Advances in Applied Mathematics 应用数学进展, 2019, 8(1), 160-169
Published Online January 2019 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2019.81018

Hans汉斯

Detecting the Early Warning Signals of Disease Progression Based on Individual Difference Time Series Network

Xiaoling Guan

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: 759966853@qq.com

Received: Jan. 2nd, 2019; accepted: Jan. 21st, 2019; published: Jan. 28th, 2019

Abstract

Pre-disease state is a critical stage of the disease state. Patients in this state can return to the normal state as long as they receive reasonable and effective treatment. Therefore, it is of great significance for medical workers and patients to detect the pre-disease state. In this paper, an algorithm is developed to establish the sampling-specific temporal differential network (SSDN) based on the individual single sample, and according to the sampling-specific temporal differential network, the signal of disease progression can be detected effectively. The validity of the method is verified by numerical simulation and two real data.

Keywords

Single Sample, Temporal Differential Network, Dynamic Network Biomarkers, Composite Variable, Critical-Point

基于个体时序列差异网络探测疾病恶化的预警信号

关小玲

华南理工大学数学学院，广东 广州
Email: 759966853@qq.com

收稿日期：2019年1月2日；录用日期：2019年1月21日；发布日期：2019年1月28日

文章引用：关小玲. 基于个体时序列差异网络探测疾病恶化的预警信号[J]. 应用数学进展, 2019, 8(1): 160-169.
DOI: 10.12677/aam.2019.81018

摘　要

前疾病状态是疾病状态的一个临界期，处于这个状态的患者,只要经过合理有效的治疗，就可以回到正常状态。所以，探测前疾病状态对于医疗工作者以及病人来说有着极其重要的意义。本文开发了一种算法，基于个体单样本建立个体时序列差异网络，并根据所建立的个体时序列差异网络，可以有效地探测疾病恶化的信号。该方法的有效性得到了数值模拟和两个真实数据的检验。

关键词

个体单样本，时序列差异网络，动态网络生物标志物，复合变量，临界点

1. 引言

考虑到人类细胞中分子组分之间的功能相互依赖性，疾病表型很少是单个生物分子发生异常表达的结果，而是由基因、蛋白质和代谢物等相关因素共同作用导致的，它们所构成的复杂相互作用网络，是我们用来研究生物疾病进展的有效依据[1] [2]。因此，了解生物分子相互作用网络的动态变化，对于全面研究复杂疾病的进展，从而更好地检测相关疾病的预警信号具有重要意义。

已有理论表明，处于前疾病状态和正常状态的生物系统在分子表达上是静态相似的，但在动力学上不同[3] [4] [5]，利用这些动态差异的特征，探索正常状态和前疾病状态的分子间的差异关联信息，可以有效地识别出处于前疾病状态的系统。本文基于单样本结合差异网络开发了一种新的检测疾病恶化的预警信号的方法，即利用时间进程数据构建一系列个体特异网络，获得相邻时间点个体特异网络的时序差异信息，进而得到个体时序列差异网络，用以整合具有时间上不同特征的生物分子，包括具有差异表达方差的基因和具有差异表达协方差的基因对之间的相互作用。由此从个体时序列差异网络的角度系统地展示了生物系统发展的动态变化，可以准确地描述个体的特定疾病状态。

为了更好的量化这些动态差异并准确地识别临界点，我们提出了一个复合变量 I，作为用于识别即将到来的临界点的特定标识符，I 的快速上升预示着前疾病状态的出现，而在正常状态或疾病状态中平稳地发展，I 几乎没有显著波动。为了验证方法的有效性，我们将之应用于一个数值仿真实验和两个从 NCBI 的 GEO 数据库(https://www.ncbi.nlm.nih.gov/geo/)下载得到真实疾病数据集，即前列腺癌数据(GSE 5345)和肝癌数据(GSE 80018)。通过计算发现，两种疾病发生恶化的早期预警信号出现的时间与实验观察结果一致。

2. 方法

2.1. 个体差异性时序网络的构造

当样本数量少甚至只有一个样本时，我们无法计算出样本的皮尔逊相关系数。而通过利用单样本动态网络标志物的方法可以解决这个问题。本文所述的用个体时序列差异网络检测复杂疾病恶化的临界信号方法的主要步骤如下(图 1)：

1) 构造个体特异网络(Sampling-specific differential network, SSN) (图 1(a))。首先，利用疾病数据，选定适量正常的样本作为参考样本，并利用样本的两两基因间的皮尔逊相关系数，作为两个基因节点的边[6]，构造参考样本网络。其次，通过把单个样本加到参考样本里，通过两两基因间的皮尔逊相关系数构造扰动样本网络。最后，对比参考网络和扰动网络，根据 z-score，保留两个网络共有的具有显著表达的边，进而得到个体特异网络[1]。

2) 构造个体时序列差异网络(图 1(b))。对比两两相邻时间点的个体特异网络，保留它们的差异边，得到个体差异性时序网络(Sampling-specific temporal differential network, SSDN)。

3) 由步骤二，得到个体时序列差异网络(图 1(c))，结合差异网络中边和网络节点的特性，用一个复合变量检测出复杂疾病发生病变的临界点。

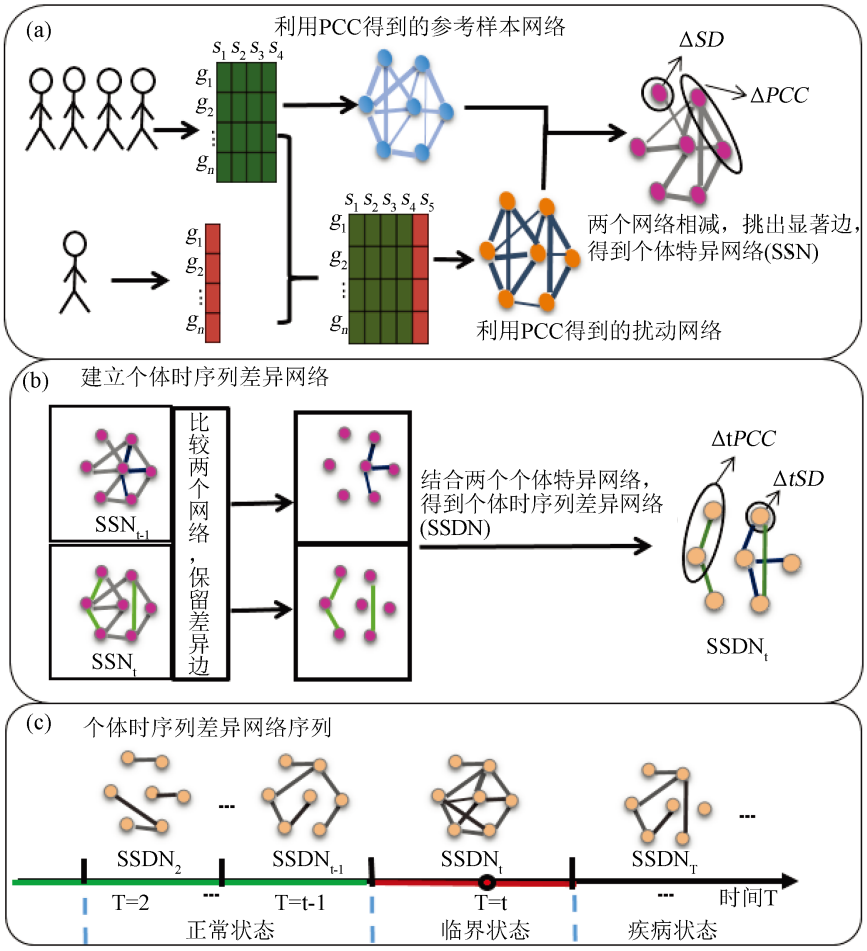

Figure 1. Flow chart of construction of sample-specific temporal differential network
图 1. 个体时序列差异网络的构造的流程图

2.2. 基于动态网络标志物的复合变量

基于 2.1，得到的单样本动态网络标志物满足以下两个条件：

1) 单样本动态网络标志物中两两元素所构成的个体时序列差异网络边的皮尔逊相关系数的绝对值均数 $\frac{\sum \Delta tPCC}{N}$ 大幅增加；

2) 单样本动态网络标志物中任意元素的表达方差的差的绝对值(ΔtSD)的大幅增加。

引入复合变量

$$I=\left(\frac{\sum \Delta tPCC}{N}\right)\left(\overline{\sum \Delta tSD}\right)$$

其中， $\sum \Delta tPCC$ 表示个体时序差异网络 SSDN 中显著边的皮尔逊相关系数的绝对值的和的， N 即表示 SSDN 中显著边的数量。 ΔtSD 表示 SSDN 中基因节点的标准差的绝对值。对以上复合变量 I 进行验算，若 I 的值在某点处急剧上升并达到峰值时，表示该点为疾病发生病变的临界点。

3. 主要结果

3.1. 八个节点的仿真数据

在这里，为了检测方法的有效性，我们使用八个基因节点网络(图 2，节点间的连线表示基因间的调控关系，其中箭头线表示正调节，钝化线表示负调节)进行数值模拟，并从理论上论证了通过复合变量 I 可以检测复杂疾病的预警信号。这种类型的基因调控网络常常应用于研究基因转录、翻译等影响基因调控的活动。下面的八个微分方程表示网络中八个基因的一般规律[7]，其中基因调控以 Michaelis-Menten 方程加入基因自身的降解率的形式表示，降解率与相应基因浓度成线性比例关系。

$$\begin{cases}
\dfrac{\mathrm{d}z_1(t)}{\mathrm{d}t}=\dfrac{(6-4|P|)z_2(t)}{15(1+z_2(t))}-\dfrac{4|P|+3}{15}z_1(t)+\varsigma_1(t)\\
\dfrac{\mathrm{d}z_2(t)}{\mathrm{d}t}=\dfrac{(3-2|P|)z_1(t)}{15(1+z_1(t))}-\dfrac{2|P|+6}{15}z_2(t)+\varsigma_2(t)\\
\dfrac{\mathrm{d}z_3(t)}{\mathrm{d}t}=-\dfrac{4|P|-8}{15}+\dfrac{4-2|P|}{15(1+z_1(t))}+\dfrac{4-2|P|}{15(1+z_2(t))}+\varsigma_3(t)\\
\dfrac{\mathrm{d}z_4(t)}{\mathrm{d}t}=-\dfrac{4|P|-10}{15}+\dfrac{5-2|P|}{15(1+z_1(t))}+\dfrac{5-2|P|}{15(1+z_2(t))}-z_4(t)+\varsigma_4(t)\\
\dfrac{\mathrm{d}z_5(t)}{\mathrm{d}t}=-\dfrac{2}{3}+\dfrac{2}{15(1+z_1(t))}+\dfrac{2}{15(1+z_2(t))}+\dfrac{2}{5(1+z_3(t))}-\dfrac{6}{5}z_5(t)+\varsigma_5(t)\\
\dfrac{\mathrm{d}z_6(t)}{\mathrm{d}t}=-\dfrac{7}{15}+\dfrac{1}{15(1+z_1(t))}+\dfrac{1}{15(1+z_2(t))}+\dfrac{1}{5(1+z_3(t))}+\dfrac{1}{5(1+z_5(t))}\\
\qquad\qquad+\dfrac{1}{5(1+z_7(t))}+\dfrac{z_8(t)}{5(1+z_8(t))}-\dfrac{7}{5}z_6(t)+\varsigma_6(t)\\
\dfrac{\mathrm{d}z_7(t)}{\mathrm{d}t}=\dfrac{z_8(t)}{10(1+z_8(t))}-\dfrac{17}{10}z_7(t)+\varsigma_7(t)\\
\dfrac{\mathrm{d}z_8(t)}{\mathrm{d}t}=\dfrac{z_7(t)}{10(1+z_7(t))}-\dfrac{17}{10}z_8(t)+\varsigma_8(t)
\end{cases}\tag{1}$$

微分方程系统(1)中的 P 表示的是标量控制参数， $\varsigma_i(t)(i=1,2,\cdots,8)$ 是均值为零，方差为 $k_{ij}=Cov(\varsigma_i,\varsigma_j)$ 的高斯噪声， $z_i(i=1,2,\cdots,8)$ 表示基因的表达浓度。八个基因的降解率分别为 $\left(\frac{4|P|+3}{15},\frac{2|P|+6}{15},\frac{4}{5},1,\frac{6}{5},\frac{7}{5},\frac{17}{10},\frac{17}{10}\right)$。从微分系统(1)可以看出，系统有一个稳定的平衡点

$\bar{z}=\left(\overline{z_1},\overline{z_2},\cdots,\overline{z_8}\right)=(0,0,0,0,0,0,0,0)$。运用欧拉序列，可以把微分方程转换成 $Z(k+1)=f\left(Z(k),P\right)$ 的形式，即

$$\begin{cases}
z_1(k+1)=z_1(k)+\left[\dfrac{(6-4|P|)z_2(k)}{15(1+z_2(k))}-\dfrac{4|P|+3}{15}z_1(k)+\varsigma_1(k)\right]\Delta t\\
z_2(k+1)=z_2(k)+\left[\dfrac{(3-2|P|)z_1(k)}{15(1+z_1(k))}-\dfrac{2|P|+6}{15}z_2(k)+\varsigma_2(k)\right]\Delta t\\
z_3(k+1)=z_3(k)+\left[-\dfrac{4|P|-8}{15}+\dfrac{4-2|P|}{15(1+z_1(k))}+\dfrac{4-2|P|}{15(1+z_2(k))}+\varsigma_3(k)\right]\Delta t\\
z_4(k+1)=z_4(k)+\left[-\dfrac{4|P|-10}{15}+\dfrac{5-2|P|}{15(1+z_1(k))}+\dfrac{5-2|P|}{15(1+z_2(k))}-z_4(k)+\varsigma_4(k)\right]\Delta t\\
z_5(k+1)=z_5(k)+\left[-\dfrac{2}{3}+\dfrac{2}{15(1+z_1(k))}+\dfrac{2}{15(1+z_2(k))}+\dfrac{2}{5(1+z_3(k))}-\dfrac{6}{5}z_5(k)+\varsigma_5(k)\right]\Delta t\\
z_6(k+1)=z_6(k)+\left[-\dfrac{7}{15}+\dfrac{1}{15(1+z_1(k))}+\dfrac{1}{15(1+z_2(k))}+\dfrac{1}{5(1+z_3(k))}+\dfrac{1}{5(1+z_5(k))}\right.\\
\qquad\left.+\dfrac{1}{5(1+z_7(k))}+\dfrac{z_8(k)}{5(1+z_8(k))}-\dfrac{7}{5}z_6(k)+\varsigma_6(k)\right]\Delta t\\
z_7(k+1)=z_7(k)+\left[\dfrac{z_8(k)}{10(1+z_8(k))}-\dfrac{17}{10}z_7(k)+\varsigma_7(k)\right]\Delta t\\
z_8(k+1)=z_8(k)+\left[\dfrac{z_7(k)}{10(1+z_7(k))}-\dfrac{17}{10}z_8(k)+\varsigma_8(k)\right]\Delta t
\end{cases}\tag{2}$$

Δt 表示微小的时间间隔，注意到 $Z(k)$ 是 $Z(t)$ 在 $k\Delta t$ 时刻的向量。我们把差分系统(1)的雅可比矩阵用 $J=\left.\dfrac{\partial f\left(Z(k);P\right)}{\partial Z}\right|_{Z=\bar{Z}}$ 表示，其中

$$J=\mathrm{e}^{\Delta t\cdot A}\tag{3}$$

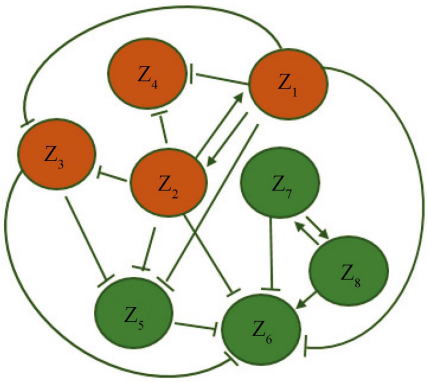

Figure 2. Eight gene regulatory networks and their differential equations
图 2. 八个基因的调控网络及其微分方程系统

这里 A 为微分方程线性系统的系数矩阵，并且矩阵 A 的八个不同的特征值组成的向量为 $E=\left(-\frac{2}{5}|P|,-\frac{3}{5},-\frac{4}{5},-1,-\frac{6}{5},-\frac{7}{5},-\frac{8}{5},-\frac{9}{5}\right)$。

对于矩阵 J，存在一个非退化矩阵

$$S=\begin{bmatrix}-2&1&0&0&0&0&0&0\\-1&-1&0&0&0&0&0&0\\1&0&-1&0&0&0&0&0\\1&0&0&-1&0&0&0&0\\0&0&1&0&-1&0&0&0\\0&0&0&0&1&-1&0&1\\0&0&0&0&0&0&-1&1\\0&0&0&0&0&0&-1&-1\end{bmatrix}$$

满足 $\Lambda=S^{-1}JS=diag\left(0.65^{|p|},0.55,0.45,0.37,0.30,0.24,0.20,0.17\right)$。明显地，当 $P\to 0$ 时，$0.65^{|p|}\to 1$，这时差分系统(1)的状态失去原有的稳定性，发生临界转变。所以，当 $P\in(-1,0]$，系统在临界点 $\overline{Z}$ 是稳定的。我们的目标是检测当控制参数 P 从 $P<0$ 接近临界值 0 的时候，系统状态发生转变的一个早期预警信号。利用公式 $Y(k)=S^{-1}\left(Z(k)-\overline{Z}\right)$，我们可以得到

$$\begin{aligned}z_1-\overline{z_1}&=-2y_1-y_2\\z_2-\overline{z_2}&=-y_1-y_2\\z_3-\overline{z_3}&=y_1-y_3\\z_4-\overline{z_4}&=y_1-y_4\\z_5-\overline{z_5}&=y_3-y_5\\z_6-\overline{z_6}&=y_5-y_6+y_8\\z_7-\overline{z_7}&=-y_7-y_8\\z_8-\overline{z_8}&=-y_7-y_8\end{aligned}$$

可知 $(z_1,z_2,z_3,z_4,z_5,z_6,z_7,z_8)$ 中的四个变量 z_1,z_2,z_3,z_4 是与 y_1 直接相关的，根据已有的理论可知，z_1,z_2,z_3,z_4 构成了八个结点仿真网络系统的主要部分，即相当于真实数据中的动态网络生物标志物。

在这里，为了检测方法的有效性，我们使用八个基因节点网络(图 2，节点间的连线表示基因间的调控关系，其中箭头线表示正调节，钝化线表示负调节)进行数值模拟，并从理论上论证了通过复合变量 I。

基于仿真模型，采集得到了 12 个样本在 41 个时间点下的八个基因节点的数据，其中容易知道，第 21 个时间点为我们所模拟的临界点。在这次的仿真数据中，参考样本选取来自于第一个时间点的任意 6 个样 本，疾病样本取自剩余的 6 个样本中的任 3 个，正常样本则为最后的 3 个样本，并且把这三个样本在第 21 个时间点的数据用第一个时间点的任意三个样本的数据替换掉，以保证每一个正常样本与疾病样本区分开来。然后根据这些模拟数据计算出个体时序列差异网络的复合变量 I，得到以图 3(a)的折线图，红色曲线代表由 3 个疾病样本计算得到的复合变量的平均值，绿色曲线表示 3 个由正常样本计算得到的的复合变量的平均值。图 3(b)的 3 个网络表示个体样本在 $p=-0.3$，$p=-0.02$ 和 $p=0.3$ 处的个体时序列差异网络，分别处于正常状态，前疾病状态和疾病状态，它们都是分别由处于三个不同状态的两个相邻时间点的个体特异网络得到的。在每个个体时序差异网络中，连接两个节点的边代表的是时间差异相关性。由图 3 可以容易得到，$p=0$ 的点为系统状态临界转换点。

由实验结果可知，此实验验证了复合变量 I 在临界状态信号传递过程中的可靠性和准确性。因此，基于个体时序列差异网络，复合变量 I 能够利用高维信息来区分前疾病的样本和正常样本。

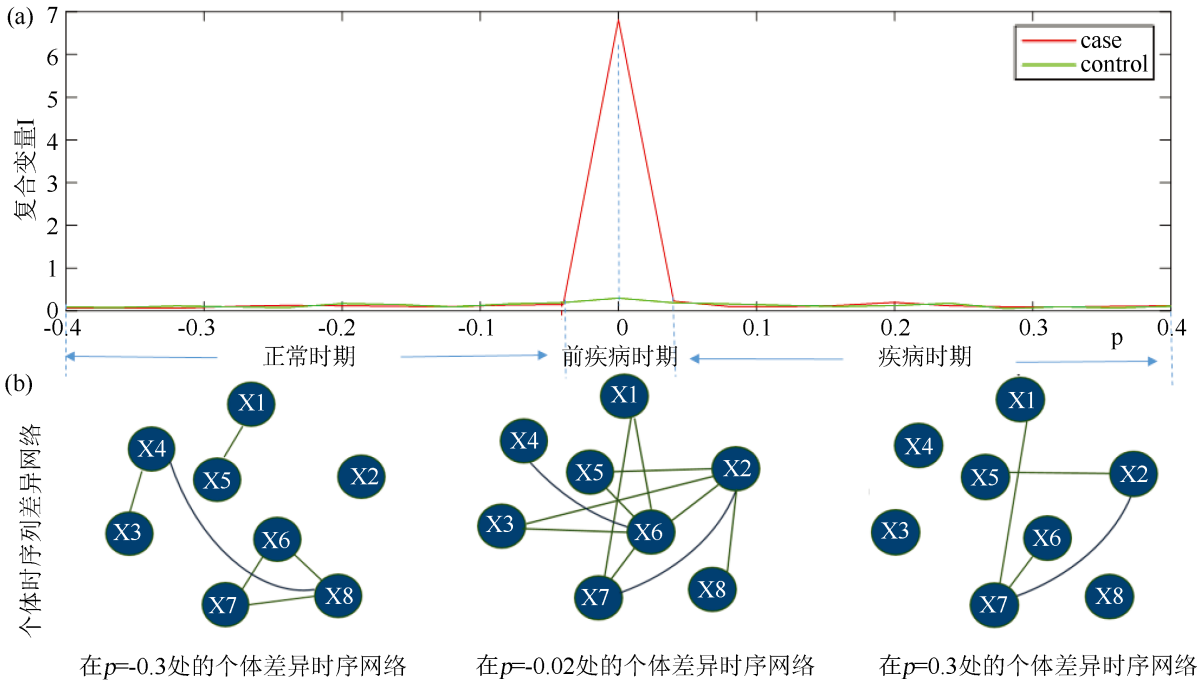

Figure 3. The line chart of numerical simulation and the sample-specific temporal differential network
图 3. 数值模拟结果折线图和个体时序列差异网络

3.2. 方法应用于两个真实的疾病数据

本节我们把这个方法运用到两个来源于 NCBI 的 GEO 数据库(https://www.ncbi.nlm.nih.gov/geo/)的真实数据，即前列腺癌数据(GSE5345)和肝癌数据(GSE80018)。通过建立个体特异时序差异网络的方法，我们把每个时间点都当作是候选的临界转化点。

前列腺癌的数据是通过前人的工作实验而得到的。实验工作者把用合成雄激素培养的前列腺癌细胞系(LNCaP)作为实验组，把用乙醇培养的前列腺癌细胞 LNCaP 作为对照组，分别在 0、6、9、12、18、24、48 小时分离 RNA 并采集基因表达数据。使用合成的雄激素 R1881 刺激人体的前列腺癌细胞系 LNCaP 的 24 小时内，部分 RNA 表达水平下调 2-3 倍直至 48 小时后恢复正常，部分上调 2~3 倍甚至 3~6 倍后恢复正常[8]。

基于前列腺癌的数据，利用 2 中的方法，计算每一个样本的复合变量 I，可以得到 4 个样本都在 24 小时处出现临界转变信号(图 4(a))，并且他们的动态网络标志物的数量分别为 246，239，217，201，其中 4 个样本之间两两共有的动态网络生物标志物的数量分别是 66，18，10，11，16，76，一共得出 164 个基因作为识别前列腺癌临界转变的动态网络标志物，图 4(c)表示这 164 个差异表达基因在临界点前后时间点的动态翻转网络。图 4(b)的圆盘图分别展示了分子网络在 0 h，9 h，12 h，18 h，24 h 和 48 h 的动态变化。每一个网络都是基于表达数据的人体前列腺基因相互作用网络构建的。节点的颜色表示基因表达的波动情况，每个网络中左上角表示 164 个差异表达最显著的基因。从圆盘图中可以看出，所选择的基因在 24 h 处表达波动最激烈，在这一时期其网络结构变化最为显著。因此，可以得出，前列腺癌病变的临界时间点在 24 h 处，与实验观测以及数值计算结果相吻合。

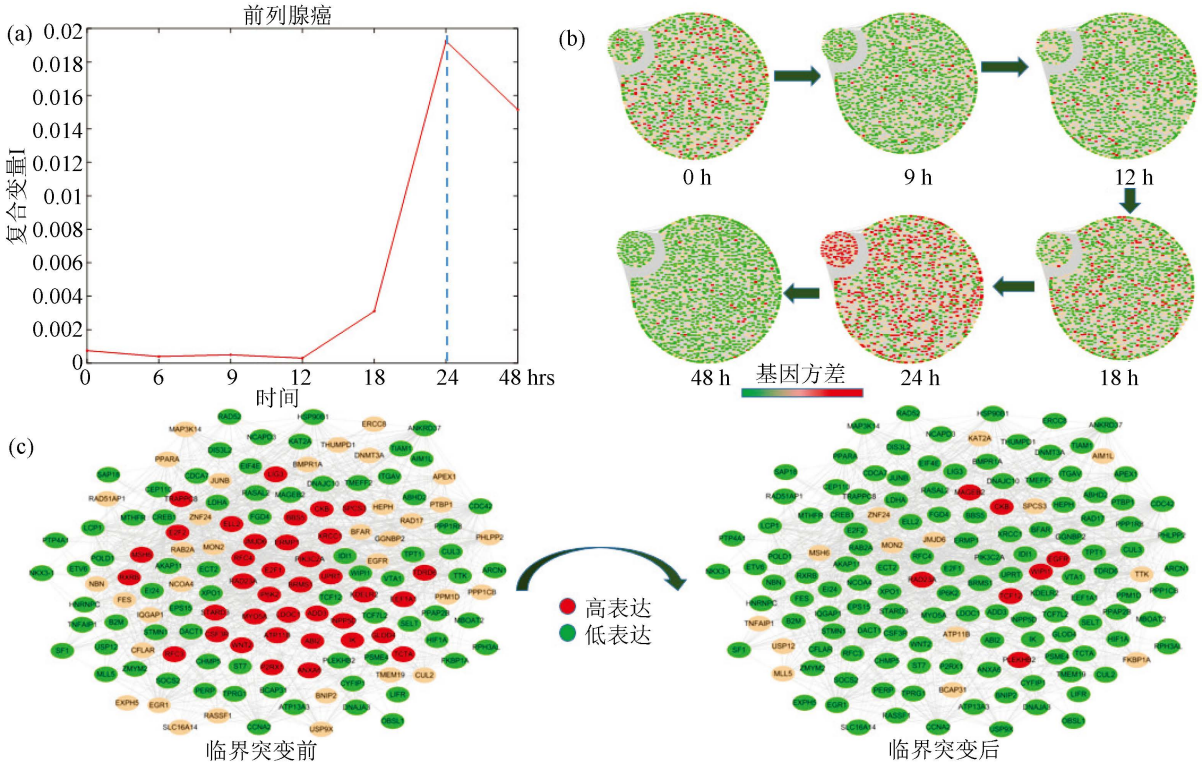

Figure 4. The line chart of early-warning signals, the dynamically changes in the network and overturn network during the progression of prostate cancer

图 4. 前列腺癌临界信号折线图，分子网络动态变化以及翻转网络图

对肝癌数据，是由实验工作人员随机给雄性小鼠分为两组，把用含有 0.05% [wt/vol]苯巴比妥(Phenobarbital, PB)饮用水喂养的那组作为实验组，把用不含 PB 的正常饮用水喂养的那组作为对照，并在第 1、3、7、14、28，57 和 91 天后对小鼠进行处理并收集小鼠的肝脏基因表达数据而得到的，以用来研究 PB 对小鼠的影响。通过实验观测可知，在 PB 处理的第 1 天，检测到肝脏切片有短暂的有丝分裂反应，而且最显著的病理异常表现为肝细胞在第 7 天后开始变得肥大，且在后期严重程度增加。PB 处理影响了肝脏中大量基因的转录，其中 PB 诱导的异型生物代谢酶，包括 CYP450 酶和 POR 还原酶，从利用 PB 处理第 1 天开始，它们的蛋白水平在整个表达过程中均发生上调[9]。

利用本文开发的方法，基于 5 个样本分别计算它们的复合变量 I，发现 5 个样本都在第 3 天出现早期疾病预警信号，由此可知，5 个样本的状态都在第 3 天发生转变(图 3(b))。并得到 5 个样本的动态网络标志物的数量分别为 181，189，178，185，204，其中 5 个样本之间两两共有的动态网络生物标志物的数量分别有 6，6，11，17，15，12，2，30，30，16，一共有 87 个基因。图 5(b)表示这 87 个差异表达基因在临界时间点前后的动态翻转网络。

4. 讨论

本文基于单样本生物动态网络的性质，从生物数据中提取生物系统中的一些动态信息，即通过对比相邻时间点的个体特异网络，提取它们之间边的相关性以及基因的表达方差等，从而获得疾病发病密切相关的主要基因，最终识别疾病进展过程中的临界状态或者是临界点，达到预测疾病的目的。基于前列腺癌数据，本文所用到的方法成功地探测出了第六个时间点为疾病恶化的临界时间点。同样，对于肝癌数据，我们也成功地探测出第二个时间为疾病恶化的临界点。这些探测结果与实验结果相一致。

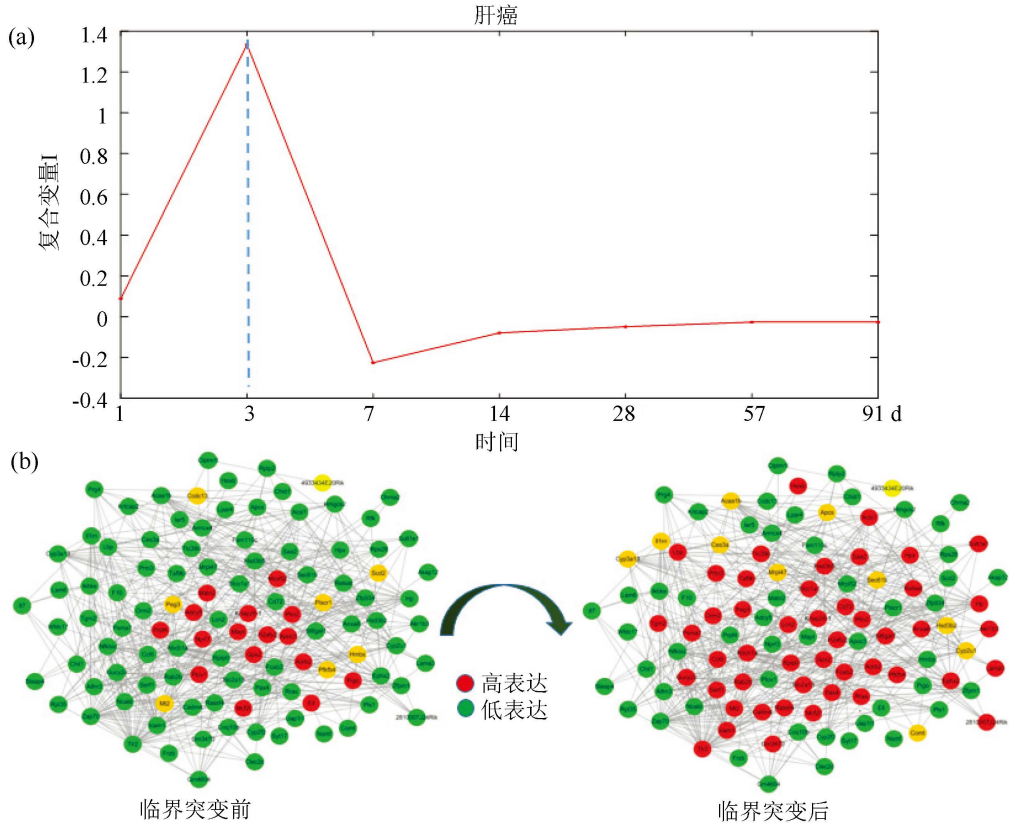

Figure 5. The line chart of early-warning signals and the overturn network during the progression of liver cancer
图 5. 肝癌临界信号折线图以及翻转网络图

基金项目

广州市科技计划珠江科技新星专项项目资助(No. 201610010029)。

参考文献

[1] Liu, X., Wang, Y., Ji, H., *et al*. (2016) Personalized Characterization of Diseases Using Sample-Specific Networks. *Nucleic Acids Research,* **44**, e164. https://doi.org/10.1093/nar/gkw772

[2] Pei, C., Yongjun, Li., Liu, X., Liu, R. and Luonan, C. (2017) Detecting the Tipping Points in a Three-State Model of Complex Diseases by Temporal Differential Networks. *Journal of Translational Medicine*, **15**, 217. https://doi.org/10.1186/s12967-017-1320-7

[3] Liu, R., *et al*. (2012) Identifying Critical Transitions and Their Leading Biomolecular Networks in Complex Diseases. *Scientific Reports*, **2**, 813. https://doi.org/10.1038/srep00813

[4] Liu, R., Yu, X., Liu, X., Xu, D., Aihara, K. and Chen, L. (2014) Identifying Critical Transitions of Complex Diseases Based on a Single Sample. *Bioinformatics*, **30**, 1579-1586. https://doi.org/10.1093/bioinformatics/btu084

[5] Chen, P. and Li, Y. (2016) The Decrease of Consistence Probability: At the Crossroad of Catastrophic Transition of a Biological System. *BMC Systems Biology*, **10**, 50. https://doi.org/10.1186/s12918-016-0295-y

[6] Liu, X.P., Liu, Z.-P., Zhao, X.-M. and Chen, L.N. (2012) Identifying Disease Genes and Module Biomarkers by Differential Interactions. *Journal of the American Medical Informatics Association*, **19**, 241-248. https://doi.org/10.1136/amiajnl-2011-000658

[7] Chen, L., Liu, R., Liu, Z., Li, M. and Aihara, K. (2012) Detecting Early-Warning Signals for Sudden Deterioration of Complex Diseases by Dynamical Network Biomarkers. *Scientific Reports*, **2**, 342. https://doi.org/10.1038/srep00342

[8] Louro, R., Nakaya, H.I., Amaral, P.P., *et al*. (2007) Androgen Responsive Intronic Non-Coding RNAs. *BMC Biology*,

5, 4. https://doi.org/10.1186/1741-7007-5-4

[9] Lempiäinen, H., Couttet, P., Bolognani, F., *et al*. (2013) Identification of Dlk1-Dio3 Imprinted Gene Cluster Noncoding RNAs as Novel Candidate Biomarkers for Liver Tumor Promotion. *Toxicological Sciences*, **131**, 375-386. https://doi.org/10.1093/toxsci/kfs303