**Introduction**

生命体系的动力学特性是由极其复杂的基因网络调控的，但这复杂的调控过程有着惊人的准确性，这其中以胚胎发育最为典型。同一生物体内内所有的细胞都由胚胎干细胞转化而来，尽管这些细胞基因型相同，但它们可以展现出不同的表型或者状态，而且在特殊的外环境或刺激作用下可以在不同表型间进行转化。细胞态转化是典型的非平衡动力学过程，它在很多其他重要的生命过程比如肿瘤转移和成纤维化中也广泛发生。研究细胞态转化的调控是生物学中最基本的问题之一。

目前单细胞测序技术的进步为研究细胞态转化中基因调控网络的动力学过程提供了大量实验数据，每个细胞的状态可以表示成高维基因矢量,其动力学过程可以用郎之万方程描述

，其中为随机项

但是受限于实验技术，单细胞测序数据只能提供静态的分布信息。因此目前发展出一些算法比如赝时间算法，RNA速率算法等方法对细胞态转化的动力学进行推测，但这些研究多是从生物信息学的角度出发，很少有研究从统计物理的角度出发分析不同细胞态转化过程的普适规律。

虽然细胞体系不能用简单序参量极性描述，但是这些基因是高度耦合的，从而把的动力学过程限制在到一个低维的流形上，这也就意味着细胞的状态可以用低维的态参量（潜变量）进行描述。如果能够定义和计算这个低维流形的度规，就可以对细胞态转化的动力学进行定量分析。但是目前这个领域的研究多局限在使用流形学习方法比如UMAP,t-SNE对细胞分布进行投影以达到可视化的效果，但这些方法不能保留数据原始的几何信息，也不能用于研究态转化的动力学过程。

基于费雪尔信息和费雪尔度规的信息几何方法为分析数据的流形结构提供了新的工具。目前在统计物理中已经开始使用该方法分析伊辛模型的临界点。

**AIM**

拟解决的科学问题一：细胞态转化的临界点分析。在外界刺激下，细胞态会在不同稳态间进行转化，转化路径要穿过临界点，这一临界点对应于化学反应中的过渡态。准确识别过渡态在基因表达空间或者态参量空间的位置不仅对于我们认识基因调控网络的拓扑性质至关重要，而且在生物和医学上有着重要意义。目前已经发展出一些不同的方法用于寻找细胞态转化或者病变的指标或者预警信号，但这些方法都是针对两种状态之间的转化，很难运用到多种状态之间的转化，比如发育过程中的分叉。在之前的研究中，我们发现细胞在不同状态间转化会伴随着基因调控网络（布尔网络模型）的阻挫瞬时升高。在本项研究中，我们会借助统计物理和信息几何的方法，发展更加普适的识别过渡态的方法，并对基因调控网络在临界点的性质进行分析。

拟解决的科学问题二：细胞态转化的流形以及路径分析。基因调控网络决定了细胞表型景观，但Waddington 提出的表型景观更接近一种比喻。尽管最近的研究提出了基于赝势的表型景观的计算方法，但这种赝势并不能用来分析细胞态转化的动力学过程。我们希望直接从数据出发，基于信息几何方法分析细胞态转化在态参量（潜变量）的流形。根据流形的测地线分析细胞态转化的路径，并希望由此分析影响转化路径的关键基因。

**Methods**

在本项研究中，我们计划采用机器学习和信息几何的方法对细胞态转化进行分析。其中关键的步骤在于从高维数据中推断具有物理意义的态参量（潜变量）及其维度。对于高维数据，需要借助机器学习对数据降维，但是常规的降维方法并不考虑数据中的动力学信息，因此我们计划高斯过程潜变量模型(Gaussian process latent variable model, GPLVM)分析其内禀自由度。

高斯过程潜变量模型可以在先验参数中加入动力学信息或者推断的动力学信息。对于高斯过程Gaussian process ，考虑输入,假设输出函数的先验满足，其中，， 。 代表核函数，作为协方差covariance。

这里并不需要知道的形式，它可以是无穷维，这里实际需要的就是核函数的形式。

对于GPLVM， 要找到输出对应的低维表示，其中.



其中是非线性函数，它的先验满足，噪声

那么边缘概率



其中是kernel核函数和噪声的超参数，是和 的函数

通过MLE(marginal likelihood estimation)找到最好的，从而得到模型的参数(1)。

通过GPLVM,可以实现系统降维，但是要得到比较好的结果，需要对的维度以及初值有比较好的估计。实际求解中需要预设的维度，初值以及核函数的形式。

而态参量（潜变量）的个数可以通过最大似然估计的方法(2-4)进行推测或者predictive coding(5)。

假设不变，通过上述分析得到态参量（潜变量）后，我们可以把映射到低维的态参量空间中，然后通过费雪尔度规来分析细胞态转化的动力学。，这里两个偏导的函数形式都是已知的。

（核函数比如rbf 实际上也依赖于距离，那么根据这个核函数计算的度规是否存在问题？？？）

另外一种方式就是用高斯混合模型 Gasusian mixture model(GMM) 拟合数据分布。GMM就是一个潜变量模型，潜变量对应了GMM里component. 但是这里我们借助eigen gene作为低维的潜变量。Eigengene 是一种非常直观的分析基因调控网络的方法，根据基因之间的相关系数，将基因分成不同的模块community,每个模块内做PCA,取PC1的值作为eigengene的值(6)。模块的数量是可调的，简单起见，可以先取2或者3,这也就是潜变量的维度。

为了简便，我们首先设置一个固定component数目，需要比较大，以保证拟合效果。

首先只考虑各个是潜变量的函数



就是各个component的权重



和都可以借助RNA velocity 得到(7-9) 。

那这样我们就可以得到关于eigengene的度规了。

**一些数学推导**

的表达式

只考虑时information velocity的计算可跳过隐变量进行

需满足的自洽条件

**Reference**

1. Li P & Chen S (2016) A review on Gaussian Process Latent Variable Models. *CAAI Transactions on Intelligence Technology* 1(4):366-376.

2. Bac J, Mirkes EM, Gorban AN, Tyukin I, & Zinovyev A (2021) Scikit-Dimension: A Python Package for Intrinsic Dimension Estimation. *Entropy (Basel, Switzerland)* 23(10).

3. Chen B*, et al.* (2022) Automated discovery of fundamental variables hidden in experimental data. *Nature Computational Science* 2(7):433-442.

4. Levina E & Bickel PJ (2004) Maximum Likelihood Estimation of Intrinsic Dimension. *NIPS*.

5. Huang Z-Y, Fan X-Y, Zhou J, & Zhou H-J (2022) Lateral predictive coding revisited: internal model, symmetry breaking, and response time. *Communications in Theoretical Physics* 74(9):095601.

6. Langfelder P & Horvath S (2007) Eigengene networks for studying the relationships between co-expression modules. *BMC Syst Biol.* 1:54. 10.1186/1752-0509-1-54.

7. La Manno G*, et al.* (2018) RNA velocity of single cells. *Nature*.

8. Qiu X*, et al.* (2022) Mapping transcriptomic vector fields of single cells. *Cell* 185(4):690-711.e645.

9. Bergen V, Lange M, Peidli S, Wolf FA, & Theis FJ (2020) Generalizing RNA velocity to transient cell states through dynamical modeling. *Nature Biotechnology* 38(12):1408-1414.