**Advances in Applied Mathematics** 应用数学进展, 2021, 10(2), 617-631
Published Online February 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.102067

Hans汉斯

# 基于网络弹性的早期疾病预警

**马　硕，刘　锐**

华南理工大学数学学院，广东　广州
Email: shea106@qq.com, scliurui@scut.edu.com

收稿日期：2021年1月23日；录用日期：2021年2月17日；发布日期：2021年2月26日

## 摘　要

**疾病的发展大致可以分为三个阶段，正常状态，前疾病状态和疾病状态。前疾病状态可以根据生物网络弹性在不同阶段的动态特性识别。然而，对于网络弹性的度量，目前的研究方法大多基于模型，并且仅适用于低维度的数据，对于高通量的基因数据并不适用。在本文中，我们提出了一个数据驱动的计算基因关联网络弹性的方法，并且使用该方法识别前疾病状态。该方法的有效性已通过在1个模拟数据集和5个真实数据集的应用中得到证实。5个真实数据集包含了小鼠急性肺部损伤的基因微阵列数据集以及4个TCGA数据库的癌症数据集(肺腺癌、胃腺癌、甲状腺癌、结肠癌)。**

## 关键词

**网络弹性，临界点，基因关联网络，交叉熵，偏最小二乘回归(PLS)**

# Early Warning of Diseases Based on Network Resilience

**Shuo Ma, Rui Liu**

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: shea106@qq.com, scliurui@scut.edu.com

Received: Jan. 23rd, 2021; accepted: Feb. 17th, 2021; published: Feb. 26th, 2021

## Abstract

**The progression of diseases can be roughly divided into three stages: normal state, pre-disease state and disease state. The pre-disease state could be identified according to the dynamic characteristics of biological network resilience at different stages. However, for the evaluation of network resilience, the current materials are mostly model-based and only applicable to low-dimensional data, rather than high-throughput genetic data. In this paper, we proposed a data-driven method**

文章引用: 马硕, 刘锐. 基于网络弹性的早期疾病预警[J]. 应用数学进展, 2021, 10(2): 617-631.
DOI: 10.12677/aam.2021.102067

**for evaluating the resilience of gene-related networks, and used this method to identify pre-disease states. The validity of this method was proved by the application of one simulated data set and five real data sets. The five real data sets included gene microarray data sets of acute lung injury in mice and four cancer data sets from TCGA database (lung adenocarcinoma, gastric adenocarcinoma, thyroid cancer, and colon cancer).**

## Keywords

**Network Resilience, Critical Stage, Gene Association Network, Cross Entropy, Partial Least Squares Regression (PLS)**



## 1. 引言

弹性意味着能够适应外部条件的改变，迅速从崩溃中恢复[1]。Ayyub 进一步将系统的弹性定义为“面对干扰等不确定情况时系统保持其功能和性能的持久性”[2]。网络弹性在许多领域都有重要应用。具体到不同网络，系统的弹性也有不同含义。比如，我们常见的社会网络中，弹性是指抵御社会行为变化的能力，它度量了特定人群和动物的社会行为的稳定性；工程网络的弹性(传统意义下的“工程弹性”)衡量的是系统受干扰之后恢复到稳定状态或平衡点的时间；生物网络中的弹性是指机体在面临外界环境参数变化时保持正常运作的能力，它能够起到阻断或延缓疾病进程的作用。

根据动力系统分岔理论，疾病的发展过程可以分为 3 个阶段/状态：a) 正常状态，它是系统远离分叉点的稳定状态，其势能低，弹性高，此时系统抵御外界干扰的能力较强；b) 前疾病状态，可视作相变前的临界点，其势能高，弹性低，此时外界微小的变化就可能使系统迅速进入疾病状态；c) 疾病状态，是系统经过分岔后的另一种稳定状态，此时系统又会回复到势能低，弹性高的情况[3] [4]。其中，正常状态和前疾病状态均为可逆过程，而疾病状态是不可逆的(图 1A)。因此，识别前疾病状态对于防止疾病恶性突变，了解疾病发生与演化至关重要。注意到弹性在正常状态和疾病状态较高，在前疾病状态较低的特点，我们可以利用生物网络弹性来识别前疾病状态。

基因关联网络是一种常见的生物网络[5]，它是以基因为节点，以基因之间的相互关联为边构成的系统。最简单的基因网络生成方法是计算基因数据集的皮尔森相关系数矩阵，如果两基因的相关系数低于某阈值，则认为它们之间没有边连接；反之，这对基因之间有边连接。由此得到的网络被称作“相关性网络”。然而，这样的网络难以区分基因间的直接相互作用和间接相互作用[6]，因为基因间的高相关系数指示的相互作用可能是直接的，也可能是间接的。为了弥补这一缺陷，许多研究材料在 PLS 的基础上计算了基因间的连通性得分，并用连通性得分来衡量基因与基因的直接相互作用强度，进而构建基因关联网络(PLS 网络) [6] [7] [8]。在本文中我们采用了 Gill *et al.*的方法[7]构建基因关联网络。

从高通量的基因数据中重构出基因网络后，我们面临的一个极具挑战性的问题是如何量化网络弹性。在度量网络弹性方面，Wang *et al.* [9]使用系统的恢复能力来度量网络弹性；Ayyub 假定破坏性事件发生服从泊松分布，构建了网络弹性的数学模型[1]；Yalda 以华盛顿地铁为案例构建交通网络[2]，在 Ayyub 提出模型的基础上，通过间接评估网络效率得到了弹性指标；Jianxi Gao *et al.* [10]研究了复杂网络中弹性的普遍特征，结果表明网络拓扑指标中网络密度、非均质性和均匀性与网络弹性密切相关。然而，这些

量化网络弹性的方法大多基于模型，并且仅适用于低维度的数据，对于高通量的基因数据并不适用。因此，找到一种可靠、高效的方法度量基因关联网络弹性的需求十分迫切。

在本文中，我们提出了一个基于交叉熵的计算网络弹性的方法。具体来说，首先要根据基因表达数据和基因间的皮尔森相关系数构造每个基因的分布，接着计算两两基因分布的交叉熵，最后结合先前构造的 PLS 网络，计算网络弹性指标(图 1B)。为了便于探测疾病恶性突变前的临界点，我们将得到的网络弹性指标取逆，得到指标 $U$，这样 $U$ 的峰值将对应着我们预测出的临界点(图 1C)。为验证方法的有效性，

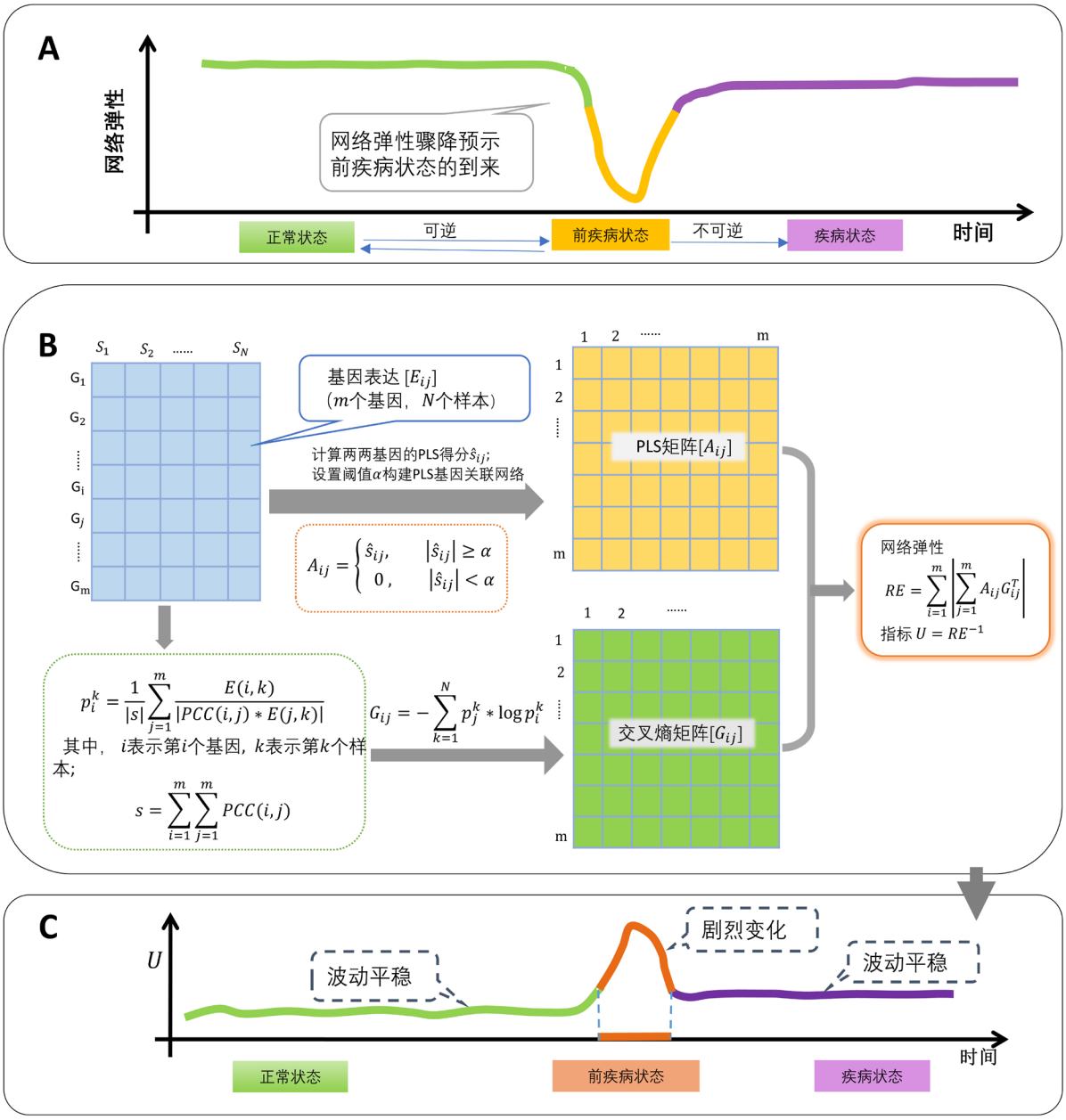



**Figure 1.** (A) The changing trend of network resilience during the progression of complex diseases (B) The algorithm of evaluating network resilience (see more details in Section 3) (C) Property of indicator $U$ during the progression of complex diseases

**图 1.** (A) 复杂疾病发展过程中的网络弹性变化趋势图；(B) 计算网络弹性的算法流程(详见第三节)；(C) 复杂疾病发展过程中指标 $U$ 的特性

我们将其应用于 1 个模拟数据集和 5 个真实数据集中，结果显示，前疾病状态可以被快速、准确地识别出来。总的来说，基于网络弹性的临界点预测方法的计算复杂度小，抗噪性能好，并且，由于它是一种无模型的方法，所以不需要对训练数据进行任何学习，这使得它对高通量基因组数据的适用范围更广。另外，我们提供了一种全新的网络弹性的度量方式，它可以用来描述疾病发展过程中的网络结构差异，为从网络视角理解疾病的发生与演化提供新的思路。

本文的组织结构安排如下：第二节是基于网络弹性的临界点预测在仿真数据集和 4 个真实数据集中的应用；第三节是总结与展望；第四节将详细介绍计算网络弹性的生物学方法。

## 2. 实验结果

### 2.1. 在模拟数据集上的应用

我们先用 16 节点的模拟网络来验证方法的有效性(图 2A)。该网络是一组 16 个基因的调控表示，其中 16 个微分方程代表 16 个基因的调控机制。此模型用 Michaelis-Menten 形式表示。这类调控网络通常被用于研究遗传调控，包括转录和翻译过程[11] [12] [13]，以及多稳定性和非线性生物过程[14] [15]。此外，Michaelis-Menten 形式的分叉经常被用来模拟基因调控网络的时期转移[16] [17]。此模型的参数 $q$ 在 −0.5 到 0.15 之间变化。

我们首先使用上述模拟数据集，通过 Cytoscape 软件(http://www.cytoscape.org/)构建 PLS 网络，图 2B 给出了部分时间点对应的 PLS 网络。从图中可以清楚地看到，正常状态的网络中的边都比较密集，网络结构更稳定；而前疾病状态的网络表现为两个局部网络相接，网络中的边也明显更稀疏，这样的结构稳定性更差。也就是说，前疾病状态与正常状态和疾病状态的 PLS 网络拓扑结构差异显著，因此 PLS 网络邻接矩阵可以成为我们衡量网络弹性的一个有力工具。

在得到每个时间点的 PLS 矩阵后，我们再结合对应时间点的交叉熵矩阵，计算出各时间点的指标 $U$。$U$ 随参数 $q$ 的变化趋势如图 2B 所示。从图中可以看到 $U$ 的峰值出现在 $q = 0$ 时，这表明 $q = 0$ 预示着相变点的来临，也表明我们预测的临界点与实际观测一致。

### 2.2. 在真实数据集上的应用

我们将基于网络弹性的临界点预测方法应用于 5 个真实的数据集。其中包括源于 NCBO 的 GEO 数据库的小鼠肺部急性损伤基因表达(GSE2625)数据集(https://www.ncbi.nlm.nih.gov/)，以及来自 TCGA 数据库的肺腺癌数据集(简称 LUAD)、胃腺癌数据集(简称 STAD)、甲状腺癌数据集(简称 THCA)、结肠癌数据集(简称 COAD) (http://cancergenome.nih.gov)。

选取富含信息的基因是降低数据复杂度、提高信噪比的第一步[18]。因此在应用网络弹性预测疾病发展过程中的临界点之前，我们要对数据集进行预处理，其目的是筛选出在不同实验条件下变化显著的基因。预处理的步骤如下：

a) 2-fold-chage：计算各基因在实验组样本中的标准差和在对照组样本中的标准差，如果前者至少是后者的 2 倍，则将该基因记录下来；

b) t-test：对每个时间点实验组和对照组的样本进行显著性水平为 0.05 的 t 检验，筛选差异程度最高的前 300~400 个基因；

c) 对每个时间点筛选出的基因取并集，形成新的数据集。

#### 2.2.1. 识别小鼠肺部急性损伤的临界点

小鼠肺部急性损伤基因表达数据集(GSE2565)中，每个时间点可以探测到的基因个数为 12,871 个，

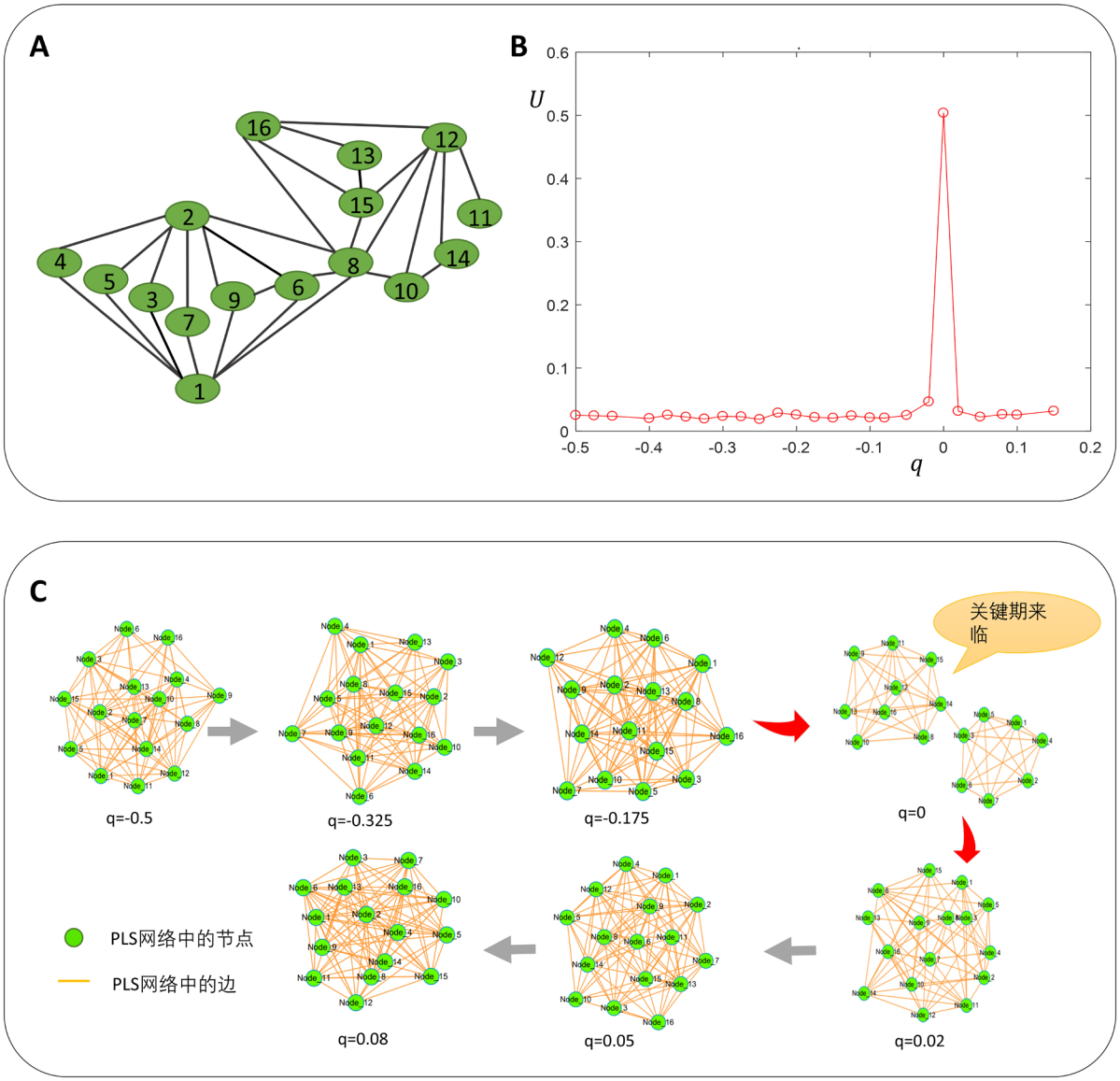



**Figure 2.** (A) 16-nodes gene regulation network based on numerical simulation (B) Curve: the indicator $U$ changing with parameter $q$ (C) 16-nodes PLS network at certain time points
**图 2.** (A)基于数值模拟的 16 节点基因调控网络图(B)指标 $U$ 随参数 $q$ 的变化趋势图(C)部分时间点的 16 节点 PLS 网络

共 9 个不同的采样点：分别是持续暴露于光气环境后的 0 小时，0.5 小时，1 小时，4 小时，8 小时，12 小时，24 小时，48 小时，72 小时。每个时间点设置的实验组和对照组中各有 6 个样本。实验结果表明，实验组的小鼠在吸入光气 8 小时后开始出现不同程度的肺部损伤，一直持续到第 12 小时。在第 12 小时，实验组的小鼠死亡率约为 50%~60%，对照组的小鼠死亡率约为 60%~70% [19]。

### 2.2.2. 识别癌症的临界点

该方法又被用于 4 个来自 TCGA 癌基因图谱的癌症数据集(肺腺癌、胃腺癌、甲状腺癌、结肠癌)。根据 TCGA 的临床资料，以上数据集由癌症样本和癌症近邻样本组成，同时各数据集中的癌症样本大部分被标记了不同的分期，这里，胃腺癌和肺腺癌总共有七个时期，甲状腺癌、结肠癌总共有四个时期。在上述 4 个数据集中，我们将未被标记分期信息的样本删去，把癌症近邻样本都看作参考样本。为了验证所识别的临界时

期，我们对临界转化前和转换临界转化后的样本进行了 Kaplan-Meier (Log-Rank)生存分析进行比较(图 3C、图 3E、图 3G、图 3J)。需要注意的是，临界转化前的样本的预后寿命通常比临界转化后的样本高。

我们识别出的 LUAD 的临界期在 IIIA 期(图 3B)，也就是说，在此之后病人的癌症状况可能会迅速恶化。对 LUAD IIIA 期前后的样本分组进行 Log-Rank 检验，会发现 IIIA 期以前的样本的生存曲线和 IIIA 期之后的生存曲线存在显著差异($p < 0.0001$，图 3C)。并且，临界期之前(IA, IB, IIA, IIB)的生存曲线没有显著差异($p = 0.096$，见附录)，临界期之后的生存曲线差异显著性也不高($p = 0.9$，见附录)。这些结果显示，LUAD 病人在 IIIA 期后面临预后寿命的锐减，也表明 IIIA 期是癌症的一个关键期。同样地，我们可以判断 STAD 的临界期——IIIA 期是一个关键期。

对于甲状腺癌(THCA)，使用我们的方法识别出的癌症迅速恶化前的临界期为 II 期(图 3D)。同时，我们发现，甲状腺癌 II 期之前和 II 期之后的生存曲线是显著不同的($p = 0.0045$，图 3E)。相应地，临界期之前样本的预计存活时间和要比临界期之后样本的预计存活时间长得多。另外，结肠癌(COAD)的临界期被确定在 II 期(图 3F)，并且 COAD 临界期之前和之后的生存曲线存在显著差异($p = 0.00059$，图 3G)。相应地，COAD 中临界期之后的样本生存时间比临界期之后少得多。

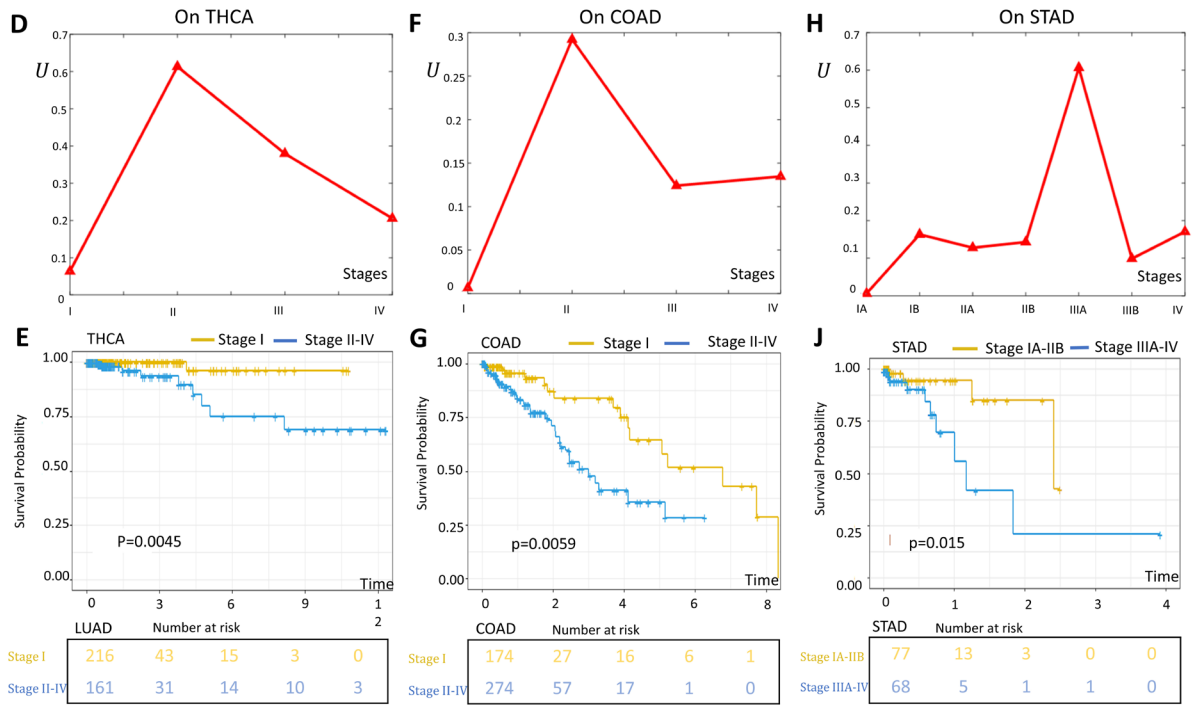
上述结果表明，基于网络弹性的临界点预测方法可以识别出癌症恶化前的临界期，并且这些关键期和预后分析紧密相关。

## 3. 主要方法

### 3.1. 预备知识

#### 3.1.1. 交叉熵

交叉熵的概念源于信息论,它可以衡量观测样本相对于参照样本的混乱程度。假定有两个概率分布：参照样本的分布为 $Q$，$Q=\{Q_i, i=1,2,\cdots,n\}$；观测样本的分布为 $P$，$P=\{P_i, i=1,2,\cdots,n\}$；那么观测样本与参照样本的交叉熵定义为[20]：



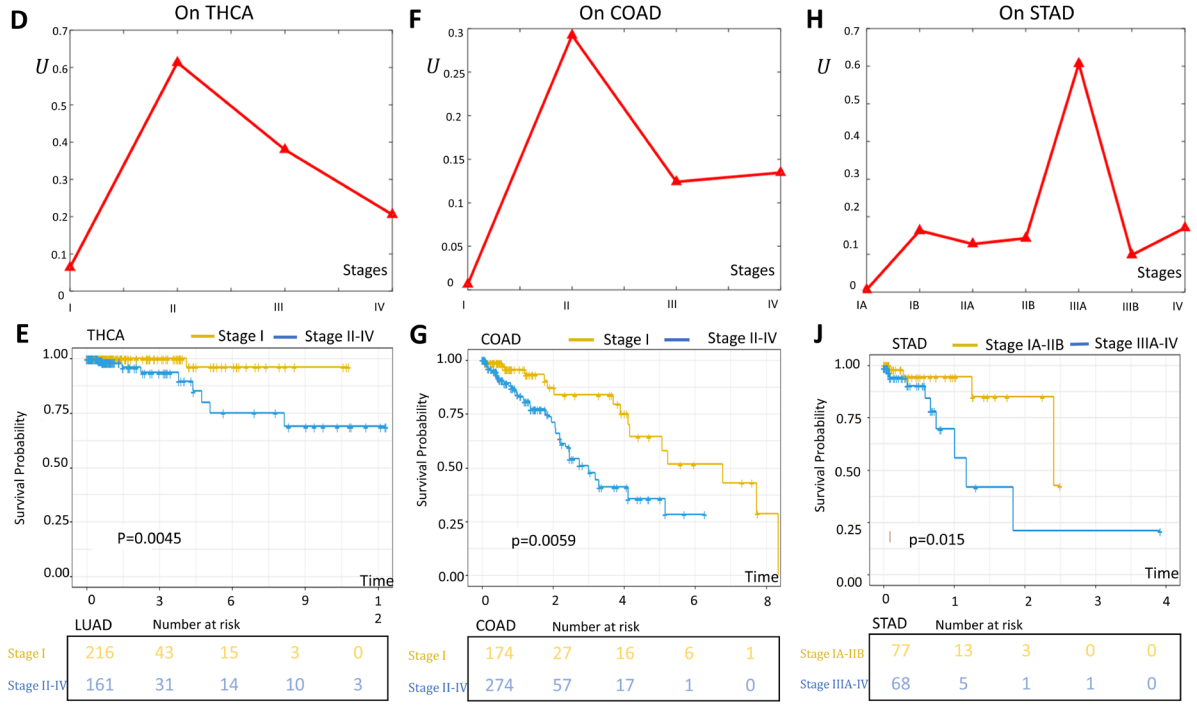



**Figure 3.** (A) Identifying critical stage on acute lung injury in mice (B) Identifying critical stage on LUAD (C) the survival curve before and after critical stage on LUAD (D) Identifying critical stage on THCA (E) the survival curve before and after critical stage on THCA (F) Identifying critical stage on COAD (G) the survival curve before and after critical stage on COAD (H) Identifying critical stage on STAD (J) the survival curve before and after critical stage on STAD

**图 3.** (A)识别小鼠肺部急性损伤的临界点 (B)识别肺腺癌(LUAD)的临界点 (C)肺腺癌临界期前后的生存曲线 (D)识别甲状腺癌(THCA)的临界点 (E)甲状腺癌临界期前后的生存曲线 (F)识别结肠癌(COAD)的临界点 (G)结肠癌临界期前后的生存曲线 (H) 识别胃腺癌(STAD)的临界点 (J) 胃腺癌临界期前后的生存曲线

$$H(P,Q) = -\sum_{i=1}^{n} P_i \times \log Q_i \tag{1}$$

### 3.1.2.PLS (Partial Least Squares Regression)

PLS 的核心思想是用每个基因的表达向量做偏最小二乘的拟合((2)式)。$x_i$ 表示标准化后的基因 $i$ 的表达，$v$ 代表 PLS 的项数，$t_i^r$ 是 $x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_m$ 的线性组合[7]。

$$x_i = \sum_{r=1}^{v} \beta_{ir} t_i^{(r)} + error \tag{2}$$

利用 PLS 计算基因间相互作用强度的算法步骤如下：

步骤 1：令 $l=1$， $X^{(1)} = [x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_m]$

步骤 2：计算 $t_i^{(l)}$

$$t_i^{(l)} = \sum_{k \neq i}^{p} c_{ik}^{(l)} X_k^{(l)}$$

其中，

$$c_{ik}^{l} = X^{(l)^{\mathrm{T}}} x_i \sqrt{x_i^{\mathrm{T}} X^{(l)} X^{(l)^{\mathrm{T}}} x_i}$$

步骤 3 $l = l+1$，迭代计算 $X^l$，如果 $l \ll v$，则返回步骤 2

$$X^l = X^{(l-1)} - t_i^{(l-1)} \left[ t_i^{(l-1)^{\mathrm{T}}} t^{l-1} \right]^{-1} t_i^{(l-1)^{\mathrm{T}}} X^{(l-1)}$$

最后，基因 $i$ 和基因 $k$ 的相互作用强度可以用下面的 PLS 得分来估计：

$$\hat{s}_{ik} = \frac{\sum_{l=1}^{v} \hat{\beta}_{il} c_{ik}^{(l)} + \sum_{l=1}^{v} \hat{\beta}_{kl} c_{ki}^{(l)}}{2}$$

记 $\hat{s}_{ik}$ 为 PLS 网络中基因 $i$ 和基因 $k$ 所连边的权，其中，

$$\hat{\beta}_{il} = t_i^{(l)^{\mathrm{T}}} x_i \Big/ t_i^{(l)^{\mathrm{T}}} t_i^{(l)}$$

## 3.2. 理论基础

Baruch Barzel *et al.*研究了网络动力学中的普遍特征[21]，发现生物网络中节点$i$的扰动对其邻域内节点的影响可以表示为：

$$I(i) = \sum_{j=1}^{m} A_{ij} G_{ij}^{\mathrm{T}} \tag{3}$$

其中 $A_{ij}$ 表示网络的邻接矩阵，$G_{ij}$ 表示基因 $j$ 的扰动对基因 $i$ 的影响程度。$I(i)$ 表示节点 $i$ 对其一阶邻居节点的影响力。

这里我们用基因 $j$ 相对于基因 $i$ 的混乱程度，即 $H(i,j)$ 来代替 $G_{ij}$ ((7)式)。$I(i)$ 较大，意味着节点 $i$ 微小的变动会对其邻域内节点产生很大影响力；那么反过来，节点 $i$ 对它邻域内其他节点的变动不敏感。因此 $I(i)$ 可以看作是节点 $i$ 维持其原有状态的能力。同时考虑到网络弹性由网络中各个节点刻画，我们将基因关联网络的弹性量化为：

$$RE = \sum_{i=1}^{m} \left| I(i) \right| \tag{4}$$

## 3.3. 算法步骤

### 3.3.1. PLS 网络构建

首先对基因表达数据 Z-Score 标准化，设置默认的 PLS 项数 $v=3$，然后根据 3.1.2 节的算法计算 PLS 得分得到基因间边的权。

考虑到基因之间的关联网络结构通常是稀疏的，也就是说，大部分的潜在联系实际上是不存在的[6]，因此我们构建网络的方法是仅保留 PLS 得分较高的边($\left|\hat{s}_{ij}\right| \geq \alpha$)。具体做法是设置阈值参数 $\alpha$，参数 $\alpha$ 的确定方式是使得初始时间点的 PLS 网络中边数约为原始网络结构中的 10% ((5)式)。

$$sign\left(\left|A_{ij}\right|\right) \approx m(m-1) \times 10\% \tag{5}$$

其中 $m$ 为节点/基因个数，$\left[A_{ij}\right]$ 是 PLS 矩阵，它按照下面的分段函数来确定：

$$A_{ij} = \begin{cases} \hat{s}_{ij}, & \left|\hat{s}_{ij}\right| \geq \alpha \\ 0, & \left|\hat{s}_{ij}\right| < \alpha \end{cases} \tag{6}$$

### 3.3.2. 计算两两基因的交叉熵

$\left[E_{ij}\right]$：$m \times N$ 的表达矩阵行为基因，列为样本

$\left[G_{ij}\right]$：$m \times m$ 的交叉熵矩阵

$p_i^k$：样本 $k$ 中基因 $i$ 的概率

$$G_{i,j} = -\sum_{k=1}^{N} p_j^k * \log p_i^k \tag{7}$$

$$p_i^k = \frac{1}{|s|} \sum_{j=1}^{m} \frac{E(i,k)}{\left|PCC(i,j) * E(j,k)\right|} \tag{8}$$

$$s=\sum_{i=1}^{m}\sum_{j=1}^{m}PCC(i,j) \tag{9}$$

$$PCC(i,j)=\frac{\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2\sum_{i=1}^{N}(y_i-\bar{y})^2}} \tag{10}$$

#### 3.3.3. 网络弹性测算

得到 PLS 矩阵和交叉熵矩阵后，按照(3)式和(4)式计算 PLS 网络的弹性。为了便于探测疾病恶性突变前的临界点，我们将得到的网络弹性指标取逆，得到不稳定性指标 $U$，这样 $U$ 的峰值将对应着我们预测出的临界点(图 1C)

$$U=1/RE \tag{11}$$

## 4. 总结与展望

网络弹性是复杂网络的关键特征之一，如何量化网络弹性一直是一个备受关注的问题。许多研究者已经在这方面做过大量工作，然而，他们方法大多基于模型，并且仅适用于低维度的数据，对于高通量的基因数据并不适用。在本文中，我们提出了一种数据驱动的衡量基因关联网络弹性的方法，并根据网络弹性在疾病发展过程中的特性来识别前疾病状态。为验证算法的有效性，我们将算法应用于 1 个模拟数据集和 5 个真实数据集中。结果表明，基于网络弹性的临界点预测方法可以快速、有效地识别复杂疾病恶性突变前的临界状态。

相对于传统的比较基因表达差异区分正常状态与疾病状态的方法，我们的方法旨在于捕捉疾病发展过程中的动态信息，识别前疾病状态。该方法有下面几个优点。首先，它是一种无模型的方法，只依赖于数据驱动，所以它对于高通量基因组数据的适用范围更广。其次，由于我们在构建基因关联网络时只保留 PLS 得分较高的边，所以使用该方法计算简单高效，而且可以减弱其受噪声的影响。最后，我们的方法可以用来描述疾病发展过程中的网络结构差异，为从网络视角理解疾病的发生与演化提供了新的思路。

## 基金项目

本文受广东省基础与应用基础研究基金资助(No. 2019B151502062)资助。

## 参考文献

[1] Ayyub, B.M. (2015) Practical Resilience Metrics for Planning, Design, and Decision Making. *Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, **1**, Article ID: 04015008. https://doi.org/10.1061/AJRUA6.0000826

[2] Saadat, Y., Ayyub, B.M., Zhang, Y.J., Zhang, D.M. and Huang, H.W. (2020) Resilience-Based Strategies for Topology Enhancement and Recovery of Metrorail Transit Networks. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, **6**, Article ID: 4020017. https://doi.org/10.1061/AJRUA6.0001057

[3] Liu, R., Chen, P. and Chen, L.N. (2020) Single-Sample Landscape Entropy Reveals the Imminent Phase Transition during Disease Progression. *Bioinformatics*, **36**, 1522-1532. https://doi.org/10.1093/bioinformatics/btz758

[4] Liu, R., Li, M.Y., Liu, Z.-P., Wu, J.R., Chen, L.N. and Aihara, K. (2012) Identifying Critical Transitions and Their Leading Biomolecular Networks in Complex Diseases. *Scientific Reports*, **2**, 813. https://doi.org/10.1038/srep00813

[5] Liu, C., Ma, Y.F., Zhao, J., Nussinov, R., Zhang, Y.-C., Cheng, F.X. and Zhang, Z.-K. (2020) Computational Network Biology: Data, Models, and Applications. *Physics Reports*, **846**, 1-66. https://doi.org/10.1016/j.physrep.2019.12.004

[6] Tenenhaus, A., Guillemot, V., Gidrol, X. and Frouin, V. (2010) Gene Association Networks from Microarray Data Using a Regularized Estimation of Partial Correlation Based on PLS Regression. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **7**, 251-262. https://doi.org/10.1109/TCBB.2008.87

[7] Gill, R., Datta, S. and Datta, S. (2010) A Statistical Framework for Differential Network Analysis from Microarray

Data. *BMC Bioinformatics*, **11**, 95. https://doi.org/10.1186/1471-2105-11-95

[8] Wang, Y.K., Wu, H. and Yu, T.W. (2017) Differential Gene Network Analysis from Single Cell RNA-seq. *Journal of Genetics and Genomics*, **44**, 331-334. https://doi.org/10.1016/j.jgg.2017.03.001

[9] Wang, J.W., Gao, F. and Ip, W.H. (2010) Measurement of Resilience and Its Application to Enterprise Information Systems. *Enterprise Information Systems*, **4**, 215-223. https://doi.org/10.1080/17517571003754561

[10] Gao, J., Barzel, B. and Barabási, A.L. (2016) Universal Resilience Patterns in Complex Networks. *Nature*, **530**, 307-312. https://doi.org/10.1038/nature16948

[11] Garcia-Ojalvo, J., Elowitz, M.B. and Strogatz, S.H. (2004) Modeling a Synthetic Multicellular Clock: Repressilators Coupled by Quorum Sensing. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 10955-10960. https://doi.org/10.1073/pnas.0307095101

[12] Sherman, M.S. and Cohen, B.A. (2012) Thermodynamic State Ensemble Models of cis-Regulation. *PLOS Computational Biology*, **8**, e1002407. https://doi.org/10.1371/journal.pcbi.1002407

[13] Cantone, I., Marucci, L., Iorio, F., Ricci, M.A., Belcastro, V. and Bansal, M. (2009) A Yeast Synthetic Network for *in Vivo* Assessment of Reverse-Engineering and Modeling Approaches. *Cell*, **137**, 172-181. https://doi.org/10.1016/j.cell.2009.01.055

[14] Chen, Y., Kim, J.K., Hirning, A.J., Josić, K. and Bennett, M.R. (2015) Emergent Genetic Oscillations in a Synthetic Microbial Consortium. *Science*, **349**, 986-989. https://doi.org/10.1126/science.aaa3794

[15] Li, C., Chen, L. and Aihara, K. (2006) Stability of Genetic Networks with SUM Regulatory Logic: Lur'e System and LMI Approach. *IEEE Transactions on Circuits and Systems I*: *Regular Papers*, **53**, 2451-2458. https://doi.org/10.1109/TCSI.2006.883882

[16] Gardner, T.S., Cantor, C.R. and Collins, J.J. (2000) Construction of a Genetic Toggle Switch in *Escherichia coli*. *Nature*, **403**, 339-342. https://doi.org/10.1038/35002131

[17] O'Brien, E.L., Van Itallie, E. and Bennett, M.R. (2012) Modeling Synthetic Gene Oscillators. *Mathematical Biosciences*, **236**, 1-15. https://doi.org/10.1016/j.mbs.2012.01.001

[18] 万江. 基于 SOM 基因聚类的基因数据组织样本聚类[D]: [硕士学位论文]. 西安: 西安电子科技大学, 2005.

[19] Sciuto, A.M., Phillips, C.S., Orzolek, L.D., Hege, A.I., Moran, T.S. and Dillman, J.F. (2005) Genomic Analysis of Murine Pulmonary Tissue Following Carbonyl Chloride Inhalation. *Chemical Research in Toxicology*, **18**, 1654-1660. https://doi.org/10.1021/tx050126f

[20] Cofta, P., Ledziński, D., Śmigiel, S. and Gackowska, M. (2020) Cross-Entropy as a Metric for the Robustness of Drone Swarms. *Entropy* (*Basel*, *Switzerland*), **22**, 597. https://doi.org/10.3390/e22060597

[21] Barzel, B. and Barabási, A.-L. (2013) Universality in Network Dynamics. *Nature Physics*, **9**, 673-681. https://doi.org/10.1038/nphys2741

## 附　录：

**1. 数值模拟方法**

首先考虑用下式表示的离散的时变动力系统：

$$Z(t+1)=f(Z(t);P) \tag{S1}$$

这里 $Z(t)=(z_1(t),\cdots,z_n(t))$ 是时刻 $k$ 表示系统特征的 $n$ 维的向量，$P(t)=(p_1,\cdots,p_s)$ 是控制系统参数缓慢变化的参数向量。 $f:R^n\times R^s\to R^n$，该映射通常是非线性的。如果下面三个条件同时成立，那么带有参数$P$的动力系统发展方程(S1)存在分叉点或关键期[3]：

1、$\bar{Z}$ 是系统的不动点，即满足 $\bar{Z}=f(\bar{Z};P)$；

2、存在 $P_c$，使得一个或一对雅可比矩阵 $J=\left.\dfrac{\partial f(Z;P_c)}{\partial Z}\right|_{Z=\bar{Z}}$ 的特征值等于 1；

3、当 $P\neq P_c$ 时，$f$ 的线性化系统的特征值不恒等于 1。

以上 3 个条件说明系统在 $\bar{Z}$ 经历了相变，或者说当 $P$ 接近阈值 $P_c$ 时系统出现分叉。当方程(S1)接近 $\bar{Z}$，$P$ 值到达 $P_c$ 前，我们假定系统处在稳定点，从而线性化系统的特征值范围都大于 0 小于 1。$P$ 取值为 $P_c$ 时，系统相变产生，这也意味着 $P_c$ 为系统的分叉参数。

Michaelis-Menten 形式调控网络通常被用于研究遗传调控，包括转录和翻译过程[11] [12] [13]，以及多稳定性和非线性生物过程[14] [15]。此外，Michaelis-Menten 形式的分叉经常被用来模拟基因调控网络的时期转移[16] [17]。本文中用到的模拟网络时用 Michaelis-Menten 形式表示的，它是一组 16 个基因的调控表示，网络中的节点表示生物分子，边表示生物分子之间正或负的调控关系。此模型的参数 $q$ 在-0.5 到 0.15 之间变化。下面 16 个微分方程代表 16 个基因的调控机制：

$$\frac{\mathrm{d}z_1(t)}{\mathrm{d}t}=\frac{(8-4q)z_2(t)}{15(1+z_2(t))}-4\left(\frac{1+q}{15}\right)+\zeta_1(t)$$

$$\frac{\mathrm{d}z_2(t)}{\mathrm{d}t}=\frac{(4-2q)z_1(t)}{15(1+z_1(t))}-2\left(\frac{4+q}{15}\right)z_2(t)+\zeta_2(t)$$

$$\frac{\mathrm{d}z_3(t)}{\mathrm{d}t}=\frac{4q-10}{15}+\frac{5-2q}{15(1+z_1(t))}+\frac{5-2q}{15(1+z_2(t))}-z_3(t)+\zeta_3(t)$$

$$\frac{\mathrm{d}z_4(t)}{\mathrm{d}t}=\frac{(6-2q)z_1(t)}{15(1+z_1(t))}+\frac{(6-2q)z_2(t)}{15(1+z_2(t))}-\frac{6}{5}z_4(t)+\zeta_4(t)$$

$$\frac{\mathrm{d}z_5(t)}{\mathrm{d}t}=\frac{4q-14}{15}+\frac{7-2q}{15(1+z_1(t))}+\frac{7-2q}{15(1+z_2(t))}-\frac{7}{5}z_5(t)+\zeta_5(t)$$

$$\frac{\mathrm{d}z_6(t)}{\mathrm{d}t}=\frac{4q-16}{15}+\frac{2(4-q)}{15(1+z_1(t))}+\frac{2(4-q)}{15(1+z_2(t))}-\frac{8}{5}z_6(t)+\zeta_6(t)$$

$$\frac{\mathrm{d}z_7(t)}{\mathrm{d}t}=\frac{(9-2q)z_1(t)}{15(1+z_1(t))}+\frac{(9-2q)z_2(t)}{15(1+z_2(t))}-\frac{9}{5}z_7(t)+\zeta_7(t)$$

$$\frac{\mathrm{d}z_8(t)}{\mathrm{d}t}=-\frac{13}{15}+\frac{2}{15(1+z_1(t))}+\frac{2}{15(1+z_2(t))}+\frac{2}{5(1+z_6(t))}+\frac{2z_{10}(t)}{5(1+z_{10}(t))}$$
$$+\frac{3z_{12}(t)}{5(1+z_{12}(t))}+\frac{z_{15}(t)}{5(1+z_{15}(t))}+\frac{1}{5(1+z_{16}(t))}-2z_8(t)+\zeta_8(t)$$

$$\frac{\mathrm{d}z_9(t)}{\mathrm{d}t}=-1+\frac{1}{5\left(1+z_1(t)\right)}+\frac{1}{5\left(1+z_2(t)\right)}+\frac{3}{5\left(1+z_6(t)\right)}-\frac{11}{5}z_9(t)+\zeta_9(t)$$

$$\frac{\mathrm{d}z_{10}(t)}{\mathrm{d}t}=\frac{3z_{12}(t)}{5\left(1+z_{12}(t)\right)}-\frac{12}{5}z_{10}(t)+\zeta_{10}(t)$$

$$\frac{\mathrm{d}z_{11}(t)}{\mathrm{d}t}=\frac{z_{12}(t)}{4\left(1+z_{12}(t)\right)}-\frac{13}{5}z_{11}(t)+\zeta_{11}(t)$$

$$\frac{\mathrm{d}z_{12}(t)}{\mathrm{d}t}=-\frac{2}{15}+\frac{2z_{15}(t)}{5\left(1+z_{15}(t)\right)}+\frac{2}{5\left(1+z_{16}(t)\right)}-\frac{14}{5}z_{12}(t)+\zeta_{12}(t)$$

$$\frac{\mathrm{d}z_{13}(t)}{\mathrm{d}t}=-\frac{24}{5}+\frac{1}{1+z_{15}(t)}+\frac{19}{5\left(1+z_{16}(t)\right)}-5z_{13}(t)+\zeta_{13}(t)$$

$$\frac{\mathrm{d}z_{14}(t)}{\mathrm{d}t}=-\frac{8}{5}+\frac{4}{5\left(1+z_{10}(t)\right)}+\frac{4}{5\left(1+z_{12}(t)\right)}-\frac{16}{5}z_{14}(t)+\zeta_{14}(t)$$

$$\frac{\mathrm{d}z_{15}(t)}{\mathrm{d}t}=\frac{z_{16}(t)}{10\left(1+z_{16}(t)\right)}-\frac{7}{2}z_{15}(t)+\zeta_{15}(t)$$

$$\frac{\mathrm{d}z_{16}(t)}{\mathrm{d}t}=\frac{z_{15}(t)}{10\left(1+z_{16}(t)\right)}-\frac{7}{2}z_{16}(t)+\zeta_{16}(t)$$

这里，$q$ 是控制参数，$\zeta_i(t)$ 是 0 均值的高斯噪声，$z_i(t)(i=1,2,\cdots,16)$ 是 mRNA-$i$ 的浓度，mRNA 的降解速率是：$\left(-4\frac{1+q}{15},-2\frac{4+q}{15},-1,-\frac{6}{5},-\frac{7}{5},-\frac{8}{5},-2,-\frac{11}{5},-\frac{12}{5},-\frac{13}{5},-\frac{14}{5},-5,-\frac{16}{5},-\frac{7}{2},-\frac{7}{2}\right)$。系统的稳定平衡点是 $\bar{Z}=(\bar{z}_1,\bar{z}_2,\cdots,\bar{z}_{16})=(0,0,\cdots,0)$。使用 Euler 函数，可以把微分方程组转化为 $Z(k+1)=f\left(Z(k),q\right)$ 的形式：

$$z_1(k+1)=z_1(k)+\left[\frac{(8-4q)z_2(t)}{15\left(1+z_2(t)\right)}-4\left(\frac{1+q}{15}\right)+\zeta_1(t)\right]\Delta t$$

$$z_2(k+1)=z_2(k)+\left[\frac{(4-2q)z_1(t)}{15\left(1+z_1(t)\right)}-2\left(\frac{4+q}{15}\right)z_2(t)+\zeta_2(t)\right]\Delta t$$

$$z_3(k+1)=z_3(k)+\left[\frac{4q-10}{15}+\frac{5-2q}{15\left(1+z_1(t)\right)}+\frac{5-2q}{15\left(1+z_2(t)\right)}-z_3(t)+\zeta_3(t)\right]\Delta t$$

$$z_4(k+1)=z_4(k)+\left[\frac{(6-2q)z_1(t)}{15\left(1+z_1(t)\right)}+\frac{(6-2q)z_2(t)}{15\left(1+z_2(t)\right)}-\frac{6}{5}z_4(t)+\zeta_4(t)\right]\Delta t$$

$$z_5(k+1)=z_5(k)+\left[\frac{4q-14}{15}+\frac{7-2q}{15\left(1+z_1(t)\right)}+\frac{7-2q}{15\left(1+z_2(t)\right)}-\frac{7}{5}z_5(t)+\zeta_5(t)\right]\Delta t$$

$$z_6(k+1)=z_6(k)+\left[\frac{4q-16}{15}+\frac{2(4-q)}{15\left(1+z_1(t)\right)}+\frac{2(4-q)}{15\left(1+z_2(t)\right)}-\frac{8}{5}z_6(t)+\zeta_6(t)\right]\Delta t$$

$$z_7(k+1)=z_7(k)+\left[\frac{(9-2q)z_1(t)}{15(1+z_1(t))}+\frac{(9-2q)z_2(t)}{15(1+z_2(t))}-\frac{9}{5}z_7(t)+\zeta_7(t)\right]\Delta t$$

$$z_8(k+1)=z_8(k)+\left[-\frac{13}{15}+\frac{2}{15(1+z_1(t))}+\frac{2}{15(1+z_2(t))}+\frac{2}{5(1+z_6(t))}+\frac{2z_{10}(t)}{5(1+z_{10}(t))}+\frac{3z_{12}(t)}{5(1+z_{12}(t))}+\frac{z_{15}(t)}{5(1+z_{15}(t))}+\frac{1}{5(1+z_{16}(t))}-2z_8(t)+\zeta_8(t)\right]\Delta t$$

$$z_9(k+1)=z_9(k)+\left[-1+\frac{1}{5(1+z_1(t))}+\frac{1}{5(1+z_2(t))}+\frac{3}{5(1+z_6(t))}-\frac{11}{5}z_9(t)+\zeta_9(t)\right]\Delta t$$

$$z_{10}(k+1)=z_{10}(k)+\left[\frac{3z_{12}(t)}{5(1+z_{12}(t))}-\frac{12}{5}z_{10}(t)+\zeta_{10}(t)\right]\Delta t$$

$$z_{11}(k+1)=z_{11}(k)+\left[\frac{z_{12}(t)}{4(1+z_{12}(t))}-\frac{13}{5}z_{11}(t)+\zeta_{11}(t)\right]\Delta t$$

$$z_{12}(k+1)=z_{12}(k)+\left[-\frac{2}{15}+\frac{2z_{15}(t)}{5(1+z_{15}(t))}+\frac{2}{5(1+z_{16}(t))}-\frac{14}{5}z_{12}(t)+\zeta_{12}(t)\right]\Delta t$$

$$z_{13}(k+1)=z_{13}(k)+\left[-\frac{24}{5}+\frac{1}{1+z_{15}(t)}+\frac{19}{5(1+z_{16}(t))}-5z_{13}(t)+\zeta_{13}(t)\right]\Delta t$$

$$z_{14}(k+1)=z_{14}(k)+\left[-\frac{8}{5}+\frac{4}{5(1+z_{10}(t))}+\frac{4}{5(1+z_{12}(t))}-\frac{16}{5}z_{14}(t)+\zeta_{14}(t)\right]\Delta t$$

$$z_{15}(k+1)=z_{15}(k)+\left[\frac{z_{15}(t)}{10(1+z_{16}(t))}-\frac{7}{2}z_{16}(t)+\zeta_{16}(t)\right]\Delta t$$

$$z_{16}(k+1)=z_{16}(k)+\left[\frac{z_{15}(t)}{10(1+z_{16}(t))}-\frac{7}{2}z_{16}(t)+\zeta_{16}(t)\right]\Delta t$$

其中，$\Delta t$ 是时间间隔，$Z(k)$ 是 $Z(t)$ 在时刻 $k$ 的向量。微分方程组的雅可比矩阵 $J=\left.\frac{\partial f(Z;q)}{\partial Z}\right|_{Z=\bar{Z}}$，用 $J=\mathrm{e}^{\Delta tA}$ 表示如下：

$$A=\begin{bmatrix}
\frac{-4-4q}{15} & \frac{8-4q}{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{4-2q}{15} & \frac{-8-2q}{15} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{-5+2q}{15} & \frac{-5+2q}{15} & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{6-2q}{15} & \frac{6-2q}{15} & 0 & -\frac{6}{5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{-7+2q}{15} & \frac{-7+2q}{15} & 0 & 0 & -\frac{7}{5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{-8+2q}{15} & \frac{-8+2q}{15} & 0 & 0 & 0 & -\frac{8}{5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
\frac{9-2q}{15} & \frac{9-2q}{15} & 0 & 0 & 0 & 0 & -\frac{9}{5} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
-\frac{2}{15} & -\frac{2}{15} & 0 & 0 & 0 & -\frac{2}{5} & 0 & -2 & 0 & \frac{2}{5} & 0 & \frac{3}{5} & 0 & \frac{1}{5} & 0 & -\frac{1}{5}\\
-\frac{1}{5} & -\frac{1}{5} & 0 & 0 & 0 & -\frac{3}{5} & 0 & 0 & -\frac{11}{5} & 0 & 0 & 0 & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{12}{5} & 0 & \frac{3}{5} & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{13}{5} & \frac{1}{4} & 0 & 0 & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{14}{5} & 0 & 0 & \frac{2}{5} & -\frac{2}{5}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -5 & 0 & -1 & \frac{19}{5}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{4}{5} & 0 & 0 & -\frac{4}{5} & 0 & -\frac{16}{5} & 0 & 0\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & -\frac{7}{2} & \frac{1}{10}\\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{10} & -\frac{7}{2}
\end{bmatrix}$$

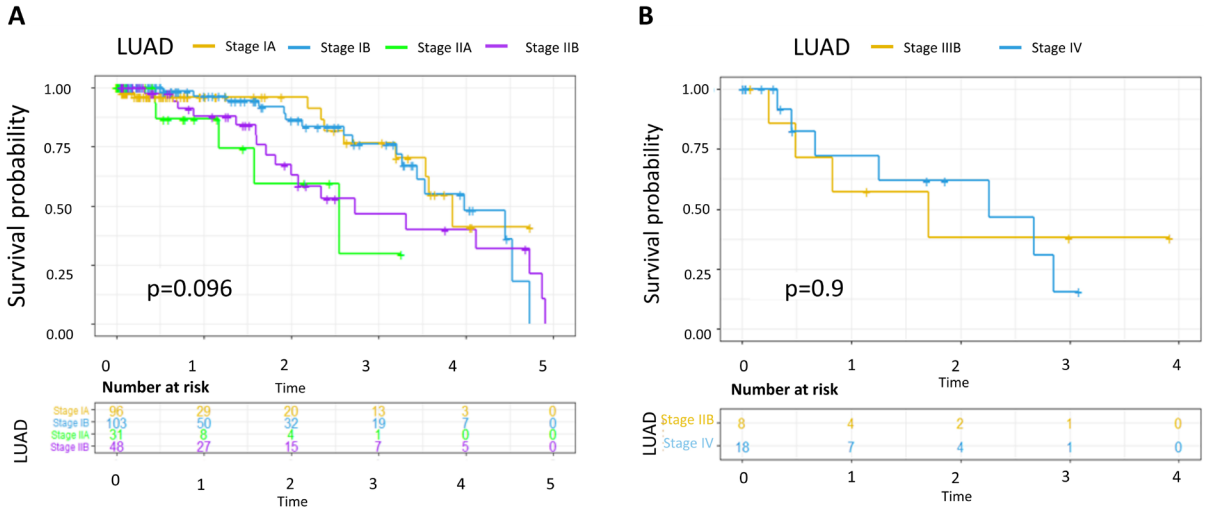
取 $\Delta t=1$，从上式中解得 16 个不同得特征值：

$$\left(0.69^{q},0.45,0.37,0.30,0.25,0.20,0.17,0.14,0.11,0.09,0.07,0.06,0.05,0.04,0.033,0.027\right)$$

显然，当 $q$ 趋于 0 时，$0.69^{q}$ 趋于 1。也就是说，$q_c=0$ 是系统开始变得不稳定的相变点，它是一个关键值点。控制参数 $q$ 的取值范围是$[-0.5,0.15]$。在每次数值模拟时，我们改变参数 $q$ 的大小，把对应生成的 10 个样本作为参照样本，然后再根据每次模拟出的样本表达计算 $U$ 的值。我们共进行了 100 次的数值模拟实验，将每次实验中得到的指标 $U$ 取平均值，作为最终得分。最后，绘制出 $U$ 和 $q$ 的变化趋势图。$U$ 的峰值对应的 $q$ 取值为 0，与系统的相变点 $q_c$ 重合，说明我们预测的临界点与实际观测一致。

**2.LUAD 临界期前后的生存曲线**

我们对 LUAD IIIA 期前后的样本分别进行 Log-Rank 检验。$p$ 可以用来衡量生存曲线的差异程度，$p$ 值越接近 0，表明差异程度越大。图 A 为肺腺癌(LUAD)临界期前的生存曲线，其中黄色，蓝色、绿色和紫色的曲线分别表示肺腺癌患者 IA 期，IB 期、IIA 期和 IIB 期的生存曲线。从图中可以看出，临界期之前(IA, IB, IIA, IIB)的生存曲线并没有显著差异($p$ = 0.096)。图 B 为肺腺癌(LUAD)临界期后的生存曲线，黄色曲线为肺腺癌患者 IIIB 期的生存曲线，蓝色曲线为肺腺癌患者 IV 期的生存曲线。生存分析结果表明，$p$ = 0.9，即肺腺癌患者 IIIB 期和 IV 的生存曲线差异程度较小。


A
LUAD
Stage IA
Stage IB
Stage IIA
Stage IIB
Survival probability
1.00
0.75
0.50
0.25
0.00
p=0.096
0
1
2
3
4
5
Number at risk
Time
LUAD
Stage IA 96 29 20 13 3 0
Stage IB 103 50 32 19 7 0
Stage IIA 31 8 4 1 0 0
Stage IIB 48 27 15 7 5 0
0
1
2
3
4
5
Time
B
LUAD
Stage IIIB
Stage IV
Survival probability
1.00
0.75
0.50
0.25
0.00
p=0.9
0
1
2
3
4
Number at risk
Time
LUAD
Stage IIB 8 4 2 1 0
Stage IV 18 7 4 1 0
0
1
2
3
4
Time