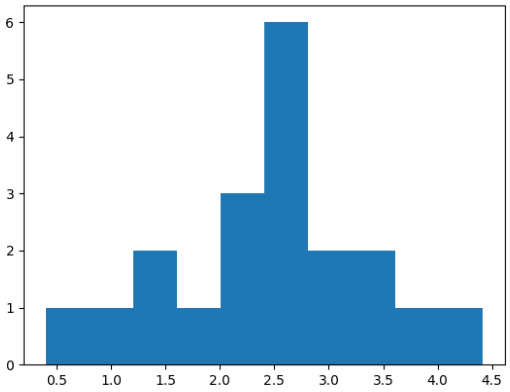
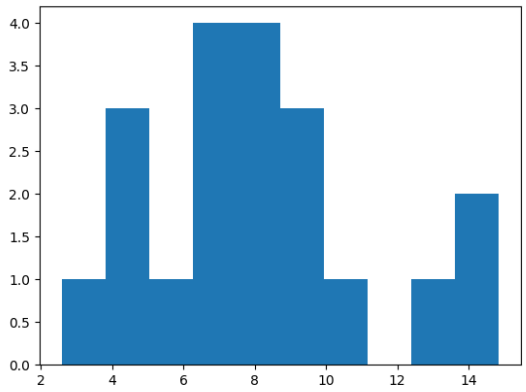
Bagging: Bootstrap Aggregating



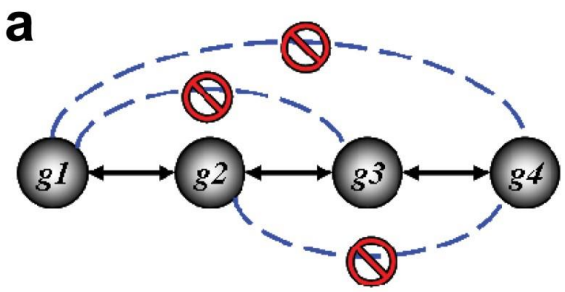
Mean



Variance

Data Processing Inequality (DPI)

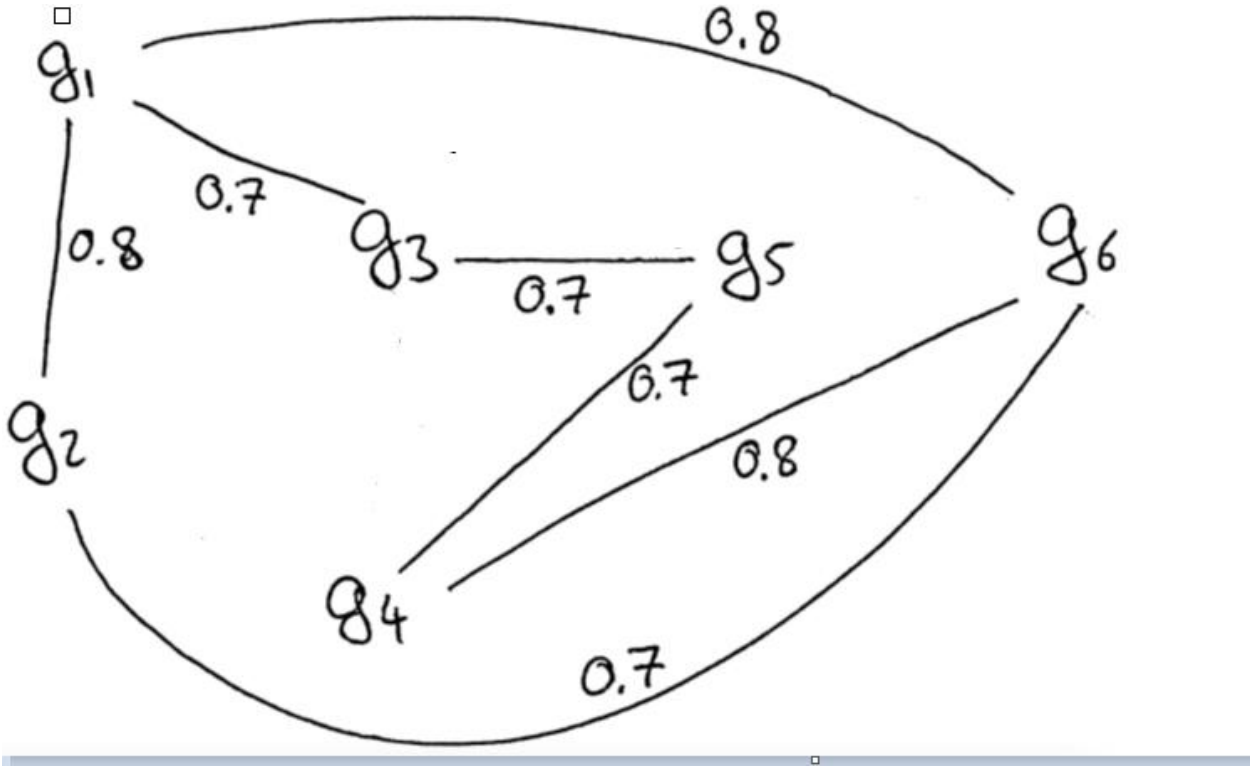
1. DPI:



Examples of the data processing inequality. (a) g1, g2, g3, and g4 are connected in a linear chain relationship. Although all six gene pairs will likely have enriched mutual information, the DPI will infer the most likely path of information flow. For example, g1 ↔g3 will be eliminated because I(g1, g2) >I(g1, g3) and I(g2, g3) >I(g1, g3). g2 ↔ g4 will be eliminated because I(g2, g3) >I(g2, g4) and I(g3, g4) >I(g2, g4). g1 ↔ g4 will be eliminated in two ways: first, because I(g1, g2) >I(g1, g4) and I(g2, g4) >I(g1, g4), and then because I(g1, g3) >I(g1, g4) and I(g3, g4) >I(g1, g4). The issue could be a nonzero DPI threshold is used, if one gene regulate a pathway, the pathway between regulator and gene should not be removed.

From: ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context

2.



Network inference challenge

1.

WGCNA: 1. Construct a gene co-expression network represented mathematically by an adjacency matrix, the element of which indicates co-expression similarity between a pair of genes. 2. Identify modules 3. Relate modules to phenotypes 4. Study inter-module relationships 5. Find key drivers in interesting modules

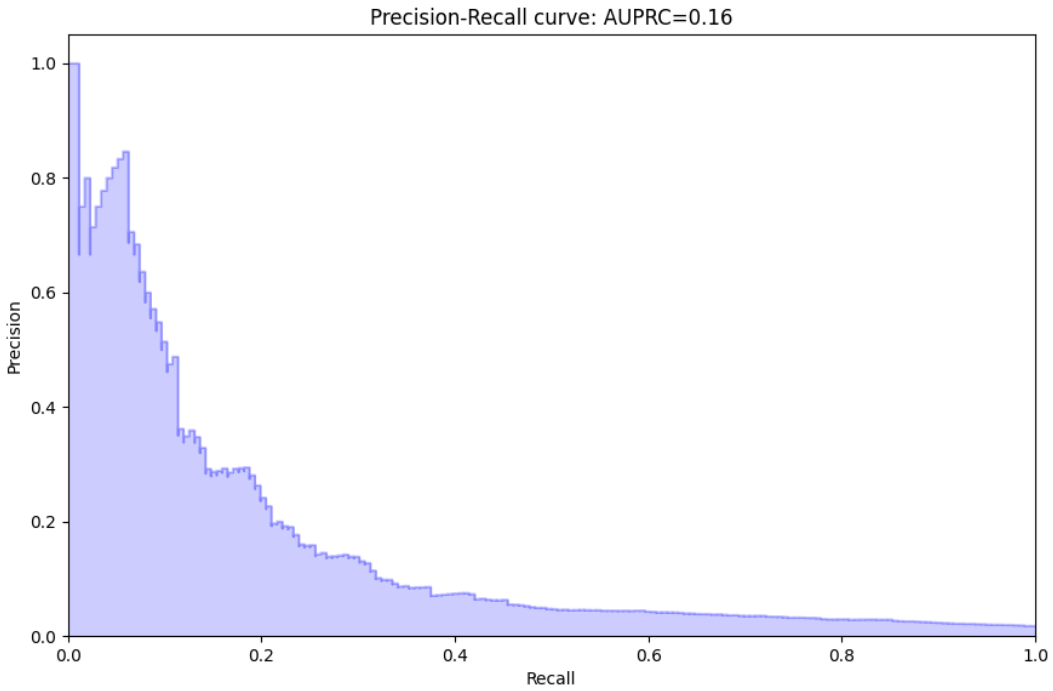
TIGRESS: Using a popular feature selection method, least angle regression (LARS) combined with stability selection for GRN inference.

GENIE3: The targeted networks are directed graphs with p nodes, where each node represents a gene, and an edge directed from one gene i to another gene j indicates that gene i (directly) regulates the expression of gene j. We only consider unsigned edges; when gene i is connected to gene j, the former can be either an activator or a repressor of the latter.

ARACNE: Identify candidate interactions by estimating pairwise gene expression profile mutual information, $I(g_i, g_j) \equiv I_{ij}$, an information-theoretic measure of relatedness that is zero iff $P(g_i, g_j) = P(g_i)P(g_j)$. Then filter MIs using an appropriate threshold, I_0, computed for a specific p-value, p_0, in the null-hypothesis of two independent genes. Thus in its second step, ARACNE removes the vast majority of indirect candidate interactions ($\phi_{ij} = 0$) using a well-known information theoretic property, the data processing inequality (DPI).

CLR: CLR filtered with negative PCC, least angle regression of time series, TF perturbation data1

2.



Network inference using single-cell data

As the ground truth for assessing accuracy, we use synthetic networks with predictable trajectories, literature-curated Boolean models and diverse transcriptional regulatory networks. A strategy to simulate single-cell transcriptional data from synthetic and Boolean networks **that avoids pitfalls of previously used methods.** Networks from multiple experimental single-cell RNA-seq datasets collected and use BEELINE frame work to evaluate.


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
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