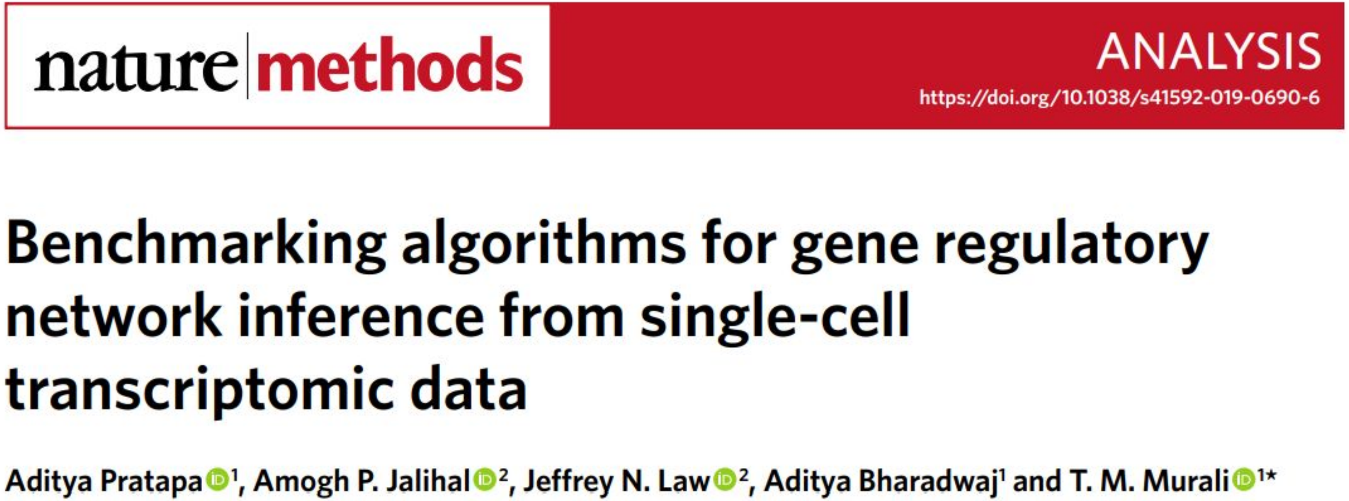
Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data

Presented by: Aji John & Johannes Linder, Yuliang Wang

UNIVERSITY *of* WASHINGTON



nature methods

ANALYSIS

https://doi.org/10.1038/s41592-019-0690-6

Benchmarking algorithms for gene regulatory network inference from single-cell transcriptomic data

Aditya Pratapa[1], Amogh P. Jalihal[2], Jeffrey N. Law[2], Aditya Bharadwaj[1] and T. M. Murali[1*]

Overview

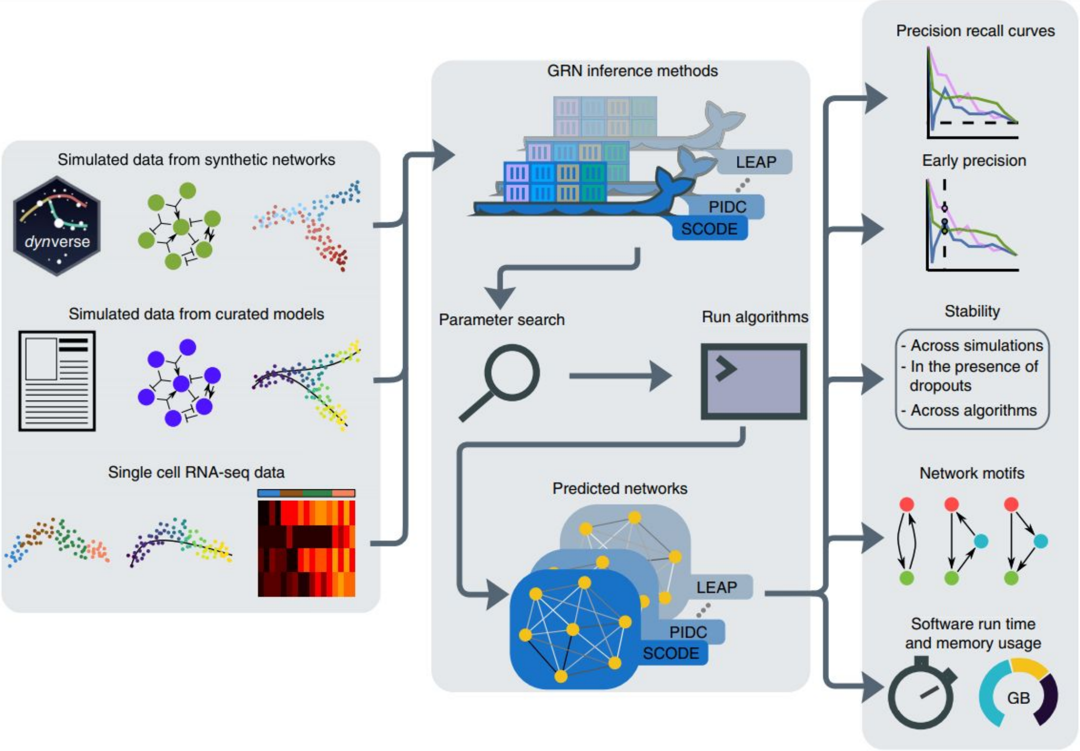


Fig. 1 | An overview of the BEELINE evaluation framework. We apply GRN inference algorithms to three types of data: datasets from synthetic networks,

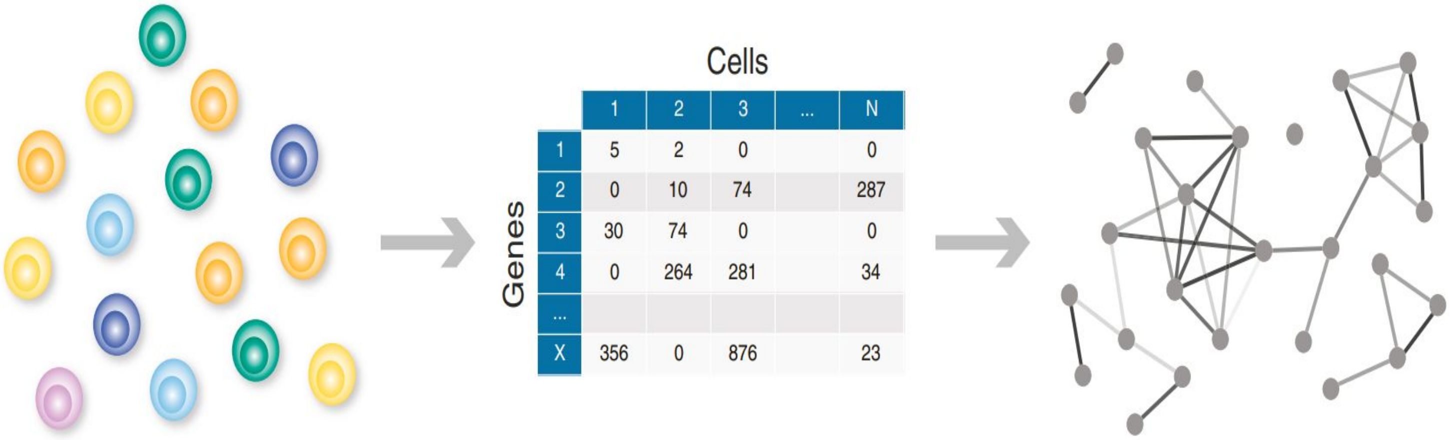
Overview

- What is a Gene Regulatory Network?
- Datasets: 400 Simulated and 5 Real sc-data
    - Synthetic, Curated (synthetic) and Real Single-cell data
    - BoolODE
- Methods: 12 Different Algorithms for GRN Inference
    - Random Forest
    - ODE and Regression
    - Correlation / Mutual Information / Causality
- Benchmarking Results

Overview

- What is a Gene Regulatory Network?
- Datasets: 400 Simulated and 5 Real sc-data
  - Synthetic, Curated (synthetic) and Real Single-cell data
  - BoolODE
- Methods: 12 Different Algorithms for GRN Inference
  - Random Forest
  - ODE and Regression
  - Correlation / Mutual Information / Causality
- Benchmarking Results

What is a Gene Regulatory Network?



	1	2	3	...	N
1	5	2	0		0
2	0	10	74		287
3	30	74	0		0
4	0	264	281		34
...					
X	356	0	876		23

Figure 1. Network Inference from Single-Cell Data

*** Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267 (2017).**

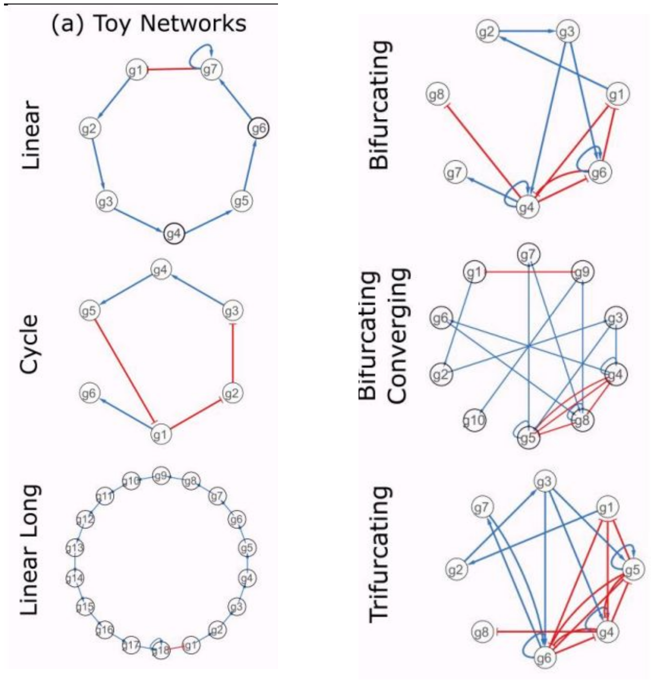
Overview

- What is a Gene Regulatory Network?
- Datasets: 400 Simulated and 5 Real sc-data
    - Synthetic, Curated (synthetic) and Real Single-cell data
    - BoolODE
- Methods: 12 Different Algorithms for GRN Inference
    - Random Forest
    - ODE and Regression
    - Correlation / Mutual Information / Causality
- Benchmarking Results

Synthetic GRN datasets

Simple in-silico single-cell gene expression datasets

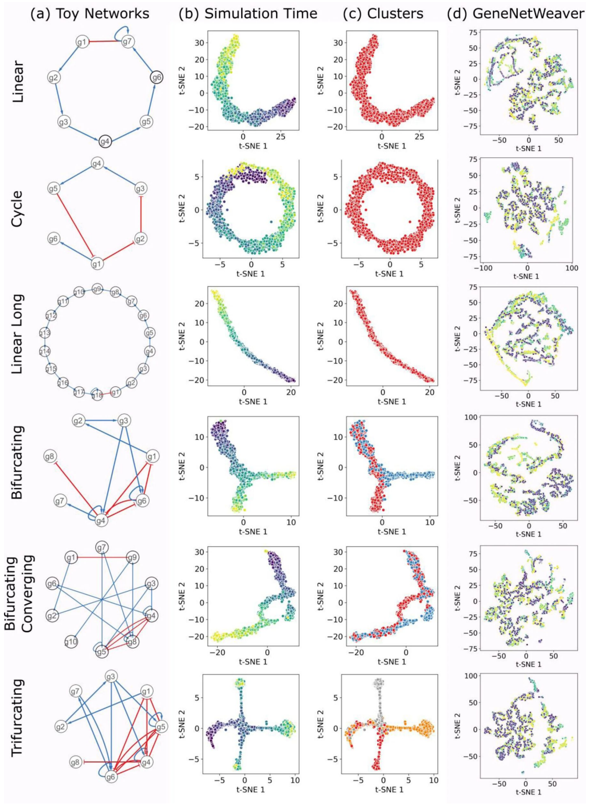
- **Not affected by pseudotime inference alg.**
- **Simple trajectories for these networks (linear, cycle ..)**
- **Used BoolODE to simulate the networks (coming up in the next set of slides)**



Synthetic GRN datasets

End product

- Used BoolODE by sampling parameters 10 times (5000 simulations per parameter set)
- 5 datasets per parameter set, one each with 100, 200, 500, 2000 and 5000 cells by sampling one cell per simulation. Finally got 50 different expression sets
- On the right is 2-D projection



Curated, simulated GRN datasets

Boolean models

- Viz. Mammalian cortical area development (mCAD), ventral spinal cord (VSC) development, hematopoietic stem cell (HSC) differentiation and gonadal sex determination (GSD)
- Used BoolODE to simulate it - 10 different sets with 2000 cells for each model
- Pseudotime using Slingshot

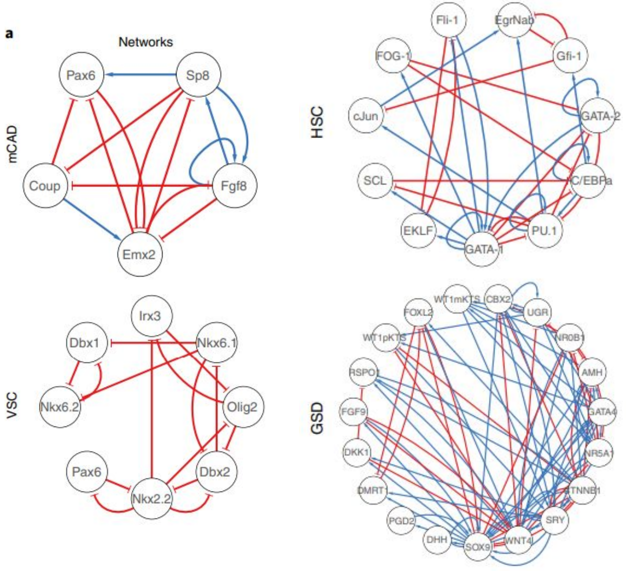


Fig. 3 | Visualization of *t*-SNE projections

Human & Mouse sc-Seq datasets

Single-cell RNA-seq datasets

- 2 in human and 3 in mouse cells (Total 7 cell types)

Dataset	Reference	Species	Starting cell type	Ending cell type(s)	#Cells	#Genes	# TFs
mHSC-E	Nestorowa *et al.*[1]	Mouse	HSCs	Erythroid	1,071	2,634	204
mHSC-L				Lymphoid	847	692	60
mHSC-GM				Granulocyte-Macrophage	889	1,595	132
mESC	Hayashi *et al.* [2]	Mouse	mESCs	Primitive endoderm	421	8,150	620
mDC	Shalek *et al.* [3]	Mouse	DCs	-	383	3,755	321
hHep	Camp *et al.* [4]	Human	iPSCs	Mature hepatocytes	425	4,336	311
hESC	Chu *et al.* [5]	Human	hESCs	Definitive endoderm	758	4,406	330

Human & Mouse sc-Seq datasets

Single-cell RNA-seq datasets - Ground truth

- **2 in human and 3 in mouse cells (Total 7 cell types)**

	Source	**#TFs**	**#Genes (incl. TFs)**	**#Edges**	**Density**	**Gene expression dataset**
Mouse	mHSC, E, L, G-M ChIP-Atlas	137	19,324	1,078,888	0.407	mHSC, Nestorowa *et al.*[1]
	mESC, ESCAPE+ ChIP-Atlas	247	25,703	6,348,394	0.154	mESC, Hayashi *et al.*[2]
	mESC, LOGOF, ESCAPE	57	18,427	104,797	0.1	mESC, Hayashi *et al.*[2]
	DC, ChIP-Atlas	36	11,092	30,658	0.077	mDC, Shalek *et al.*[3]
	TRRUST + RegNetwork	1,455	17,852	100,139	0.004	All mouse datasets
	STRING	1,350	7,771	157,134	0.015	
Human	HEPG2, ChEA + ChIP-Atlas	84	16,822	342,862	0.243	Camp *et al.*[4]
	hESC, ChEA + ChIP-Atlas	130	18,104	436,563	0.186	Chu *et al.*[5]
	TRRUST + RegNetwork + DoRothEA	2,165	23,566	386,293	0.008	All human datasets
	STRING	1,489	8,806	198,285	0.015	

Overview

- What is a Gene Regulatory Network?
- Datasets: 400 Simulated and 5 Real sc-data
    - Synthetic, Curated (synthetic) and Real Single-cell data
    - BoolODE
- Methods: 12 Different Algorithms for GRN Inference
    - Random Forest
    - ODE and Regression
    - Correlation / Mutual Information / Causality
- Benchmarking Results

Single-cell generation (BoolODE)

To sample a dataset of N cells:

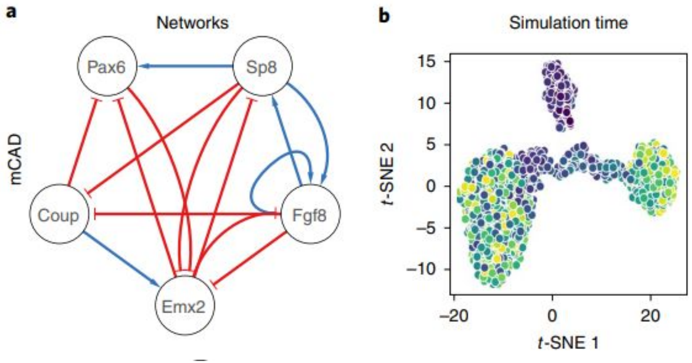
1. Initialize gene and protein concentrations $[x_i]$, $[p_i]$ at time 0 for every gene i.

2. Simulate the system ODE.

3. Sample state at an end time t.

$$\frac{\mathrm{d}[x_i]}{\mathrm{dt}} = mf(R_i) - l_x[x_i]$$

$$\frac{\mathrm{d}[p_i]}{\mathrm{dt}} = r[x_i] - l_p[p_i]$$



Single-cell generation (BoolODE)

To sample a dataset of N cells:

1. Initialize gene and protein concentrations $[x_i]$, $[p_i]$ at time 0 for every gene i.

2. Simulate the stochastic system ODE.

3. Sample state at an end time t.

$$\frac{\mathrm{d}[x_i]}{\mathrm{dt}} = mf(R_i) - l_x[x_i]$$

$$\frac{\mathrm{d}[p_i]}{\mathrm{dt}} = r[x_i] - l_p[p_i]$$

$$\Big\downarrow$$

$$\frac{\mathrm{d}[x_i]}{\mathrm{dt}} = mf(R_i) - l_x[x_i] + s\sqrt{[x_i]}\Delta W_t$$

$$\frac{d[p_i]}{dt} = r[x_i] - l_p[p_i] + s\sqrt{[p_i]}\Delta W_t$$

$$\Delta W_t = \mathcal{N}(0, h)$$

Single-cell generation (BoolODE)

- The regulatory function f(*R_i*) is constructed from the (known) GRN (regulator set *R_i*).

E.g. Gene X is governed by activator proteins P (or) Q, and inhibited by R.

$$X = (P \vee Q) \wedge \neg(R)$$

$$\frac{\mathrm{d}[x_i]}{\mathrm{dt}} = mf(R_i) - l_x[x_i]$$

Single-cell generation (BoolODE)

- The regulatory function f(R_i) is constructed from the (known) GRN (regulator set R_i).

=> f is a Hill function with P and Q in numerator and P, Q and R in denominator.

Parameters globally shared across genes.

Set to achieve target steady states.

$$X = (P \vee Q) \wedge \neg(R)$$

$$\Big\downarrow$$

$$\frac{d[X]}{dt} = m\left(\frac{[P]+[Q]+[P][Q]}{\begin{array}{c}1+[P]+[Q]+[R]+[P][Q]\\ +[P][R]+[Q][R]+[P][Q][R]\end{array}}\right) - l_x[X]$$

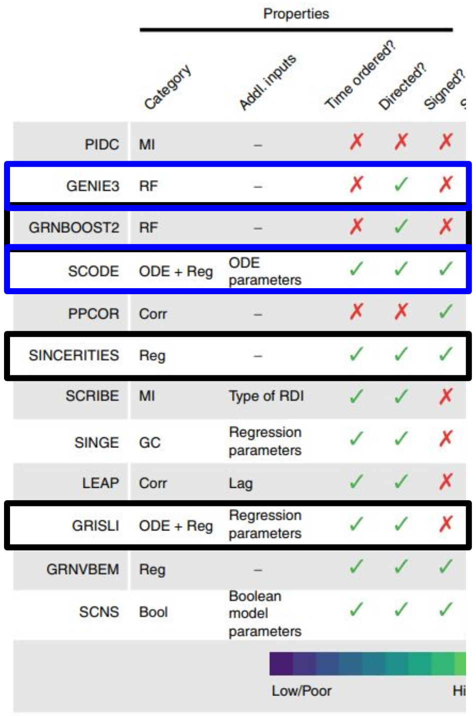
Overview

- What is a Gene Regulatory Network?
- Datasets: 400 Simulated and 5 Real sc-data
    - Synthetic, Curated (synthetic) and Real Single-cell data
    - BoolODE
- Methods: 12 Different Algorithms for GRN Inference
    - Random Forest
    - ODE and Regression
    - Correlation / Mutual Information / Causality
- Benchmarking Results

Prediction-based Algorithms

Algorithms based on predicting or modeling gene expression based on other gene profiles.

- **GENIE3 (Random Forest)**
- **GRNBOOST2 (Random Forest)**
- **SCODE (ODE + Regression)**
- **SINCERITIES (Sparse Regression)**
- **GRISLI (ODE + Regression)**



	Properties				
	Category	Addl. inputs	Time ordered?	Directed?	Signed?
PIDC	MI	–	✗	✗	✗
GENIE3	RF	–	✗	✓	✗
GRNBOOST2	RF	–	✗	✓	✗
SCODE	ODE + Reg	ODE parameters	✓	✓	✓
PPCOR	Corr	–	✗	✗	✓
SINCERITIES	Reg	–	✓	✓	✓
SCRIBE	MI	Type of RDI	✓	✓	✗
SINGE	GC	Regression parameters	✓	✓	✗
LEAP	Corr	Lag	✓	✓	✗
GRISLI	ODE + Reg	Regression parameters	✓	✓	✗
GRNVBEM	Reg	–	✓	✓	✓
SCNS	Bool	Boolean model parameters	✓	✓	✓

Low/Poor Hi

Fig. 6 | Summary of properties of GRN inference algorithms a

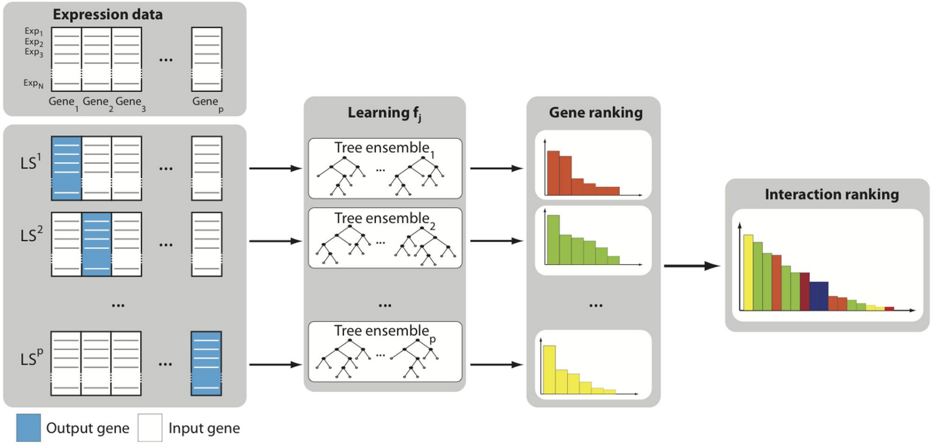
UNIVERSITY *of* WASHINGTON

GENIE3 (Random Forest)

Idea (for Genes *j* = 1 to *p*)

➢ Generate the learning sample of input-output pairs for gene j:

$$LS^j = \{(\mathbf{x}_k^{-j}, x_k^j), k = 1, \ldots, N\}.$$



Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5(9): e12776.

GENIE3 (for "GEne Network Inference with Ensemble of trees")

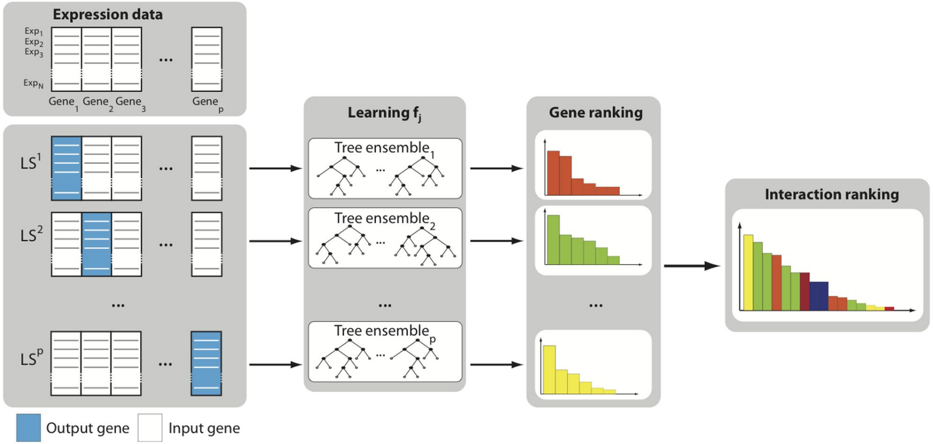
GENIE3 (Random Forest)

Idea (for Genes *j* = 1 to *p*)

- ➢ Generate the learning sample of input-output pairs for gene j:

$$LS^j = \{(\mathbf{x}_k^{-j}, x_k^j), k = 1, \dots, N\}.$$

- ➢ Use feature selection technique on LS^j to get confidence intervals $w_{i,j}$



Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5(9): e12776.

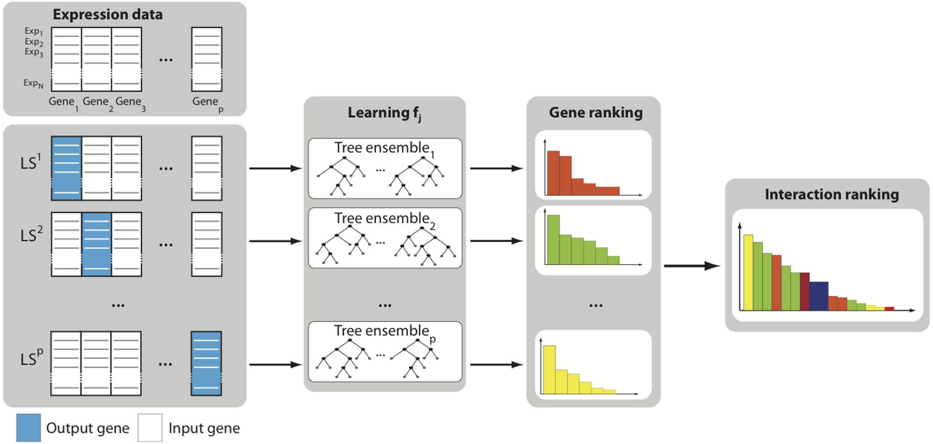
GENIE3 (Random Forest)

Idea (for Genes *j* = 1 to *p*)

- Generate the learning sample of input-output pairs for gene j:

$$LS^j = \{(\mathbf{x}_k^{-j}, x_k^j), k = 1, \ldots, N\}.$$

- Use feature selection technique on LS^j to get confidence intervals $w_{i,j}$
- Aggregate the *p* individual gene rankings to get global rankings



Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P (2010) Inferring Regulatory Networks from Expression Data Using Tree-Based Methods. PLoS ONE 5(9): e12776.

ODE+ Regression

TF expression dynamics throughout differentiation with linear ordinary differentiation equations (ODEs) :

dx = **A**xdt

where x is a vector of length G (G is the number of TFs) that denotes the expression of TFs and **A** corresponds to a square matrix with dimensions equal to G that denotes the regulatory network among TFs.

- **Infer TF regulatory network by optimizing A such that the ODE can successfully describe the observed expression data at a time point.**
- **Pseudotime data also required as input.**

Matsumoto H, Kiryu H, Furusawa C, Ko MSH, Ko SBH, Gouda N, Hayashi T, Nikaido I. SCODE: an efficient regulatory network inference algorithm from single-cell RNA-Seq during differentiation. Bioinformatics. 2017

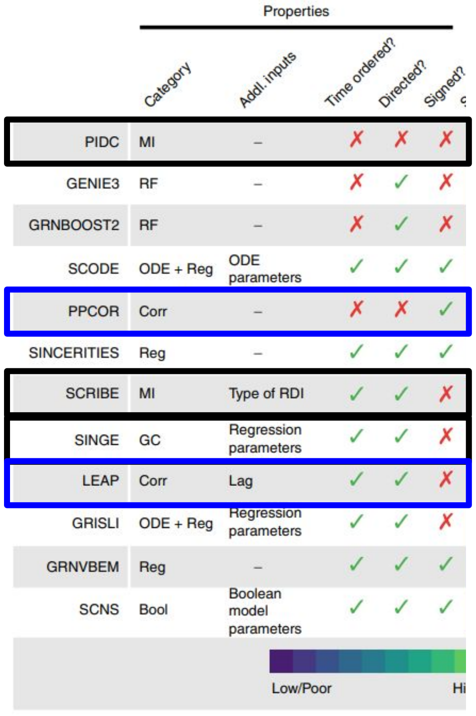
Overview

- What is a Gene Regulatory Network?
- Datasets: 400 Simulated and 5 Real sc-data
    - Synthetic, Curated (synthetic) and Real Single-cell data
    - BoolODE
- Methods: 12 Different Algorithms for GRN Inference
    - Random Forest
    - ODE and Regression
    - Correlation / Mutual Information / Causality
- Benchmarking Results

Correlation / Info. Theory Algorithms

Algorithms based on measuring amount of information one gene provides of another gene.

- PIDC (Mutual Information)
- PPCOR (Correlation)
- SCRIBE (Mutual Information)
- SINGE (Granger Causality)
- LEAP (Correlation)



	Properties				
	Category	Addl. inputs	Time ordered?	Directed?	Signed?
PIDC	MI	–	✗	✗	✗
GENIE3	RF	–	✗	✓	✗
GRNBOOST2	RF	–	✗	✓	✗
SCODE	ODE + Reg	ODE parameters	✓	✓	✓
PPCOR	Corr	–	✗	✗	✓
SINCERITIES	Reg	–	✓	✓	✓
SCRIBE	MI	Type of RDI	✓	✓	✗
SINGE	GC	Regression parameters	✓	✓	✗
LEAP	Corr	Lag	✓	✓	✗
GRISLI	ODE + Reg	Regression parameters	✓	✓	✗
GRNVBEM	Reg	–	✓	✓	✓
SCNS	Bool	Boolean model parameters	✓	✓	✓

Fig. 6 | Summary of properties of GRN inference algorithms a

PPCOR (Correlation)

Partial Correlation

- Degree of association between two random variables, with the effect of a set of controlling random variables removed.

$$\rho_{XY\cdot\mathbf{Z}} = \frac{\rho_{XY\cdot\mathbf{Z}\setminus\{Z_0\}} - \rho_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}\,\rho_{Z_0Y\cdot\mathbf{Z}\setminus\{Z_0\}}}{\sqrt{1-\rho^2_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}}\,\sqrt{1-\rho^2_{Z_0Y\cdot\mathbf{Z}\setminus\{Z_0\}}}}$$

*** Kim, S. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. Commun. Stat. Appl. Methods 22, 665–674 (2015).**

PPCOR (Correlation)

Partial Correlation

- Degree of association between two random variables, with the effect of a set of controlling random variables removed.

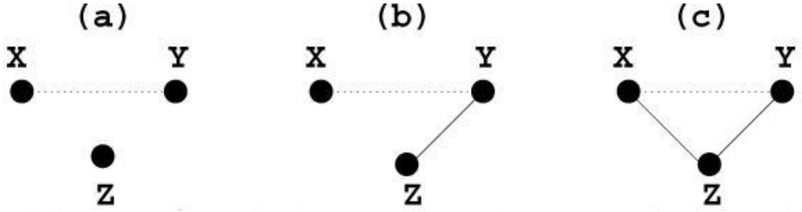
$$\rho_{XY\cdot\mathbf{Z}} = \frac{\rho_{XY\cdot\mathbf{Z}\setminus\{Z_0\}} - \rho_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}\rho_{Z_0Y\cdot\mathbf{Z}\setminus\{Z_0\}}}{\sqrt{1-\rho^2_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}}\sqrt{1-\rho^2_{Z_0Y\cdot\mathbf{Z}\setminus\{Z_0\}}}}$$

Example: Variables X, Y, Z

(a) $p_{XY*Z} = p_{XY}$

(b) p_{XY*Z} != p_{XY}

(c) p_{XY*Z} != p_{XY}



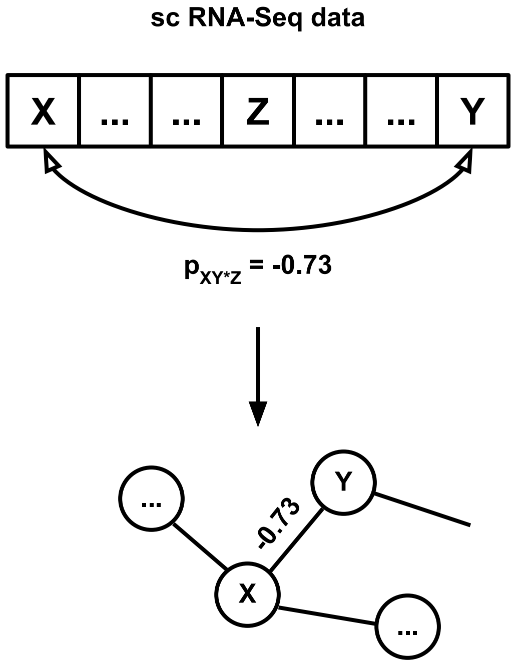
* Kim, S. ppcor: An R package for a fast calculation to semi-partial correlation coefficients. Commun. Stat. Appl. Methods 22, 665–674 (2015).

PPCOR (Correlation)

<u>PPCOR Algorithm for GRNs:</u>

1. Calculate p_{XY*Z} for all pairs of genes X and Y, removing the effects of all other genes Z.

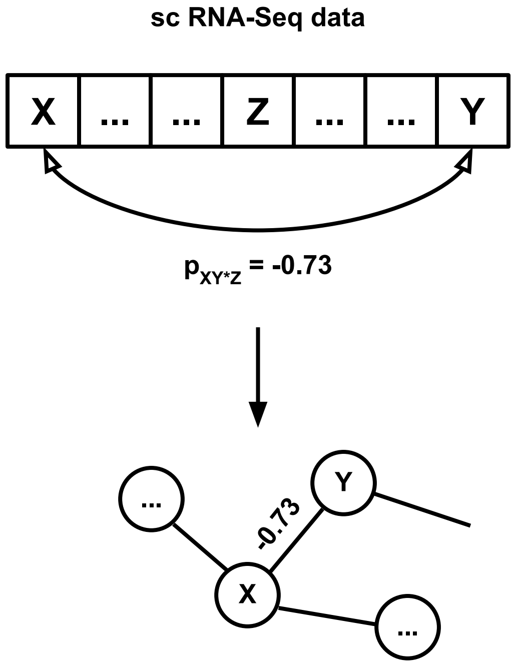
2. Use coefficients p_{XY*Z} as interaction weights in the GRN. Sign = Activation / Inhibition.



PPCOR (Correlation)

PPCOR Algorithm Properties:

- Does not rely on pseudo-time.
- Undirected GRN graph.
- Signed GRN graph.



PIDC (Mutual Information)

Mutual Information

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

- **Entropy H(X): Degree of uncertainty in X**

*** Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267 (2017).**

PIDC (Mutual Information)

Mutual Information

- Entropy H(X): Degree of uncertainty in X

- Mut. Information I(X, Y): Amount of information that X provides about Y

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$
$$= H(X) + H(Y) - H(X,Y)$$

*** Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267 (2017).**

PIDC (Mutual Information)

Mutual Information

- Entropy H(X): Degree of uncertainty in X

- Mut. Information I(X, Y): Amount of information that X provides about Y

- PID (Partial Inf. Decomp.) I(Z; X, Y): How much information X, Y provide about Z.

$$H(X) = -\sum_{x \in X} p(x) \log p(x)$$

$$I(X;Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$
$$= H(X) + H(Y) - H(X,Y)$$

$$I(X;X,Y) = \mathrm{Synergy}(Z;X,Y) + \mathrm{Unique}_Y(Z;X) + \mathrm{Unique}_X(Z;Y) + \mathrm{Redundancy}(Z;X,Y)\ ,$$

*** Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267 (2017).**

PIDC (Mutual Information)

Mutual Information

- Entropy H(X): Degree of uncertainty in X

- Mut. Information I(X, Y): Amount of information that X provides about Y

- PID (Partial Inf. Decomp.) I(Z; X, Y): How much information X, Y provide about Z.

- There is a relationship between I(X, Z) and components of the PID.

$$H(X) = -\sum_{x\in X} p(x)\log p(x)$$

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right)$$
$$= H(X) + H(Y) - H(X,Y)$$

$$I(X;X,Y) = \text{Synergy}(Z;X,Y) + \text{Unique}_Y(Z;X) + \text{Unique}_X(Z;Y) + \text{Redundancy}(Z;X,Y)\ ,$$

$$I(X;Z) = \text{Unique}_Y(Z;X) + \text{Redundancy}(Z;X,Y)$$

*** Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267 (2017).**

PIDC (Mutual Information)

The PUC (Prop. Unique Contribution):

- Computed between two genes X and Y as the sum of the ratio Unique$_Z$(X; Y) / I(X; Y) for every other gene Z in a network.

- The mean proportion of MI between two genes X and Y that is accounted for by their unique information only (information from other genes has been removed).

$$u_{X,Y} = \sum_{Z \in S \setminus \{X,Y\}} \frac{\mathrm{Unique}_Z(X;Y)}{I(X;Y)} + \sum_{Z \in S \setminus \{X,Y\}} \frac{\mathrm{Unique}_Z(Y;X)}{I(X;Y)}$$

*** Chan, T. E., Stumpf, M. P. H. & Babtie, A. C. Gene regulatory network inference from single-cell data using multivariate information measures. Cell Syst. 5, 251–267 (2017).**

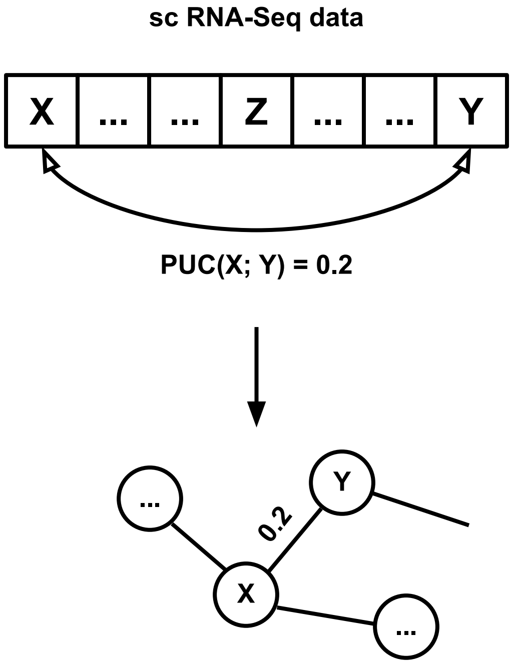
PIDC (Mutual Information)

PIDC Algorithm for GRNs:

1. Calculate PUC(X; Y) for all pairs of genes X and Y, removing the effects of all other genes Z.

2. Calculate per-gene thresholds to keep only the most significant PUC’s.

3. Use PUC(X; Y) as the interaction strength between genes X and Y.

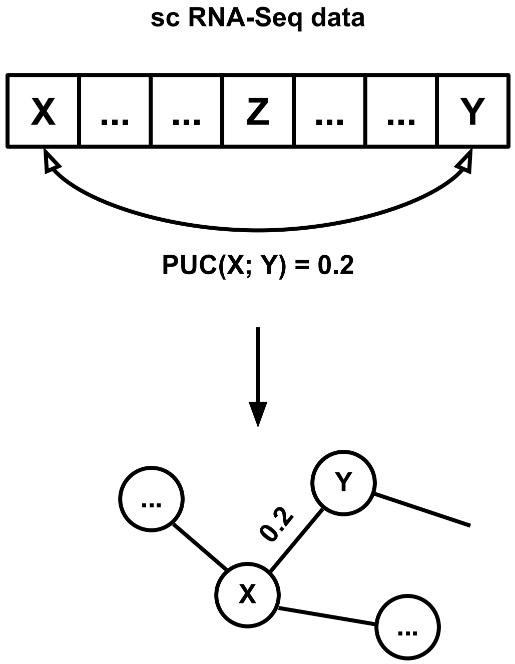


PIDC (Mutual Information)

<u>PIDC Algorithm for GRNs:</u>

- **Does not rely on pseudo-time.**
- **Undirected GRN graph.**
- **Unsigned GRN graph.**

Comment from authors: Fewer false positives (indirect connections) compared to PPCOR



LEAP (Correlation)

<u>Pseudo-time ordering</u>

- Pre-processing step

- In: Gene-cell count matrix

- Out: Gene-cell count matrix
  sorted such that cell i “precedes” j, i < j

- LEAP internally uses Monocle for this

* Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics 33, 764–766 (2017).

LEAP (Correlation)

<u>Pseudo-time ordering</u>

- Pre-processing step
- In: Gene-cell count matrix
- Out: Gene-cell count matrix
  sorted such that cell i “precedes” j, i < j
- LEAP internally uses Monocle for this

Genes

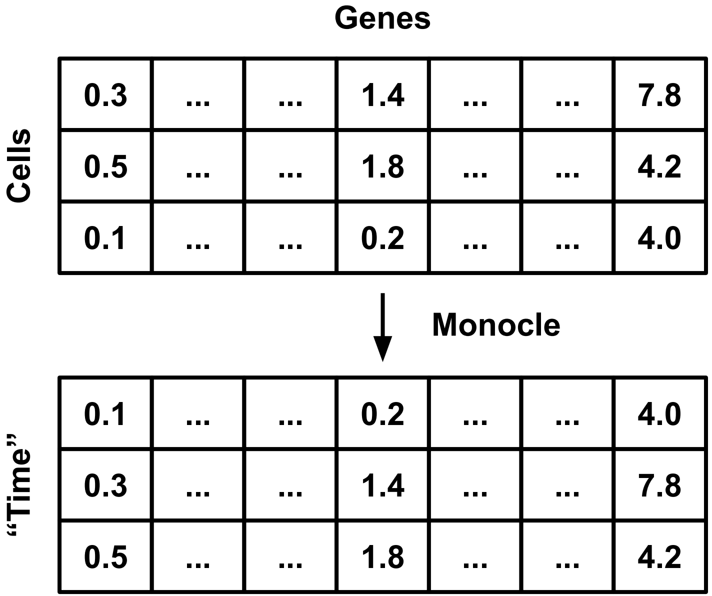
Cells	0.3	...	...	1.4	...	...	7.8
	0.5	...	...	1.8	...	...	4.2
	0.1	...	...	0.2	...	...	4.0

* Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics 33, 764–766 (2017).

LEAP (Correlation)

__Pseudo-time ordering__

- Pre-processing step
- In: Gene-cell count matrix
- Out: Gene-cell count matrix sorted such that cell i “precedes” j, i < j
- LEAP internally uses Monocle for this



0.3	...	...	1.4	...	...	7.8
0.5	...	...	1.8	...	...	4.2
0.1	...	...	0.2	...	...	4.0

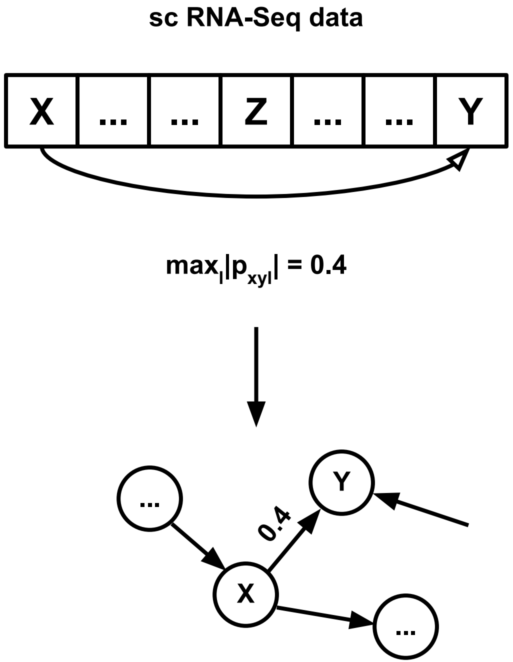
0.1	...	...	0.2	...	...	4.0
0.3	...	...	1.4	...	...	7.8
0.5	...	...	1.8	...	...	4.2

* Specht, A. T. & Li, J. LEAP: constructing gene co-expression networks for single-cell RNA-sequencing data using pseudotime ordering. Bioinformatics 33, 764–766 (2017).

LEAP (Correlation)

Lag-based correlation testing

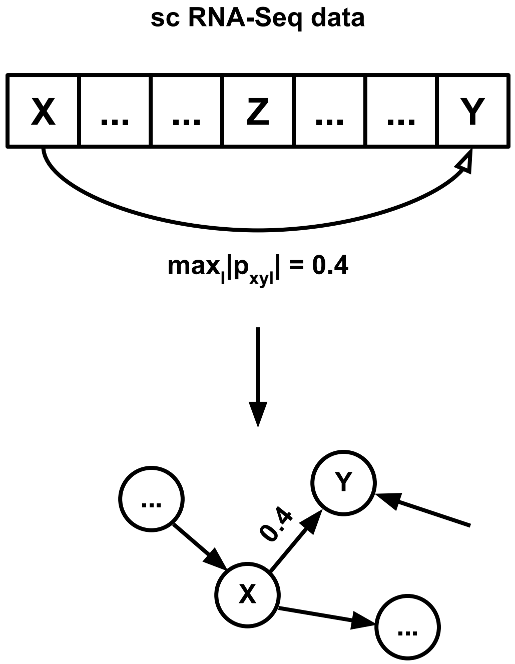
- Tests for correlation between gene i and j, but over windows of pseudo-time delay.
- Time series for gene i: $X_{i,\ 1}, \ldots, X_{i,\ s}$
- Time series for lagged gene j: $X_{i,\ l+1}, \ldots, X_{i,\ l+s}$
- Interaction strength = $\max_l |p_{ijl}|$



LEAP (Correlation)

LEAP Algorithm for GRNs:

- Relies on pseudo-time.
- Directed GRN graph.
- Unsigned GRN graph.



sc RNA-Seq data

X	...	...	Z	...	...	Y

$\max_l |p_{xyl}| = 0.4$

Overview

- What is a Gene Regulatory Network?
- Datasets: 400 Simulated and 5 Real sc-data
    - Synthetic, Curated (synthetic) and Real Single-cell data
    - BoolODE
- Methods: 12 Different Algorithms for GRN Inference
    - Random Forest
    - ODE and Regression
    - Correlation / Mutual Information / Causality
- Benchmarking Results

Benchmarking Results

- Each algorithm tested on each of the six types of synthetic datasets (Linear, Cycle, Long Linear, Bifurc., Bicurfc. Converg., Trifurc.).
- Test the ability to infer edges in the GRN.

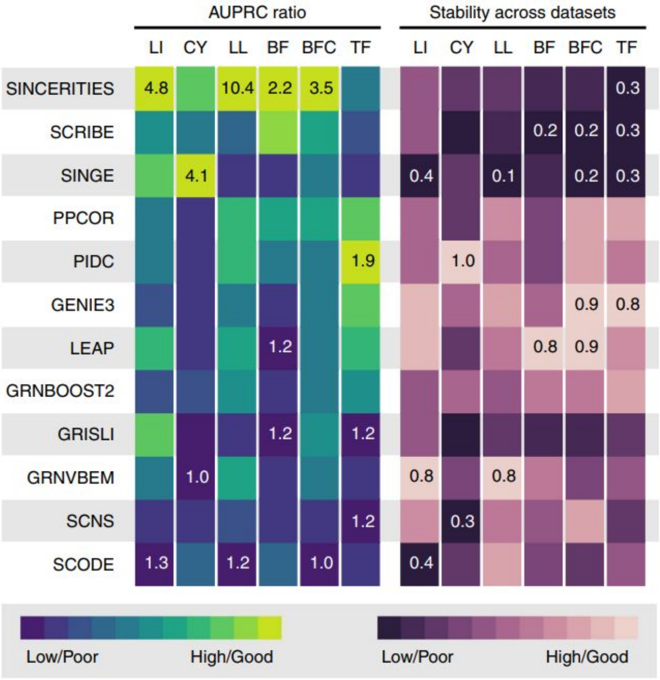


Fig. 2 | Summary of results for datasets from synthetic networks. The first

Benchmarking Results

- Each algorithm tested on each of the six types of synthetic datasets (Linear, Cycle, Long Linear, Bifurc., Bicurfc. Converg., Trifurc.).
- Test the ability to infer edges in the GRN.
- AUPRC Ratio: Divide area under precision-recall curve by that of random predictor.
- Stability: For all datasets of a single type, compare the Top-K-edges (Jaccard Index).

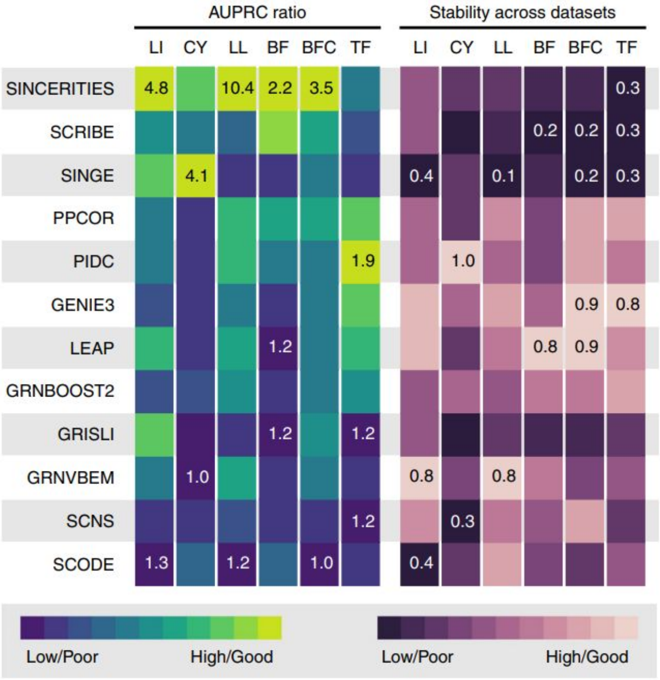
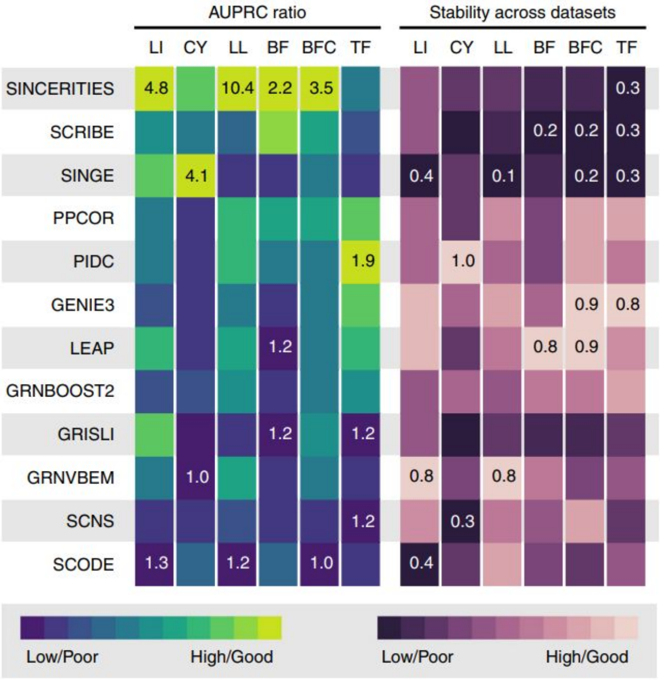


Fig. 2 | Summary of results for datasets from synthetic networks. The first

Benchmarking Results

- Not shown: Varying the number of cells (100 - 5,000) had no significant effect on GENIE3, GRNVBEM, LEAP, SCNS and SCODE.



	AUPRC ratio						Stability across datasets					
	LI	CY	LL	BF	BFC	TF	LI	CY	LL	BF	BFC	TF
SINCERITIES	4.8		10.4	2.2	3.5							0.3
SCRIBE										0.2	0.2	0.3
SINGE		4.1					0.4		0.1		0.2	0.3
PPCOR												
PIDC						1.9		1.0				
GENIE3											0.9	0.8
LEAP				1.2						0.8	0.9	
GRNBOOST2												
GRISLI				1.2		1.2						
GRNVBEM		1.0					0.8		0.8			
SCNS						1.2		0.3				
SCODE	1.3		1.2		1.0		0.4					

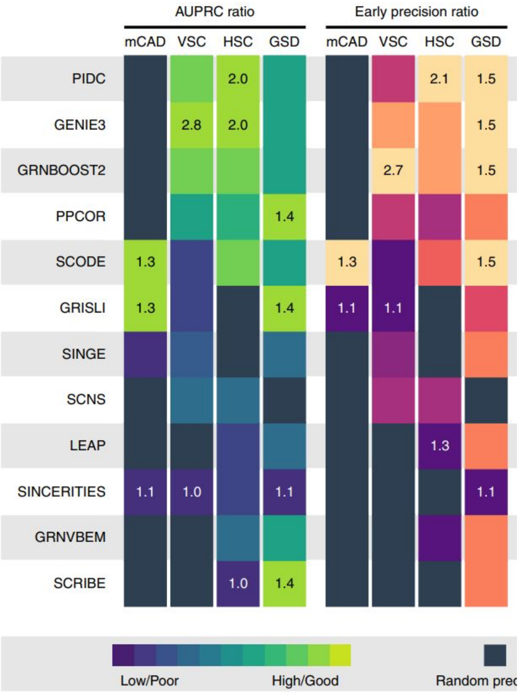
Fig. 2 | Summary of results for datasets from synthetic networks. The first

Benchmarking Results

- Each algorithm tested on each of the four curated GRN datasets.
- Best algorithms on synthetic GRNs: SINCERITIES, SCRIBE and SINGE

  Close to random perf. on curated GRNs.
- Possible causes:

  Denser sub-networks.



	AUPRC ratio				Early precision ratio			
	mCAD	VSC	HSC	GSD	mCAD	VSC	HSC	GSD
PIDC			2.0				2.1	1.5
GENIE3		2.8	2.0					1.5
GRNBOOST2						2.7		1.5
PPCOR				1.4				
SCODE	1.3				1.3			1.5
GRISLI	1.3			1.4	1.1	1.1		
SINGE								
SCNS								
LEAP							1.3	
SINCERITIES	1.1	1.0		1.1				1.1
GRNVBEM								
SCRIBE			1.0	1.4				

Fig. 4 | Summary of results for ten datasets without dropouts from curated models. Rc

Benchmarking Results

- Not shown: Most methods had significant drops in AUPRC ratio with either 50% or 70% dropout.
- The four methods not affected by dropout: GRNVBEM, LEAP, SCRIBE and SINCERITIES Had worse than random AUPRC on on the mCAD and VSC datasets.

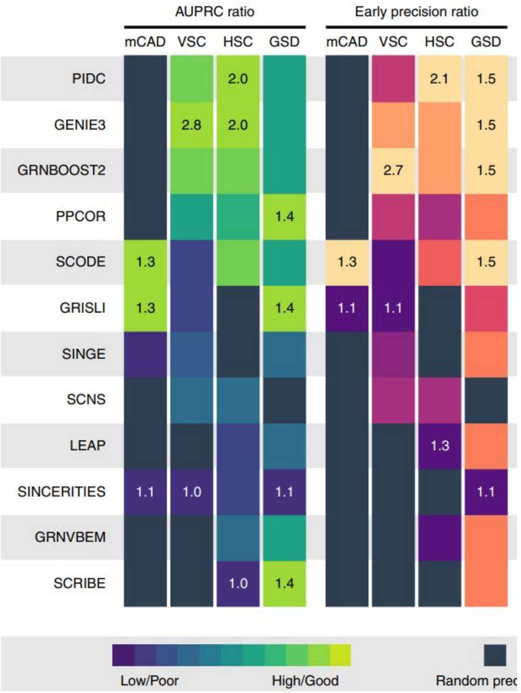


Fig. 4 | Summary of results for ten datasets without dropouts from curated models. Rc

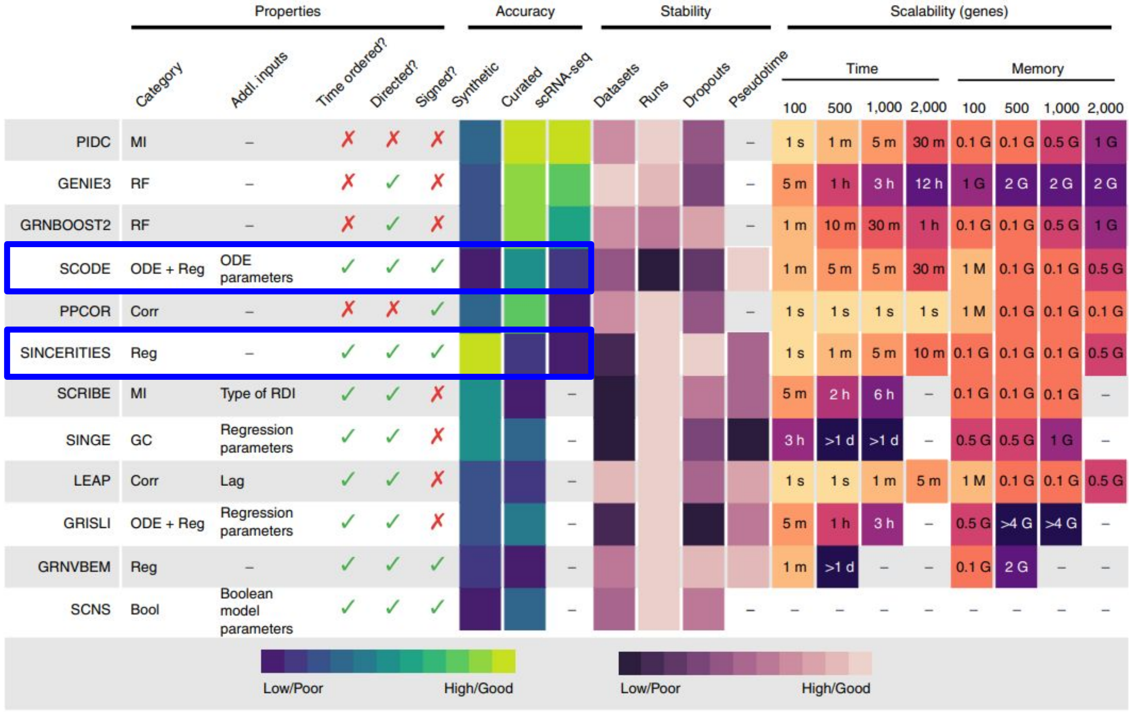
Benchmarking Results

- Pseudo-time ordered algorithms are much worse on real sc RNA-Seq data

  E.g.

  SCODE, SINCERITIES

- Possible causes: Noise pseudo-time.



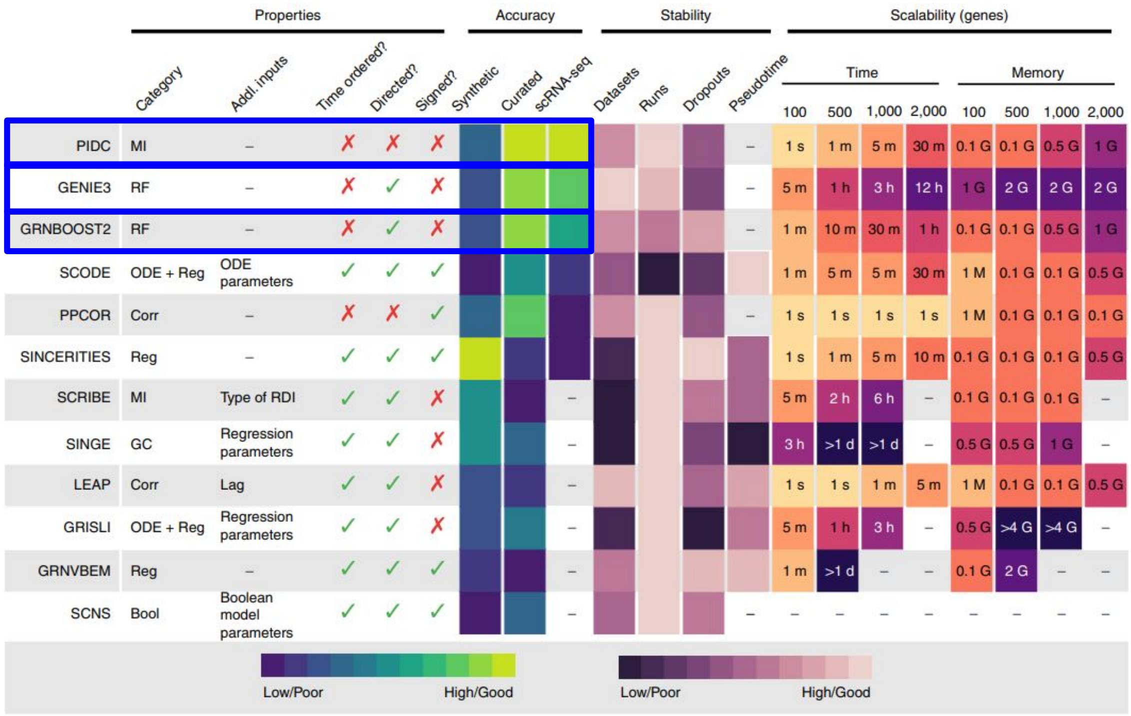
	Properties					Accuracy			Stability				Scalability (genes)							
													Time				Memory			
	Category	Addl. inputs	Time ordered?	Directed?	Signed?	Synthetic	Curated	scRNA-seq	Datasets	Runs	Dropouts	Pseudotime	100	500	1,000	2,000	100	500	1,000	2,000
PIDC	MI	–	✗	✗	✗							–	1 s	1 m	5 m	30 m	0.1 G	0.1 G	0.5 G	1 G
GENIE3	RF	–	✗	✓	✗							–	5 m	1 h	3 h	12 h	1 G	2 G	2 G	2 G
GRNBOOST2	RF	–	✗	✓	✗							–	1 m	10 m	30 m	1 h	0.1 G	0.1 G	0.5 G	1 G
SCODE	ODE + Reg	ODE parameters	✓	✓	✓								1 m	5 m	5 m	30 m	1 M	0.1 G	0.1 G	0.5 G
PPCOR	Corr	–	✗	✗	✓							–	1 s	1 s	1 s	1 s	1 M	0.1 G	0.1 G	0.1 G
SINCERITIES	Reg	–	✓	✓	✓								1 s	1 m	5 m	10 m	0.1 G	0.1 G	0.1 G	0.5 G
SCRIBE	MI	Type of RDI	✓	✓	✗			–					5 m	2 h	6 h	–	0.1 G	0.1 G	0.1 G	–
SINGE	GC	Regression parameters	✓	✓	✗			–					3 h	>1 d	>1 d	–	0.5 G	0.5 G	1 G	–
LEAP	Corr	Lag	✓	✓	✗			–					1 s	1 s	1 m	5 m	1 M	0.1 G	0.1 G	0.5 G
GRISLI	ODE + Reg	Regression parameters	✓	✓	✗			–					5 m	1 h	3 h	–	0.5 G	>4 G	>4 G	–
GRNVBEM	Reg	–	✓	✓	✓			–					1 m	>1 d	–	–	0.1 G	2 G	–	–
SCNS	Bool	Boolean model parameters	✓	✓	✓			–				–	–	–	–	–	–	–	–	–

Fig. 6 | Summary of properties of GRN inference algorithms and results obtained from BEELINE. Each row corresponds to one of the algorithms included

Benchmarking Results

Recommendation:

- **PIDC, GENIE3 and GRNBoost2.**
- **GENIE3 and PIDC had better stability.**
- **GRNBoost2 faster than GENIE3.**



	Properties					Accuracy			Stability				Scalability (genes)							
													Time				Memory			
	Category	Addl. inputs	Time ordered?	Directed?	Signed?	Synthetic	Curated	scRNA-seq	Datasets	Runs	Dropouts	Pseudotime	100	500	1,000	2,000	100	500	1,000	2,000
PIDC	MI	–	✗	✗	✗							–	1 s	1 m	5 m	30 m	0.1 G	0.1 G	0.5 G	1 G
GENIE3	RF	–	✗	✓	✗							–	5 m	1 h	3 h	12 h	1 G	2 G	2 G	2 G
GRNBOOST2	RF	–	✗	✓	✗							–	1 m	10 m	30 m	1 h	0.1 G	0.1 G	0.5 G	1 G
SCODE	ODE + Reg	ODE parameters	✓	✓	✓								1 m	5 m	5 m	30 m	1 M	0.1 G	0.1 G	0.5 G
PPCOR	Corr	–	✗	✗	✓							–	1 s	1 s	1 s	1 s	1 M	0.1 G	0.1 G	0.1 G
SINCERITIES	Reg	–	✓	✓	✓								1 s	1 m	5 m	10 m	0.1 G	0.1 G	0.1 G	0.5 G
SCRIBE	MI	Type of RDI	✓	✓	✗			–					5 m	2 h	6 h	–	0.1 G	0.1 G	0.1 G	–
SINGE	GC	Regression parameters	✓	✓	✗			–					3 h	>1 d	>1 d	–	0.5 G	0.5 G	1 G	–
LEAP	Corr	Lag	✓	✓	✗			–					1 s	1 s	1 m	5 m	1 M	0.1 G	0.1 G	0.5 G
GRISLI	ODE + Reg	Regression parameters	✓	✓	✗			–					5 m	1 h	3 h	–	0.5 G	>4 G	>4 G	–
GRNVBEM	Reg	–	✓	✓	✓			–					1 m	>1 d	–	–	0.1 G	2 G	–	–
SCNS	Bool	Boolean model parameters	✓	✓	✓			–				–	–	–	–	–	–	–	–	–

Low/Poor High/Good

Low/Poor High/Good

Fig. 6 | Summary of properties of GRN inference algorithms and results obtained from BEELINE. Each row corresponds to one of the algorithms included

Discussion

- Surprisingly, classical algorithms designed for bulk transcriptomics data (GENIE3, PPCOR) outperformed specialized algorithms such as LEAP or SINCERITIES.

- Is Inference (GRN) based on already inferred data (pseudo-time) the culprit?

- What other data could be incorporated to increase performance?

- Do we believe in the BoolODE-simulated single-cell datasets?