Hans Journal of Computational Biology 计算生物学, 2018, 8(4), 59-69
Published Online December 2018 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2018.84008

Hans汉斯

Detecting Critical Point of Complex Disease Network by Constructing Individual-Specific Anomaly Index

Yulin Huang, Quandi Wang

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: 418691809@qq.com, qdwang@scut.edu.cn

Received: Dec. 21st, 2018; accepted: Jan. 3rd, 2019; published: Jan. 10th, 2019

Abstract

Detecting critical points of complex diseases is very important for early diagnosis of diseases. By exploring information of high-throughput data, we combine individual-specific network and hidden Markov model to construct individual-specific abnormal indicators in order to detect the critical points between relative health period and disease critical period. To verify the validity of the method, it was applied to simulated data sets, lung acute injury data and prostate cancer data. The critical points were successfully found before malignant mutation. The validity and sensitivity of signal genes were verified by biological function analysis.

Keywords

Complex Disease Network, Critical Point, Hidden Markov Model, Individual Specific Network

构建个体特异性异常指标探测复杂疾病网络临界点

黄煜林，王全迪

华南理工大学，数学学院，广东 广州
Email: 418691809@qq.com, qdwang@scut.edu.cn

收稿日期：2018年12月21日；录用日期：2019年1月3日；发布日期：2019年1月10日

摘　要

探测复杂疾病临界点对疾病早期诊断至关重要，通过对高通量的生物分子数据的挖掘，本文提出一种结

文章引用: 黄煜林, 王全迪. 构建个体特异性异常指标探测复杂疾病网络临界点[J]. 计算生物学, 2018, 8(4): 59-69.
DOI: 10.12677/hjcb.2018.84008

合个体特异性网络与隐马尔科夫模型的方法，构建个体特异性异常指标，以探测从相对健康期到疾病临界期的临界点。为验证该方法的有效性，将该方法分别应用在模拟数据集、肺部急性损伤数据、前列腺癌数据中，均成功在疾病恶性突变前找到其各自的临界点。信号基因的有效性和敏感性都通过生物功能分析得到了验证。

关键词

复杂疾病网络，临界点，隐马尔科夫模型，个体特异性网络

1. 引言

研究临界点的方法，主要是通过非线性规划、构造特征变量等手段与方法。在一部分求解临界点的问题中，将临界点问题转换为计算有约束条件的非线性规划的最优解，并利用牛顿法、罚函数法等方法求出临界点。而另外一部分求解临界点的问题中，则是通过研究系统本身的性质，构造其独特的特征统计量，以其统计量的突破以判定其为临界点。

研究表明[1] [2] [3] [4]，许多复杂疾病存在临界点。在疾病发生前疾病可能会潜伏数年，但发生异变却只仅仅在疾病爆发前几个月时间内。换言之，在疾病爆发前存在一个疾病过渡的时期，在该时期中，生物系统从健康状态过渡到疾病发生状态。而疾病过渡期就是我们所需要探寻的复杂疾病临界点。探测复杂疾病临界点至关重要，一方面对疾病的早期治疗有极大的帮助[5]，另外一方面，通过针对临界点时的研究也进一步促进对疾病发生机理的了解[6] [7]。

疾病的发展过程一般可以分为三个过程：1) 健康状态。在该状态下系统对扰动具有较强的鲁棒性与回复能力。2) 疾病过渡期。该状态是发生相变之前的临界状态，该状态下系统表现为对扰动非常敏感、回复能力弱。3) 疾病爆发期。在该状态下系统表现为对扰动又具有强鲁棒性与回复能力。现有研究表明，对比健康状态与疾病过渡期的患者，尽管在一些静态特征上并无显著性差异，但一些动态特征则存在显著性差异，具体为：1) 影响疾病动态网络标志物基因之间相关性提高。2) 动态网络标志物基因与其他基因之间相关性减少。3) 动态网络标志物基因的标准差大幅增加[2]。

本文结合差异化网络的方法与隐马尔科夫模型，以疾病发展过程的三个过程作为隐马尔科夫模型的隐藏状态，以构造个体差异化网络后与参考基因网络的皮尔逊相关系数差值作为观测序列，将之前所有时刻数据为训练样本，训练当前模型，再根据当前模型计算得到下一时刻进入疾病过渡期的概率，并以此作为探测临界点的基本指标。通过结合构建个体差异化网络的方法后，探测临界点时对所需要的实验样本数量更少。

为了检验方法有效性，将该方法用到九个节点的模拟数据集、肺部急性损伤数据(GSE2565)、前列腺癌数据(GSE5345)当中，均成功在其疾病爆发期之前探测到其临界点，动态网络标志物基因在临界点转为高表达，在其临界点前后基因表达发生较大的翻转。另外，对指标产生较明显影响的基因分别进行功能分析与生存，通过功能分析可以得到，这些基因均来自一些对疾病有直接关系的通路，同时这些有明显影响的基因通过生存分析看到对前列腺患者的存活率有显著影响，上述结果均说明方法稳定可靠。

2. 方法

2.1. 单样本个性化网络

传统探测临界点方法均利用多个实验样本以提供多组实验基因，继而能够计算实验组中基因与基因之间相关性，探测的是一个普遍性临界点。而 2016 年有研究提出了构建基于个人特异性的网络方法[8]。其个人特异性网络的构建方法(图 1(A))：

取出 N 个正常人的样本，不妨假设其中与探测疾病可能相关的基因对共有 m 对，计算每对基因之间的皮尔逊相关系数(Pearson Correlation Coefficient，下文均用 *PCC* 代替)，得到 m 个 *PCC* 值，不妨按一定顺序命名为 $PCC\_ref_i, i=1,2,\cdots,m$ 。

将特定个人(实验样本)的基因组加入 N 个正常人样本中，即共 $N+1$ 个样本，这时重新计算每两个基因之间的 PCC，不妨记为 $PCC\_ind_i, i=1,2,\cdots,m$ 。计算 $\Delta PCC_i = PCC\_ref_i - PCC\_ind_i$ 。

每一对基因之间的均可计算Δ*PCC*，这就组成了单样本的个性化网络，并且当 n 较大时，Δ*PCC* 符合以下分布：

$$\mu_{\Delta PCC_i} = 0,\ \sigma_{\Delta PCC_i} = \frac{1}{N-1}\left(1-PCC_i^2\right) \tag{1}$$

通过 z 检验之后，得到 z 值：

$$z_i = \frac{\Delta PCC_i}{\left(1-PCC_i\right)/\left(n-1\right)} \tag{2}$$

那么此时 z_i 符合正态分布。

但实际上生物数据中通常样本量极少，所以 z 并非正态分布，而Δ*PCC* 是服从火山分布[8]，而该种分布只知道与 n 相关，无法写出表达式。

2.2. 隐马尔科夫模型

其正常无差异的基因对的Δ*PCC* 应服从“火山分布”。“火山分布”在 n 较大时，根据文章中的测试：$n > 50$ 能够通过 Kolmogorov-Smirnov 测试，以证明火山分布与标准正态分布是相似的。但其在 n 较小时，并无一个表达式或是一个相似的分布，而只能确定其与 n 相关。而在 n 较小时，利用 z-test 或是 t-test 进行显著性检验，显然也是不合理的。

但在确定 n 之后，在同一的状态下，其Δ*PCC* 是符合一定的分布。但当状态发生改变时，根据复杂疾病临界点的理论，在同一组中的基因对的 *PCC* 变为趋于 1，为组内基因与组外基因的 *PCC* 则变为趋于 0。换一句话说，当状态从疾病发生前转变到疾病过渡期时，与疾病相关的基因对的 *PCC* 值与正常状态下的 *PCC* 值发生显著性变化，其Δ*PCC* 不再服从上一时刻中的分布。故利用隐马尔科夫模型，即可通过训练得到疾病发生前的Δ*PCC* 的分布，进而计算下一个时刻已知观测情况下，其为疾病发生前、疾病过渡期的概率。

我们将疾病发生前、疾病过渡期、疾病爆发期作为隐马尔科夫模型中无法直接观测到的三个状态，但由于我们研究的仅仅是探测疾病过渡期，故实际仅考虑疾病发生前、疾病过渡期，隐马尔科夫模型变量设定如表 1。

Table 1. Variable setting of hidden Markov model
表 1. 隐马尔科夫模型变量设定

	变量意义
w_1, w_2, w_3	隐藏状态，分别代表疾病发生前、疾病过渡期、疾病爆发期。
$a_{ij}, i=1,2,3; j=1,2,3$	状态转移矩阵，代表从状态 j 到状态 i 的概率。

Continued

$b_{jk}(t), j=1,2,3,4,5; k=1,2,3$	由于隐马尔科夫模型中仅能使用离散变量作为观测序列。这里将ΔPCC均匀切分为等长的5份，依次标序1，2，3，4，5。这里$b_{jk}(t)$代表状态k下观测到j的概率。
$\theta_t=(A,B,\pi)$	t时刻及其以前所有训练所得模型。
$O_t=\{o_1,o_2,\cdots,o_t\}$	t时刻及其以前的所有观测序列。
o_t	t时刻的观测序列，$o_t=\{\Delta pcc_1,\Delta pcc_2,\cdots,\Delta pcc_l\}$，$l$为基因对对数。

那么，当我们探测临界点T时，T满足以下条件：(1) $T-1$时刻为状态w_1，(2) T时刻进入到状态w_2。即求得T，使得$t=T$时，公式(3)达到最大

$$P(t)=P_t(s_t=w_2\mid s_{t-1}=w_1,s_{t-2}=w_1,\cdots,s_1=w_1,\theta_{t-1},O) \tag{3}$$

研究中，由于不考虑状态w_3的情况，并且三种状态是有序发生的，所以有

$$I(t)=1-P_t(s_t=w_1\mid s_{t-1}=w_1,s_{t-2}=w_1,\cdots,s_1=w_1,\theta_{t-1},O) \tag{4}$$

由于t时刻的状态跳转仅与$t-1$时刻状态有关，所以(4)可以写为：

$$\begin{aligned} I(t)&=1-P_t(s_t=w_1\mid s_{t-1}=w_1,\theta_{t-1},O)\\ &=1-\frac{P_t(s_t=w_1,s_{t-1}=w_1\mid\theta_{t-1},O)}{P_t(s_{t-1}=w_1\mid\theta_{t-1},O)} \end{aligned} \tag{5}$$

并以此判定t时刻是否是临界点T，记$I(t)$为个体特异性异常指标，当指标越大，则说明此时样本发生状态跳转的可能性越大，反之则为样本发生状态跳转的可能性越小。结合复杂疾病临界点理论，则该指标在疾病过渡期会发生突增。

2.3. 算法设计

步骤一：取$N_{\text{reference}}$个正常人样本，计算与疾病相关的所有基因对之间的PCC，记为PCC_i^*，$i=1,2,\cdots,m$，其中m是相关基因对数量。

步骤二：记实验样本总数为N_{case}，对每一个实验样本分别进行以下操作：结合$N_{\text{reference}}$个正常人样本，形成共$N_{\text{reference}}+1$个样本，对时间$T-1$以前的所有基因对，求其PCC，并计算$\left|\Delta pcc_i^j(t)\right|$，其中$j=1,2,\cdots,N_{\text{case}}$。记对照组样本数共$N_{\text{control}}$，对所有对照样本做相同的处理，得$\left|\Delta pcc_i^j(t)\right|$，$j=1,2,\cdots,N_{\text{control}}$，并对各个$\left|\Delta pcc\right|$进行分组。

步骤三：以$T-1$时刻以前的这些基因对(不分对照组或实验组)作为训练样本，以其$o_j^t=\left\{\Delta pcc_1^j(t),\Delta pcc_2^j(t),\cdots,\Delta pcc_m^j(t)\right\}$，$j=1,2,\cdots,N_{\text{case}}+N_{\text{control}}$作为观测序列，并认为这些样本均处于$w_1$状态。得到$T-1$时刻的隐马尔科夫模型$\theta_{T-1}$(图1(B))。

步骤四：对第j个实验样本，$j=1,2,\cdots,N_{\text{case}}$，分别进行以下操作：用训练得到的$\theta_{T-1}$，计算实验组在$T$时刻观测序列为$o_j^T$不一致性指标：

$$I_j(T)=1-\frac{P_t(s_t=w_1,s_{t-1}=w_1\mid\theta_{t-1},o_j^T)}{P_t(s_{t-1}=w_1\mid\theta_{t-1},o_j^T)} \tag{6}$$

另一方面，对对照组实验样本j_2，$j_2=1,2,\cdots,N_{\text{control}}$，计算对照组在$T$时刻观测序列$o_j^T$下的不一致性指标：

$$J_j(T)=1-\frac{P_t(s_t=w_1,s_{t-1}=w_1\mid\theta_{t-1},o_j^T)}{P_t(s_{t-1}=w_1\mid\theta_{t-1},o_j^T)} \tag{7}$$

步骤五：计算 $I_*(T)=\frac{1}{N_{\text{case}}}\sum_{j=1}^{N_{\text{case}}} I_j(T)-\frac{1}{N_{\text{control}}}\sum_{j=1}^{N_{\text{control}}} J_j(T)$，以此作为个体特异性异常指标。如果 $I_*(T)$ 发生急剧上升的话，则认为 T 为临界点。否则进入下一个时间点，回到步骤二。其每一个时间点所有样本的 $I_j(T)$ 或 $J_j(T)$ 做成箱型图，将个体特异性异常指标连接得到图 1(C)。

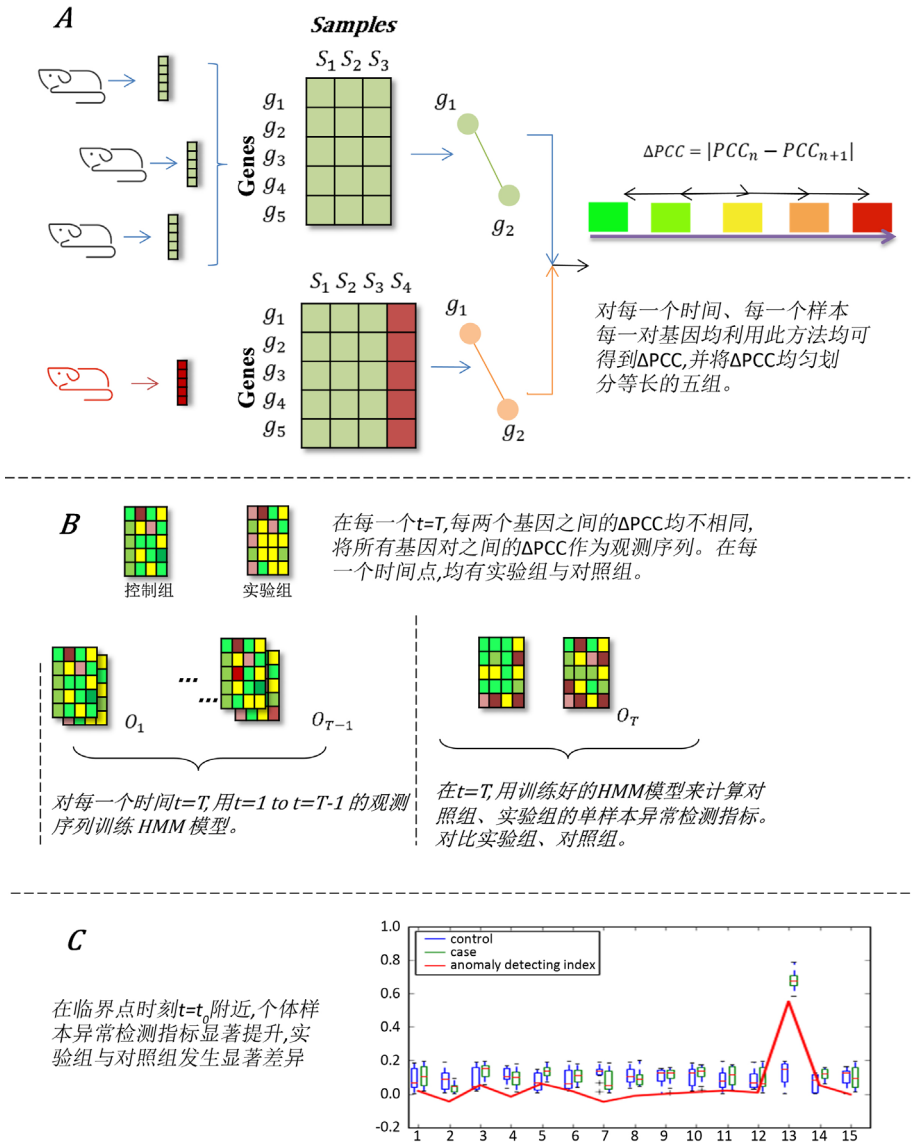

Figure 1. Using hidden Markov model to construct individual-specific anomaly index to detect complex disease network critical point

图 1. 利用隐马尔科夫模型构建个体特异性异常指标并探测临界点方法概述

3. 主要结果

3.1. 探测 9 节点模拟数据临界点

该部分本文通过构造 9 节点基因调控网络，生成疾病仿真数据，并将本文中算法应用在模拟数据中探测疾病临界点。

复杂疾病发展过程的动态过程无论在恶化前或者是恶化后都是非常复杂的，由于复杂疾病有超过数千个因素影响(如基因因素、遗传因素等)，因此状态方程通常具有大量变量与参数。但过去研究表示[9] [10] [11] [12]，可以将复杂疾病动态过程简化为：

$$Z(k+1)=f(Z(k);Q) \tag{8}$$

其中，$Z(k)=(z_1(k),\cdots,z_n(k))$ 代表在第 k 个时间点的 n 维基因状态，同时也是基因或蛋白质浓度，$Q=(q_1,\cdots,q_s)$ 是 s 维参数向量，也代表多个疾病状态影响因素。若系统同时满足以下条件：

1) $\bar{Z}$ 是系统(8)的平衡点，即 $\bar{Z}=f(\bar{Z};Q)$。

2) 存在一个 q_c 使得雅各比矩阵 $J=\left.\dfrac{\partial f(Z(k),q_c)}{\partial Z}\right|_{Z=\bar{Z}}$ 存在一个特征值，该特征值或其复共轭对的模等于 1。

3) 当 $q\neq q_c$ 时，J 的特征值的模均不等于 1。

那么当 q 达到 q_c 时，系统在 $\bar{Z}$ 产生分岔。

根据以上特征，本文利用以 Michaelis-Menten 方程结合基因自身的降解速率，构建出 9 个节点的基因调控网络，公式(9)中 9 个微分方程体现了节点之间调控关系，其基因网络如图 2(A)所示。

$$\begin{cases}
\dfrac{dz_1(t)}{dt}=\dfrac{(8-4|Q|)z_2(t)}{15(1+z_2(t))}-\dfrac{4+4|Q|}{15}z_1(t)+\tau_1(t)\\
\dfrac{dz_2(t)}{dt}=\dfrac{(4-2|Q|)z_1(t)}{15(1+z_2(t))}-\dfrac{8+2|Q|}{15}z_2(t)+\tau_2(t)\\
\dfrac{dz_3(t)}{dt}=\dfrac{4|Q|-10}{15}+\dfrac{5-2|Q|}{15(1+z_1(t))}+\dfrac{5-2|Q|}{15(1+z_2(t))}-z_3(t)+\tau_3(t)\\
\dfrac{dz_4(t)}{dt}=\dfrac{4|Q|-12}{15}+\dfrac{(6-2|Q|)z_1(t)}{15(1+z_1(t))}+\dfrac{(6-2|Q|)z_2(t)}{15(1+z_2(t))}-\dfrac{6}{5}z_4(t)+\tau_4(t)\\
\dfrac{dz_5(t)}{dt}=\dfrac{4|Q|-14}{15}+\dfrac{(7-2|Q|)z_1(t)}{15(1+z_1(t))}+\dfrac{(7-2|Q|)z_2(t)}{15(1+z_2(t))}-\dfrac{7}{5}z_5(t)+\tau_5(t)\\
\dfrac{dz_6(t)}{dt}=-\dfrac{11}{5}+\dfrac{1}{15(1+z_1(t))}+\dfrac{1}{15(1+z_2(t))}+\dfrac{z_3(t)}{5(1+z_3(t))}+\dfrac{1}{5(1+z_5(t))}\\
\qquad\qquad+\dfrac{1}{5(1+z_7(t))}+\dfrac{1}{5(1+z_8(t))}-\dfrac{8}{5}z_6(t)+\tau_6(t)\\
\dfrac{dz_7(t)}{dt}=\dfrac{z_8(t)}{10(1+z_8(t))}-\dfrac{19}{10}z_7(t)+\tau_7(t)\\
\dfrac{dz_8(t)}{dt}=\dfrac{z_7(t)}{10(1+z_7(t))}-\dfrac{19}{10}z_8(t)+\tau_8(t)\\
\dfrac{dz_9(t)}{dt}=\dfrac{z_7(t)}{10(1+z_7(t))}+\dfrac{3z_7(t)}{10(1+z_7(t))}-\dfrac{11}{5}z_8(t)+\tau_9(t)
\end{cases} \tag{9}$$

在公式(9)中，参数 Q 是标量控制变量，$\tau_i(t)$ 是均值为 0 的高斯噪音。$z_i(t)$ 代表 mRNA-i 浓度。其稳定平衡点为

$$\bar{Z}=(\bar{z}_1,\bar{z}_2,\bar{z}_3,\bar{z}_4,\bar{z}_5,\bar{z}_6,\bar{z}_7,\bar{z}_8,\bar{z}_9)=(0,0,0,0,0,0,0,0,0) \tag{10}$$

并且可以将微分方程转变为 $Z(k+1)=f(Z(k),q)$ 的形式。

我们设 $Z(k+1)$ 的雅各比矩阵为 $J=\left.\frac{\partial f(Z(k),q)}{\partial Z}\right|_{Z=\bar{Z}}$，这里

$$J=\mathrm{e}^{\Delta t\cdot A} \tag{11}$$

A 为公式(9)的系数矩阵。令 $\Delta t=1$，从 J 中我们可以得到 9 个互不相同的特征值 $\left(0.67^{|Q|},0.45,0.37,0.3,0.25,0.2,0.17,0.14,0.11\right)$。当 $Q\to 0$ 时，$0.67^{|Q|}\to 1$，此方程在该点为一个分岔点，即当 $Q\in(-1,0]$ 时，平衡点 $\bar{Z}$ 是稳定，当 Q 从小于 0 接近 $P_c=0$ 时，系统变得不稳定，即 P_c 为该微分方程一个分岔点。

将本文方法应用于上述构造的模拟数据集中，其个体特异性异常指标如图 2(B)所示，当参数接近 $Q=0$ 的时候，个体特异性异常指标剧烈的增长，而在当 Q 远离 0 点时，指标接近于 0。另外，在 $Q=0$ 的时，实验样本的指标所形成的“箱体”变大，即趋于不稳定。

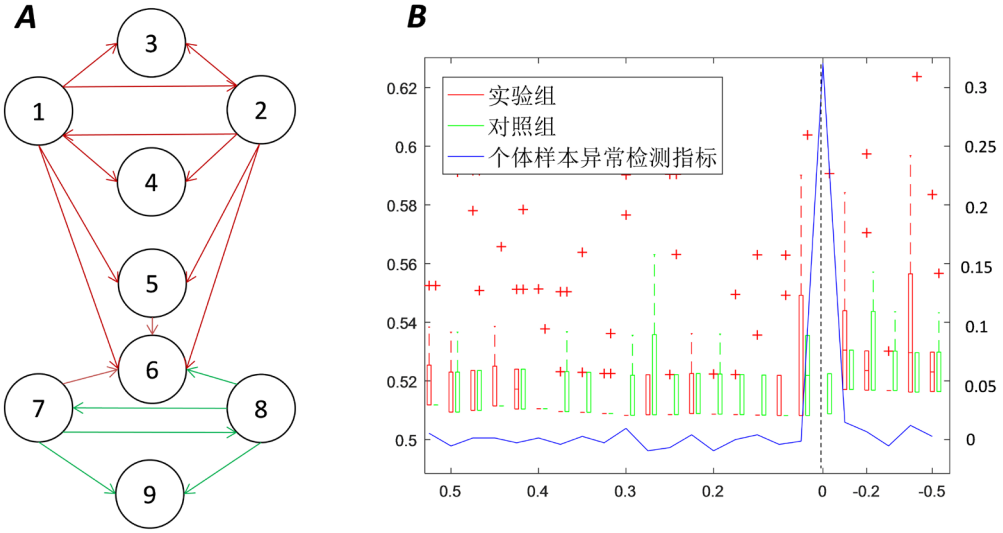

Figure 2. Simulated data set network architecture (A) and verification results (B)
图 2. 模拟数据集网络结构(A)与验证结果(B)

值得注意的是，虽然模拟数据集中已知基因连接结构，但使用本文方法时并未使用其网络结构，即探测临界点时不再需要知道原基因之间的连接结构，仅仅需要把可能相关的基因多个时间节点的值放入模型当中即可。

3.2. 探测实际数据集的临界点

将该种方法应用于肺癌数据(GSE2565)与前列腺癌数据(GSE5345)当中，数据可以从 GEO 数据库(https://www.ncbi.nlm.nih.gov/geo)中获得。

GSE2565 数据是将小鼠暴露在氯化碳(光气)与空气当中得来的，这会导致小鼠在 24 个小时内发生不可逆转急性肺损伤和肺水肿，在暴露后的 0.5、1、4、8、12、24、48、72 小时后分别搜集对照组与实验组小鼠的 RNA，而对照组、实验组在每一个时刻分别取 6 个小鼠样本，以判断肺癌发病的时间。

根据肺癌数据，利用本文构建的个体特异性异常指标的方法，可以得到在时间点 $t=8$ h 时个体特异性异常指标明显增长并达到峰值(图 3(A)、图 3(B))。其 DNB (由临界点变化基因表达最显著的前 200 个

基因组成，下同)中 55.4%基因对的$|\Delta PCC|$经过 z 检验其 p 值小于 0.05，于此相对比的在同一时刻对照组的基因对 p 值小于 0.05 的占 25.21%。

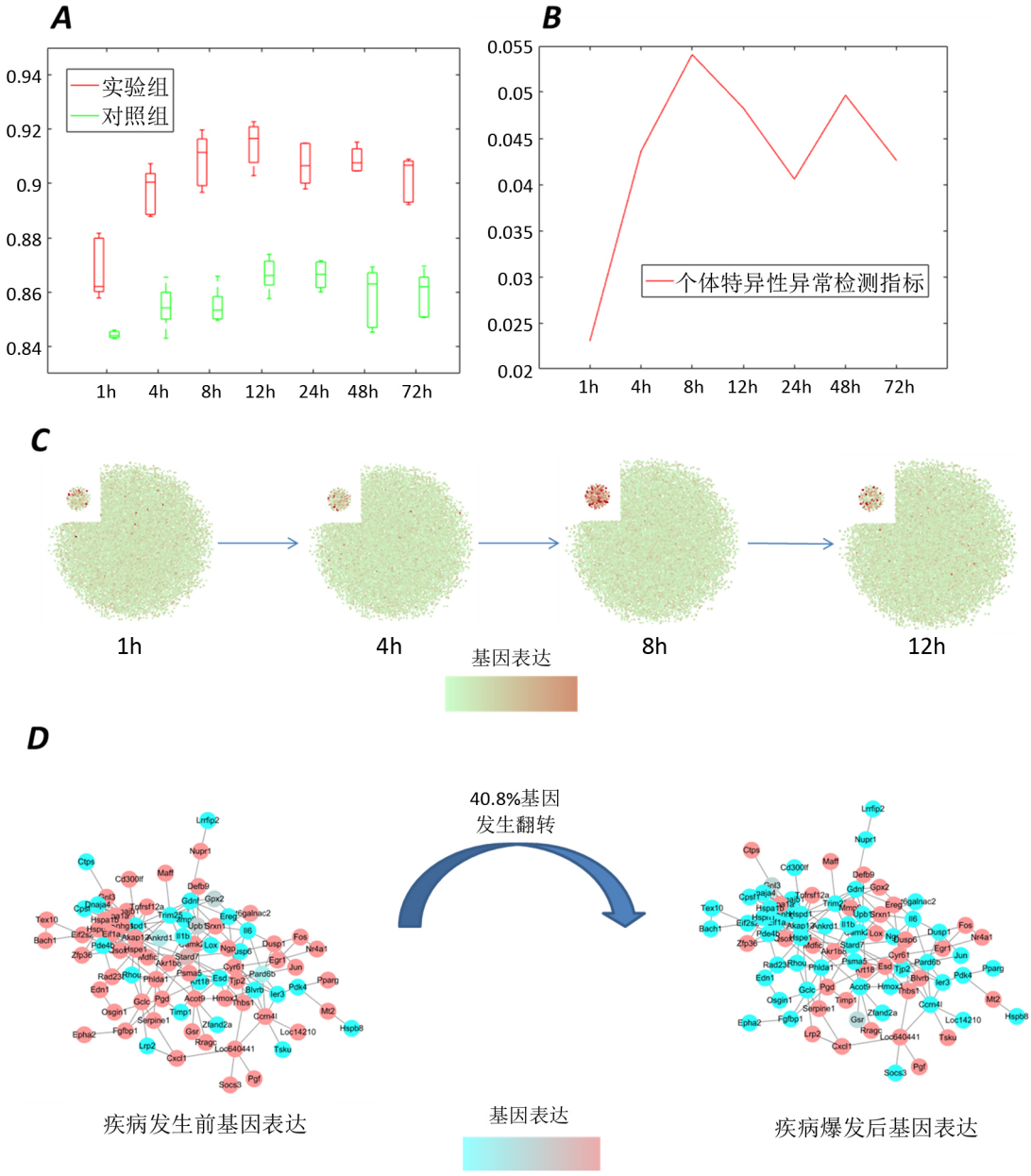

Figure 3. Individual specific abnormal indicator method for lung injury data
图 3. 个体特异性异常指标在急性肺损伤数据中应用

利用 Cytoscape 画出其基因表达与网络动态变化图(图 3(C))，其中 DNB 基因放于左上角的圆形当中。在 $t=8$ 时 DNB 基因表达明显，而在其他时间表达并无明显变化。另外，个体特异性异常指标突增前后基因表达由低表达(高表达)变为高表达(低表达)的基因有 40.8% (图 3(D))，与其相对比，所有基因平均翻转为 10.2%。

上面都指明在临界点前后，基因发生了显著的变化。这与实验结论相一致[13]：暴露在光气后 8 小时，BALF 蛋白水平升高、肺水肿增加，这导致了存活率降低，在 12 小时观察到 50%~60%的死亡率、24 小时后观察到 60%~70%的死亡率。这恰好说明暴露于光气 8 小时进入了疾病爆发期。

研究指出[14]，雄激素对前列腺癌细胞系 LNCaP 有显著的调控作用，GSE5345 数据就是通过将前列腺癌细胞系 LNCaP 暴露在雄激素(以在乙醇作为对照)的实验中，分别在 0、6、9、12、18、24 和 48 小时采集 RNA 得来，每个时间点对照组、实验组分别各有 4 个数据。

前列腺癌数据中，通过数值实验可以看到，个体特异性异常指标在 $t = 24$ h 时候明显增长并达到峰值(图 4(A)、图 4(B))，其 DNB 中 $|\Delta PCC|$ 经过 z 检验后 p 值小于 0.05 的有 48.18%，于此相对比的在同一时刻对照组的基因对 p 值小于 0.05 的占 33.36%。

针对数据中的 DNB 数据利用 SurvExpress 进行生存分析(图 4(C))，当 DNB 基因异常时，前列腺癌病人存活率显著下降，存活时间显著缩短。而针对其 DNB 中所有基因进行单个生存分析，其对患者存活率有显著影响的基因分别有 POLD1，B2M，MBOAT2，DNMT3A，EGR1，SYNPO2，SORBS1，ETV6，ALDH1A2，CD47，EGFR，BCL2L2，STX12，ZNF273，EIFAY，TSTA3，共 16 个基因。其基因表达与网络动态变化图如图 4(D)所示，DNB 基因放于左上角，可以看到在 $t = 24$ h 基因表达明显，而在其他时间所有基因并无明显表达。DNB 基因发生反转的比率占 42.3%，而所有基因反转的比率占 13.6%。

过去对该数据集的研究中[15]指出：在 24 h 时，雄激素调节的内含子 RNA 分为三种类型，第一组 RNA 在 24 小时内水平下降 2~3 倍；第二组 RNA 在 24 小时内增加 2~3 倍；第三组 RNA 在 24 内增加 3~6 倍。而三组 RNA 均在 24 小时后逐渐恢复。这说明了在 24 小时为前列腺癌重要临界点。

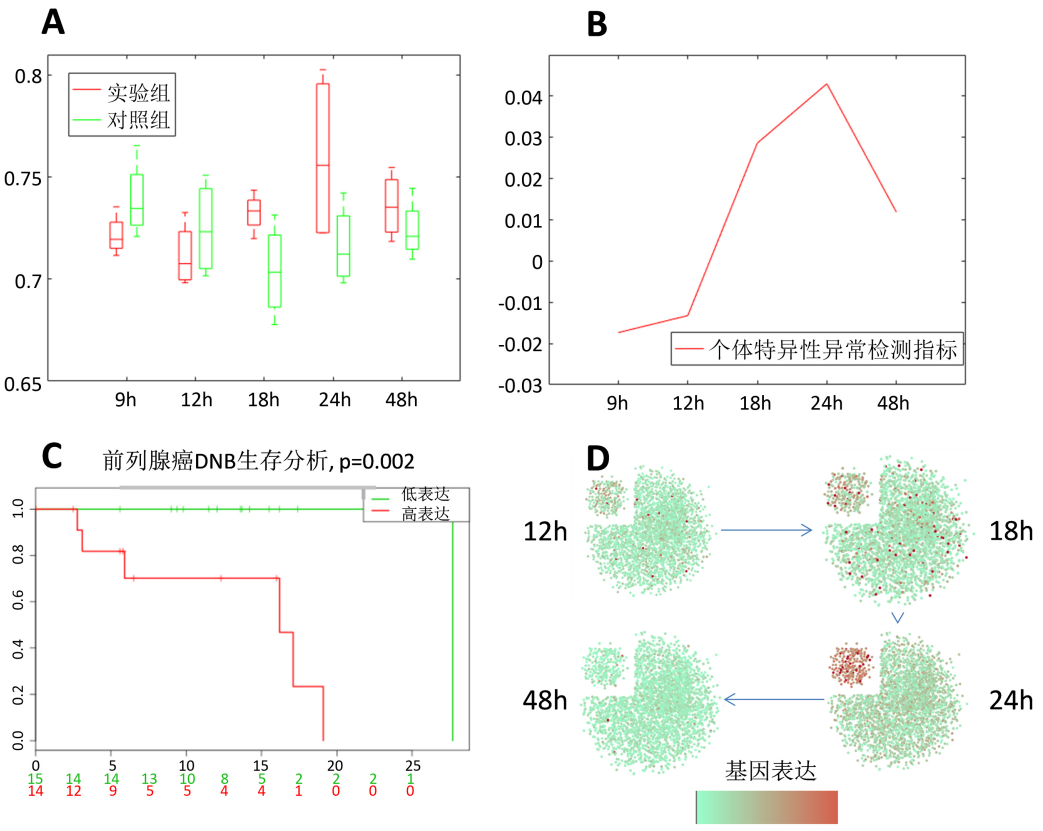

Figure 4. Individual specific abnormal indicator method for prostate data
图 4. 个体特异性异常指标在前列腺癌数据中应用

3.3. 功能分析

利用 KEGG (https://www.kegg.jp/) [15]对影响个体特异性异常指标最显著的 200 个基因进行功能

分析(如表 2，仅列出部分)，其对异常指标影响最显著的 200 个基因中，所属于通路均与疾病发展联系密切。

针对鼠类肺部损伤数据，KEGG 分析结果为主要影响了代谢途径通路、癌症通路、MAPK 信号通路等，MAPK 与 PI3K-AKT 信号通路影响细胞的繁衍与凋亡过程，均为受其氧化反应影响的通路。针对前列腺癌数据，最显著的基因中与影响了癌症及相关通路有关(如癌症通路、癌症中蛋白聚糖)，同时也与影响 MAPK 通路等关于细胞增殖与凋亡的通路有关。

Table 2. Result of KEGG
表 2. KEGG 分析结果

疾病名称	通路 ID	通路名称	p value
急性肺损伤	mmu04060	细胞因子－细胞因子受体相互作用(9 个基因)	1.5E−2
	mmu05200	TNF 信号通路(10 个基因)	1.3E−5
	mmu04010	MAPK 信号通路(12 个基因)	4.6E−4
	mmu05166	HTLV-1 感染通路(12 个基因)	1.0E−3
	mmu04151	PI3K-AKT 信号通路(9 个基因)	9.4E−2
前列腺癌	hsa05200	癌症通路(14 个基因)	3.3E−3
	hsa04015	RAP1 信号通路(10 个基因)	2.8E−3
	hsa05205	癌症中蛋白聚糖(13 个基因)	2.2E−5
	hsa05166	HTLV-1 感染通路(11 个基因)	3.9E−3
	hsa04810	肌动蛋白细胞骨架的调控(10 个基因)	2.7E−3
	hsa04010	MAPK 信号通路(9 个基因)	2.5E−2

4. 讨论

本文通过结合隐马尔科夫模型与构建个体特异性网络的方法构建了一种新的探测临界点的指标，通过提取个体特异性网络中的信息，能够更有效利用实验样本与正常样本具有差异性的信息，聚焦实验样本的特殊性。此外，通过结合机器学习的方法，学习实验样本在正常状态下的表现，更能够分辨下一时刻是否发生变异。也正如此，本文所提供方法不需要提前知道原生物基因网络结构，适用范围更广。

该方法在 9 个节点理论数据集、急性肺部损伤数据、前列腺癌数据中均通过了验证(图 2、图 3、图 4)，在疾病恶化前都发现了异常信号，说明该方法对于探测复杂疾病临界点是有效的。在找到疾病临界点后，一方面对早期治疗有一定帮助，另一方面通过对影响指标的基因的分析，可以找到对疾病影响显著的基因、通路等。对于肺部损伤、前列腺癌中，表 2 所示通路对其疾病影响显著，结合生存分析，找到 POLD1，B2M，MBOAT2，DNMT3A，EGR1，SYNPO2，SORBS1，ETV6，ALDH1A2，CD47，EGFR，BCL2L2，STX12，ZNF273，EIFAY，TSTA3 共 16 个基因对前列腺癌有显著影响。

这项工作有一定的局限性。首先，本文所使用方法与分析具体的病理机制仍然依赖于实验后数据的搜集与临床研究的支持。其次，该方法精度依赖于实验采集的时间点密度、每个时间的样本数量，对时间的精确程度仍然依赖于实验设置时间点，实验时间点之间开始发生异变的时间无法估计(如小鼠肺部损伤实验中，实验得到 $t = 4$ h 未进入临界点、$t = 8$ h 已经进入临界点，但再具体的精度仍需要实验支撑)，另外由于指标是一个统计性指标，其精度随着实验样本的数量增大而提高，在实验样本量过小的时候这可能是不准确的。

参考文献

[1] Chen, L., *et al*. (2009) Biomolecular Networks: Methods and Applications in Systems Biology. John Wiley & Sons, Hoboken, NJ. https://doi.org/10.1002/9780470488065

[2] Achiron, A., *et al*. (2010) Microarray Analysis Identifies Altered Regulation of Nuclear Receptor Family Members in the Pre-Disease State of Multiple Sclerosis. *Neurobiology of Disease*, **38**, 201-209. https://doi.org/10.1016/j.nbd.2009.12.029

[3] He, D., *et al*. (2012) Coexpression Network Analysis in Chronic Hepatitis B and C Hepatic Lesion Reveals Distinct Patterns of Disease Progression to Hepatocellular Carcinoma. *Journal of Molecular Cell Biology*, **4**, 140-152. https://doi.org/10.1093/jmcb/mjs011

[4] Litt, B., *et al*. (2001) Epileptic Seizures May Begin Hours in Advance of Clinical Onset: A Report of Five Patients. *Neuron*, **30**, 51-64. https://doi.org/10.1016/S0896-6273(01)00262-8

[5] Liu, J.K., *et al*. (2001) Pituitary Apoplexy. *Seminars in Neurosurgery*, **12**, 315-320. https://doi.org/10.1055/s-2001-33622

[6] Liu, R., *et al*. (2012) Identifying Critical Transitions and Their Leading Networks for Complex Diseases. *Scientific Reports*, **2**, 1-9. https://doi.org/10.1038/srep00813

[7] Liu, X.P., *et al*. (2013) Detecting Early-Warning Signals of Type 1 Diabetes and Its Leading Biomolecular Networks by Dynamical Network Biomarkers. *BMC Medical Genomics*, **6**, S8.

[8] Liu, X., Wang, Y., Ji, H., *et al*. (2016) Personalized Characterization of Diseases Using Sample-Specific Networks. *Nucleic Acids Research*, **44**, gkw772. https://doi.org/10.1093/nar/gkw772

[9] Liu, R., *et al*. (2015) Identifying Early-Warning Signals of Critical Transitions with Strong Noise by Dynamical Network Markers. *Scientific Reports*, **5**, Article ID: 17501. https://doi.org/10.1038/srep17501

[10] Liu, R., *et al*. (2014) Early Diagnosis of Complex Diseases by Molecular Biomarkers, Network Biomarkers, and Dynamical Network Biomarkers. *Medicinal Research Reviews*, **34**, 455-478. https://doi.org/10.1002/med.21293

[11] Chen, L., *et al*. (2012) Detecting Early-Warning Signals for Sudden Deterioration of Complex Diseases by Dynamical Network Biomarkers. *Scientific Reports*, **2**, 1-8. https://doi.org/10.1038/srep00342

[12] Liu, R., Chen, P., Chen, L., *et al*. (2016) Detecting Critical State before Phase Transition of Complex Biological Systems by Hidden Markov Model. *Bioinformatics*, **32**, 2143-2150. https://doi.org/10.1093/bioinformatics/btw154

[13] Sciuto, A.M., Phillips, C.S., Orzolek, L.D., *et al*. (2005) Genomic Analysis of Murine Pulmonary Tissue Following Carbonyl Chloride Inhalation. *Chemical Research in Toxicology*, **18**, 1654-1660. https://doi.org/10.1021/tx050126f

[14] Verjovski-Almeida, S., da Silva Aline, M., Sogayar Mari, C., *et al*. (2007) Androgen Responsive Intronic Non-Coding RNAs. *BMC Biology*, **5**, 4. https://doi.org/10.1186/1741-7007-5-4

[15] Sherman, B.T., Lempicki, R.A. and Huang, D.W. (2009) Systematic and Integrative Analysis of Large Gene Lists Using DAVID Bioinformatics Resources. *Nature Protocols*: *Recipes for Researchers*, **4**, 44-57. https://doi.org/10.1038/nprot.2008.211