# Comparing Statistical Approaches for Network Inference in Noisy Biological Time-Series Data

## Problem

Gene regulatory network inference is commonly used to generate biological hypotheses rather than definitive conclusions. In many real-world experimental settings, researchers must work with short, noisy time-series datasets due to cost, experimental complexity, or biological constraints.

This project evaluates how different statistical modeling choices affect the reliability of inferred networks under these realistic data limitations, helping clarify which conclusions are supported by the data and which are not.

## Skills demonstrated

- Time-series data preprocessing and normalization
- Sparse model fitting and regularization
- Directed vs undirected network inference
- Model evaluation with ROC/AUC
- Critical interpretation under data limitations

## Tools and methods

Python, NumPy, SciPy, scikit-learn, matplotlib, pandas and sklearn

## Data

- Time-series gene expression data sampled every 10 minutes
- **5 genes**, short time series
- A small sized regulatory network was used as a reference for evaluating inferred gene–gene interactions
- A small sample size reflects common challenges in biological experiments

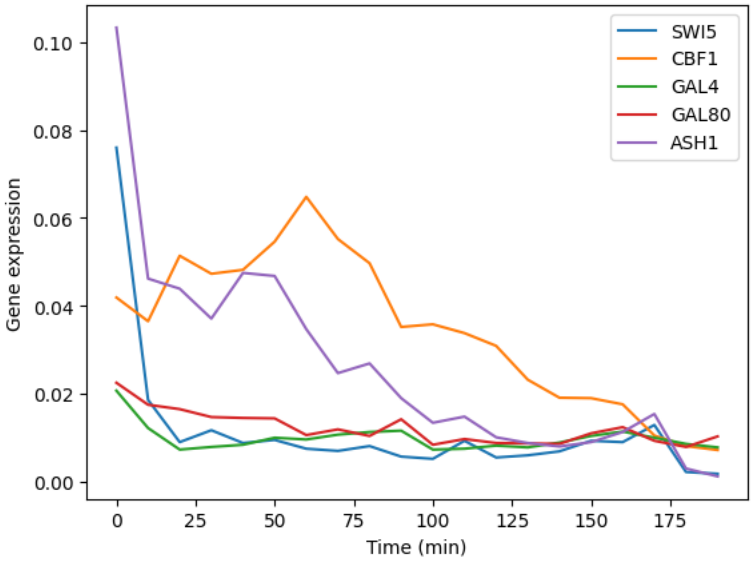


*Figure 1. Gene expression time-series profiles sampled at 10-minute intervals*

**Approach**

Rather than optimizing a single model, this analysis focuses on comparing modeling assumptions. The Gaussian Graphical Model serves as a baseline that captures conditional dependence without explicit temporal structure, while the VAR(1) model explicitly incorporates time-lagged effects. This comparison highlights the impact of temporal modeling choices on inferred regulatory interactions.

1. **Gaussian Graphical Model (Graphical Lasso)**

- Modeled gene expression as a multivariate Gaussian distribution
- Used **L1-regularized precision matrix estimation** to infer conditional dependencies
- Extended to a **time-lagged formulation** to introduce directionality
- Ranked inferred interactions using absolute partial correlations

**Goal:** Establish a baseline model for gene–gene dependency inference

2. **Vector Autoregressive Model (VAR(1) with L1 Regularization)**

- Modeled gene expression dynamics as:

$$x(t+1) = Ax(t) + \varepsilon(t)$$

- Each coefficient represents a **directed lagged influence**
- Used L1 regularization to encourage sparsity and interpretability
- Ranked inferred edges by coefficient magnitude

**Goal:** Explicitly model temporal causality and improve directed inference

**Evaluation**

- Compared predicted edges against the known regulatory network
- Constructed **ROC curves** and computed **AUC** values

| Model | AUC |
|---|---|
| Graphical Lasso | ~ 0.44 |
| VAR(1) Lasso | ~ 0.58 |

*Figure 2. In the context of small gene networks and short time-series data, AUC values should be interpreted comparatively rather than absolutely. The goal is not high predictive accuracy, but to assess whether incorporating temporal structure leads to more informative rankings of candidate regulatory interactions under constrained data conditions.*

## Results & Insights

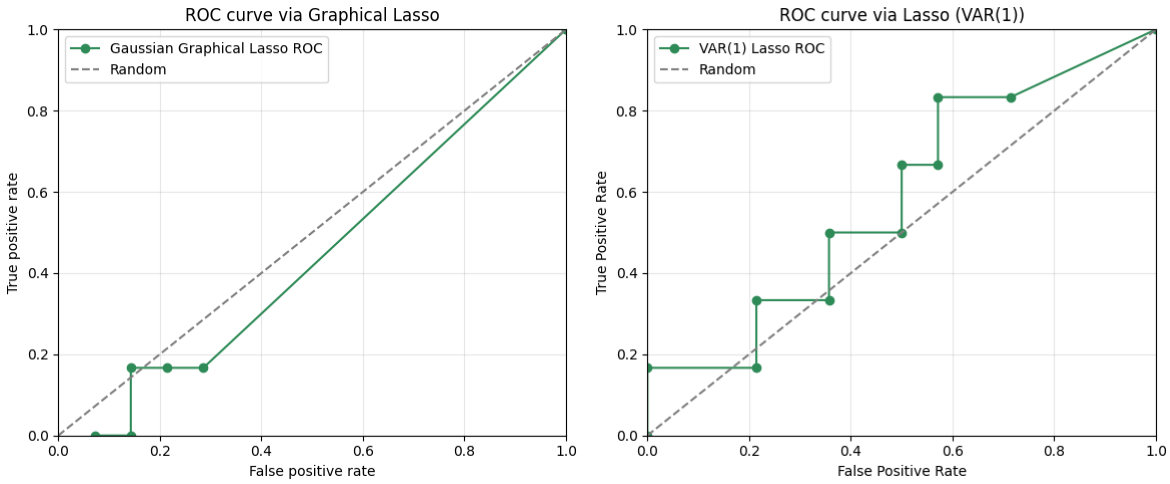


*Figure 3. The VAR(1) model consistently outperformed the Gaussian Graphical Lasso, achieving higher AUC and demonstrating improved recovery of directed regulatory interactions.*

- The Gaussian Graphical Model showed limited ability to recover directed regulatory relationships, highlighting the limitations of undirected partial correlations for causal inference.
-
- The VAR(1) model demonstrated improved performance by explicitly incorporating temporal structure. Despite improvement, performance remained limited due to:
    - Small network size
    - Short time series
    - Measurement noise

- These results demonstrate that **model choice alone cannot overcome data limitations**

## Potential extensions

With larger datasets or longer time series, several extensions could further improve inference quality:

- Evaluating higher-order VAR models to capture slower regulatory dynamics
- Assessing edge stability via bootstrapping or resampling
- Incorporating prior biological knowledge (e.g., known transcription factors) to reduce false positives
- Comparing against additional baselines such as correlation-based or Granger causality methods

These extensions highlight how modeling strategy and data availability jointly influence network inference performance.

***Key takeaway*** - This project demonstrates how model selection, evaluation strategy, and data limitations interact in real-world time-series analysis, emphasizing the importance of critical interpretation alongside statistical modeling.