CS-E5885 Modeling biological networks project report- Nguyen The Long (102896277)

IDENTIFICATION OF GENE REGULATORY NETWORK FROM GENE EXPRESSION TIME

I. INTRODUCTION

The goal of this project is to infer the structure of a 5-gene network using gene expression time-course data from a switch-off experiment. The activity of galactose is decreased at the beginning of the experiment, and gene expression is measured at 10-minute intervals from 0 to 190 minutes. The known network structure is used as a reference to evaluate the performance of various computational methods.

II. DATA DESCRIPTION

The dataset consists of time-series gene expression measurements for five genes, collected at 10-minute intervals over 190 minutes. The data can be figured in Figure 1.

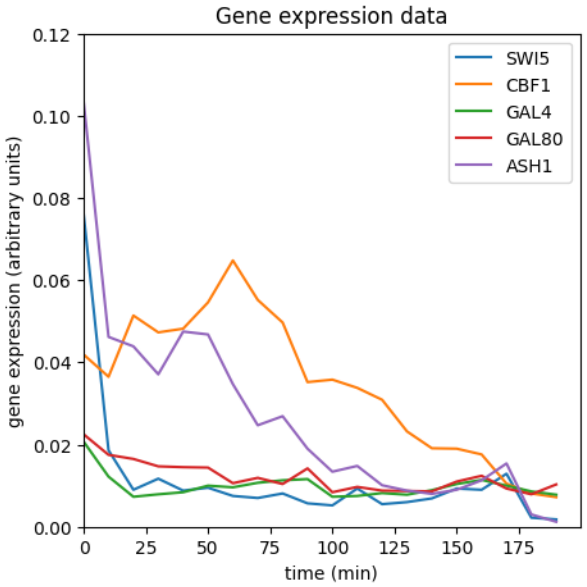

Figure 1: Gene expression data of 5 network

This data captures dynamic changes in gene expression following the switch-off of galactose. Moreover, the biological network structure can be presented in Figure 2 and the generated ground truth values of the five genes relation presented in Table 1 can be referred from the original paper by Cantone et al. (2009) [1].

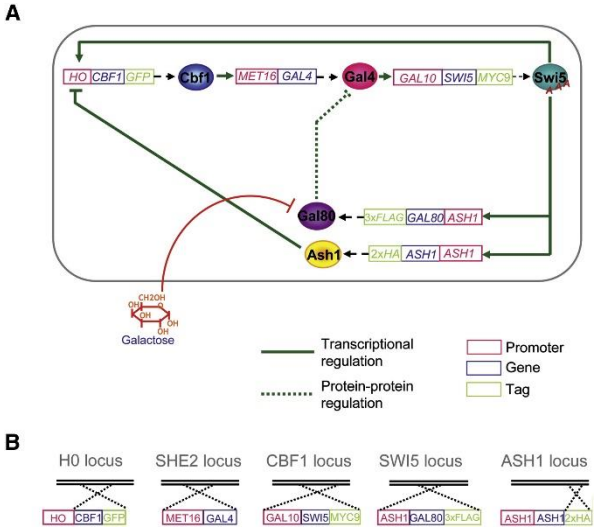

Figure 2: Construction of IRMA, a Synthetic Network in Yeast [1]

Table 1: Ground truth values of the relationship between transcriptional regulation of 5 genes

	ASH1	CBF1	GAL4	GAL80	SWI5
ASH1	0	1	0	0	0
CBF1	0	0	1	0	0
GAL4	0	0	0	0	1
GAL80	0	0	0	0	0
SWI5	1	1	0	1	0

III. MODEL IMPLEMENTATION AND EVALUATION

1) Model Selection

There are multiple approaches for this biological network, which have been researched and succeeded. However, regarding this report, multiple models have been implemented for this project with the fundamental knowledge of mathematical progress. They are mainly random models, Ordinary Differential Equations (ODEs) models, Stochastic Differential Equations (SDEs) models with/without minimizing objective functions, and Linear Regression models. The reason for choosing is because of the characteristics of these models, they can capture the continuous changes in gene expression over time, making them suitable for time-series data. However, since the general form of the model is simple when we compare it with the complexity of the biological network structure, the fitting will not be satisfied. Therefore, the main purpose of this report is to reclaim that idea and motivate us not to choose these models, or should have practical improvement based on these theoretical models for building a better model in the future.

1.1. Random Model:

The random model generates a network structure by randomly assigning interactions between genes. However, the random model here is based on the random function from the NumPy library [2], which is independent of time-series data of the gene expression. The generated random variables are proportional to uniform distribution ~ U(0,1). The reason for choosing that is because the ground truth values are only 0 or 1 (even though in practice, it will be in decimal numbers, but for figure reference, we can only know that they are connected). Therefore, this model serves as a baseline for comparison.

1.2. ODEs Model with and without optimization:

The ODE model uses a system of differential equations to describe the interactions between genes. The general form of the ODE system is: $\frac{dy}{dt} = A \cdot y$, where y is the vector of gene expressions, A is the matrix of interaction parameters, defined by the random algorithm, which is also the target output. Each element A_{ij} represents the influence of gene j on gene i. The goal is to estimate the matrix A from the gene expression data by using parameter fitting, which can be fitted by solving the least squared problem: $A_{ij} = arg\min||\text{sol}[i,j] \cdot \beta + \epsilon - \text{data}[:,i]||$, where $\text{sol}[i,j]$ is the solution of the ODE for gene j, β is the parameter to be estimated, and ϵ is the error term.

Due to the original A here being generated by a random function, to improve the accuracy of the ODE model, we need to optimize the parameters A to minimize the difference between the observed and predicted gene expressions, which is $Error = \sum(y_{observed} - y_{predicted})^2$. The minimize function is based on the Trust-Region algorithm [3]. The reason for this is because of the robustness and efficiency of the algorithm since we can adapt the step size based on the quality of the model approximation.

1.3. SDEs Model with and without optimization:

The SDE model extends the ODE model by incorporating stochastic noise to account for random fluctuations in gene expression. The general form of the SDE system is: $y = A \cdot ydt + \sigma dW$ where σ is the noise intensity, dW is the Wiener process representing the Gaussian noise. Similar to the ODEs with optimization, we optimize the parameters A in the SDEs model to minimize the error between the observed and predicted gene expressions.

1.4. Linear Regression (LR) Model:

The linear regression model uses linear regression to infer the interactions between genes. The model assumes a linear relationship between the gene expressions:

$$y_i = \sum_j A_{ij} \cdot y_j + \epsilon,$$

where y_i is the expression of gene i, A_{ij} Is the interaction parameter between gene j and gene i, ϵ is the error term (Here, $\epsilon = 0$).

Mathematical Formulation: The parameters A are estimated using ordinary least squares (OLS) regression. The goal is to minimize the sum of squared errors as the previous idea of ODEs.

2) Evaluation Metrics

To evaluate the performance of each model, we use the Receiver Operating Characteristic (ROC) curve and the confusion matrix. These metrics help us understand how well the models predict the gene interactions compared to the ground truth.

2.1. Receiver Operating Characteristic (ROC) Curve

The ROC curve is a graphical representation of a model's diagnostic ability. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR) at various threshold settings. TPR (also known as sensitivity or recall) is the ratio of correctly predicted positive observations to all actual positives, while FPR is the ratio of incorrectly predicted positive observations to all actual negatives.

Mathematically, the TPR and FPR are defined as:

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{FN}},$$

Where TP is True Positives, FN is False Negatives, FP is False Positives, and TN is True Negatives.

The area under the ROC curve (AUC) provides a single measure of the model's performance. A model with an AUC of 1.0 is perfect, while an AUC of 0.5 indicates no discriminative power.

2.2. Confusion Matrix

The confusion matrix is a table used to describe the performance of a classification model. It compares the predicted classifications to the actual classifications. The matrix includes four outcomes:

- Edge Found, True: True Positive (TP)
- Edge Found, False True Negative (TN)
- No Edge, False: False Positive (FP)
- No Edge, False: False Negative (FN)

The confusion matrix helps in calculating various performance metrics such as accuracy, precision, recall, and F1 score. The confusion matrix here is built based on the threshold value of the model, which can be calculated by maximizing the difference between the TPR and FPR, which can be described as

$$\mathrm{index} = \mathrm{argmax}(\mathrm{TPR} - \mathrm{FPR}), \mathrm{threshold} = \mathrm{thresholds}(\mathrm{index}),$$

In case the threshold cannot be defined, we set it to a default value of 0.5.

Here, since the result is based on the ground truth expectation, which is the best result that we can expect from the model. However, in case we do not follow this, one of the following methods that we can use to decide the threshold is based on the middle value of the normalization (0.5 is suitable for this task) or the average diagonal of the matrix for the linear regression, since there is no actual connection between the same gene.

IV. RESULTS

1) Model results

The ROC curve (Figure 3) and confusion matrices (Figure 4) highlight the performance of the implemented models: Random, ODEs, ODEs with optimization, SDEs, SDEs with optimization, and Linear Regression.

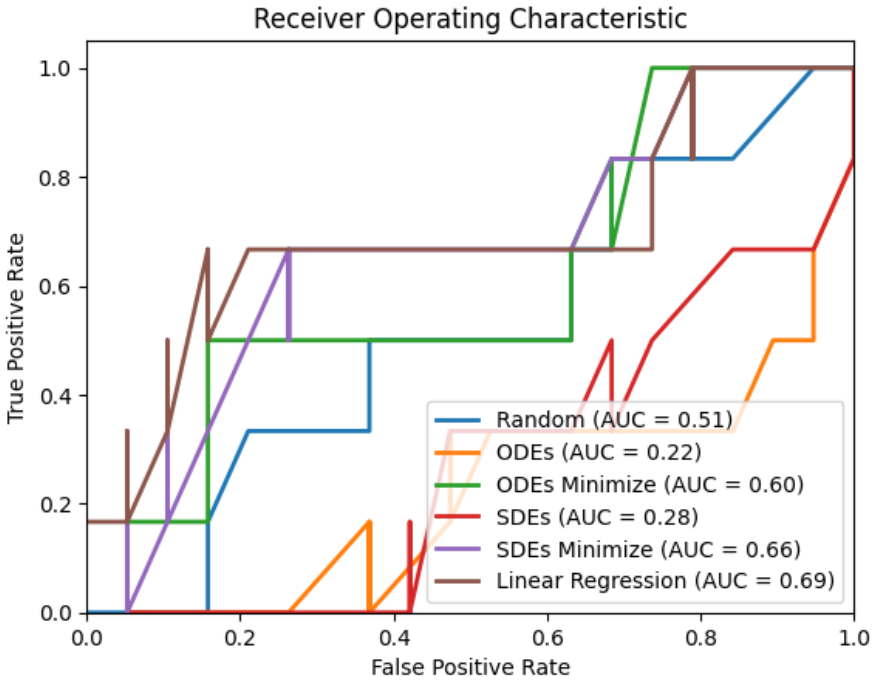

Figure 3. ROC plot for model performance.

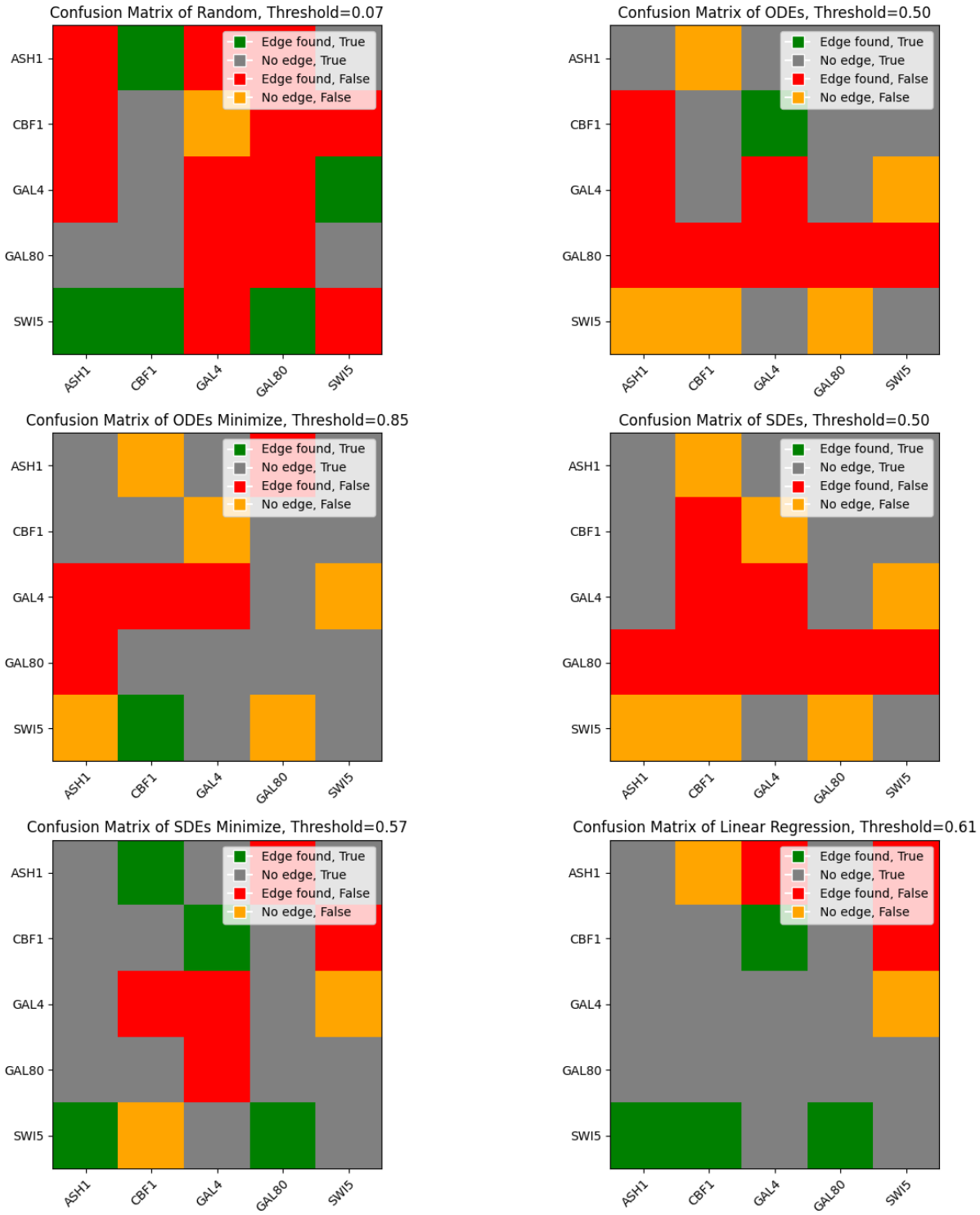

Figure 4. Confusion Matrix of all models

2) Analysis

The Linear Regression model demonstrated the best performance among all the methods, achieving an AUC of 0.69. This result indicates its superior ability to infer gene relationships accurately, making it the most reliable model in this study. Following closely, the SDEs model with optimization attained an AUC of 0.66, highlighting the significant benefits of incorporating stochastic noise along with parameter optimization. The ODEs model with optimization achieved an AUC of 0.60, which, while not as strong as the Linear Regression or optimized SDEs, still outperformed the basic versions of the ODEs and SDEs models without optimization. This emphasizes the importance of parameter tuning in improving model performance.

In contrast, the Random model produced an AUC of 0.51, which is close to random chance, as expected due to its lack of structure and reliance on randomly assigned interactions. The ODEs and SDEs models without optimization performed poorly, with AUC values of 0.22 and 0.28, respectively. These low scores reflect the limitations of the basic models when optimization is not applied, as they fail to accurately capture the dynamics of gene interactions.

3) Discussion

The results demonstrate several key insights into the strengths and limitations of the models for inferring the 5-gene network:

3.1. Importance of Optimization:

A critical observation is that both the ODE and SDE models benefited significantly from parameter optimization. Without optimization, their predictive performance was considerably suboptimal, as reflected by the low AUC values of 0.22 and 0.28, respectively. The optimization process, which involves minimizing the error between predicted and observed values, enables these models to better capture the underlying dynamics of gene expression. This highlights that while the theoretical frameworks of these models are well-suited for time-series data, parameter tuning is essential for achieving meaningful performance.

3.2. Linear Regression Superiority:

The Linear Regression model emerged as the best-performing method in this study, achieving the highest AUC of 0.69. Its success can be attributed to its ability to approximate linear relationships within the synthetic 5-gene network. However, it is essential to acknowledge that Linear Regression assumes a strictly linear structure, which may not generalize effectively to more complex or nonlinear biological systems. While it performs well under the current setup, its limitations may become evident when applied to more realistic biological networks with intricate regulatory interactions. Moreover, 0.69 in AUC can be defined as poor discrimination, which is a limitation of the linearity properties of the model.

3.3. Random Model as a Baseline:

The Random model, which achieved an AUC of 0.51, performed no better than chance. This result is unsurprising given its reliance on randomly assigned interactions, emphasizing the need for informed, data-driven approaches when modeling biological systems. While the Random model serves as a useful baseline for comparison, it further underscores the importance of structured methods that leverage the available time-series gene expression data.

3.4. Stochastic Effects:

The inclusion of stochastic noise in the SDE models adds a layer of biological realism to the simulations, as real-world biological systems are inherently noisy. However, the performance of the SDE models without optimization remained poor, indicating that the added model complexity requires careful parameter tuning to achieve improved results. The optimized SDE model, with an AUC of 0.66, demonstrated that stochastic effects can enhance predictive accuracy when combined with robust optimization. This highlights the trade-off between model complexity and fitting accuracy.

3.5. Practical Challenges:

The results also shed light on the practical limitations of the implemented models. The ODEs, SDEs, and Linear Regression approaches used in this study simplify the biological system, which inherently restricts their predictive accuracy. Real-world biological networks are far more complex, involving nonlinear interactions, time delays, and additional factors that these models do not fully capture. To address these challenges, more sophisticated modeling approaches - such as regularization techniques, applying probability techniques such as Bayesian inference, or advanced machine learning methods like neural networks - could be explored in future studies. These techniques have the potential to improve performance by capturing more intricate relationships within the data.

V. CONCLUSION

Among the tested models**,** Linear Regression provided the best results for this specific 5-gene network, followed closely by SDEs with optimization and ODEs with optimization. This study highlights the importance of parameter optimization for differential equation models and underscores the need for advanced techniques to handle the complexity of gene regulatory networks. Future work should focus on incorporating nonlinear dynamics, regularization, and machine learning approaches to further enhance model accuracy.

VI. REFERENCES

[1] Cantone, I., et al. (2009). A Yeast Synthetic Network for In Vivo Assessment of Reverse-Engineering and Modeling Approaches. *Cell*, 137(1), 172-181.
[2] *Numpy.random.rand — NumPY v2.2 Manual*. (n.d.).
https://numpy.org/doc/stable/reference/random/generated/numpy.random.rand.html
[3] *minimize(method='trust-constr') — SciPy v1.14.1 Manual*. (n.d.).
https://docs.scipy.org/doc/scipy/reference/optimize.minimize-trustconstr.html

APPENDIX

Since the code is too long to attach, I will attach the code.zip file, including the ipynb PDF result and the code.