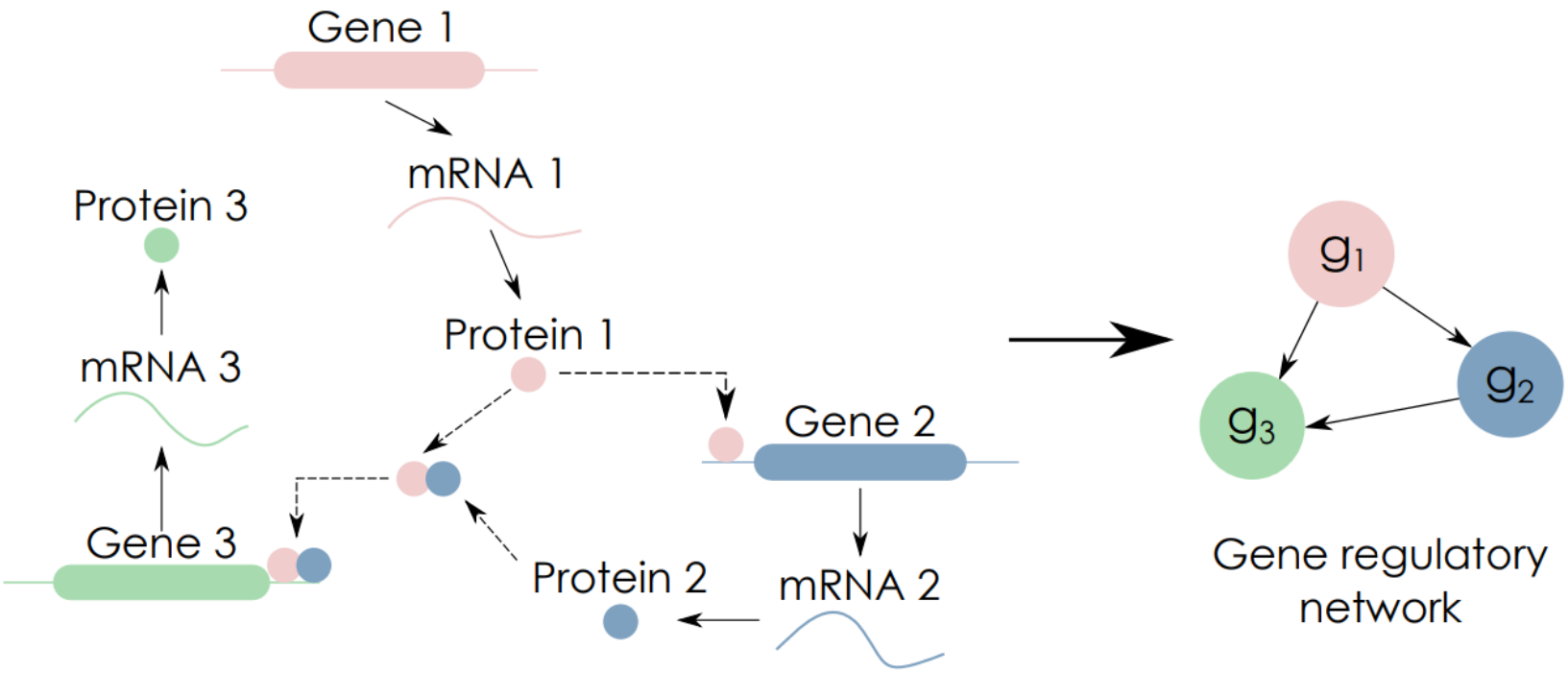
- Introduction to gene regulatory network (GRN).
- Types of data that can be used to infer GRN structures.
- Introduce a novel GRN inference method, WENDY, which is based on the dynamics of covariance matrices.

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/or proteins).
- Some regulators are small molecules, such as oxygen, sugars and vitamins. Some regulators are proteins. We focus on regulations between genes.

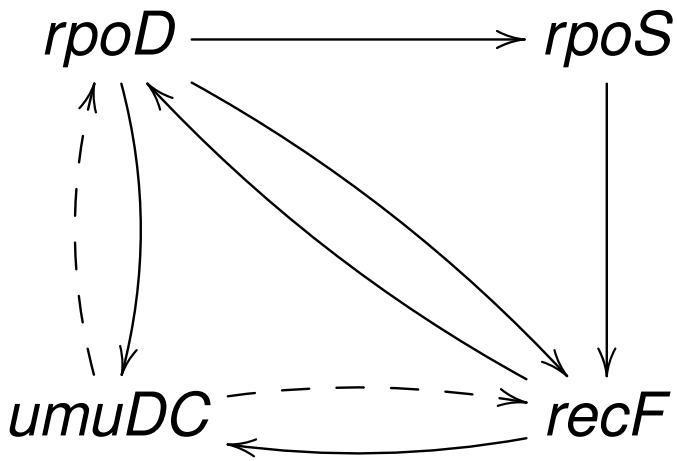
$$\text{gene} \xrightarrow{\text{expression}} \text{protein} \xrightarrow{\text{regulation}} \text{another gene}$$

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, the measurement of 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 recently.
- 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.

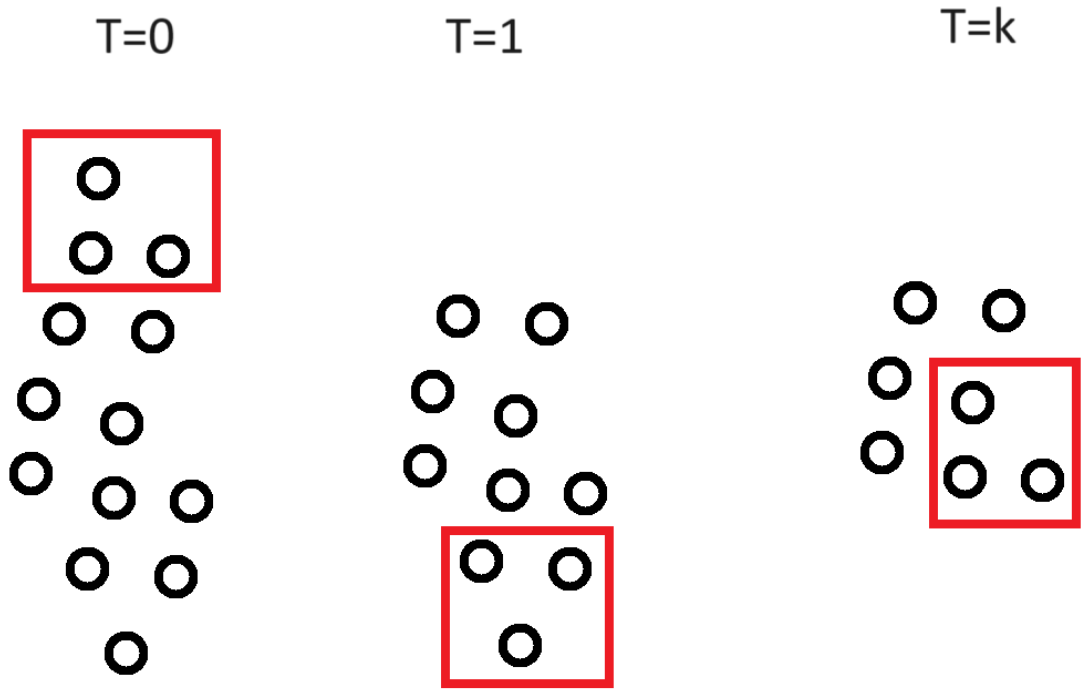
Data types: Single-cell vs Bulk

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

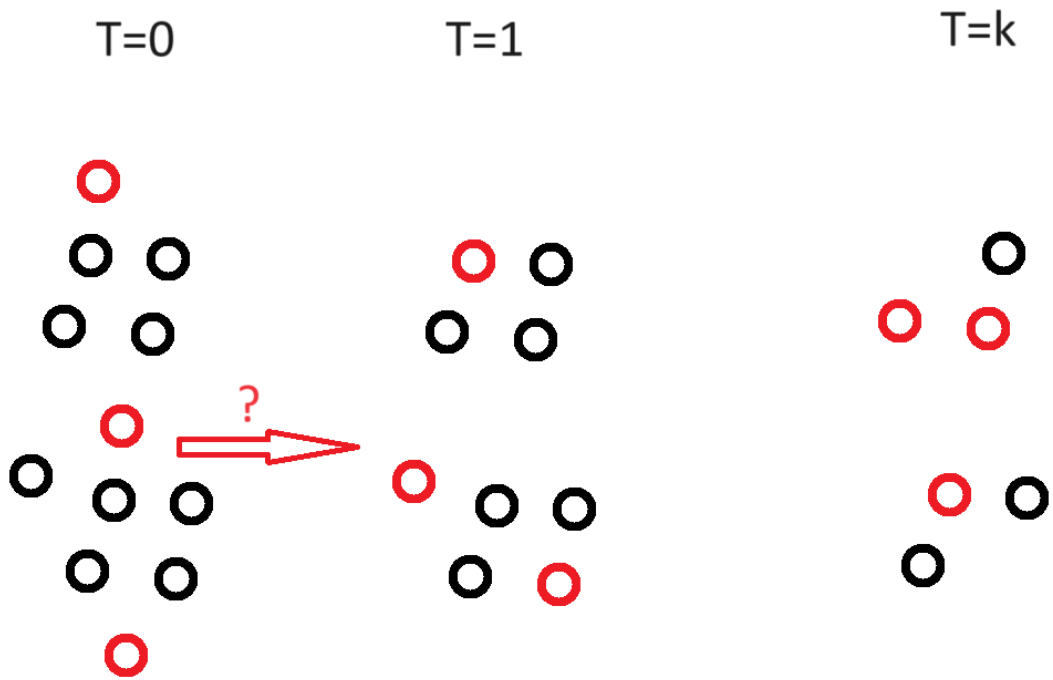


Data types: One-time vs. Time series

- 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

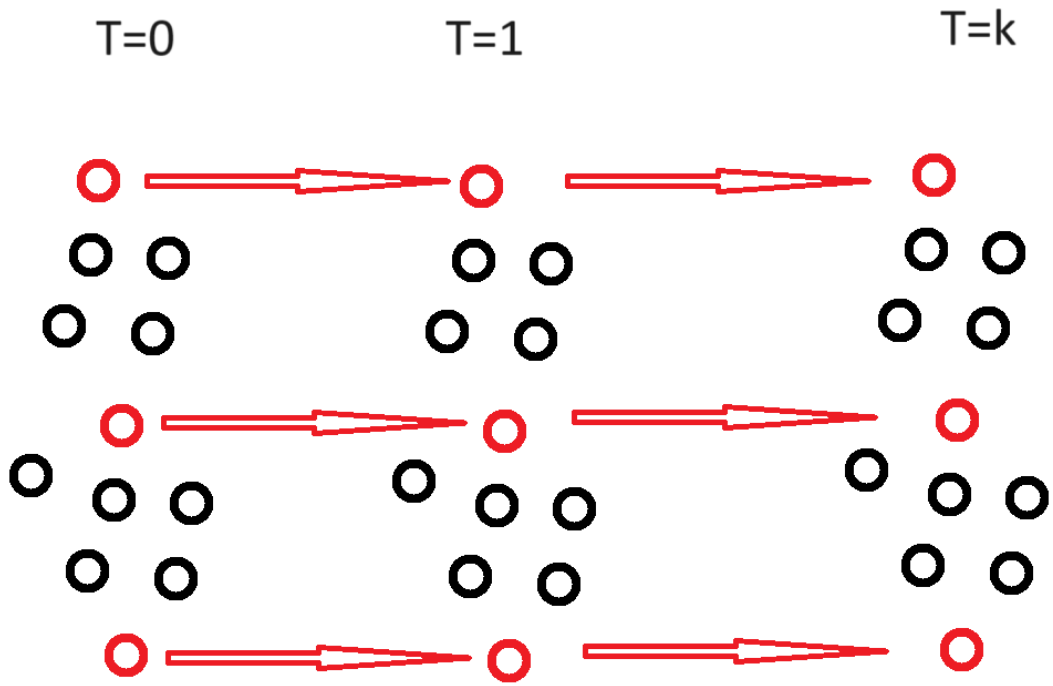
- At single-cell level, if the same cell can be measured only once, we can only obtain the marginal distribution for each time point, $\mathbb{P}[X_i(0) = c_0], \mathbb{P}[X_i(1) = c_1], \mathbb{P}[X_i(2) = c_2], \ldots$.



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

Data types: Joint distribution vs. Marginal distribution

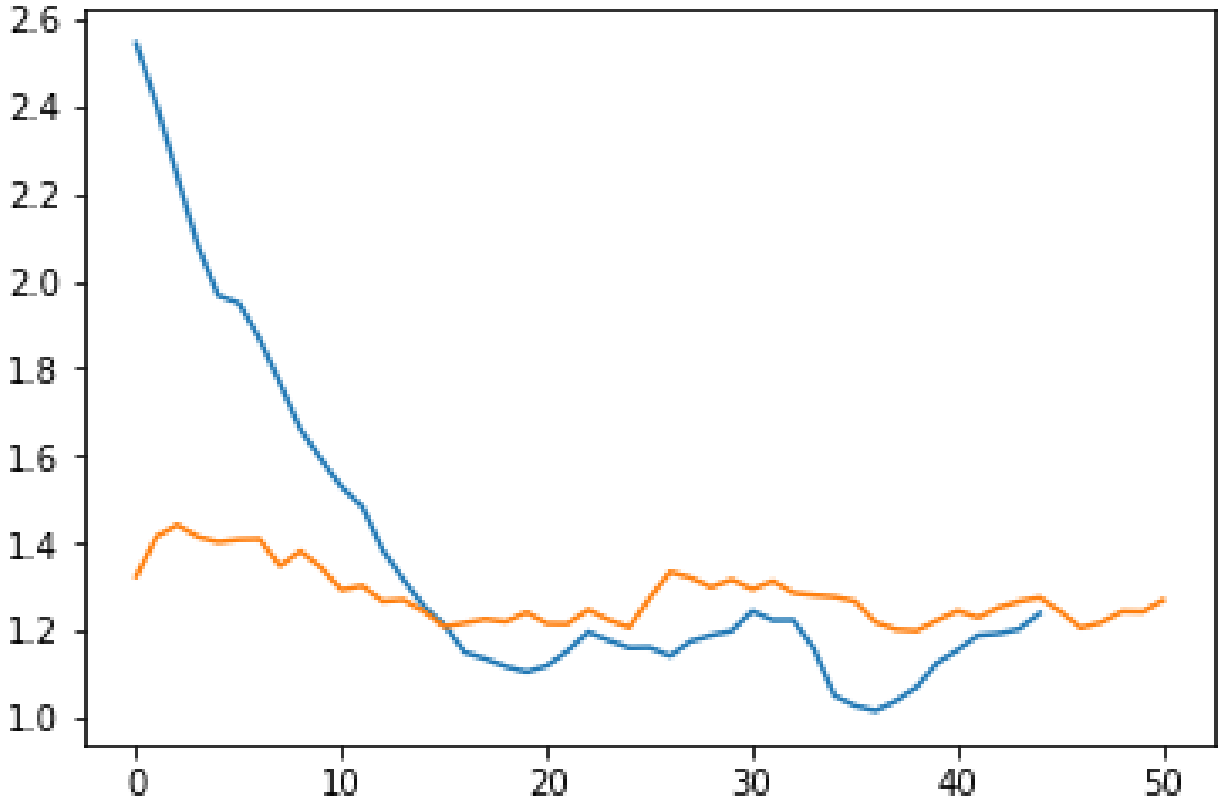
- If the same cell can be measured multiple times, we obtain the joint distribution for multiple time points, $\mathbb{P}[X_i(0) = c_0, X_i(1) = c_1, X_i(2) = c_2, \ldots]$.



- 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 related downstream 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.

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.
- Some methods cannot infer all regulatory relations. Some methods can infer all regulatory relations, but the directions are unknown (X regulates Y or Y regulates X). Some methods can infer all regulatory relations including the directions.

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.
- GRN inference methods generally cannot infer autoregulation (one gene regulates itself).

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: Yes?	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.
- In practice, data types 2, 8, 9 are more common.
- Data type 9 has only one specific method, SINCERITIES, which is inefficient in utilizing data (requires at least 6 time points to function well).
- We want to develop a new inference method for data type 9.

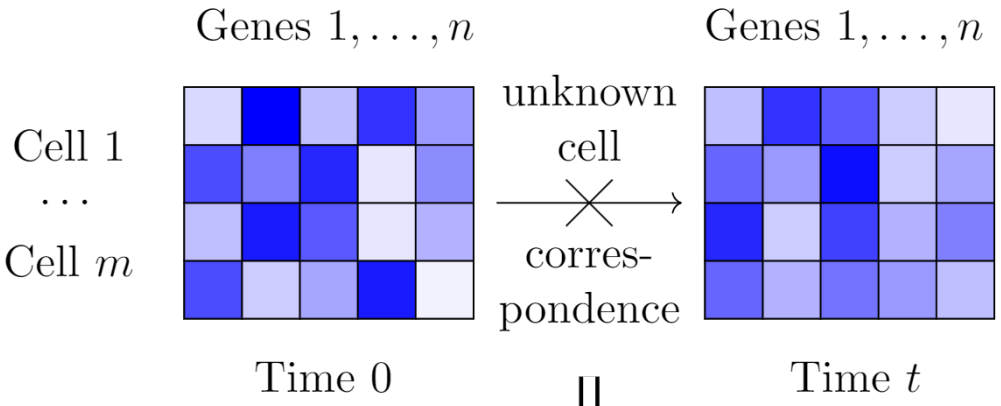
- Questions?

WENDY method

- For data type 9, we develop an algorithm of netWork infErence by covariaNce DYnamics (WENDY), which only requires data at 2 time points.
- WENDY can infer all regulatory relations, including their directions, but not autoregulation.
- Data type 9 means single-cell level time series data without joint distribution of different time points, measured after general interventions.
- We first sketch the workflow of WENDY, and then derive the mathematics behind this method.

WENDY method

single-cell gene expression data



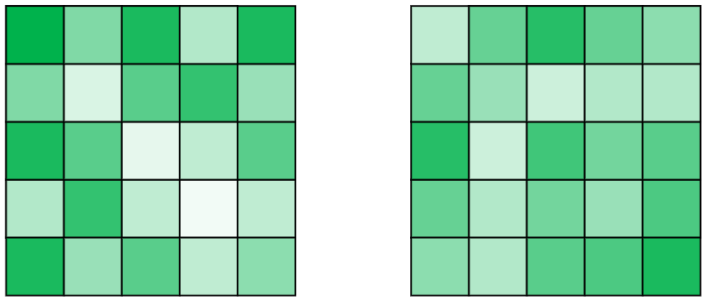
optimization problem

$$\arg\min_A \quad \tfrac{1}{2}\textstyle\sum_{i\neq j}\{[K(t)$$

$$-(I+tA^{\mathrm{T}})K(0)(I+tA)]^2_{i,j}+\lambda A^2_{i,j}\}$$

solve A, the GRN

covariance matrices of genes



gene expression model

$$K(t)=(I+tA^{\mathrm{T}})K(0)(I+tA)+D$$

WENDY method

- After adding drugs or other interventions, we measure the expression levels of n genes at time 0 and t as $n \times 2$ random variables: $X(0) = [X_1(0), \ldots, X_n(0)]$, $X(t) = [X_1(t), \ldots, X_n(t)]$.
- Since we add interventions to drive the system away from stationary, $X(0)$ and $X(t)$ are not identically distributed.
- We take expectations for $X(0)$ and $X(t)$ to obtain deterministic expression levels $x(0) = [x_1(0), \ldots, x_n(0)]$ and $x(t) = [x_1(t), \ldots, x_n(t)]$.

WENDY method

- Assume that $x(\tau)$ satisfies a linear ODE system:

$$\mathrm{d}x(\tau)/\mathrm{d}\tau = x(\tau)A + b.$$

- Here A is an invertible $n \times n$ matrix, representing the GRN we want: $A_{i,j} > 0$ means gene i activates gene j; $A_{i,j} < 0$ means gene i inhibits gene j; and $A_{i,j} = 0$ means gene i does not regulate gene j directly.
- However, $A_{i,i}$ is the combination of degradation and possibly autoregulation of gene i, and $A_{i,i} \neq 0$ does not necessarily mean autoregulation of gene i.
- Besides, b is a $1 \times n$ vector, representing the base synthesis rate.

WENDY method

- The solution is

$$x(t) = [x(0) + bA^{-1}]e^{tA} - bA^{-1}.$$

- Its first-order approximation is

$$x(t) = x(0)(I + tA) + tb.$$

- Inspired by this deterministic relation, we can write down a similar trajectory/cell-wise relation for the random variables:

$$X(t) = X(0)(I + tA) + tb + X(0) \odot \epsilon(t).$$

- Here $\epsilon(t) = [\epsilon_1(t), \ldots, \epsilon_n(t)]$ is an n-dimensional normal random noise with mean $(0, \ldots, 0)$ and a diagonal covariance matrix.
- $\odot$ is the entrywise (Hadamard) product: $X(0) \odot \epsilon(t) = [X_1(0)\epsilon_1(t), \ldots, X_n(0)\epsilon_n(t)]$.

WENDY method

Given the equation of $X(t)$, we want to derive the dynamics of $K(t)$.

$$
\begin{aligned}
K(t) =& \mathbb{E}\{[X(t)^{\mathrm{T}} - x(t)^{\mathrm{T}}][X(t) - x(t)]\} \\
=& \mathbb{E}\{(I + tA^{\mathrm{T}})[X(0)^{\mathrm{T}} - x(0)^{\mathrm{T}}][X(0) - x(0)](I + tA)\} \\
&+ \mathbb{E}\{(I + tA^{\mathrm{T}})[X(0)^{\mathrm{T}} - x(0)^{\mathrm{T}}][X(0) \odot \epsilon(t)]\} \\
&+ \mathbb{E}\{[X(0) \odot \epsilon(t)]^{\mathrm{T}}[X(0) - x(0)](I + tA)\} \\
&+ \mathbb{E}\{[X(0) \odot \epsilon(t)]^{\mathrm{T}}[X(0) \odot \epsilon(t)]\} \\
=& (I + tA^{\mathrm{T}})\mathbb{E}\{[X(0)^{\mathrm{T}} - x(0)^{\mathrm{T}}][X(0) - x(0)]\}(I + tA) \\
&+ \mathbb{E}\{(I + tA^{\mathrm{T}})[X(0)^{\mathrm{T}} - x(0)^{\mathrm{T}}][X(0) \odot \mathbb{E}\epsilon(t)]\} \\
&+ \mathbb{E}\{[X(0) \odot \mathbb{E}\epsilon(t)]^{\mathrm{T}}[X(0) - x(0)](I + tA)\} \\
&+ \mathbb{E}[X(0)^{\mathrm{T}}X(0)] \odot \mathbb{E}[\epsilon(t)^{\mathrm{T}}\epsilon(t)] \\
=& (I + tA^{\mathrm{T}})K(0)(I + tA) + D.
\end{aligned}
$$

Here D is an unknown diagonal matrix.

WENDY method

- We want to find A to (approximately) satisfy

$$K(t) - (I + tA^{\mathrm{T}})K(0)(I + tA) = D.$$

- Since the diagonal matrix D is unknown, we only want to match off-diagonal elements of $K(t)$ and $(I + tA^{\mathrm{T}})K(0)(I + tA)$.
- Now we have an optimization problem:

$$\min_A f_\lambda(A) := \frac{1}{2}\sum_{i\neq j}\{[K(t) - (I + tA^{\mathrm{T}})K(0)(I + tA)]_{i,j}\}^2 + \lambda A_{i,j}^2.$$

- This non-convex optimization problem can be solved numerically with BFGS or other methods.
- In simulations, we find that the regularization term $\lambda A_{i,j}^2$ does not help much. We can set $\lambda = 0$ in practice.

WENDY method

- In practice, it is possible that the cell number m is smaller than the gene number n.
- Directly calculating the covariance matrix will lead to a degenerate matrix.
- We can use graphical lasso or other methods to estimate an invertible covariance matrix.
- Overall procedure of WENDY: Use graphical lasso to calculate $K(0), K(t)$. Then use BFGS to solve

$$\min_A f_\lambda(A) := \frac{1}{2}\sum_{i\neq j}\{[K(t) - (I + tA^{\mathrm{T}})K(0)(I + tA)]_{i,j}\}^2 + \lambda A_{i,j}^2.$$

- A is the GRN we want.

- Questions?

Theoretical comparison

- For n genes, the GRN matrix A (ignoring diagonal entries) has $n^2 - n$ independent values.
- WENDY uses data at $T = 2$ time points to extract $n^2 + n$ independent values.
- The other method for data type 9, SINCERITIES, needs data at $T = n$ time points to extract $n^2 - n$ independent values. This explains why SINCERITIES needs data at more time points.

Performance on synthetic data

- We compare five GRN inference methods.
- WENDY works for data type 9 (also 10).
- SINCERITIES works for data type 9 (also 10).
- GENIE3 works for data type 2 (also 4, 5, 7, 9, 10).
- NonlinearODEs works for data type 8 (also 9, 10).
- dynGENIE3 works for data type 5 (also 10).
- Notice that dynGENIE3 does not work for data type 9.

Performance on synthetic data

- We test them on two synthetic data sets of data type 10 (so that all methods can be applied): DREAM4 data sets (10/100 genes) and SINC data sets (10/20 genes, 10/30/100 cells).
- We use two quantities to measure the performance: AUROC and AUPR.

		WENDY	SINCERITIES	Nonlinear ODEs	GENIE3	dynGENIE3
DREAM4 (10 genes)	AUROC	0.64	0.61	0.64	0.63	0.76
	AUPR	0.29	0.21	0.28	0.24	0.46
DREAM4 (100 genes)	AUROC	0.62	0.55	0.58	0.64	0.74
	AUPR	0.05	0.03	0.03	0.04	0.14
	Total	1.60	1.40	1.53	1.55	2.10

Performance on synthetic data

- WENDY has satisfactory performances on both data sets.

		WENDY	SINCERITIES	Nonlinear ODEs	GENIE3	dynGENIE3
SINC $m = 10$ (10 genes)	AUROC	0.59	0.44	0.47	0.46	0.43
	AUPR	0.14	0.11	0.11	0.11	0.10
SINC $m = 10$ (20 genes)	AUROC	0.76	0.42	0.51	0.52	0.40
	AUPR	0.16	0.05	0.08	0.08	0.04
SINC $m = 30$ (10 genes)	AUROC	0.67	0.46	0.49	0.50	0.44
	AUPR	0.19	0.12	0.12	0.15	0.10
SINC $m = 30$ (20 genes)	AUROC	0.81	0.47	0.53	0.47	0.40
	AUPR	0.21	0.07	0.08	0.07	0.04
SINC $m = 100$ (10 genes)	AUROC	0.69	0.48	0.50	0.55	0.42
	AUPR	0.22	0.13	0.12	0.19	0.10
SINC $m = 100$ (20 genes)	AUROC	0.82	0.48	0.54	0.54	0.39
	AUPR	0.23	0.08	0.08	0.09	0.05
	Total	5.49	3.31	3.63	3.73	2.91

Performance on synthetic data

- Time costs of different methods.

		WENDY	SINCERITIES	Nonlinear ODEs	GENIE3	dynGENIE3
SINC	$m = 10$	0.22 ± 0.04	0.09 ± 0.01	0.19 ± 0.02	4.27 ± 0.06	7.28 ± 0.08
$n = 10$	$m = 30$	0.15 ± 0.03	0.11 ± 0.01	0.18 ± 0.02	4.83 ± 0.05	15.86 ± 0.17
genes	$m = 100$	0.14 ± 0.03	0.12 ± 0.01	0.19 ± 0.03	7.32 ± 0.09	51.79 ± 0.63
SINC	$m = 10$	0.38 ± 0.10	0.42 ± 0.01	0.35 ± 0.03	8.55 ± 0.05	16.24 ± 0.13
$n = 20$	$m = 30$	0.34 ± 0.03	0.45 ± 0.01	0.36 ± 0.08	9.95 ± 0.07	38.21 ± 0.28
genes	$m = 100$	0.27 ± 0.02	0.64 ± 0.01	0.34 ± 0.04	16.32 ± 0.17	131.89 ± 0.76

- WENDY has a satisfactory time cost.

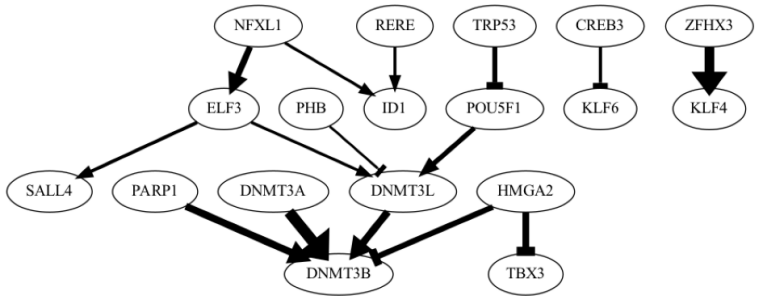
Application to experimental data

- mESC data set: Single-cell expression levels of mouse embryonic stem cells, measured at $t = 0, 12, 24, 48, 72$h.
- MEF data set: Single-cell expression levels of mouse embryonic fibroblast cells, measured at $t = 0, 2, 5, 20, 22$d.
- hESC data set: Single-cell expression levels of human embryonic stem cells, measured at $t = 0, 12, 24, 36, 72, 96$h.
- Since these cells are developing, the GRN might change along time.
- For each data set, we apply WENDY to each pair of neighboring time points, and compare the evolution of GRNs.

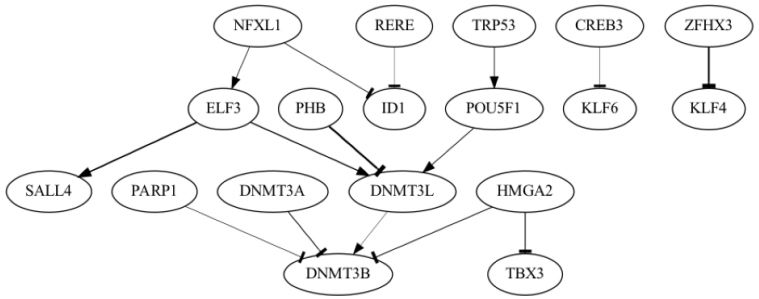
Application to experimental data

Width of an arrow means the strength of this regulatory relationship.

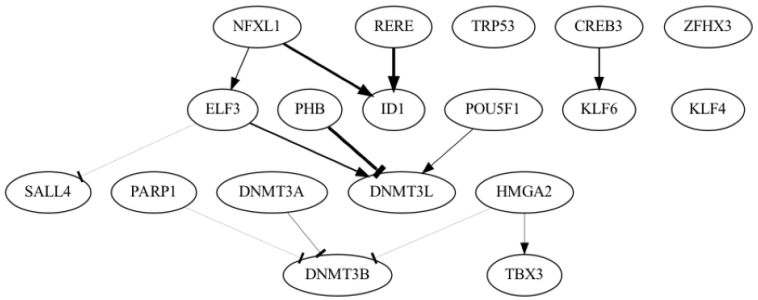
from t=0h to t=12h



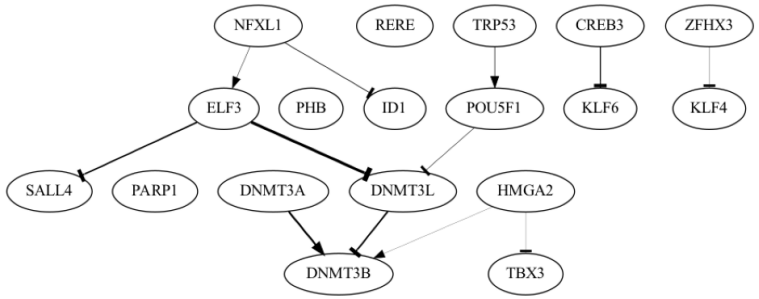
from t=12h to t=24h



from t=24h to t=48h



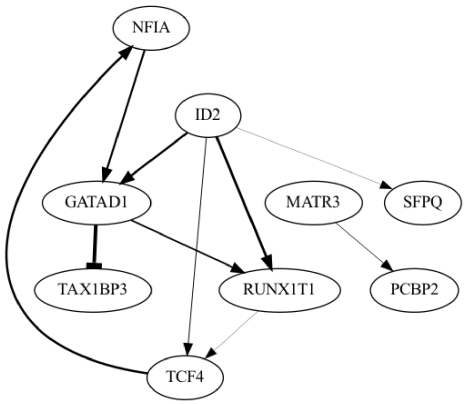
from t=48h to t=72h



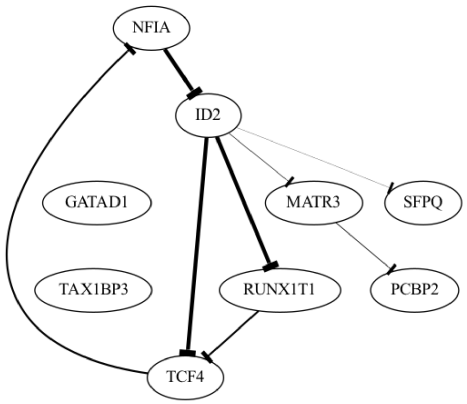
mESC data set: regulation strength decreases.

Application to experimental data

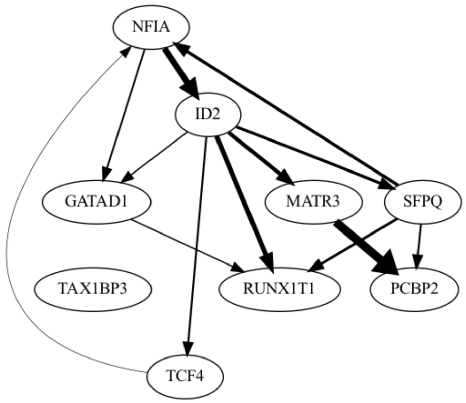
from t=0d to t=2d



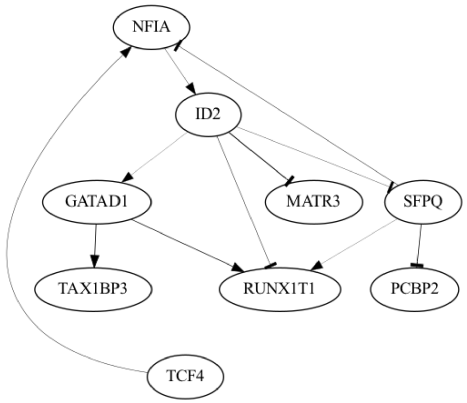
from t=2d to t=5d



from t=5d to t=20d



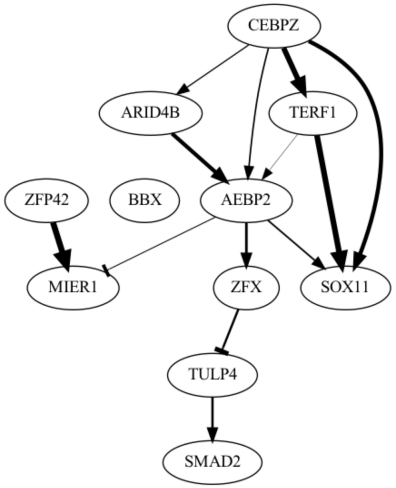
from t=20d to t=22d



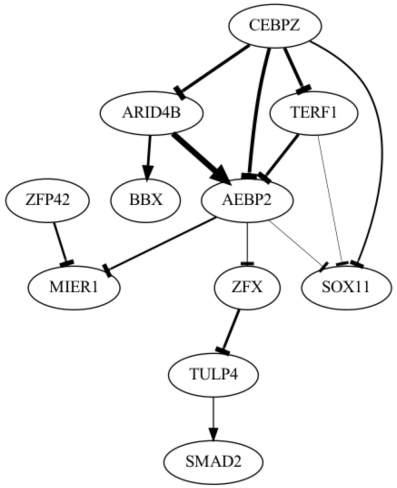
MEF data set: fluctuation of regulation strength.

Application to experimental data

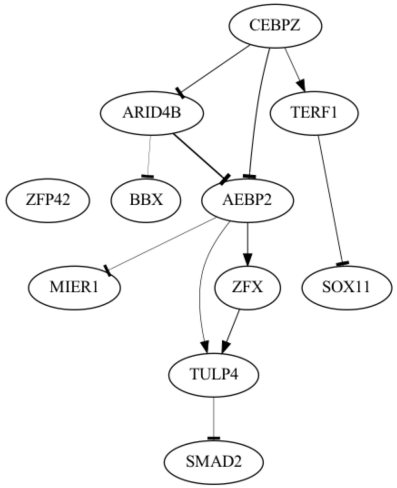
from t=0h to t=12h



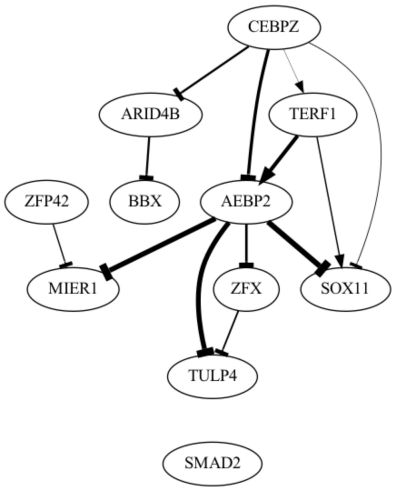
from t=12h to t=24h



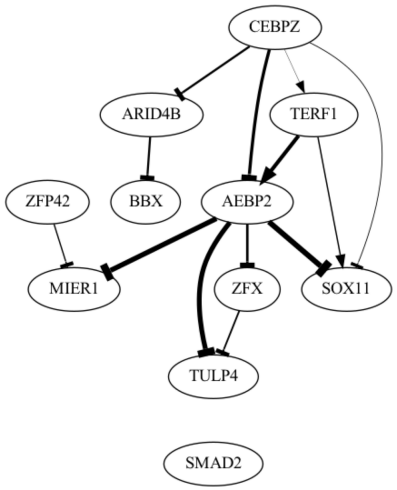
from t=24h to t=36h



from t=36h to t=72h



from t=72h to t=96h



hESC data set: center of regulation moves downstream.

Summary

- Introduce the GRN structure inference problem.
- Classify the inference problem into 15 subproblems by data types.
- For data type 9 (single-cell expression level data at multiple time points, measured after general interventions, where the joint distribution is unknown), introduce the WENDY method, which can infer the full GRN with data at 2 time points.

References

Yue Wang, Peng Zheng, Yu-Chen Cheng, Zikun Wang, and Aleksandr Aravkin. (2024). "WENDY: Gene Regulatory Network Inference with Covariance Dynamics." bioRxiv: 10.1101/2024.04.04.588131.

Code files: https://github.com/YueWangMathbio/WENDY