Several mathematical and computational methods for analysing genetic regulatory networks in hematopoiesis

PhD Final Review

Siyuan Wu

Supervisor: A/Prof. Tianhai Tian
Dr. Tiangang Cui

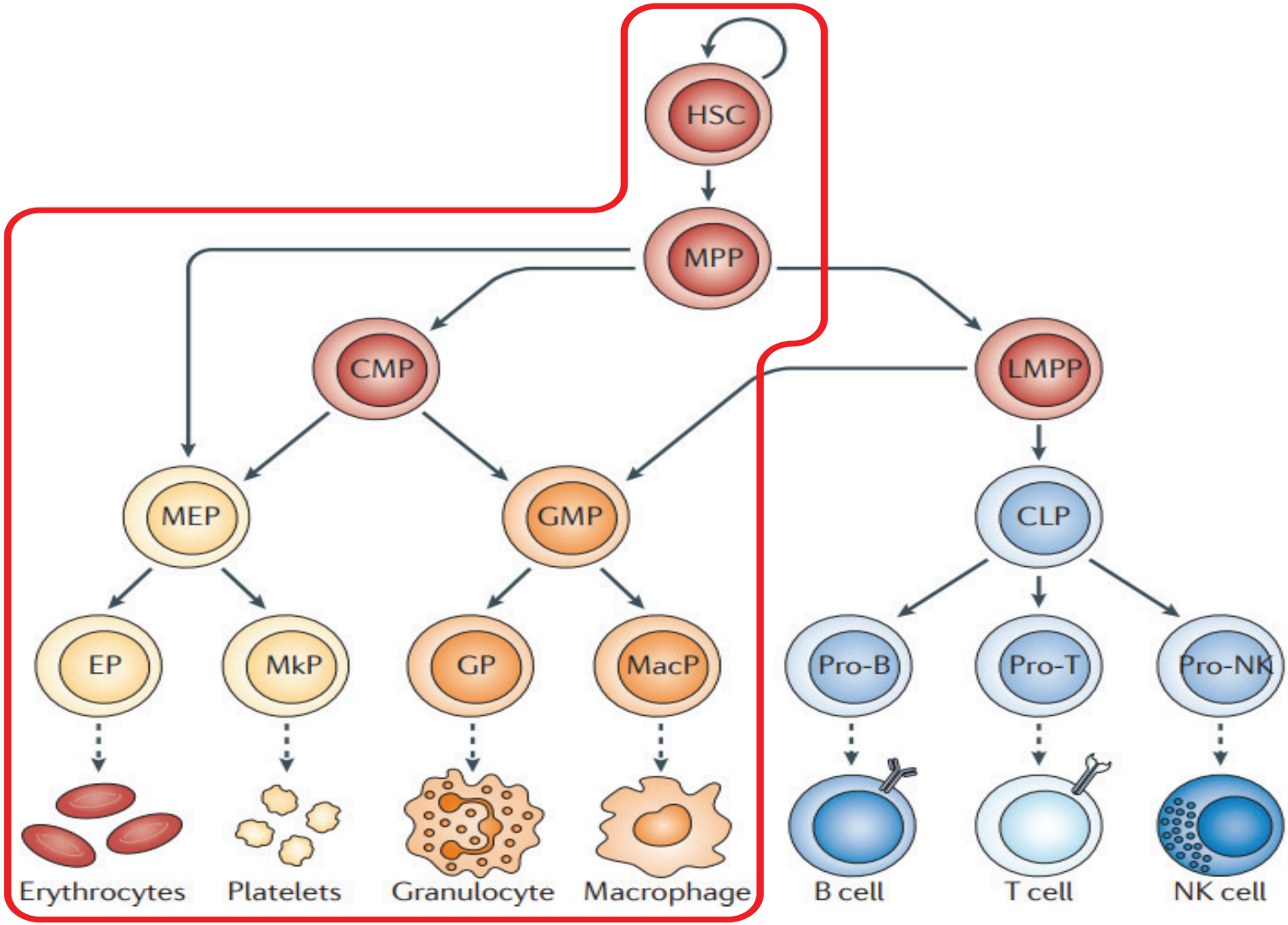
School of Mathematics
Monash University

30-06-2021

Outline

Outline

Introduction to hematopoiesis



Outline

Motivation

- The detailed regulatory mechanisms for fate determination of HSCs are still unravelled
- The effect of possible protein heterodimers and/or synergistic effects in genetic regulation have not been discussed in most mathematical models.
- A key challenge in current inference methods is the large number of unknown parameters compared with the relatively small amount of data.

Objectives

1. A general approach by combining both the top-down approach (for reducing the complexity of the network structure) and bottom-up approach (for derive the detailed dynamic property).
2. Protein dimers and cooperative binding (PDTCBs) are considered into our proposed mathematical model.

Background

- Microarray dataset GSE49991 from NCBI GEO database.
  1. Data of gene expression levels of differentiation into erythrocyte and neutrophil respectively
  2. Data of 3 representative groups at 29 time points over a week after differentiation started.
  3. Based on Go enrichment analysis and literature, the following eleven genes have been choosen into our study:

$$\boldsymbol{x} = \{Gata1, Gata2, Runx1, Spfi1, Cbfa2t3,$$
$$Ets1, Notch1, Tal1, Ldb1, Gfi1b, Gfi1\}.$$

Top-down approach: probabilistic graphical model

According to the Gaussian graphical model, we proposed the Extended Forward Search Algorithm to infer the topological structure of regulatory networks that includes both genes and PDTCBs. In this work it is assumed that a system includes genes $\{g_1, \ldots, g_m\}$ with expression levels x_{ij} for gene g_i at time point j.

Extended Forward Search Algorithm

We want to construct the regulatory network with m genes and n PDTCBs ($m = 11, n = 66$ in our case). Based on the expression levels of each genes, we calculate the following matrix:

1. A: covariance matrix with m-dimension of m genes.
2. B: covariance matrix from m genes to n PDTCBs which is a $m \times n$ matrix.
3. C: covariance matrix with n-dimension of n PDTCBs.

Moreover, we do not study the dynamics of PDTCBs and the regulation from single gene to PTDCBs due to lack of biological evidences.

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B'} & \boldsymbol{C} \end{bmatrix} \tag{1}$$

Extended Forward Search Algorithm I

Algorithm:

1. An initial empty graph (matrix) $\boldsymbol{G}$ is built by the N-dimensional identity matrix ($N = m + n$).

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_1 & \boldsymbol{G}_2 \\ \boldsymbol{G}_3 & \boldsymbol{G}_4 \end{bmatrix} \tag{2}$$

2. Substitute all covariance values from the diagonal positions of sub-matrix $\boldsymbol{A}$ into the corresponding positions of sub-matrix $\boldsymbol{G}_1$, then, use the Iterative Maximum Likelihood Estimates Algorithm to compute the new covariance matrix.
3. Add an undirected edge E_{ij}^1 into $\boldsymbol{G}_1$. Then, compute a new covariance matrix. Based on the deviance difference between the new covariance matrix and that before addition, test and record the significance of the added edge E_{ij}^1, then remove it from $\boldsymbol{G}_1$.
4. Repeat the computation in Step 3 for all possible undirected edges. Add the edge with the smallest p-value into the sub-graph $\boldsymbol{G}_1$ permanently.

Extended Forward Search Algorithm II

5. Go back to step 3, add the second edge in the updated sub-graph $\boldsymbol{G}_1$. Repeat the computation in steps 3 and 4 until the smallest p-value of an added edge is larger than the cutoff p-value.
6. Based on the last updated undirected graph $\boldsymbol{G}_1$, the graph orientation rules are applied to transform the undirected graph into a directed acyclic graph (DAG).
7. Add an undirected edge E_{ij}^2 between the i^{th} gene and the j^{th} PDTCB into the latest graph $\boldsymbol{G}$. Then, compute a new covariance matrix. Repeat the computation in steps 3 to 5 with an undirected edge E_{ij}^2.
8. The last updated sub-graph $\boldsymbol{G}_1$ and $\boldsymbol{G}_3$ with m_1 and n_1 directed edges, denoted as $\boldsymbol{A_s}$ and $\boldsymbol{B'_s}$, is the predicted directed regulatory network among m genes and from n PDTCBs to m genes, respectively. The result matrix is given as follows,

$$\boldsymbol{G_s} = \begin{bmatrix} \boldsymbol{A_s} \\ \boldsymbol{B'_s} \end{bmatrix}. \tag{3}$$

Bottom-up approach: mathematical model

A number of mathematical formalisms have been proposed to describe the dynamical interactions between different genes in the network, such as the models with linear functions,

$$F_i(t, \boldsymbol{x}) = \sum_{j=1, j\neq i}^{n} a_{ij} x_j - k_i x_i \tag{4}$$

or the models with non-linear functions,

$$F_i(t, \boldsymbol{x}) = \frac{\sum_{j=1}^{n} a_{ij} x_j}{1 + \sum_{j=1}^{n} b_{ij} x_j} - k_i x_i. \tag{5}$$

Bottom-up approach: mathematical model I

We proposed the following model by applying second truncated Taylor series approximation for the non-linear model (5)

$$F_i(t, \boldsymbol{x}) = \sum_{j=1, j\neq i}^{m} \alpha_j^i x_j + \sum_{1\leq j<k\leq n} \beta_{jk}^i x_j x_k - k_i x_i \tag{6}$$

- x_i : Concentration of TF of single gene.
- $x_i x_j$: Concentration of PDTCBs.
- k_i : self-degradation rate of gene x_i.
- α^i (β^i) : the corresponding coefficient of the target gene x_i.

Interpretation of corresponding coefficient:

- If α_j^i (β_{jk}^i) is positive/nagetive, it means that the gene x_j ($x_j x_k$ dimer or cooperative binding) will active/suppress the expression of gene x_i.
- If the α_j^i (β_{jk}^i) is zero, it means that there is no regulatory relationship exists.

Bottom-up approach: mathematical model II

Assumptions for the mathematical model:

1. The regulations from different genes to a particular gene are additive. Similarly, the regulations from PDTCBs to a particular gene are also additive.
2. α_j^i is the regulation strength from the j-th gene to the i-th gene.
3. β_{jk}^i consists of the regulation strength from the PDTCB $\{x_j - x_k\}$ to the i-th gene and equilibrium constant in the chemical reaction.
4. The auto-regulation is not considered, namely $\alpha_i^i = 0$, to avoid confusion between auto-regulation term $\alpha_i^i x_i$ and degradation term $k_i x_i$.

Parameter inference I

In our case, we have 11 genes and 66 PDTCBs.

Full connection network: 858 unknown parameters in our model for both two differentiation pathways.

After EFSA: 103 unknown parameters (92 directed edges and 11 self-degradation terms) for the erythroid differentiation and 91 unknown parameters (80 directed edges and 11 self-degradation terms) for the neutrophil differentiation.

Parameter inference II

Estimating by simple Genetic Algorithm.

Setup:

1. 1000 generations and 300 individuals per generation.
2. Initial rate constants, $(\alpha_j^i, \beta_{jk}^i, k_i)$, are uniformly distributed within $[W_{min}, W_{max}]$.
3. Set 200 different random seeds for parameter estimation (Different random seed cause different final estimate of rate constants.)
4. Simulation error is calculated by

$$E = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{M}(x_i(t_j) - x_i^*(t_j))^2}. \tag{7}$$

We selected the top ten sets with the minimal estimated errors for further analysis and comparison

Robustness analysis

In this perturbation test, the perturbed parameter is generated by

$$\overline{m_i} = m_i \times (1 + \mu \times \varepsilon), \tag{8}$$

where $\varepsilon \sim N(0, 1)$ and μ is the variation controlling parameter.

- Measurement for the robustness property:

$$E^{(k)} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{M} (x_{ij}^{(k)}(p) - x_{ij}^{(k)})^2} \tag{9}$$

- Robust average:

$$RA = \frac{1}{N} \sum_{k=1}^{N} E^{(k)}, \tag{10}$$

- Robust standard deviation:

$$RSTD = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (E^{(k)} - RA)^2} \tag{11}$$

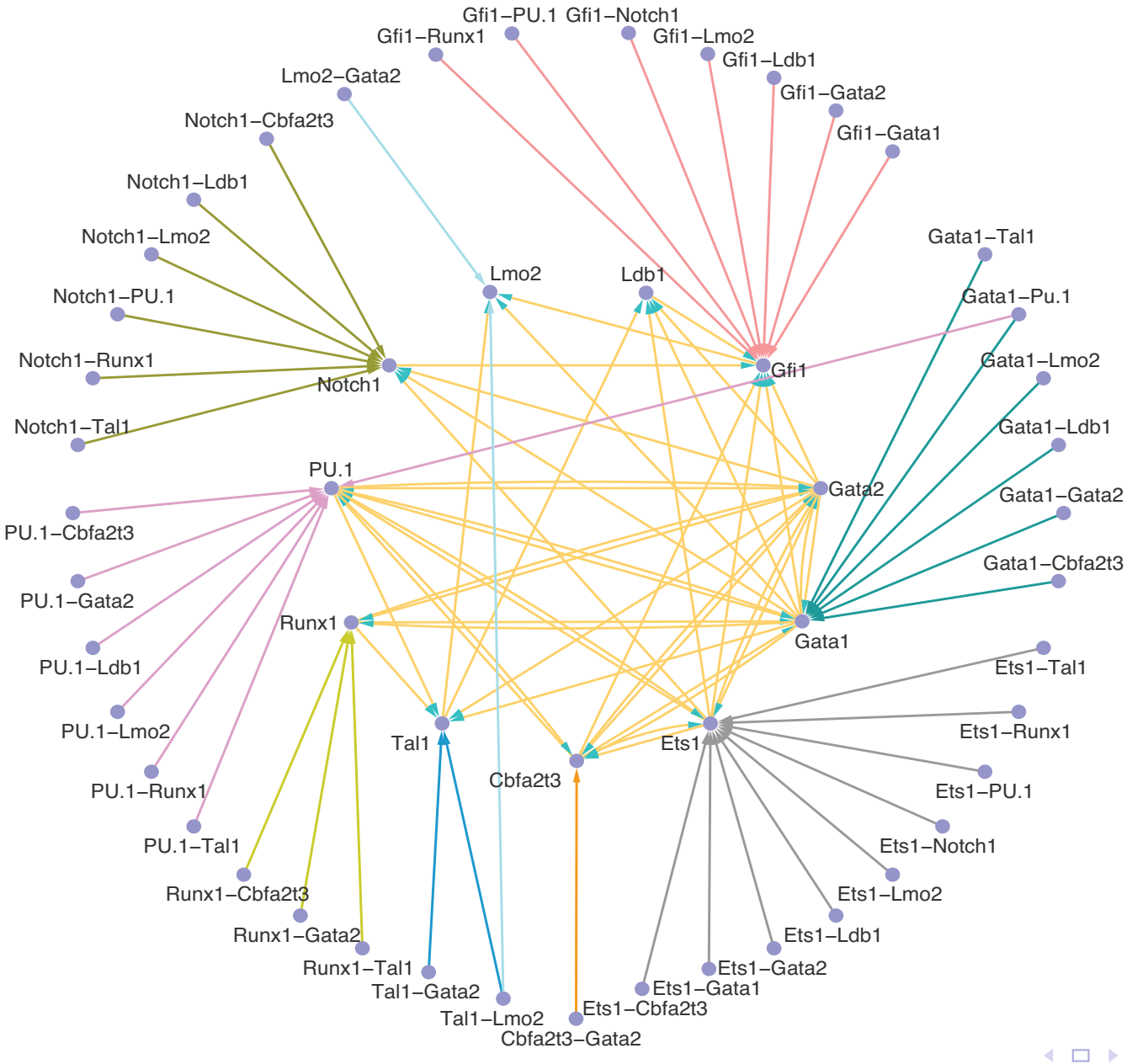
Reduction of Network Model - Edge Deletion

Next, we also test the possibility to delete the potential insignificant edges from our predicted regulatory networks.

1. Test the deletion of regulations from PDTCBs to genes.
2. Test the deletion of regulations between different genes

Criteria: Remove one specific edge permanently if the corresponding new system has the minimal change in simulation error and robustness property.

Inferred regulatory network for the erythroid differentiation



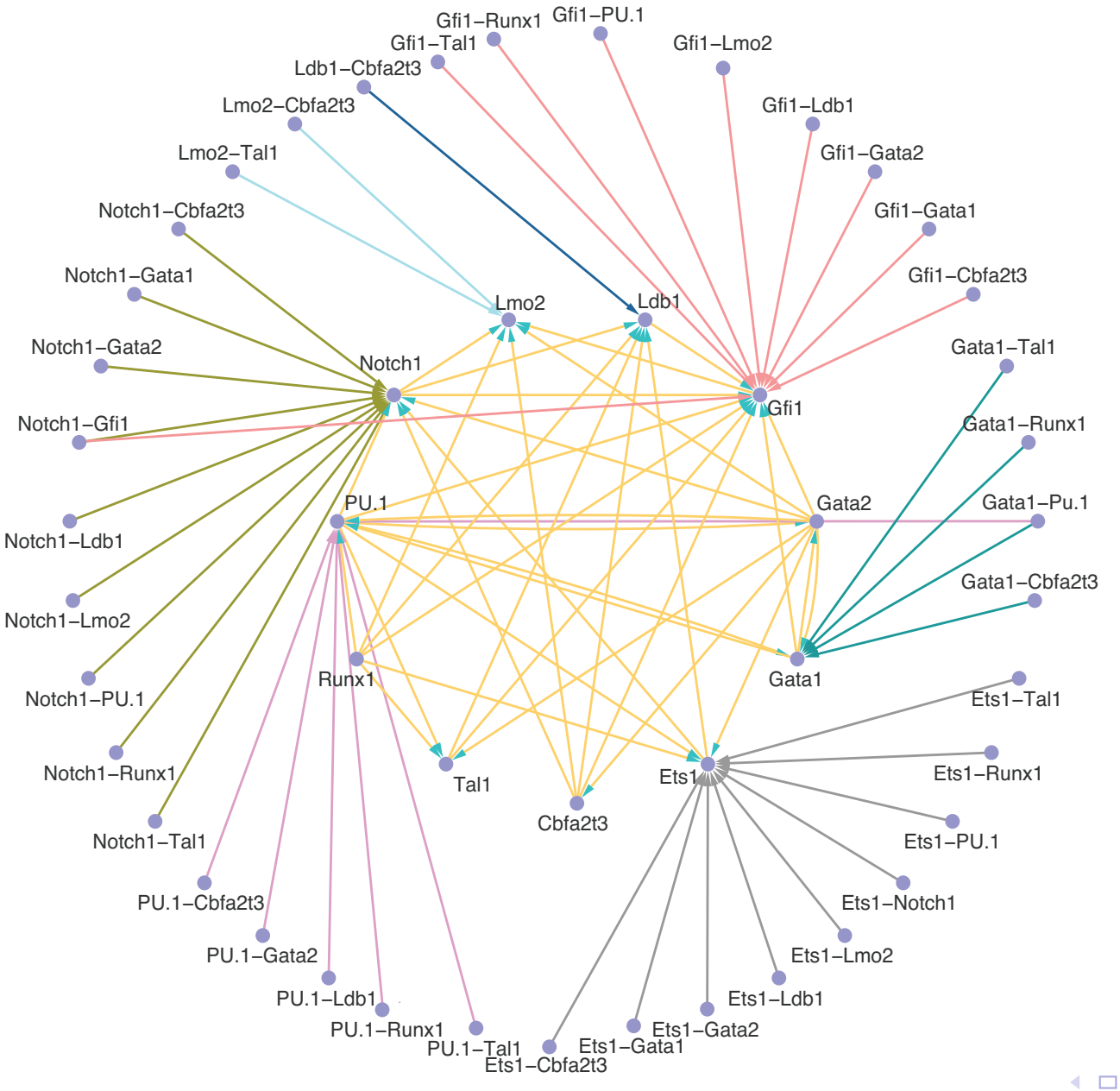
Color Reference Table

- Regulation among genes
- Regulation from NLTs to Gfi1
- Regulation from NLTs to Gata1
- Regulation from NLTs to Ets1
- Regulation from NLTs to Cbfa2t3
- Regulation from NLTs to Tal1
- Regulation from NLTs to Runx1
- Regulation from NLTs to PU.1
- Regulation from NLTs to Notch1
- Regulation from NLTs to Lmo2

Isolated NLTs Table

Gata1–Runx1	Tal1–Ldb1
Gata1–Notch1	Tal1–Gfi1
Gata2–Notch1	Cbfa2t3–Lmo2
Gata2–Ldb1	Cbfa2t3–Ldb1
Runx1–Lmo2	Cbfa2t3–Gfi1
Runx1–Ldb1	Lmo2–Ldb1
Tal1–Cbfa2t3	Gfi1–Ets1

Inferred regulatory network for the neutrophil differentiation



Color Reference Table

- Regulation among genes
- Regulation from NLTs to Gfi1
- Regulation from NLTs to Gata1
- Regulation from NLTs to Ets1
- Regulation from NLTs to PU.1
- Regulation from NLTs to Notch1
- Regulation from NLTs to Lmo2
- Regulation from NLTs to Ldb1

Isolated NLTs Table

- Gata1–Gata2
- Gata1–Lmo2
- Gata1–Ldb1
- Gata2–Runx1
- Gata2–Tal1
- Gata2–Cbfa2t3
- Gata2–Lmo2
- Gata2–Ldb1
- Runx1–Tal1
- Runx1–Cbfa2t3
- Runx1–Lmo2
- Runx1–Ldb1
- PU.1–Lmo2
- Tal1–Cbfa2t3
- Tal1–Ldb1
- Lmo2–Ldb1
- Gfi1–Ets1

Simulation result of the regulatory network for erythroid differentiation

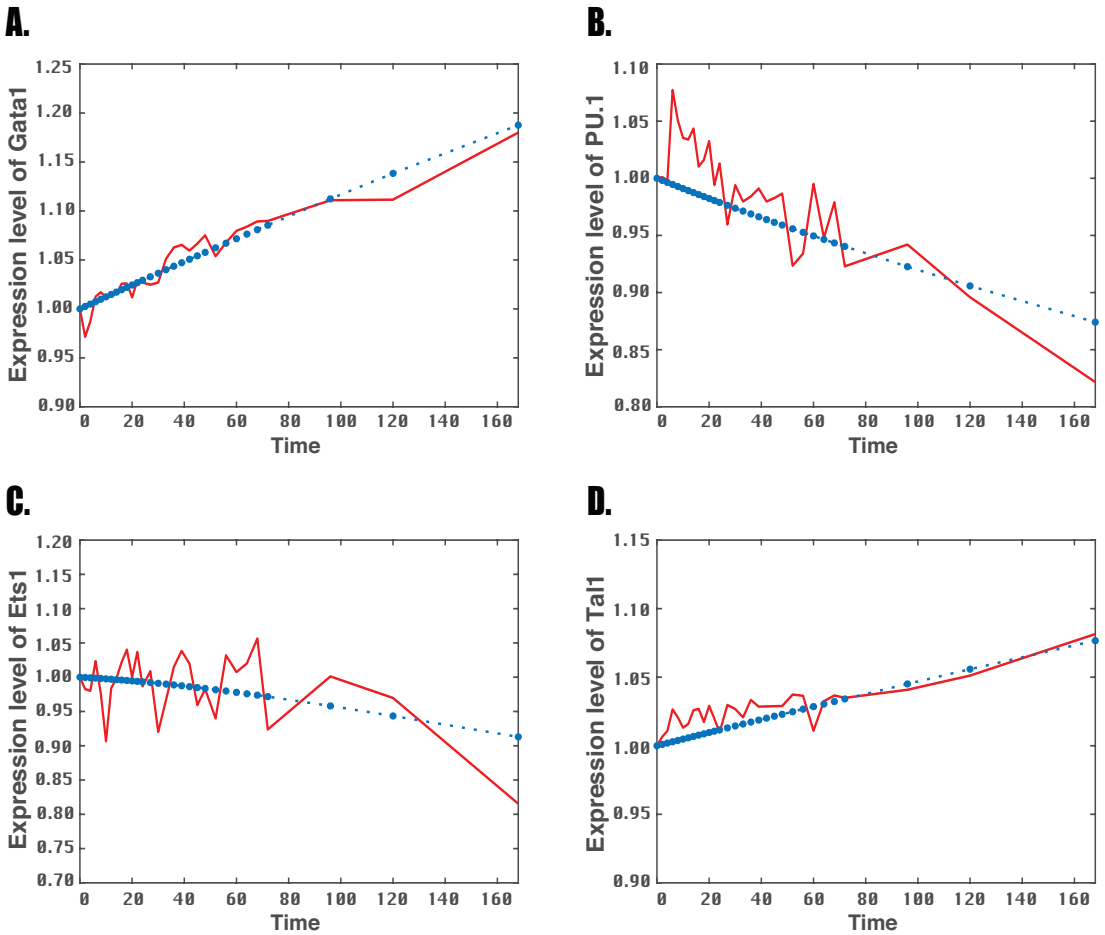


Figure: Red dash line: microarray data; Blue solid line: simulation of the regulatory network.

Simulation result of the regulatory network for neutrophil differentiation

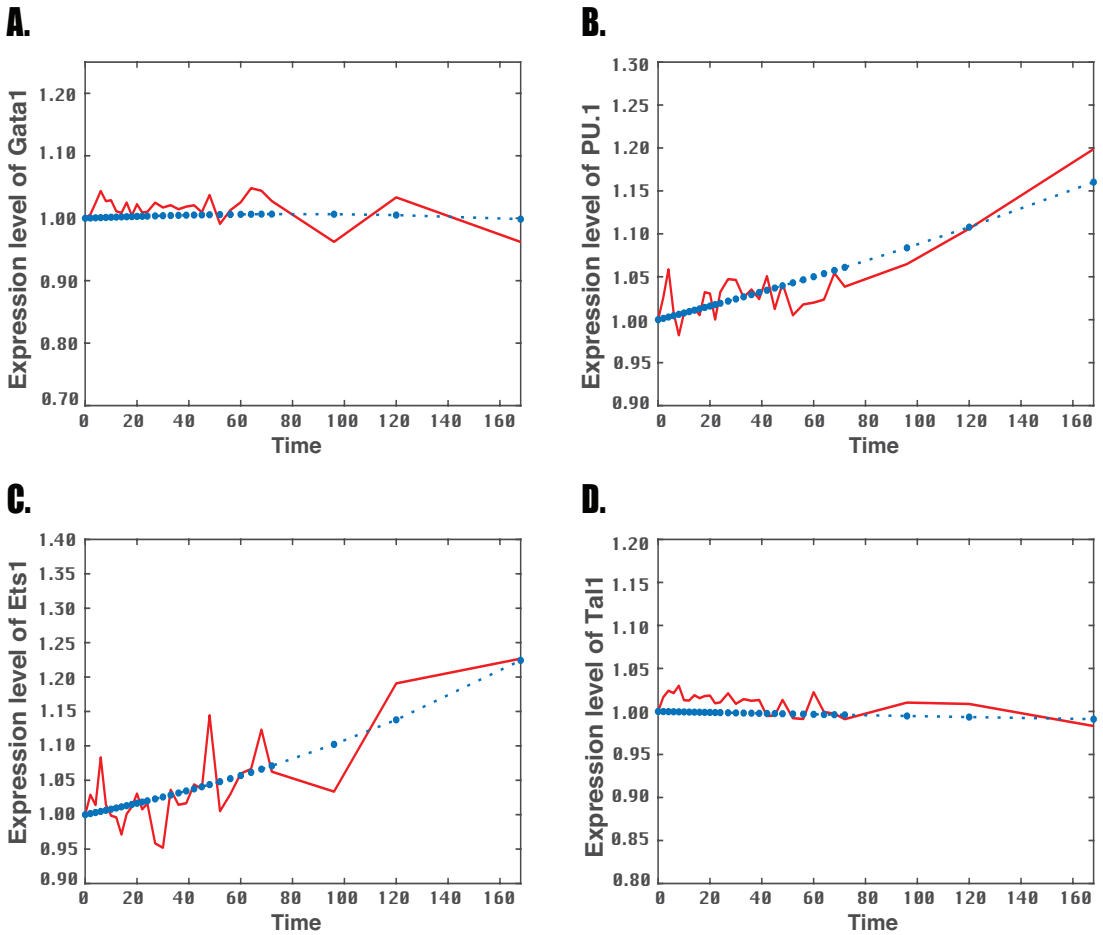


Figure: Red dash line: microarray data; Blue solid line: simulation of the regulatory network.

Potential issue

We need to consider the following issue in future research:

1. Improve the mathematical model to fit the gene expression data.
2. Our probabilistic graphical model only capture the linear relationship between two different components.
3. We ignored the effect of upstream regulatory factors and different binding sites on the DNA sequence.

Outline

Background

A multistable system is a dynamic system with multiple stable states. Multistability is a common dynamic behaviour in the biological system. For example,

- biochemical reactions
- cell signalling systems
- genetic regulatory networks (GRN)

Motivation

Although much progress has been made to describe the stability in the cells' developmental processes, constructing a multistable system with mathematical modellings to describe its mechanism is still a major challenge. We found that most of the studies are still based on bistability. It is still hard to formalize multistability mathematically without high cooperativity coefficient within the biological system.

Formulation of the embeddedness I

For a regulatory network with n genes, the expression level of the i-th gene at time t is denoted as $x_i(t)$. We can use the following a set of n-coupled ordinary differential equations (ODE) to describe the dynamics of the network

$$\frac{d\boldsymbol{X}}{dt} = \boldsymbol{F}(\boldsymbol{X}, \boldsymbol{\Theta}, t) \tag{12}$$

where $\boldsymbol{X} = (X_1, X_2, \cdots, X_n) \in \{\mathbb{R}^+\}^n$ denotes the vector of expression levels of n genes in the n-dimensional non-negative real space, $\boldsymbol{F}(\boldsymbol{x}, \boldsymbol{\Theta}, t)$ is the non-linear vector field and $\boldsymbol{\Theta} \in \{\mathbb{R}^+\}^s$ is the parameter space of s parameters.

Formulation of the embeddedness II

We firstly consider the following two systems with bistability,

$$\text{System 1: } \frac{dX_i}{dt} = \mathcal{F}_i(X_1, \cdots, X_n, X_{n+1}, \cdots, X_{n+N}, \mathbf{\Theta}_1, t) \text{ for } i = 1, \cdots, n+N. \tag{13}$$

$$\text{System 2: } \frac{dY_j}{dt} = \mathcal{G}_j(Y_1, Y_2, \cdots, Y_m, \mathbf{\Theta}_2, t) \text{ for } j = 1, 2, \cdots, m. \tag{14}$$

Then, we consider the following embeddedness

$$X_{n+k} = \mathcal{H}_k(Y_1, Y_2, \cdots, Y_m) \text{ and } \boldsymbol{W} = (X_1, X_2, \cdots, X_n, Y_1, Y_2, \cdots, Y_m). \tag{15}$$

Formulation of the embeddedness III

Thus, we can obtain a embedded system with the parameter space $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}_1 \cup \boldsymbol{\Theta}_2$, which is defined as

$$\frac{d\boldsymbol{W}}{dt} = \boldsymbol{F}(\boldsymbol{W}, \boldsymbol{\Theta}^*, t). \tag{16}$$

This embedded system consists of two components:

$$\text{C1: } \frac{dX_i}{dt} = \mathcal{F}_i(X_1, \cdots, X_n, \mathcal{H}_k(Y_1, Y_2, \cdots, Y_m), \boldsymbol{\Theta}^*, t). \tag{17}$$

$$\text{C2: } \frac{dY_j}{dt} = \mathcal{R}_j(X_1, \cdots, X_n, Y_1, Y_2, \cdots, Y_m, \boldsymbol{\Theta}^*, t). \tag{18}$$

where $i = 1, \cdots, n$, $k = 1, \cdots, N$ and $j = 1, \cdots, m$.

Verification I

To verify whether the framework can realise the tristability, we applied our proposed method to the two bistable toggle switch modules. The first module, named as $Z - U$, is modelled by the following equations with parameter space $\mathbf{\Theta}_1 = \{a = 0.2, b = 4, c = 3\}$, given by

$$\begin{aligned} \frac{dz}{dt} &= \mathcal{F}_1(z, u, \mathbf{\Theta}_1, t) = 0.2 + \frac{4}{1 + u^3} - z, \\ \frac{du}{dt} &= \mathcal{F}_2(z, u, \mathbf{\Theta}_1, t) = 0.2 + \frac{4}{1 + z^3} - u. \end{aligned} \tag{19}$$

The second, named as $X - Y$, module satisfies the same model with the same parameter space.

$$\begin{aligned} \frac{dx}{dt} &= \mathcal{G}_1(x, y, \mathbf{\Theta}_1, t) = 0.2 + \frac{4}{1 + y^3} - x, \\ \frac{dy}{dt} &= \mathcal{G}_2(x, y, \mathbf{\Theta}_1, t) = 0.2 + \frac{4}{1 + x^3} - y. \end{aligned} \tag{20}$$

Verification II

Let consider the $u = \mathcal{H}(x, y) = x + y$, $\boldsymbol{W} = (z, x, y)$ and $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}_1$. Then the non-linear vector fields $\mathcal{G}_{1,2}(x, y, \boldsymbol{\Theta}_1, t)$ are transformed to new non-linear vector fields $\mathcal{R}_{1,2}(x, y, z, \boldsymbol{\Theta}_1, t)$, respectively, which include both genes x, y and z from two sub-systems with the negative regulations from gene z to genes x and y.

$$\begin{aligned}
\frac{dx}{dt} &= \mathcal{R}_1(x, y, z, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{(1 + y^3)(1 + z^3)} - x, \\
\frac{dy}{dt} &= \mathcal{R}_w(x, y, z, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{(1 + x^3)(1 + z^3)} - y, \\
\frac{dz}{dt} &= \mathcal{F}_1(z, u = x + y, \boldsymbol{\Theta}_1, t) = 0.2 + \frac{4}{1 + (x + y)^3} - z.
\end{aligned} \tag{21}$$

Verification III

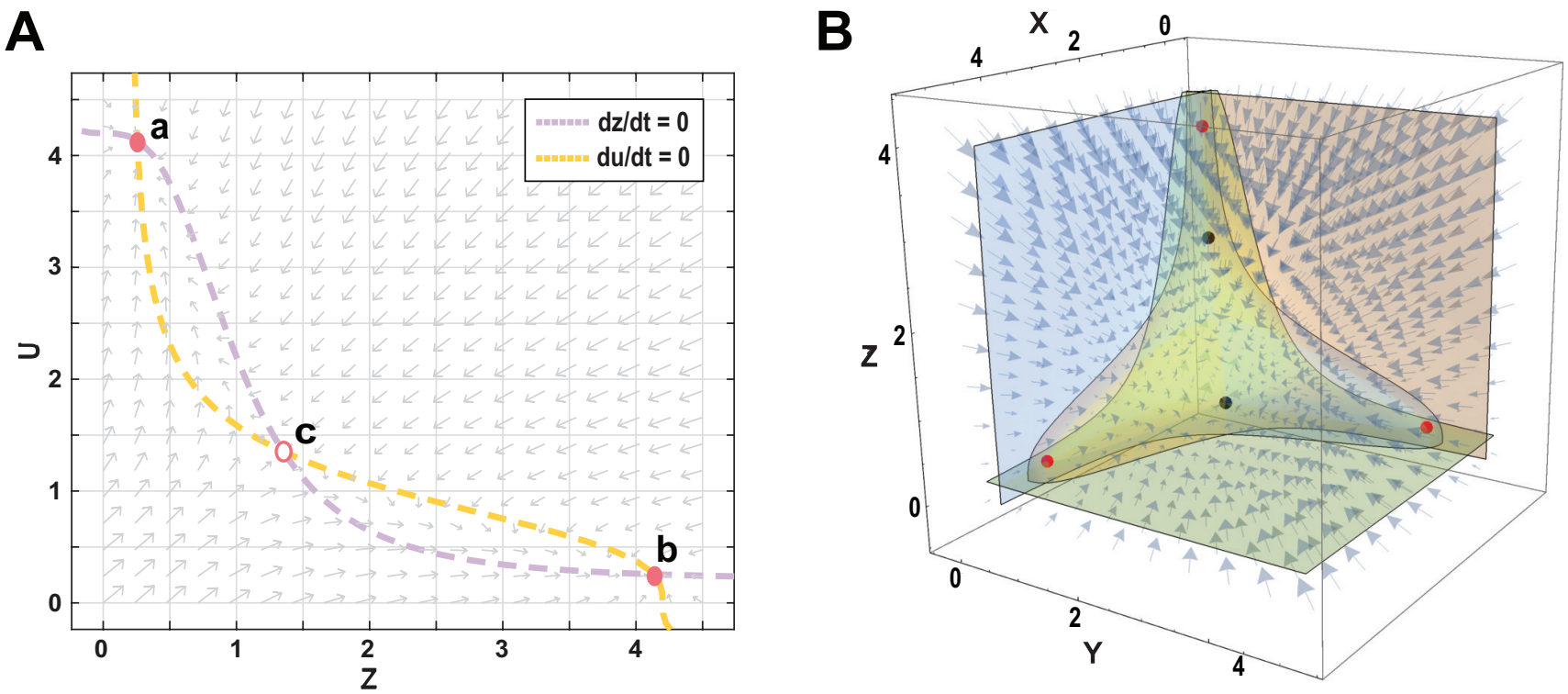


Figure: **Realization of tristability by embedding two bistable sub-systems.** (A) The phase plane of the toggle switch sub-system (19) with bistability (a and b: stable steady states, c: saddle state). (B) The 3D phase portrait of the embedded system (21) with tristability (Three red points: stable steady states; two black points: saddle states)

Embedding method in HSC genetic regulatory network I

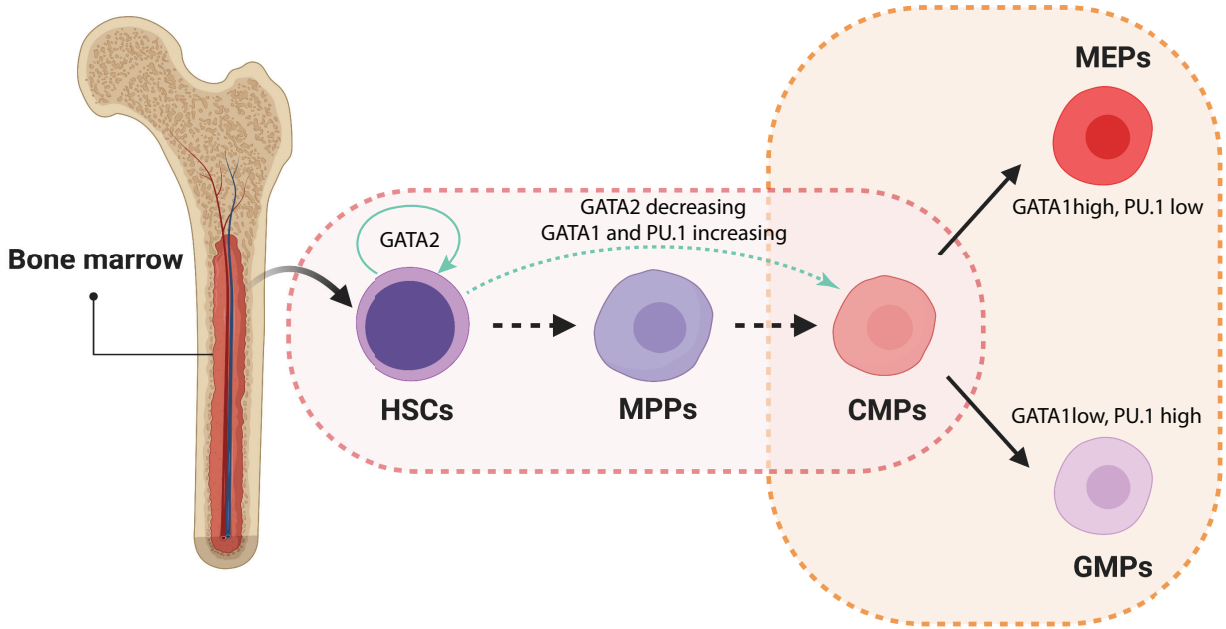


Figure: Brief flowchart for differentiation of HSCs to MEPs and GMPs, respectively. Created with BioRender.com.

Embedding method in HSC genetic regulatory network II

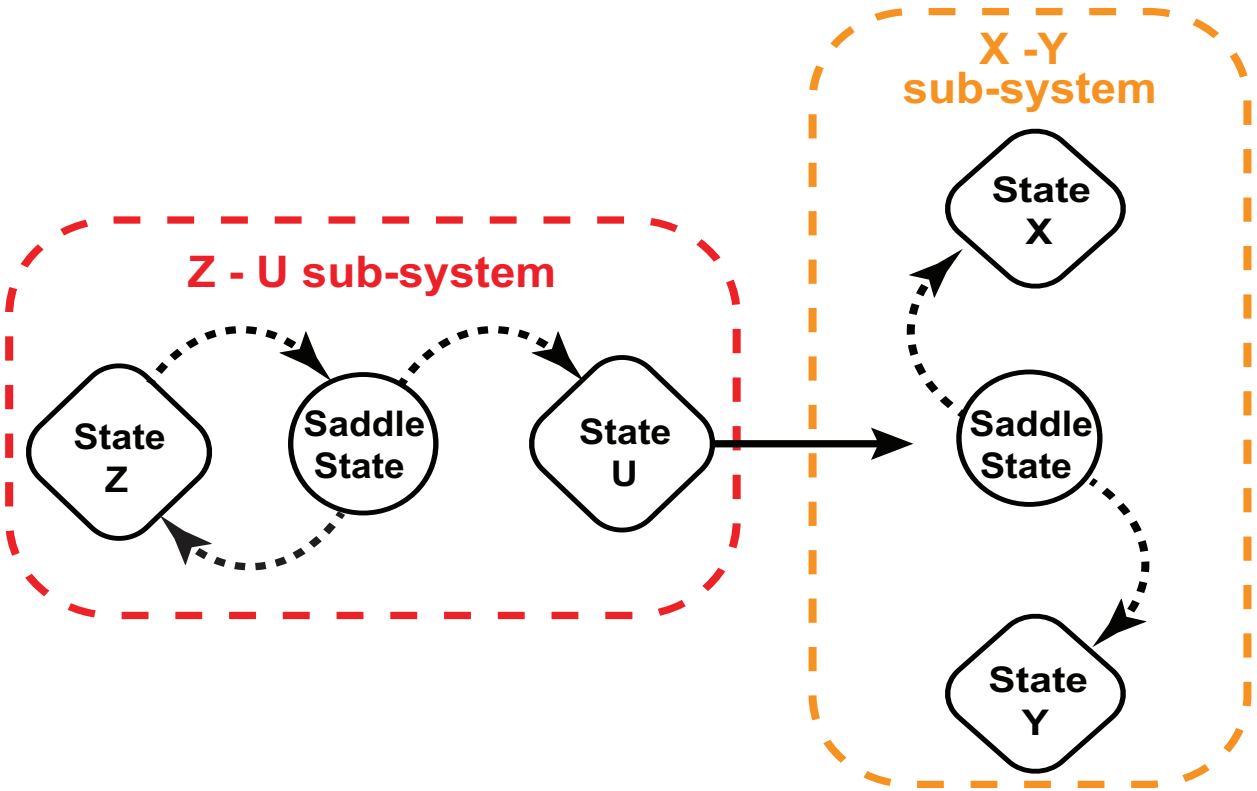
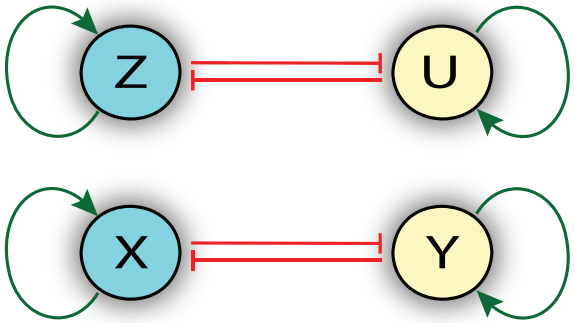


Figure: The illustrative diagram of embeddedness: The principle of embeddedness: Z-U module is the first bistable sub-system. Once this module crosses the saddle point from state Z to state U, it enters the X-Y sub-system that has two stable steady states X and Y, reaching either state X or state Y via the imaginary state U.

Modelling of two-nodes GRNs I



Formulation of Z-U sub-system and X-Y sub-system

$$\begin{aligned}
\frac{dz}{dt} &= \mathcal{F}_1(z,u,\Theta_1,t) = \frac{a_1 z}{1+b_1 z}\frac{1}{1+b_2 u} - k_1 z, \\
\frac{du}{dt} &= \mathcal{F}_2(z,u,\Theta_1,t) = \frac{c_1 u}{1+d_1 u}\frac{1}{1+d_2 z} - k_2 u.
\end{aligned} \tag{22}$$

$$\begin{aligned}
\frac{dx}{dt} &= \mathcal{G}_1(x,y,\Theta_2,t) = \frac{\alpha_1 x}{1+\beta_1 x}\frac{1}{1+\beta_2 y} - k_3 x, \\
\frac{dy}{dt} &= \mathcal{G}_2(x,y,\Theta_2,t) = \frac{\gamma_1 y}{1+\sigma_1 y}\frac{1}{1+\sigma_2 x} - k_4 y.
\end{aligned} \tag{23}$$

Modelling of two-nodes GRNs II

Theorem 1

For the X-Y system, there are at most five sets of non-negative equilibrium points exist. Three equilibrium states are $(0,0)$, $(x_e,0)$ and $(0,y_e)$, where $x_e = \frac{\alpha_1 - k_3}{k_3\beta_1}$ and $y_e = \frac{\gamma_1 - k_4}{k_4\sigma_1}$ for $\alpha_1 > k_3$ and $\gamma_1 > k_4$. The other two possible equilibrium states are denoted as (x_1^*, y_1^*) and (x_2^*, y_2^*), where x_1^* and x_2^* are the positive real solutions of the following equation:

$$\mathcal{A}m^2 + \mathcal{B}m + \mathcal{C} = 0, \tag{24}$$

if $-\frac{\mathcal{B}}{\mathcal{A}} > 0$, $\frac{\mathcal{C}}{\mathcal{A}} > 0$ and $\mathcal{B}^2 - 4\mathcal{A}\mathcal{C} \geq 0$. Where $m = \beta_1 x, \mathcal{A} = A_1B_1 - B_1, \mathcal{B} = A_1 - B_1 - 1 + A_1B_1 - A_1B_2 + A_2B_1$ and $\mathcal{C} = A_1 + A_2 - 1 - A_1B_2$. In addition, $A_1 = \frac{\beta_2}{\sigma_1}, A_2 = \frac{\alpha_1}{k_3}, B_1 = \frac{\sigma_2}{\beta_1}$ and $B_2 = \frac{\gamma_1}{k_4}$. Furthermore, to have the positive value of y_1^* and y_2^*, there is one more condition:

$$x_{1,2}^* < \frac{A_2 - 1}{\beta_1} \text{ or } x_{1,2}^* < \frac{B_2 - 1}{\sigma_2}. \tag{25}$$

Modelling of two-nodes GRNs III

Theorem 2

The X-Y system has three equilibrium states: $(0,0)$, $(x_e,0)$ and $(0,y_e)$.

1. The equilibrium state $(0,0)$ is unstable if $\alpha_1 > k_3$ and $\gamma_1 > k_4$.
2. The equilibrium state $(x_e,0)$ is stable if $\frac{\gamma_1}{1+\sigma_2 x_e} < k_4$.
3. The equilibrium state $(0,y_e)$ is stable if $\frac{\alpha_1}{1+\beta_2 y_e} < k_3$.

Theorem 3

The positive equilibrium states (x_1^*, y_1^*) and (x_2^*, y_2^*) are stable if the following condition is satisfied.

$$\beta_1\sigma_1\eta_y\xi_x - \beta_2\sigma_2\theta_x\rho_y > 0. \tag{26}$$

where $\theta_x = 1+\beta_1 x$, $\eta_y = 1+\beta_2 y$, $\rho_y = 1+\sigma_1 y$ and $\xi_x = 1+\sigma_2 x$.

The two-nodes GRN - GATA Switching I

To study the cell fate commitment of HSCs, we begin with the GATA-switching, which is a main driver of hematopoiesis. During the process of GATA-switching, the displacement of GATA2 and binding of GATA1 at the same *cis-element* will lead cells to leave the HSCs state. Then, after unbinding of GATA2, the dynamical behaviours of GATA1-PU.1 will decide the final cell fate.

The two-nodes GRN - GATA Switching II

GATA Switching

Assume the GATA switching happens over a time interval $[t_1, t_2]$, we introduce the extra term k^*, where $k^* = k_0^* > 0$ for $t \in [t_1, t_2]$, which is the displacement rate of GATA2 protein at the binding side during GATA-switching process. Otherwise, $k^* = 0$.

$$\begin{aligned} \frac{dz}{dt} &= \frac{a_1 z}{1 + b_1 z} \frac{1}{1 + b_2 u} - k_1 z - k^* z, \\ \frac{du}{dt} &= \frac{c_1 u}{1 + d_1 u} \frac{1}{1 + d_2 z} - k_2 u + \psi k^* z, \end{aligned} \tag{27}$$

where the additional term of ψ denotes the synthesis rate constant of GATA1.

The two-nodes GRN - GATA Switching III

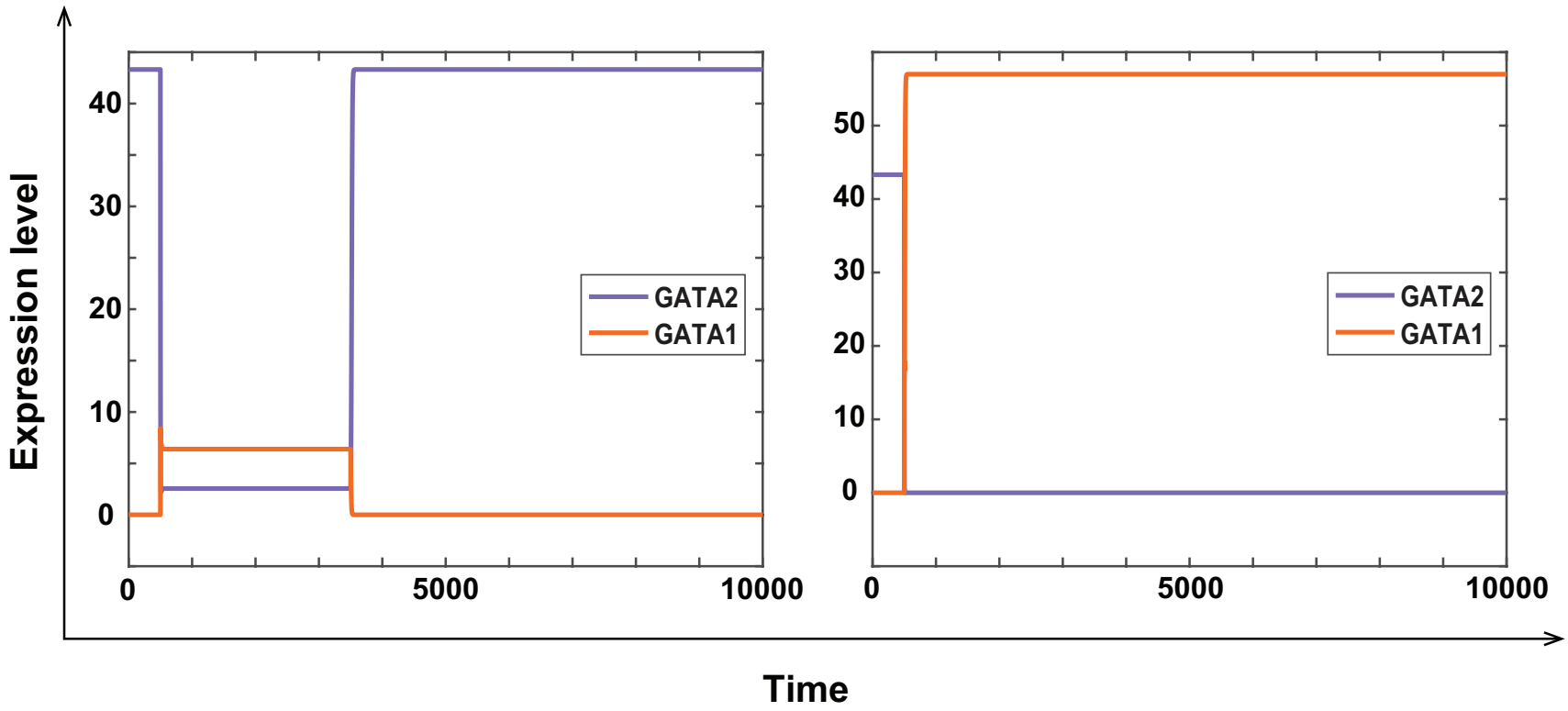


Figure: **Simulations of GATA-switching of the deterministic model** (27). Left panel: An unsuccessful switching with small value of k_0^* due to the displacement of *GATA2* is not enough for cells leave the HSCs state (Z state); Right panel: A successful switching with the enough displacement of *GATA2* by using large value of k_0^*. Cells leave the HSCs state and enter the U state.

The two-nodes GRN - GATA1-PU.1 Module I

While the expression level of GATA2 decreasing, the system approaches state U and triggers the X-Y system. This system also determines the final stable state of the whole embedded system. To determine this system, we simulated based on

1. Single-cell data from the latest biological experiment results.
2. Approximate Bayesian computational rejection-sampling algorithm, where the distance function is defined as

$$\rho(\mathrm{X}, \mathrm{X}^*) = \sum_{i=1}^{m} [|x_i - x_i^*| + |y_i - y_i^*|], \tag{28}$$

where (x_i, y_i) and (x_i^*, y_i^*) are the observed data and simulated data of the model at time point t_i for genes (X, Y), respectively.

The two-nodes GRN - GATA1-PU.1 Module II

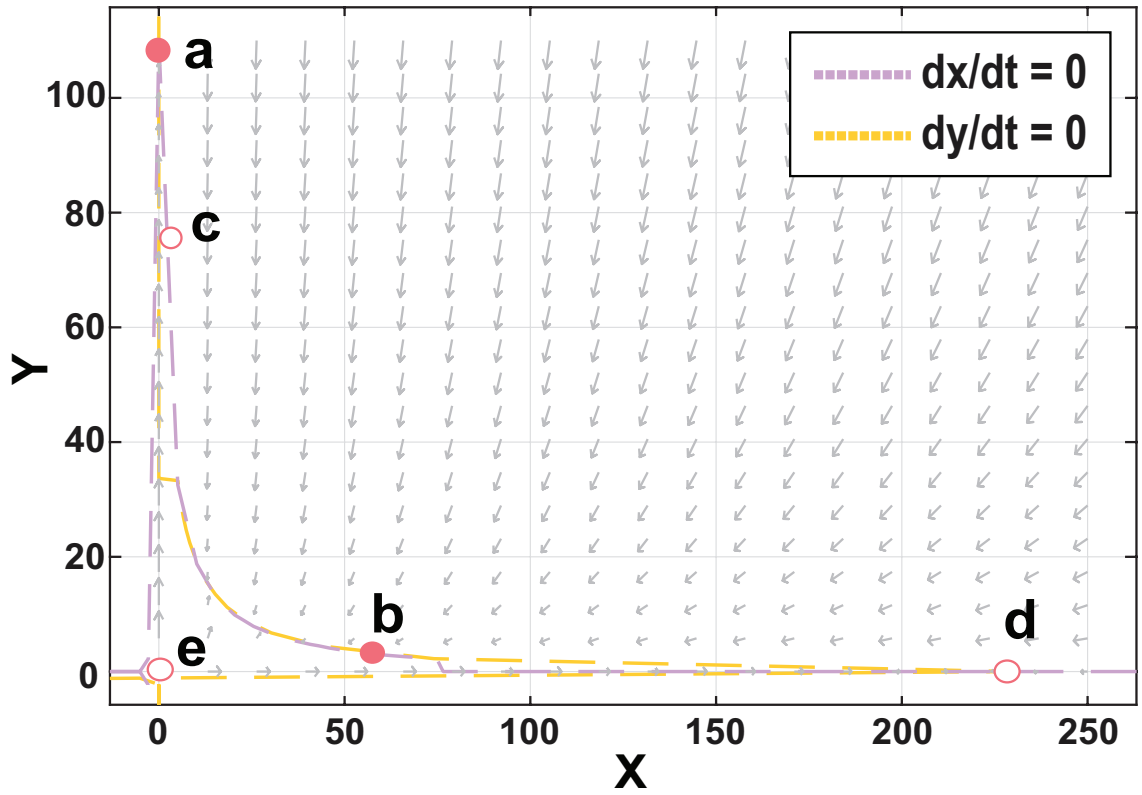
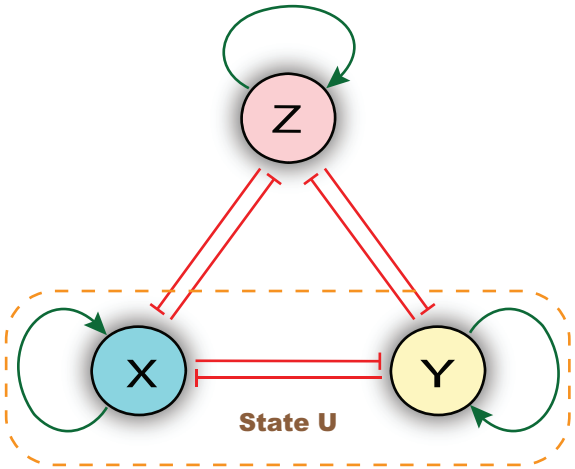


Figure: Phase plane of the *GATA1-PU.1* module shows the bistable property of the proposed model, where a and b are stable steady states

Modelling of the three-nodes embedded GRN I

Let consider the $u = \mathcal{H}(x, y) = \mu x + \delta y$, where μ and δ are two positive control parameters.



Modelling of the three-nodes embedded GRN II

Formulation of the embedded X-Y-Z system

$$\begin{aligned}
\frac{dx}{dt} &= \frac{\alpha_1 x}{1+\beta_1 x}\frac{1}{1+\beta_2 y}\frac{1}{1+d_2 z} - k_3 x, \\
\frac{dy}{dt} &= \frac{\gamma_1 y}{1+\sigma_1 y}\frac{1}{1+\sigma_2 x}\frac{1}{1+d_2 z} - k_4 y, \\
\frac{dz}{dt} &= \frac{a_1 z}{1+b_1 z}\frac{1}{1+b_2(\mu x+\delta y)} - k_1 z.
\end{aligned} \tag{29}$$

Theorem 4

If $(x_e, 0)$ and $(0, y_e)$ are the equilibrium states of X-Y system and $(z_e, 0)$ is a equilibrium state of Z-U system, where $x_e = \frac{\alpha_1 - k_3}{k_3\beta_1}$, $y_e = \frac{\gamma_1 - k_4}{k_4\sigma_1}$ and $z_e = \frac{a_1 - k_1}{k_1 b_1}$. Then $(x_e, 0, 0)$, $(0, y_e, 0)$ and $(0, 0, z_e)$ are three equilibrium states of system. Moreover, if (x_1^*, y_1^*) and (x_2^*, y_2^*) are two positive equilibrium states of X-Y system as stated in **Theorem 1**, then $(x_1^*, y_1^*, 0)$ and $(x_2^*, y_2^*, 0)$ are still two equilibrium states of X-Y-Z system.

Modelling of the three-nodes embedded GRN III

Theorem 5

If $(x_e, 0)$ and $(0, y_e)$ are both stable states of X-Y system and $(z_e, 0)$ is a stable state of Z-U system.

1. The equilibrium state $(x_e, 0, 0)$ is stable if $\frac{a_1}{1+b_2 x_e} < k_1$.
2. The equilibrium state $(0, y_e, 0)$ is stable if $\frac{a_1}{1+b_2 y_e} < k_1$.
3. The equilibrium state $(0, 0, z_e)$ is stable if $\frac{\alpha_1}{1+d_2 z_e} < k_3$ and $\frac{\gamma_1}{1+d_2 z_e} < k_4$.

Theorem 6

Suppose (x^*, y^*) is a stable state of X-Y system, then the equilibrium state $(x^*, y^*, 0)$ is also a stable state of the X-Y-Z system if

$$\frac{a_1}{1 + b_2(x^* + y^*)} < k_1 \tag{30}$$

The embedded three-nodes GRN with GATA switching I

To realise the GATA-switching, the system is updated as the follows

The three-nodes GRN with GATA Switching

$$\begin{aligned}
\frac{dx}{dt} &= \frac{\alpha_1 x}{1+\beta_1 x}\frac{1}{1+\beta_2 y}\frac{1}{1+d_2 z} - k_3 x + \psi k^* z, \\
\frac{dy}{dt} &= \frac{\gamma_1 y}{1+\sigma_1 y}\frac{1}{1+\sigma_2 x}\frac{1}{1+d_2 z} - k_4 y, \\
\frac{dz}{dt} &= \frac{a_1 z}{1+b_1 z}\frac{1}{1+b_2(\mu x+\delta y)} - k_1 z - k^* z.
\end{aligned} \tag{31}$$

The embedded three-nodes GRN with GATA switching II

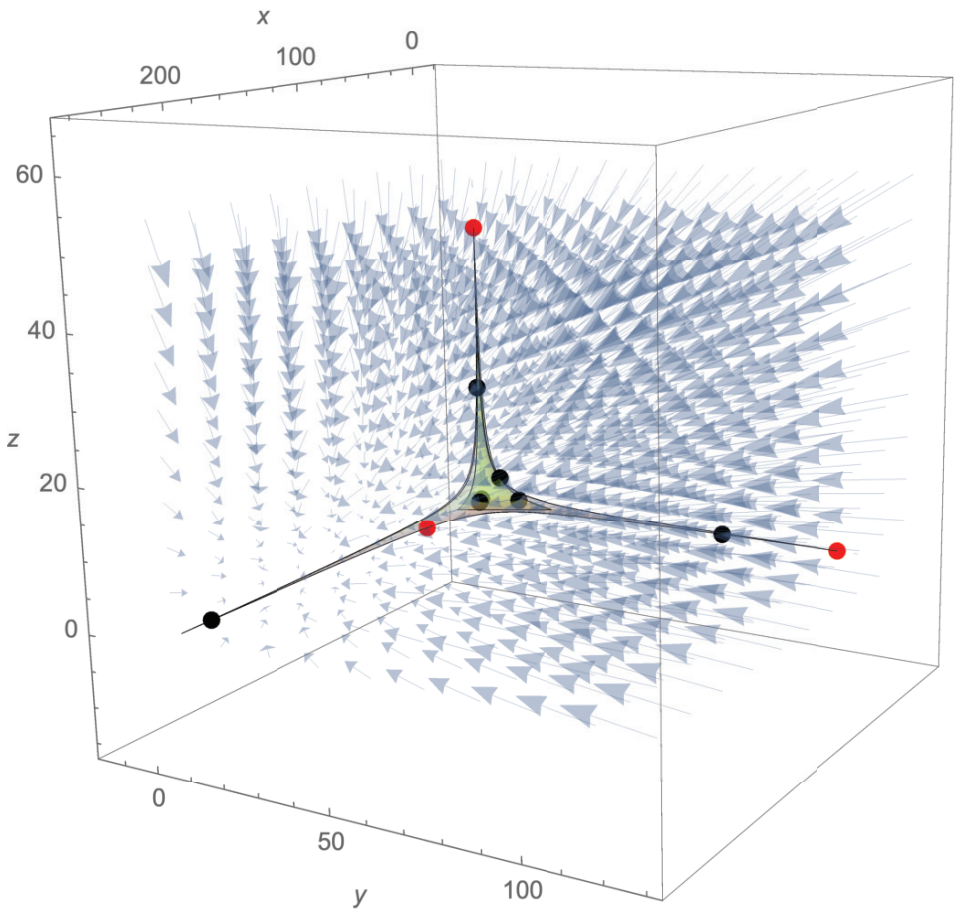


Figure: **The 3D phase portrait of the embedded system.** Based on the experimental data, the proposed model successfully realise the tristability properties, with the same parameter values. Red points: stable steady states; Black points: saddle states.

The embedded three-nodes GRN with GATA switching III

In fact, experimental studies suggest that *GATA2* moderately simulates the expression of gene *GATA1*. Thus, we make a modification to model (31) by adding the term d^*z in the first equation to represent the weak positive regulation from *GATA2* to *GATA1*. In addition, to avoid zero basal gene expression levels, we also add a constant to each equation of the proposed model.

Modified three-nodes GRN with GATA switching

$$\begin{aligned}
\frac{dx}{dt} &= \frac{\alpha_0 + \alpha_1 x}{1+\beta_1 x}\frac{1}{1+\beta_2 y}\frac{1+d^*z}{1+d_2 z} - k_3 x + \psi k^* z, \\
\frac{dy}{dt} &= \frac{\gamma_0 + \gamma_1 y}{1+\sigma_1 y}\frac{1}{1+\sigma_2 x}\frac{1}{1+d_2 z} - k_4 y, \\
\frac{dz}{dt} &= \frac{a_0 + a_1 z}{1+b_1 z}\frac{1}{1+b_2(x+y)} - k_1 z - k^* z.
\end{aligned} \tag{32}$$

The embedded three-nodes GRN with GATA switching IV

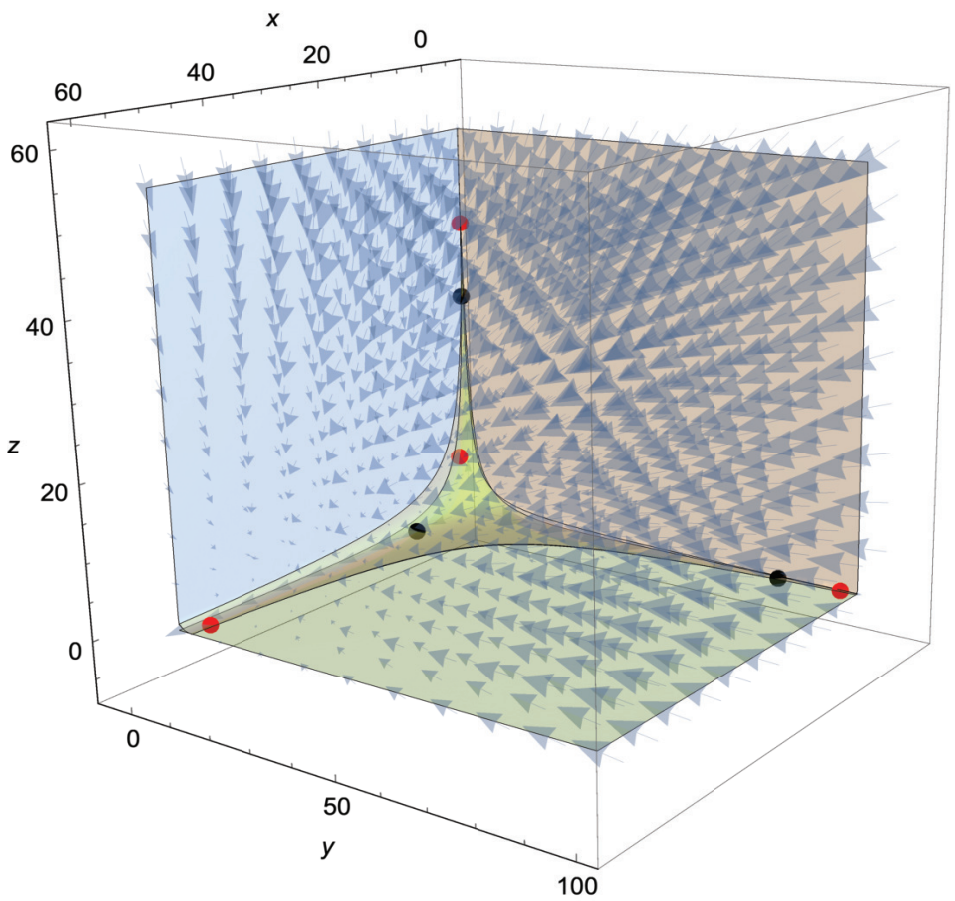


Figure: **The 3D phase portrait of the modified embedded system.** Red points: stable steady states; Black points: saddle states.

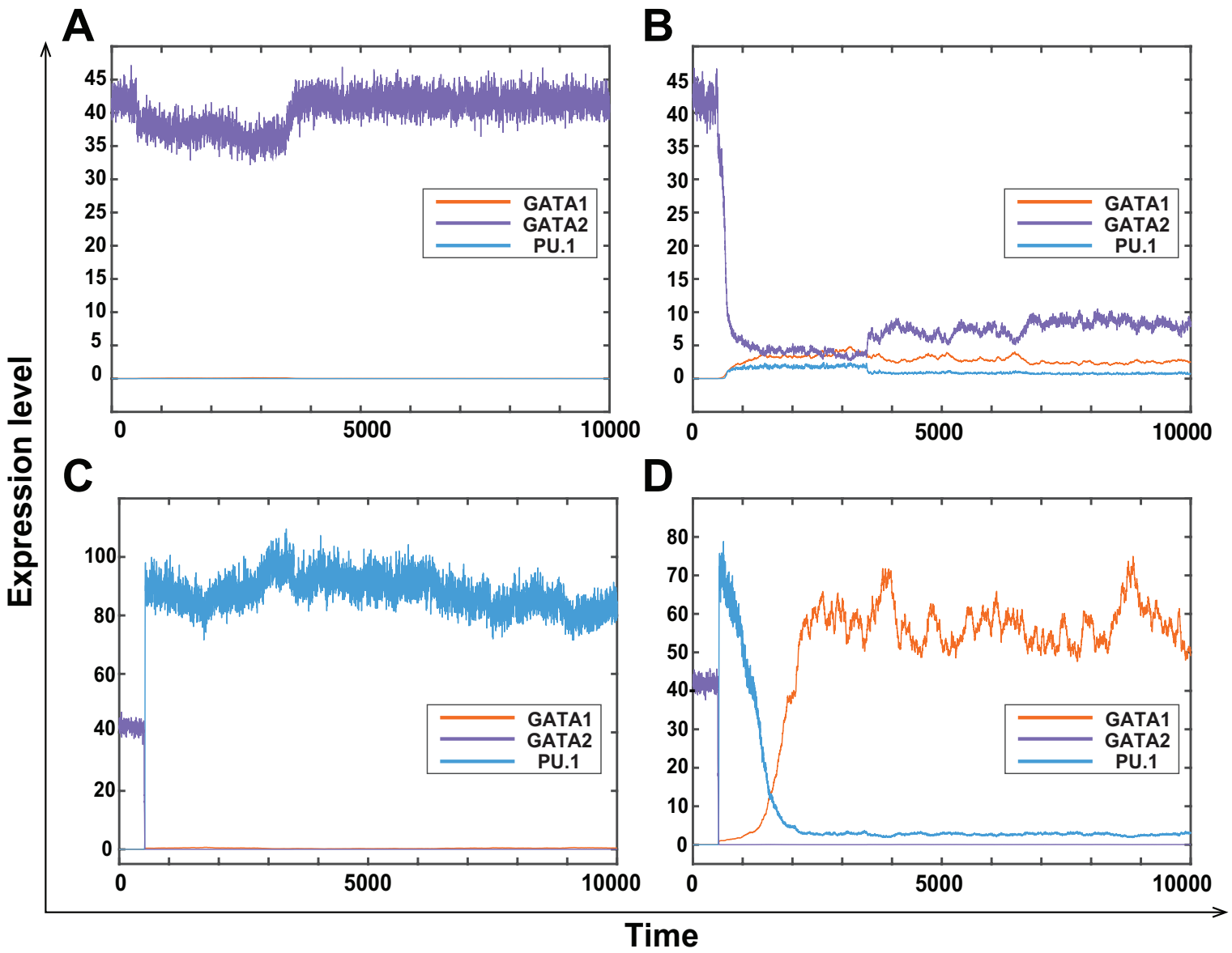
Stochastic dynamics I

Deterministic model always fail to describe the heterogeneity in the mechanism of cell fate commitment. To solve this issue, we proposed the following stochastic differential equations (SDE) model to describe the noise during the developmental process, given by

Stochastic dynamics

$$
\begin{aligned}
dX(t) &= \left[\frac{\alpha_0+\alpha_1 X(t)}{1+\beta_1 X(t)}\frac{1}{1+\beta_2 Y(t)}\frac{1}{1+d_2 Z(t)} - k_3 X(t) + \psi k^* Z(t)\right] dt \\
&\quad + \omega_1[k_3 X(t) + \psi k^* Z(t)] dW_t^1, \\
dY(t) &= \left[(\frac{\gamma_0+\gamma_1 Y(t)}{1+\sigma_1 Y(t)}\frac{1}{1+\sigma_2 X(t)}\frac{1}{1+d_2 Z(t)} - k_4 Y(t)\right] dt \\
&\quad + \omega_2 k_4 Y(t) dW_t^2, \\
dZ(t) &= \left[(\frac{a_0+a_1 Z(t)}{1+b_1 Z(t)}\frac{1}{1+b_2(\mu X(t)+\delta Y(t))} - k_1 Z(t) - k^* Z(t)\right] dt \\
&\quad + \omega_3(k_1+k^*) Z(t) dW_t^3.
\end{aligned}
\tag{33}
$$

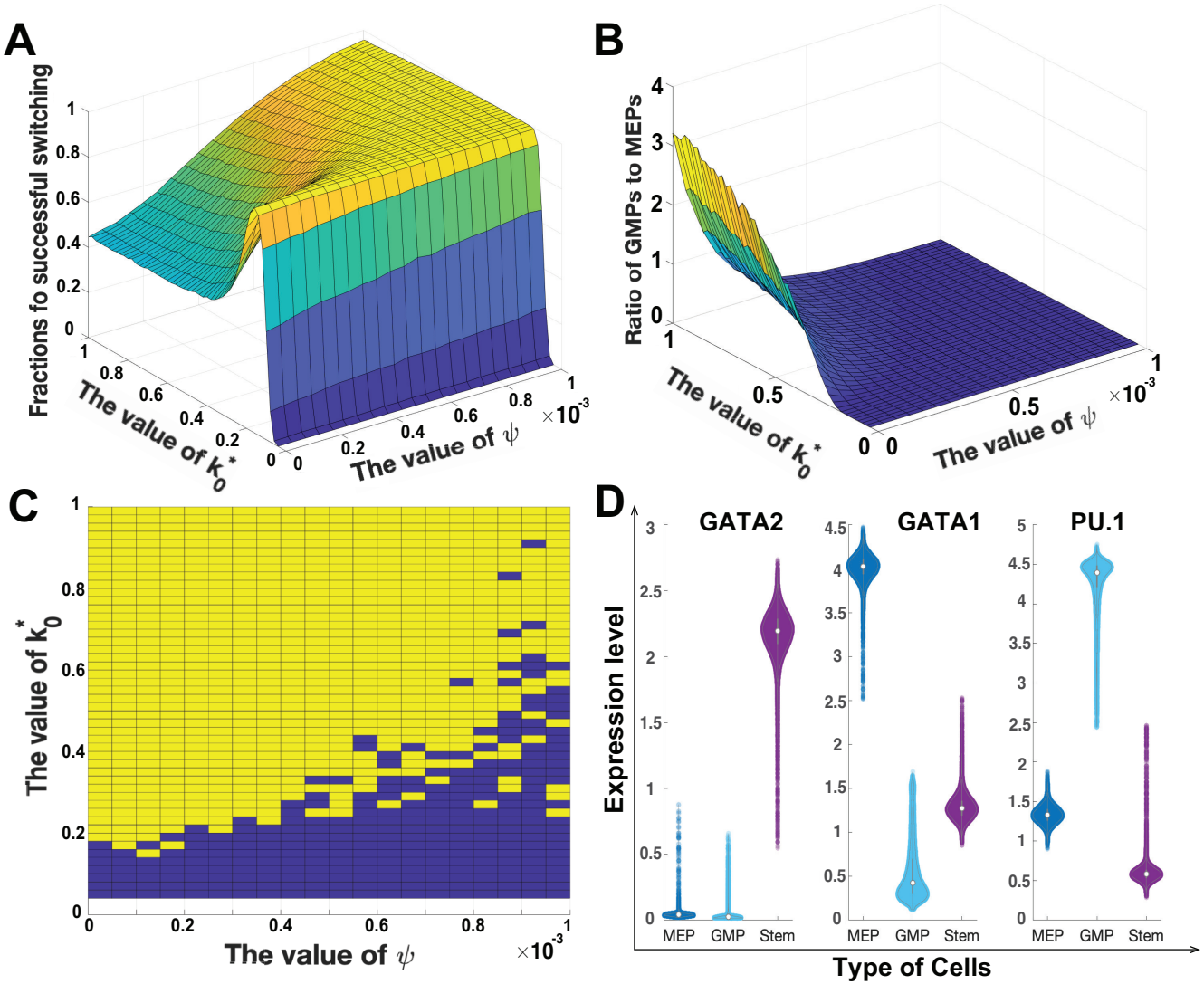
Stochastic dynamics III



Stochastic dynamics IV

Stochastic simulations shows four stable states that correspond to the experimentally observed four different states. (A) Simulation of unsuccessful GATA switching which makes the cell stays at the HSCs state. (B) Simulation of unsuccessful GATA switching but the cell enters the state with low expression of all three genes. (C) Simulation of successful switching which leads to the GMP state with high expression levels of *PU.1*. (D) Simulation of successful switching which leads to the MEP state with high expression levels of *GATA1*.

Stochastic dynamics V



Stochastic dynamics VI

Distributions of different cell types derived from stochastic simulations. (A) Frequencies of cells having successful switching for each set of parameters (k_0^*, ψ). (B) Ratios of GMP cells to MEP cells when the cells have successful switching in (A) for each set of parameters (k_0^*, ψ). (C) Parameter sets of (k_0^*, ψ) that generate stochastic simulations with four steady states. (yellow part) or with two or three states (blue part). (D) Violin plots of the natural log normalised (expression level per cell + 1) distributions for three genes in different cell states derived from stochastic simulations with parameters $k_0^* = 0.52$ and $\psi = 0.0002$.

Open questions

1. How to embed more modules with more transcriptional factors to develop mathematical models with more stable steady states.
2. How to determine the conditions for realizing the multistable properties in stochastic models with transcriptional bursting.

Outline

Motivation

Because of the belief that gene activity should be discontinuous, transcriptional bursting is a considerable idea within the biological system. Genes are restricted in their expression ability with very low expression rates in "off" states and highly expressed with stochastic bursts in "on" states. We aim to develop a stochastic model to describe transcriptional bursting and also attempt to capture both intrinsic and extrinsic noise that affects the transcriptional process.

Objectives

1. How to determine the time length between two bursting processes?
2. How to determine the bursting size for each process?

Leaky telegraph model

To achieve the multistability with the transcriptional bursting, we decide to develop the method based on the leaky telegraph model

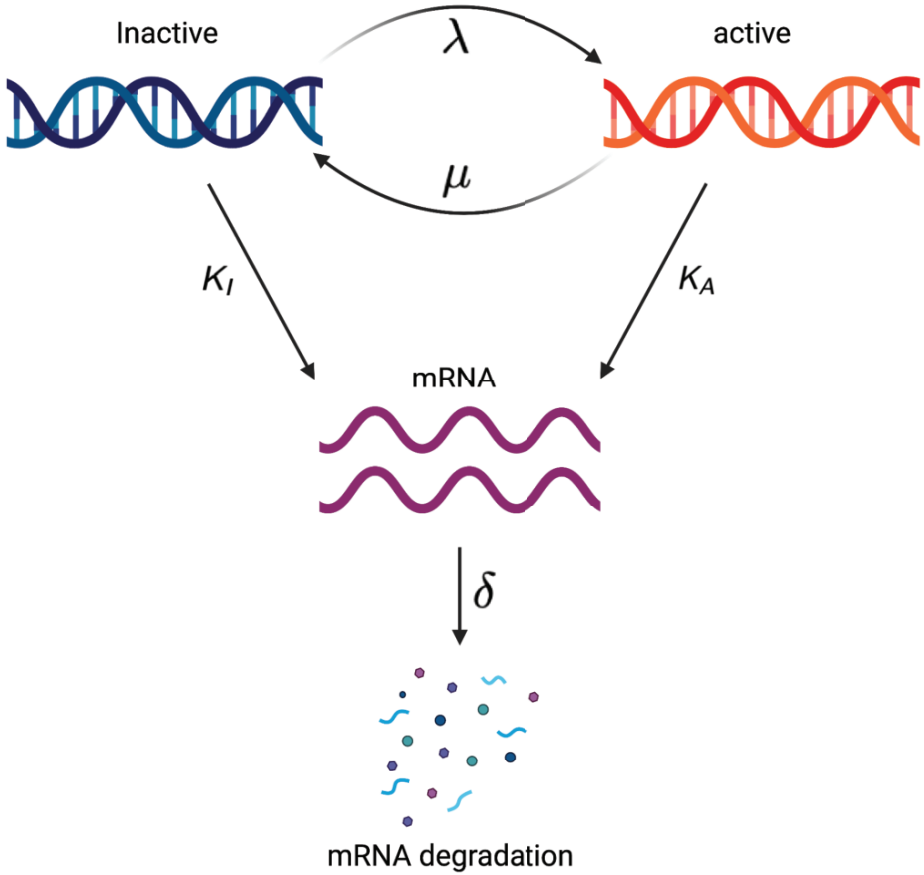


Figure: Created with Biorender.com

Next step

1. Use exponential distribution to describe the waiting time for the next bursting arrival.
2. Use Poisson distribution with constant rate to describe the size for each bursting process.
3. Use Poisson distribution with varies rate (setup as a function) to describe the size for each bursting process.

Outline

Progress to date

During my 3-year research life, I have attended two academic conferences. Moreover, I have published one conference paper and one journal paper. I also submitted another journal paper. The new project is still in progress

Publication:

1. **Wu S.**, Cui T. and Tian T. Mathematical Modelling of Genetic Network for Regulating the Fate Determination of Hematopoietic Stem Cells, Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2018), 2167-2173, IEEE Press.
2. **Wu S.**, Cui T, Zhang X and Tian T. 2020. A non-linear reverse-engineering method for inferring genetic regulatory networks. PeerJ 8:e9065
3. **Wu S.**, Zhou T. and Tian T. A robust method for designing multistable systems by embedding bistable subsystems. (Journal submitted)

Conference:

1. 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2018). Madrid Spain.
2. 2019 CSIAM 1st Annual Conference on Mathematical Life Science. Guangzhou China.

Outline of thesis chapters

I prefer to finish the thesis based on my publication. The framework and possible chapters of my final thesis include but are not limited to:

- **Chapter 1** - Introduction
- **Chapter 2** - Inference for genetic regulatory networks (**P.S: Based on the BIBM conference paper and published paper on PeerJ**)
- **Chapter 3** - Embedded method for realisation of tristability (**P.S: Based on the submitted paper**)
- **Chapter 4** - Application of embedded method on the GATA1-GATA2-PU.1 regulatory complex. (**P.S: Based on the submitted paper**)
- **Chapter 5** - Stochastic model for transcriptional bursting. (**P.S: Based on my current project**)
- **Chapter 6** - Limitation of study and Open questions
- **Chapter 7** - Conclusion

Plan to thesis completion

- **Jul. 2021:**
  1. Methods development.
  2. Draft the paper.
- **Aug. 2021 to Sep. 2021:**
  1. Finalise the current paper and submit to publisher/conference.
  2. Draft the thesis and format the contents from the published/submitted manuscripts.
  3. Write the introduction chapter and proof-read the thesis.
  4. Submit the thesis for PhD examination.

Thanks!

Thanks for your time!