Advances in Applied Mathematics 应用数学进展, 2020, 9(2), 231-237
Published Online February 2020 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2020.92026

Hans汉斯

Identifying Critical Transition with PageRank Algorithm in a Biological Process

Yangkai Wang, Rui Liu

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: wykkyle@outlook.com, scliurui@scut.edu.cn

Received: Jan. 30th, 2020; accepted: Feb. 12th, 2020; published: Feb. 19th, 2020

Abstract

With the belief that high-throughput datasets hold all the necessary information we want, a problem of information retrieval confronts us. As PageRank algorithm achieves a great success in dealing with such a problem in the field of Internet, we adapt it for high-throughput datasets in combination with the theory of dynamical network biomarker, and try to identify a critical transition in the biological processes. Our adapted PageRank algorithm successfully identifies the designated critical points in data simulations and it also produces the same results with the earlier works when applied to experimental datasets.

Keywords

Pagerank Algorithm, Dynamical Network Biomarker (DNB), Critical Phenomena, High Throughput Gene Expression Profiling

采用PageRank算法探测生物过程中的临界点

王阳开，刘　锐

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

收稿日期：2020年1月30日；录用日期：2020年2月12日；发布日期：2020年2月19日

摘　要

生物高通量表达数据包含了海量信息，因此在以之为基础的研究中所面对的问题，可以视作一种信息提取问题。受此驱动，我们对互联网领域中久负盛名的PageRank算法进行改造，并将其与动态网络标志物理论相结合，来探测生物过程中的临界点。我们的算法通过了随机生成的模拟数据的检验，并在实验

文章引用: 王阳开, 刘锐. 采用 PageRank 算法探测生物过程中的临界点[J]. 应用数学进展, 2020, 9(2): 231-237.
DOI: 10.12677/aam.2020.92026

数据中得到与文献中已发表方法一致的结果。

关键词

PageRank算法，动态网络标志物(DNB)，临界现象，高通量表达谱

1. 引言

临界现象在各领域中广泛存在，它描述了复杂系统在演化过程中的突变行为。对于复杂系统，往往难以建立解析模型或进行重复实验；而系统的静态状态参量在突变前也难以观察到显著改变。所以前人建立起了一套基于系统的动态状态参量来预测系统突变的理论，即临界点理论[1]。临界点理论背后的数学机理是微分方程的分叉，临界点理论给出的典型临界信号包括系统扰动的方差增大，自相关时间增长等等。

为了适应生物信息领域中的高通量表达技术的发展，由临界点理论演变而产生了动态网络生物标志物(DNB)理论[2] [3] [4]。经典的临界点理论要求基于长时间序列来进行分析；相反地，在高通量表达数据集合中，每次测量取得大量变量值，而测量次数相对较少。一个典型的高通量表达数据集合，在几个时间点上各自进行数次重复，总共包含几十到几百次测量，而每次测量同时取得上万个变量值。

在 DNB 理论中，所有基因被视作节点来构成一个网络，基因表达水平之间的相关程度结合外部给出的先验信息确定网络的边。DNB 理论认为如此大的一个网络不可能同步越过临界点，即总会有一个子网络率先越过临界点，而后带动整个网络翻转，这个子网络就称为 DNB。可以通过识别 DNB 来探测网络的临界状态。DNB 子网络在临界时刻表现出三项普遍特征：1) DNB 中基因表达水平的波动普遍增大，理想情形中将会增大到无穷；2) DNB 中基因表达水平的波动相关性普遍增强，理想情形中皮尔逊相关系数(PCC)的绝对值将趋近 1；3) DNB 中的节点与 DNB 之外节点间的波动相关性普遍减弱，理想情形中 PCC 将会减小为 0。DNB 在其它时刻不表现出显著特征。在 DNB 子网络的临界特征中，我们关注后两项描述相关性的特征。

PageRank 是一个受到广泛研究的算法，在不同应用情景下有着多样的具体形式[5]。基本地，PageRank 算法就是求取如下迭代的平衡点：

$$\boldsymbol{\pi} \leftarrow \alpha \boldsymbol{H}_n^{\mathrm{T}} \boldsymbol{\pi} + \alpha \left(\boldsymbol{d}^{\mathrm{T}} \boldsymbol{\pi} \right) v_n + \left(1 - \alpha \right) v_n, \tag{1}$$

其中 $\boldsymbol{\pi}$ 是向量，作为迭代变量，它的平衡点 $\boldsymbol{\pi}_0$ 就是 PageRank 算法的输出，称为 PageRank 向量；$\boldsymbol{H}_n$ 是标准化之后的邻接矩阵，记标准化之前的邻接矩阵为 $\boldsymbol{H}_r$，它包含了网络信息，它的 $[i, j]$ 元 $\boldsymbol{H}_r[i, j]$ 表示了节点 i 到节点 j 之间的连接强度；$\boldsymbol{d}$ 是悬挂向量，它是标准化邻接矩阵时的副产品，用于对全零行进行异常处理；$\boldsymbol{v}$ 是个性化向量，没有约束，$\boldsymbol{v}$ 除以它的各分量之和得到 $\boldsymbol{v}_n$，使 $\boldsymbol{v}_n$ 满足各分量之和为 1；α 是阻尼系数，传统上取 0.85，本文也遵从该传统。

我们期望通过较高的 PageRank 值来识别 DNB 子网络。DNB 中的节点两两紧密相关，这一特征有利于获得较高的 PageRank 值。另一方面，DNB 子网络与网络剩余部分相关减弱，这一点不利于我们确定

DNB，因为理论上我们不能从网络上定位一个脱离网络的小集团。综合这两点，我们的方法对 DNB 理论模型提出了一个额外要求：DNB 节点的度大于全网络平均值。考虑理想的 DNB，即 DNB 中节点两两连接，与 DNB 之外的节点没有连接；此时 DNB 中节点的度就是 DNB 的大小。这一要求等价于 DNB 在全网络的占比不能小于邻接矩阵密度。这意味着我们的算法受制于 DNB 的大小，无法定位过小的 DNB。但在实际应用中，这一要求通常是得以满足的。

2. 方法

2.1. 从 PCC 矩阵提取邻接矩阵

在每一测量时间点，由此刻的数次重复测量，可以直接计算基因表达水平之间的 PCC 矩阵 $\boldsymbol{R}$。考虑两个正态分布的无关随机变量，它们之间的皮尔逊相关系数 r 具有如下分布：

$$\sqrt{r^2(n-2)/(1-r^2)} \sim t_{n-2} \tag{2}$$

这里 n 是重复数目，t_{n-2} 是 $n-2$ 维的 T 分布；我们取 r 的单边概率 0.050 分位点为 T_c。$\boldsymbol{R}$ 映射为邻接矩阵的过程中，$\boldsymbol{R}$ 中绝对值小于 T_c 的元素归 0，绝对值大于 T_c 的元素映射到[0,1]区间：

$$\boldsymbol{H}_r[i,j] = \left(\left|\boldsymbol{R}[i,j]\right| - T_c\right)/(1-T_c) * \left|\boldsymbol{R}[i,j]\right|^n \tag{3}$$

对于实验数据，我们引入 STRING 网络[6]为外部参考，只采用 STRING 网络中存在的边；即 $\boldsymbol{H}_r$ 中元素若在 STRING 网络中没有对应边，则令为 0。

2.2. 抑制网络中的背景结构

由于 DNB 的特征仅在临界时刻表现出来，我们需要尝试排除网络中所有时间点均存在的背景结构的影响：

$$\begin{aligned} \boldsymbol{H}_m[i,j] &= \boldsymbol{H}_r[i,j]\left(1-\boldsymbol{S}[i,j]\right)/\left(1+\sum_i \boldsymbol{S}[i,j]\right) \\ \boldsymbol{\pi}_m[i] &= \boldsymbol{\pi}_o[i]\left(1+\sum_i \boldsymbol{S}[i,j]\right) \end{aligned} \tag{4}$$

这里 $\boldsymbol{S}[i,j]$ 是 $\boldsymbol{H}_r[i,j]$ 在所有参考时间点上的平均值；如果数据包含对照组，参考时间点就是对照组中的所有时间点，如果数据没有给出对照组，参考时间点就是全体时间点；$\boldsymbol{H}_m$ 以及 $\boldsymbol{\pi}_m$ 分别代替 $\boldsymbol{H}_r$ 以及 $\boldsymbol{\pi}_o$ 参与 PageRank 运算以及作为 PageRank 输出。取 $\boldsymbol{\eta}[i,j]$ 为中间变量，这一调节可以分为两部分来看待；$\boldsymbol{\eta}[i,j] = \boldsymbol{H}_r[i,j]\left(1-\boldsymbol{S}[i,j]\right)$ 用于抑制背景边：如果某一条边总是存在，那么对应的 $\boldsymbol{S}[i,j]$ 比较大，此时我们降低它的强度；$\boldsymbol{H}_m[i,j] = \boldsymbol{\eta}[i,j]/\left(1+\sum_i \boldsymbol{S}[i,j]\right)$ 以及 $\boldsymbol{\pi}_m[i] = \boldsymbol{\pi}_o[i]\left(1+\sum_i \boldsymbol{S}[i,j]\right)$ 一起抑制背景中的中心节点：如果一个节点的度始终比较大，那么对应的 $\sum_i \boldsymbol{S}[i,j]$ 较大，此时我们降低它的 PageRank 值输出。

2.3. 弥合悬挂节点的不连续性

邻接矩阵 $\boldsymbol{H}_m$ 是按行归一化的。普遍的方法是，用每一行的元素之和去除该行所有元素，这样归一化之后矩阵的行和为 1。作为异常处理，全 0 行保持不变，对应节点标记为悬挂节点。悬挂向量 $\boldsymbol{d}$ 中，对应悬挂节点的分量取 1，其余取 0。这样，当节点在是否悬挂的状态之间切换时，会导致 PageRank 向量出现不连续变动。针对这个问题，文献中提出了邻接矩阵的另一种归一化方法[7]：取 ε 为一个小的常数，用 $\varepsilon + \sum_j \boldsymbol{H}_m[i,j]$ 去除 $\boldsymbol{H}_m$ 的第 i 行，对应的悬挂向量分量取为 $\varepsilon/\left(\varepsilon + \sum_j \boldsymbol{H}_m[i,j]\right)$。我们将采用后者，因为我们在连续变化的 PCC 上划定了一个阈值来提取邻接矩阵，故而希望弥合这一阈值带来的不连续性。

我们取 ε 为式(2)中分布的单边 0.049 分位点，再经式(3)映射得到的值。

2.4. DNB 子网络中的 PageRank 值

根据 DNB 理论，非临界时刻，模型网络是一个大而稀疏的整体，而临界时刻的模型网络可以分为两部分：一部分是小而紧密相关的 DNB 网络，另一部分是稀疏的非 DNB 网络，两者间相对独立。所以我们所要做的就是把一个小而紧密相关的子网络从大而稀疏的背景网络之中识别出来。

考察一个子网络中各节点的 PageRank 值之和 E，有如下关系[8]：

$$E=\left(P_G+\left(P_I-P_o\right)-\left(P_d-P_t\right)\right)/\left(1-\alpha\right) \tag{5}$$

P_G 是由个性化向量产生的 PageRank 值，P_I 和 P_o 分别是通过边输入和输出的 PageRank 值，P_d 和 P_t 分别是子网络中经由悬挂节点向全网逸散的 PageRank 值以及由全体悬挂节点逸散而来的 PageRank 值。对 DNB 子网络：由于内部紧密相关，所以 P_d 是 0；又由于 DNB 相对独立，P_I 和 P_o 都小，理想情况下是 0。P_t 实践中发现也不大，虽然 P_t 使得 E 增大，对我们的算法有利。这样 $E\approx P_G/\left(1-\alpha\right)=\sum_i^{\mathrm{DNB}}\boldsymbol{v}_n\left[i\right]$，$i$取遍 DNB 中所有节点。

我们希望 DNB 网络中的节点具有较高的 PageRank 值，从而可以将其挑选出来。因此，我们将个性化向量 $\boldsymbol{v}$ 取为 $\boldsymbol{v}\left[i\right]=\varepsilon+\sum_j\boldsymbol{H}_m\left[i,j\right]$。一方面我们要求 DNB 中节点的度大于全网络平均值，另一方面 DNB 节点中相关性也较大。这样 $\boldsymbol{v}$ 中对应 DNB 的分量较大，从而使得 DNB 子网络中 PageRank 值的均值较全网络高。

2.5. 渐进 PageRank 方法

我们仍然不能简单地通过选取 PageRank 值最高的节点来确定 DNB。对于较大的稀疏网络，一个普遍成立的经验规律是，PageRank 值分布在高端有幂次衰减的长尾[9]。这意味着尽管 DNB 节点具有高于平均的 PageRank 值，但是仍会被非 DNB 网络的 PageRank 值分布的长尾所淹没，因为 DNB 只占全网络的一小部分。因此，我们采取了另一种思路：循环地运行 PageRank 算法，不断舍去 PageRank 值低的节点来最终确定 DNB。DNB 节点的 PageRank 值总体偏大，虽然不足以直接选出，但是可以保证在筛选中幸存。DNB 中的节点，$\boldsymbol{v}$ 中对应分量较大，相邻节点的 PageRank 值也较大，所以具有较高 PageRank 值。非 DNB 网络中 PageRank 值较高的节点，来源于同时与大量低 PageRank 值节点的偶然连接。所以，当低 PageRank 值节点不断被舍去时，前者 PageRank 值依然较高，而后者不再具有较高的 PageRank 值。这样最终剩下来的节点就是 DNB。

最为稳妥的方式是，每次丢掉 PageRank 值最低的一个节点。但是出于计算上的实践，我们选取了一个阈值 $T_p=0.8$。每次运行 PageRank 算法后舍去 PageRank 值低于 $T_p/\min\left(n,T_n\right)$ 的节点；这里 n 是当前剩余节点数目，T_n 是为了确保循环过程能够停下来而设定的停止阈值，取为最初节点数目的 0.05。T_n 的作用是：当 n 小于 T_n 后逐渐放松筛选条件，直到没有节点被舍去时终止运算。

2.6. DNB 评价指标

通过上面的渐进 PageRank 算法，我们在每个时刻挑选出了一个 DNB。而后，我们需要建立一个 DNB 评价指标来选出最显著的 DNB 作为有效 DNB。我们选择 DNB 子网络中边的平均强度作为 DNB 评价指标：

$$I\overset{\mathrm{def}}{=}\sum_{i\neq j}^{\mathrm{DNB}}\boldsymbol{H}_r\left[i,j\right]/\left(n\left(n-1\right)\right) \tag{6}$$

这里 n 是 DNB 的节点数目。有效 DNB 所对应的时刻就是我们所探测到的临界时刻。

3. 结果

3.1. 数值模拟

我们首先用数值模拟来检验我们的方法。一般的动力学系统可以表述为 $\mathrm{d}\boldsymbol{x}/\mathrm{d}t = f(\boldsymbol{x};\beta)$；假设它初始收敛于 0，而后在参数 β 逐渐改变时失稳。考虑它在奇点 0 附近的稳定性，我们只需要讨论它在奇点 0 处的线性化系统 $\mathrm{d}\boldsymbol{x}/\mathrm{d}t = \boldsymbol{A}(\beta)\boldsymbol{x}$：当 $\boldsymbol{A}(\beta)$ 负定时系统稳定；当最大特征值接近 0 时，系统逐渐失稳。考虑一个有限的时间间隔，令为 1，有 $\boldsymbol{x}_{t+1} = \exp(\boldsymbol{A}(\beta))\boldsymbol{x}_t$；差分方程的形式便于生成模拟数据。再假设 $\boldsymbol{A}$ 可以对角化为 $\boldsymbol{T}\boldsymbol{D}(\beta)\boldsymbol{T}^{-1}$。其中 $\boldsymbol{T}$ 是随机生成的稀疏矩阵，与 β 无关。而对角矩阵 $\boldsymbol{D}(\beta)$ 中也仅有第一个元素 $\lambda(\beta)$ 依赖于 β；其余元素取随机生成的不太接近 0 的负值。在数值模拟的一系列时间点中，$\lambda(\beta)$ 逐渐趋近于 0。这样，在临界点时，$\lambda(\beta)$ 是主特征值，它无限趋近于 0；$\boldsymbol{T}$ 的第一列是主特征向量，它的非零元对应 DNB。我们通过如下迭代来生成模拟数据集合：

$$\boldsymbol{x} \leftarrow \exp\left(\boldsymbol{T}\boldsymbol{D}(\beta)\boldsymbol{T}^{-1}\right)\boldsymbol{x} + \boldsymbol{\xi} \tag{7}$$

$\boldsymbol{x}$ 是迭代变量，初值为全 0；$\boldsymbol{\xi}$ 是随机向量，各分量独立服从高斯分布，用于在迭代过程中引入随机性。在模拟数据集合中，我们的渐进 PageRank 算法成功定位了 DNB 并给出了临界预警信号。在图 1 中我们展示了 DNB 评价指标随主特征值λ的变化；在绝大多数模拟数据中，DNB 评价指标在 λ 趋于 0 时趋近于 1，成功地指示了临界点。

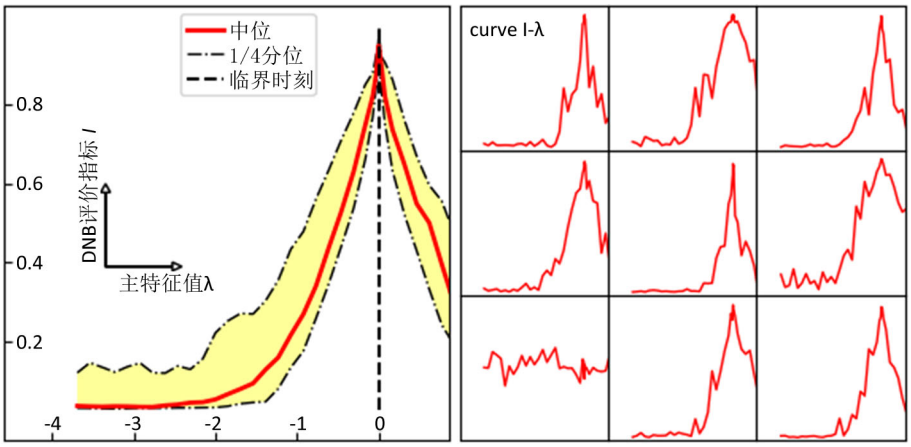

Figure 1. DNB indication for simulations (left: a synthesis result, right: specific curves)
图 1. 数值模拟的 DNB 评价指标曲线(左：综合展示，右：部分具体曲线)

3.2. 实验数据

接下来，我们将算法应用于两组实验数据中，复现了文献中报道的方法给出的结果[2] [3] [4]。第一个数据集合来源于光气引起小鼠肺部损伤的实验(GEO accession number: gse2565) [10]。在实验中，小鼠首先在光气中暴露 20 分钟，然后每间隔一段时间进行高通量表达测量。12 h 时刻观察到小鼠的死亡率在 50%~60%之间，24 h 时刻观察到小鼠的死亡率在 60%~70%之间，这意味着绝大多数死亡事件发生在前 12 h。我们的渐进 PageRank 算法确定的临界时刻是 8 h，见图 2。我们的结果与实验现象相符。

第二组数据源于 HRG 诱导 MCF-7 人体乳腺癌细胞分化的实验(GEO accession number: gse13009) [11] [12]。HRG 诱导的分化过程，与 EGF 诱导的增殖过程，在前 90 m 很相似，在 90 m 之后表现出显著不同。我们确定的临界时刻是 90 m，与实验现象相符，见图 3。

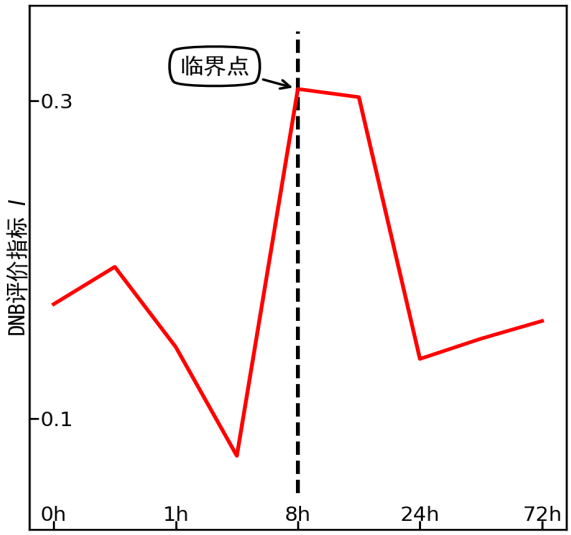

Figure 2. DNB indication for the experiment of mouse lung injury
图 2. 小鼠肺部损伤实验的 DNB 评价指标曲线

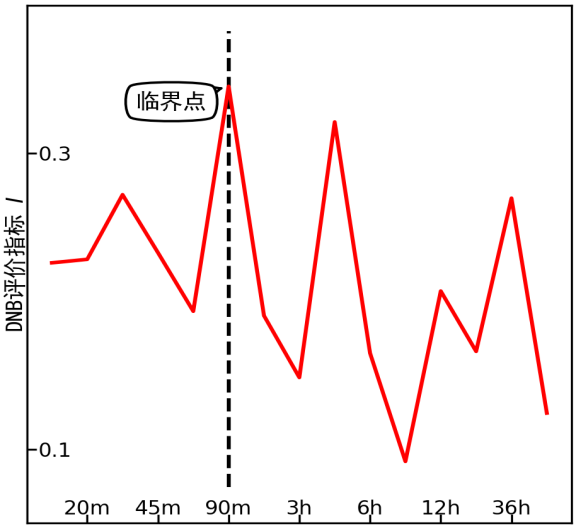

Figure 3. DNB indication for the experiment of HRG induced differentiation on MCF-7 human breast cancer cells
图 3. HRG 诱导 MCF-7 人体乳腺癌细胞分化实验的 DNB 评价指标曲线

4. 讨论

相比于文献中报道的基于 DNB 理论的临界点探测方法：在提取 DNB 的步骤中，我们用渐进 PageRank 算法替代了聚类算法；在评价 DNB 有效性的步骤中，我们用单纯基于波动相关性的指标，替代了同时基于波动相关性以及波动幅度的指标。因而在算法性能上，我们取得了三点显著改进。

首先是在 DNB 识别能力上的改进。DNB 理论中并没有对 DNB 的大小做具体规定，但是 DNB 占整个网络比值太小以至于不能识别，是实践中不得不面对的问题。文献中的解决方案是，在应用 DNB 模型之前，先依据差异表达以及波动幅度增大倍数等非网络方法来对高通量表达数据中的基因表达水平进行预筛选，以使得 DNB 模型涵盖的节点数目适合。文献中的预筛选非常激进，基因芯片给出的数目超过一

万的基因表达水平，经过预筛选后剩余不足一千，这意味着信息的大量流失。文献中 DNB 占模型网络比值普遍超过百分之二十，这意味着聚类方法的识别能力很弱。在我们的渐进 PageRank 方法中，DNB 占模型网络比值小于百分之十，这意味着我们在预筛选步骤中保留了更多的节点。这是相当大的进步，虽然距离我们舍去预筛选步骤，直接应用网络算法的目标还有相当大的距离。

第二方面的改进在于 DNB 评价指标中排除了对波动幅度的依赖。文献中的 DNB 评价指标是：$I_0 = \mathrm{SD} * \mathrm{PCC}_{\mathrm{in}} / \left(\varepsilon + \mathrm{PCC}_{\mathrm{out}}\right)$。其中，SD 是 DNB 中所节点波动标准差的均值，$\mathrm{PCC}_{\mathrm{in}}$ 是 DNB 中节点两两 PCC 的均值，$\mathrm{PCC}_{\mathrm{out}}$ 是 DNB 中的节点与 DNB 之外节点 PCC 的平均，ε 是一个防止分母为 0 的小量。由于每个基因本身的背景波动幅度不同；一个高通量表达数据中的基因表达水平 x，在参与运算之前，往往要经过一个 $x_m = \left(x - a\right)/b$ 形式的标准化，然后用 x_m 代替 x 参与后续运算，a 和 b 是随基因而异的。显然这样的标准化会直接影响 SD 的取值，但是不会影响 PCC 的取值，所以我们认为在 DNB 评价指标中排除了对波动幅度的依赖是一项改进。

第三，DNB 的特征只会在临界时刻表现出来；该文的方法对 DNB 网络中各个时间点都存在的固有结构进行了抑制；而在文献中基于聚类的方法中，每个时刻的数据是独立参与运算的。

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)。

参考文献

[1] Scheffer, M., Bascompte, J., Brock, W.A., *et al.* (2016) Early-Warning Signals for Critical Transitions. *Nature*, **461**, 53-59. https://doi.org/10.1038/nature08227

[2] Chen, L., Liu, R., Liu, Z.P., *et al.* (2012) Detecting Early-Warning Signals for Sudden Deterioration of Complex Diseases by Dynamical Network Biomarkers. *Scientific Reports*, **2**, Article No. 342. https://doi.org/10.1038/srep00342

[3] Liu, R., Wang, X., Aihara, K. and Chen, L. (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

[4] Chen, P., Li, Y., Liu, R. and Chen, L. (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

[5] Langville, A.N. and Meyer, C.D. (2011) Google's PageRank and Beyond: The Science of Search Engine Rankings. Princeton University Press, Princeton.

[6] Szklarczyk, D., Morris, J.H., Cook, H., *et al.* (2017) The STRING Database in 2017: QUALITY-Controlled Protein-Protein Association Networks, Made Broadly Accessible. *Nucleic Acids Research*, **45**, D362-D368. https://doi.org/10.1093/nar/gkw937

[7] Wang, X., Tao, T., Sun, J., *et al.* (2010) DirichletRank: Ranking Web Pages Against Link Spams.

[8] Bianchini, M., Gori, M. and Scarselli, F. (2005) Inside PageRank. *ACM Transactions on Internet Technology*, **5**, 92-128. https://doi.org/10.1145/1052934.1052938

[9] Becchetti, L. and Castillo, C. (2006) The Distribution of PageRank Follows a Power-Law Only for Particular Values of the Damping Factor. *Proceedings of the* 15*th International Conference on World Wide Web*, Edinburgh, 23-26 May 2006, 941-942. https://doi.org/10.1145/1135777.1135955

[10] 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

[11] Saeki, Y., Endo, T., Ide, K., *et al.* (2009) Ligand-Specific Sequential Regulation of Transcription Factors for Differentiation of MCF-7 Cells. *BMC Genomics*, **10**, 545. https://doi.org/10.1186/1471-2164-10-545

[12] Nagashima, T., Shimodaira, H., Ide, K., *et al.* (2006) Quantitative Transcriptional Control of ErbB Receptor Signaling Undergoes Graded to Biphasic Response for Cell Differentiation. *Journal of Biological Chemistry*, **282**, 4045-4056. https://doi.org/10.1074/jbc.M608653200