Advances in Applied Mathematics 应用数学进展, 2018, 7(1), 56-73
Published Online January 2018 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2018.71008

Hans汉斯

Data Mining and Early-Warning Model for the Sudden Deterioration of Complex Disease

Xin Lv, Rui Liu

School of Mathematics, South China University of Technology, Guangzhou Guangdong
Email: 1159155736@qq.com, 1622468926@qq.com

Received: Dec. 19th, 2017; accepted: Jan. 18th, 2018; published: Jan. 25th, 2018

Abstract

Data mining and early-warning signals of prostate cancer and liver cancer are by dynamic network biomarker method based on multi-samples or single-samples. It's vital to detect the critical point and signals of sudden deterioration, so as to diagnose the disease more accurately and put forward appropriate therapeutic plan in time. With time-course high-throughout biomolecular data, dynamic network biomarkers method based on multi-samples detected that the critical points of prostate cancer samples and liver cancer samples are the 6th time point and 2rd time point respectively, which agrees with the experiment data. In addition, 264,139 dynamical network biomakers including transcription factors were found. In fact, actual data are insufficient and the size of samples is small, and then dynamic network biomarkers method based on single-samples can be used to detect the early-warning of sudden deterioration. Also, the critical points of prostate cancer samples and liver cancer samples are the 6th time point and 2rd time point respectively based on single-samples. Finally, it shows that the found dynamic network biomakers based on multi-samples or single-samples could reflect the early-warning of sudden deterioration better after genes function analysis.

Keywords

High-Throughout Biomolecular Data, Network Analysis and Computation, Dynamic Network Biomarkers, Single-Samples Analysis

复杂疾病恶性突变的数据挖掘及预警模型

吕　欣，刘　锐

华南理工大学，数学学院，广东 广州
Email: 1159155736@qq.com, 1622468926@qq.com

文章引用：吕欣，刘锐．复杂疾病恶性突变的数据挖掘及预警模型[J]．应用数学进展, 2018, 7(1): 56-73.
DOI: 10.12677/aam.2018.71008

收稿日期：2017年12月19日；录用日期：2018年1月18日；发布日期：2018年1月25日

摘　要

利用基于多样本和基于单样本动态网络生物标志物法研究前列腺癌和肝癌两个时序列数据，检测疾病恶性突变的临界信号，确定动态网络生物标志物，进一步帮助医学工作者研究复杂疾病的发展变化机制，更高效准确诊断病情，及时提出合理的治疗方案。基于高通量生物分子数据，通过多样本动态网络生物标志法我们发现前列腺癌和肝癌样本分别在第6、2个时间点发生突变，与实验观测吻合，且分别有264，139个生物标记物。而实际样本数据并不完整且样本量少，此时需采用基于单样本动态网络生物标志物法检测疾病恶性突变信号，得到前列腺癌和肝癌样本分别在第6、2个时间点发生突变。最后对生物标记物进行生存分析等功能分析，发现这些标志物能较好的反映疾病临界变化信号。

关键词

高通量生物分子数据，网络分析与计算，动态网络生物标志物，单样本分析

Copyright © 2018 by authors and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

系统状态突变情况普遍存在于气候系统、经济系统、金融系统等，这样的突变往往发生在一个阈值，即临界点，此时系统突然从一种状态变化到另一种状态，这就是系统发生变化的动力系统分岔理论。最近，已有证据表明，在临床医学中存在类似的现象，如在很多复杂的疾病，类似癌症的慢性疾病的发展过程中恶化是突然的，疾病状态急剧转变。为了描述复杂疾病的潜在的动力学机制，它们的演化常常被建模为依赖于时间的非线性动力学系统，其中系统的突变被看作是分叉点上的相变，例如前列腺癌、哮喘发作和癫痫发作。我们对疾病的突然恶化期或临界转变点特别感兴趣。根据疾病的进展程度，我们将这一过程分为三个阶段：正常状态-临界状态-疾病状态。1) 在正常状态系统稳定，具有较强的回复力和鲁棒性；2) 临界状态是正常状态的临界情况，此时系统不稳定，回复力和鲁棒性弱，病人经过合理治疗可以恢复到正常状态阶段；3) 系统在疾病状态处于另一稳定状态，具有较强的回复力和鲁棒性，表示病人处于重病阶段，即便大力治疗也难以恢复到正常状态[1]。

对于复杂疾病而言，由于正常状态和临界状态的区别不大及疾病自身的复杂性，基于差异表达数据和其他统计指标，检测其临界信息是比较困难的但又是主要目的。通过建立网络，从动态变化视角观察复杂疾病的分子机制，对复杂疾病做出早期临床诊断以便提出及时有效的治疗方法。样本在每个时间点下的数据是高维的且有噪声，主成分分析法等传统统计学方法无法准确确定临界点。临界慢化法(*critical slowing-down*, *CSD*)为一般系统提供了检测临界点早期预警信息的方法之一，在金融，物理，生物等系统中已经成为研究热点，其应用范围愈来愈广[2]。*CSD* 法要求样本具有足够多的时序列数据，且变量需要包含反映动态临界慢化的变量。但是对于生物系统样本数据很难满足上述要求。基于高通量时序列样本数据，动态网络生物标记物法(*dynamical network biomarker*, *DNB*)有效地解决了 *CSD* 法的问题[1]。当样本量少时，基于单样本动态网络生物标志物法(*single-samples DNB*, *sDNB*)，能够准确找到临界点和生物标记物[3]。

2. 方法

2.1. 基于多样本动态网络生物标志物法

对于复杂疾病而言，由于正常状态和临界状态的区别不大及疾病自身的复杂性，检测其临界信息是比较困难的。基于多样本动态网络生物标记物法的提出解决了这些问题。一般系统达到临界点或分岔点时，至少有一组变量(动态网络生物标志物)与其他变量存在明显区别，即这组变量间的相关性强且极不稳定，满足以下三个条件[1]：

1) 动态网络生物标志物中元素的标准差均值(SD)大幅增加；

2) 动态网络生物标志物中元素间的皮尔逊相关系数的绝对值均值(PCC_{in})大幅增加；

3) 动态网络生物标志物中元素与非动态网络生物标志物中元素间的皮尔逊相关系数的绝对值均值 PCC_{out} 减少。

引入综合变量 CI ：

$$CI = \frac{SD \times PCC_{in}}{PCC_{out} + \varepsilon} \text{(} \varepsilon \text{ 是一个充分小的正数)}$$

显然， CI 值达到最大或突然大幅增加时即系统发生突变，达到临界状态，挑选出动态网络生物标志物。通过 T-检验， FDR， $fold\text{-}change$，聚类，显著性分析等步骤筛选一组相关性强且极不稳定的基因并确定临界点，具体步骤请详见附录 1。

2.2. 基于单样本动态网络生物标志物法

样本数据少甚至只有一个样本数据时，无法计算皮尔逊相关系数。基于单样本动态网络生物标志物法(*single-samples DNB*, *sDNB*)解决了这个问题。类似动态网络生物标志物法，此方法要求挑选出的动态网络生物标志物满足以下三个条件[3]：

1) 动态网络生物标志物中元素的表达偏差的绝对值均值(ΔED)大幅增加；

2) 动态网络生物标志物中元素间的皮尔逊相关系数差的绝对值均值(ΔPCC_{in})大幅增加；

3) 动态网络生物标志物中元素与非动态网络生物标志物中元素间的皮尔逊相关系数差的绝对值均值(ΔPCC_{out})大幅增加。

引入综合变量 ΔCI ：

$$\Delta CI = \frac{\Delta ED \times \Delta PCC_{in}}{\Delta PCC_{out} + \varepsilon} \text{(} \varepsilon \text{ 是一个充分小的正数)}$$

显然， ΔCI 值达到最大或突然大幅增加时即系统发生突变，达到临界状态，挑选出动态网络生物标志物。具体步骤请详见附录 2。

3. 主要结果

3.1. 基于多样本动态网络生物标志物法有关结果

将两种方法分别应用在前列腺癌数据(*GSE*5345)和肝癌数据(*GSE*80018)两种数据，数据下载自 *NCBI* 的 *GEO* 数据库(https://www.ncbi.nlm.nih.gov/geo/)。数据具体信息及处理过程请详见附录 3。

根据前列腺癌数据，利用基于多样本动态网络生物标志物法，得到综合指标 CI 的值在第 24 小时显著增加并达到峰值，动态网络生物标志物是第 164 类，共 264 个基因(含 47 个上游转录因子)，临界点是第 24 小时(图 1(a))，数据处理过程请详见附录 3.1。另外，根据前列腺癌的动态网络生物标志物及由 *STRING* 得到的蛋白质交互网络利用 *Cytoscape* 画出疾病基因动态变化过程和翻转网络，其中发生翻转即在疾病恶

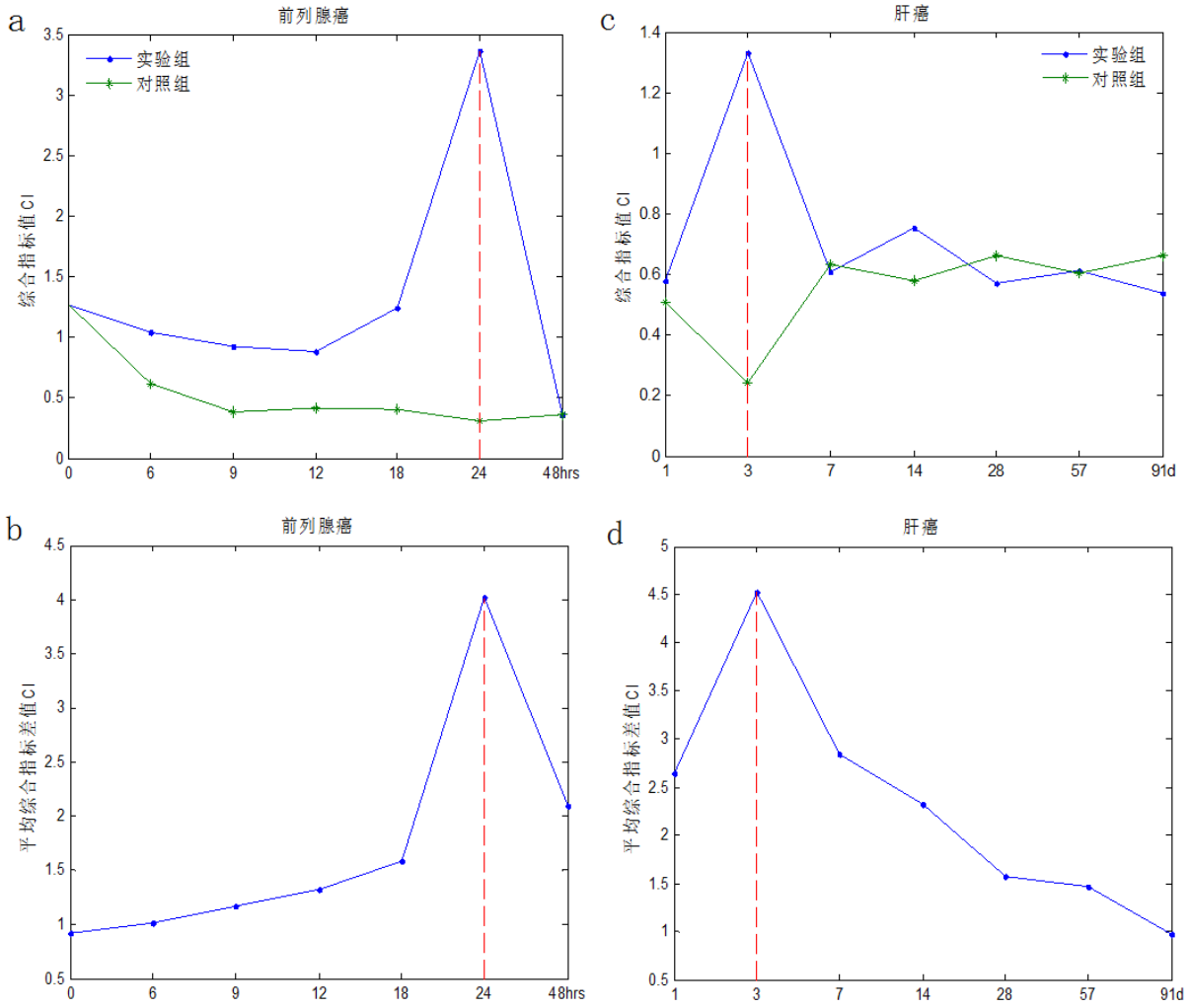

Figure 1. Line chart: early-warning signals based on multiple and single samples
图 1. 多样本和单样本临界信号折线图

性突变前高表达或低表达而在恶性突变后低表达或高表达的基因有 51.2%左右(图 2)。结果与实验观测吻合，用合成雄激素 *R*1881 刺激人类前列腺癌细胞系 *LNCap* 24 小时后，部分 *RNA* 表达水平下降 2~3 倍，部分表达水平上调 2~3 倍，甚至 3~6 倍[4]。

类似的，根据肝癌数据，计算综合指标 *CI*，发现其值在第 3 天显著增加并达到峰值，动态网络生物标志物是第 164 类，共 139 个基因(含 12 个上游转录因子)，临界点是第 3 天(图 1(c))，数据处理过程请详见附录 3.2。结果与实验观测吻合，给小鼠喂食含 0.05% [*wt*/*vol*]苯巴比妥饮用水 1 天后，检测到因细胞周期/有丝分裂的相关基因瞬时上调使得大部分的基因的转录水平发生变化，并且与细胞周期/干细胞调节机制的相关基因发生下调。与转录分析一致，给小鼠喂食含 0.05% [*wt*/*vol*]苯巴比妥饮用水 1 天后，观察到包括酶 *CYP*450 和还原酶 *POR* 在内的蛋白质表达水平都发生上调。最显著的组织病理学异常是第 7 天开始，观察到肝细胞肥大，并在以后的时间点更加严重[5]。

3.2. 基于单样本动态网络生物标志物法有关结果

当样本数据少，无法利用动态网络生物标志物法确定疾病恶性突变的临界点时，基于单样本动态网络生物标志物法能够依据少量样本数据检测临界信号确定临界点，即对两个数据的单样本综合指标差值

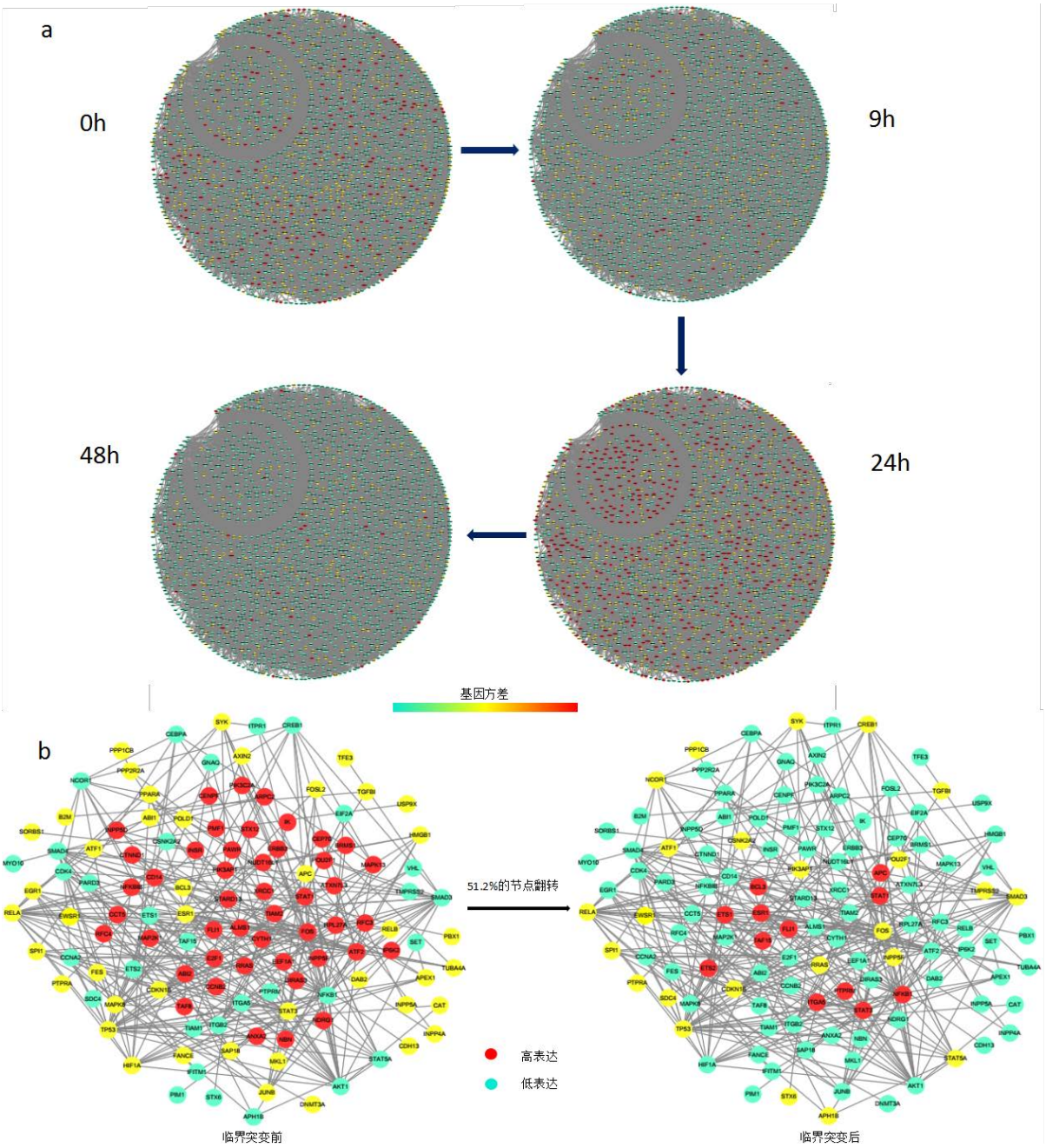

Figure 2. Dynamically changes in the network including the DNB and overturn network during the progression of prostate cancer

图 2. 前列腺癌圆盘图和翻转网络图

ΔCI 分别取均值确定临界点。根据前列腺癌数据，利用基于单样本动态网络生物标志物法检测 4 个样本都在第 24 小时发生突变(图 1(b))，及分别有 262，327，168，180 个生物标记物，其中包括上游转录因子，它们两两间交集分别有 103，38，37，37，84，45 个基因，并对这 183 基因做生存分析等功能分析。根据肝癌数据，利用基于单样本动态网络生物标志物法得到 5 个样本都在第 3 天发生突变(图 1(d))，及分别有 190，138，186，131，170 个生物标记物，它们两两间交集分别有 6，9，14，19，15，9，5，28，11 个基因，并对这 83 个基因做生存分析等功能分析，数据处理过程请详见附录 4。

3.3. 功能分析

用 *DAVID*6.8 (https://david.ncifcrf.gov/summary.jsp)对动态网络生物标志物做 *KEGG* 富集分析，前列腺癌和肝癌的动态网络生物标志物中的基因与癌症发展机制密切相关(表 1，只列出部分)。基于前列腺癌数据，*KEGG* 富集分析结果表明动态网络生物标志物与疾病的发展变化密切相关，如癌症通路、癌症转录失调、*PI3K-Akt* 信号通路、*MAPK* 信号通路、*ErbB* 信号通路、*Wnt* 信号通路等，其中大多与细胞生长、分化、增殖、凋亡，调节转录和翻译有关，*Wnt* 信号通路异常可以导致前列腺癌。肝癌的动态网络生物标志物中的基因与代谢通路、*MAPK* 信号通路、癌症通路、癌症转录失调等。*MAPK* 和 *PI3-K*/*Akt* 通路与细胞增殖和凋亡有关，并且 *MAPK* 信号通路，*ErbB* 信号，*Wnt* 信号通路等通路已被证实与多种癌症的发病机理相关。*KEGG* 富集分析表明利用动态网络生物标志物得到的结果是符合系统生物学的。

3.4. 生存分析

用 *SurvExpress* (http://bioinformatica.mty.itesm.mx:8080/Biomatec/SurvivaX.jsp)对由基于多样本动态网络标志物法得到的生物标志物做生存分析，可知当生物体内 *ANXA*6, *CEP*70, *CCNA*2, *C-MYC*, *E2F*-2, *FES*, *HOXB*6, *IFITM*1, *NCOR*1, *POLD*1, *RAB*11*FIP*1, *TP*53, *TRAK*1, *UBXN*11 这 14 个基因表达异常时，前列腺癌患者存活率下降，生存时间变短(图 3，请详见附录图 A1)；当 *AKAP*8, *COL*15*A*1, *ENO*2, *FUNDC*2, *HMGCS*2, *HOMER*1, *IHH*, *JUN*, *NDUFA*1, *PAM*, *RRAS*2, *TMEM*106*C*, *WIZ*, *WNT*1 这 14 个基因表达异常时，肝癌患者存活率下降，生存时间变短(图 4，请详见附录图 A2)。并且两个数据各自的动态网络标志物的生存分析的结果表明由基于多样本动态网络标志物法找到的生物标记物能够较好地反映各自癌症的临界突变信号。

由单样本得到的生物标志物，对每个单样本的生物标志物及所有两两间公共的生物标志物做生存分析，发现当 *AIM*1*L*, *ANXA*6, *CCNA*2, *C-MYC*, *E2F*-2, *ECT*2, *POLD*1, *RAD*51*AP*1, *TMEM*19, *TP*53 这 10 个基因表达异常时，前列腺癌患者存活率下降，存活时间变短(图 5，请详见附录图 A3)；当 *APOE*, *COL*15*A*1, *HMGCS*2, *JUN*, *OLIG*1, *RFK*, *UAP*1*L*1 这 7 个基因表达异常时，肝癌患者存活率下降，存活时间变短(图 6，请详见附录图 A4)。这些分析结果证实了我们得到的动态网络生物标志物和病人的生存时间存在很强的关联，即当动态网络生物标志物所含基因表达异常时，病人的预后更差。

Table 1. Enrichment analysis by KEGG
表 1. KEGG 富集分析

疾病	通路 ID	P 值	通路名称
肝癌	mmu05202	0.018	癌症转录失调
	mmu04010	0.018	MAPK 信号通路
	mmu04620	0.023	Toll 样受体信号通路
	mmu01100	0.038	代谢通路
	mmu04621	0.045	NOD 样受体信号通路
前列腺癌	hsa05202	6.48E−13	癌症转录失调
	hsa05200	2.74E−11	癌症通路
	hsa04668	3.38E−07	TNF 信号通路
	hsa05215	1.01E−04	前列腺癌
	hsa04068	1.29E−04	FoxO 信号通路

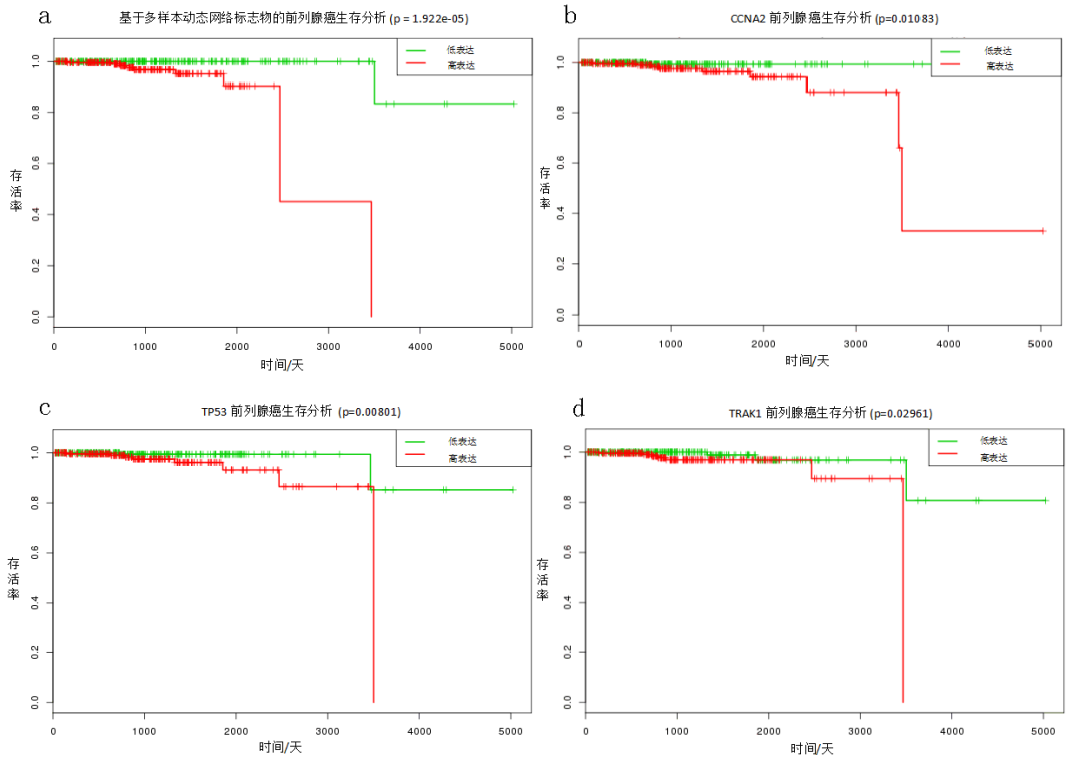

Figure 3. Survival analysis based on multi-samples of prostate cancer
图 3. 基于多样本动态网络生物标志物法的前列腺癌生存分析

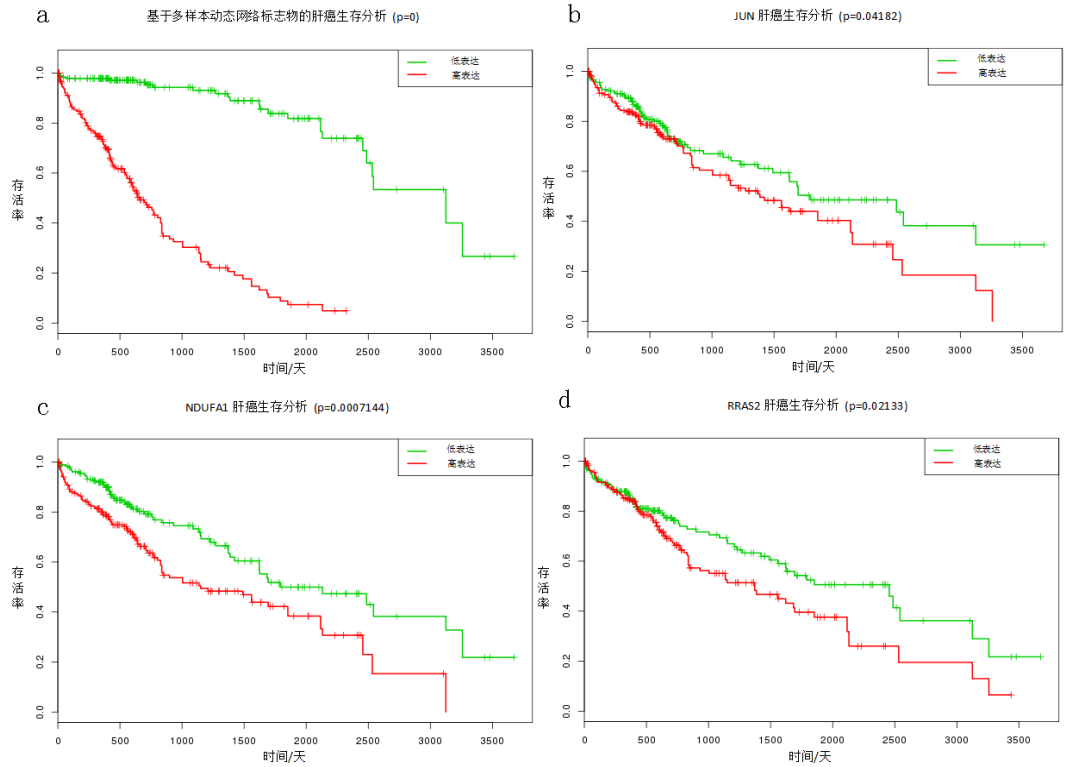

Figure 4. Survival analysis based on multi-samples of prostate cancer
图 4. 基于多样本动态网络生物标志物法的前列腺癌生存分析

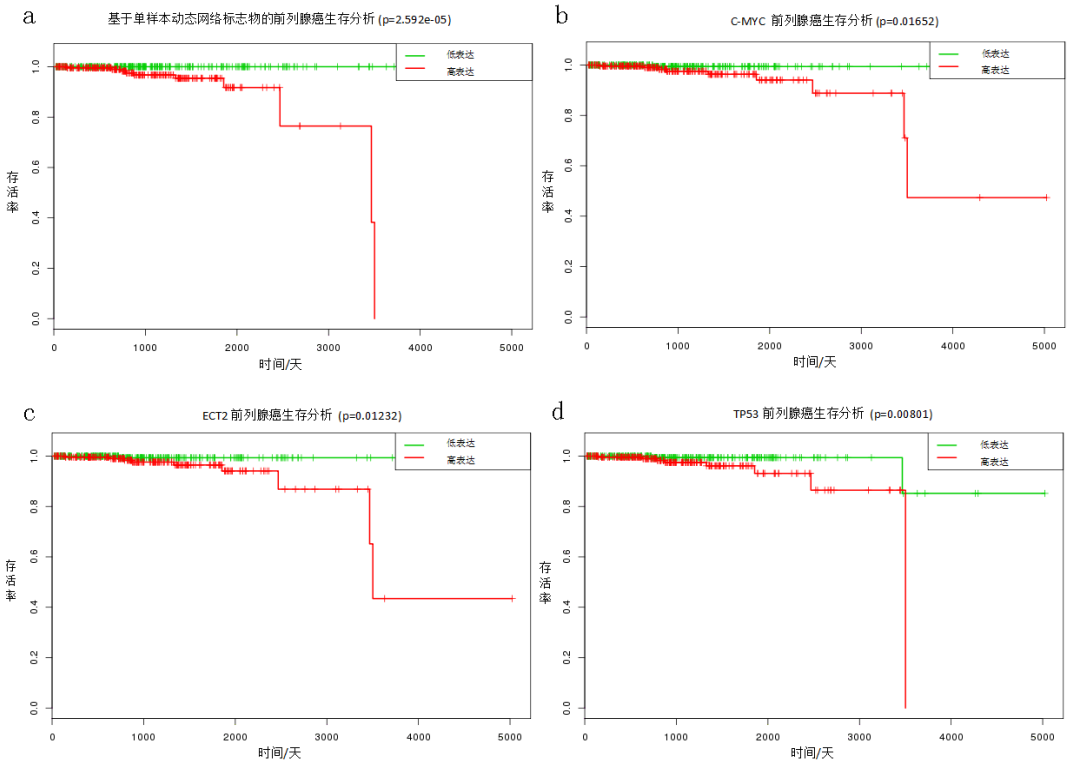

Figure 5. Survival analysis based on single-samples of prostate cancer
图 5. 基于单样本动态网络生物标志物法的前列腺癌生存分析

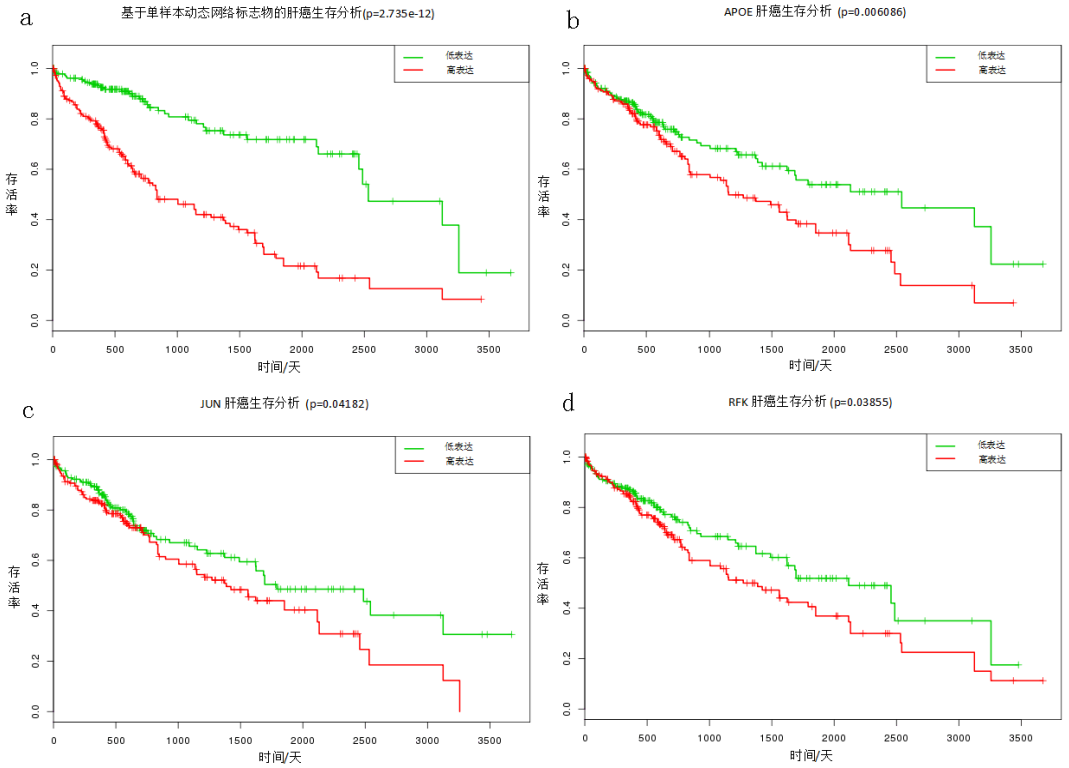

Figure 6. survival analysis based on single-samples of liver cancer
图 6. 基于单样本动态网络生物标志物法的肝癌生存分析

4 讨论

复杂疾病严重损害人类健康，我们通过检测疾病恶性突变的预警信号可以更好预防和治疗复杂疾病。虽然关键是通过检测疾病恶性突变的临界信号来采取适当的干预措施防止疾病继续恶化，但是在疾病的发展过程中，由于临界突变前系统可能没有显著变化，准确检测疾病的临界信号仍富有挑战。这也是基于传统生物标志物方法可能无法诊断疾病临界信号的原因。而动态网络生物标志物方法基于生物网络的动态性质，揭示了时序列数据在接近突变点时的临界信号，这有助于我们确定许多复杂疾病的新治疗方案的有效时间窗。

基于前列腺癌数据，利用动态网络生物标志物和生存分析等方法我们一共找到了 18 个基因(*AIM1L*, *ANXA*6, *CEP*70, *CCNA*2, *C-MYC*, *E2F*-2, *ECT*2, *FES*, *HOXB*6, *IFITM*1, *NCOR*1, *POLD*1, *RAB*11*FIP*1, *TP*53, *TRAK*1, *UBXN*11, *RAD*51*AP*1, *TMEM*19)，当它们表达过低或过高时，会使得患者存活率下降，生存时间变短。已有研究表明这 18 个基因多数与细胞增殖、分裂、凋亡等有关，当它们异常表达时可能会引起细胞癌变进而诱发癌症。如 *C-MYC* 能使细胞无限增殖，促进细胞分裂，引发细胞凋亡。*C-MYC* 与多种肿瘤发生发展有关，关于前列腺癌的大量研究表明多达 30%的癌症甚至在癌前病变阶段的 *C-MYC* 基因的拷贝数已经增加了[6] [7]，在前列腺肿瘤细胞中 *EGR* 过度表达可能激活 *C-MYC* 促进肿瘤的发展[8]。*TP*53 是抑癌基因，在许多人类癌症中发生突变，*TP*53 的错义突变、插入或缺失造成的失活突变非常常见，基因突变频率确实很高，尤其在前列腺细胞中[9]。*NCOR*1 与多种人类恶性肿瘤相关，它在调节各种核受体以及染色体重塑方面发挥重要作用，与膀胱癌及前列腺癌的发生、发展、预后及治疗的敏感性密切相关，*NCOR1* 在前列腺癌组织中的表达较癌旁组织明显升高[10]。

基于肝癌数据，利用动态网络生物标志物和生存分析等方法一共找到了 18 个基因(*AKAP*8, *APOE*, *COL*15*A*1, *ENO*2, *FUNDC*2, *HMGCS*2, *HOMER*1, *IHH*, *JUN*, *NDUFA*1, *OLIG*1, *PAM*, *RFK*, *RRAS*2, *TMEM*106*C*, *UAP*1*L*1, *WIZ*, *WNT*1)，当它们过高或过低表达时会使得患者存活率下降，生存时间变短。这 18 个基因已有多数被证实与癌症有关。如 *COL15A1* 的缺少与肌肉和微血管恶化有关，与癌旁组织相比，它在肿瘤中表达显著增加[11]。*ENO2* 是一个重要的肺癌肿瘤标志物，能作为生物标志物来帮助识别乳腺癌神经内分泌分化，与人类恶性肿瘤有关[12] [13]。*HMGCS*2 在肝脏中的表达与肝脏癌前病变和肝癌发展有关，细胞分化时 *HMGCS*2 表达增加，是 *C-MYC* 的直接靶节点，*C-MYC* 能抑制 *HMGCS*2 转录活性[14] [15]。*C-JUN* 通过拮抗 *P*53 的活性来抑制 *TNF-α* 诱导的细胞凋亡，进而促进肝肿瘤的发展[16]。

参考文献 (References)

[1] 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**, 342. https://doi.org/10.1038/srep00342

[2] Strogatz, S.H. (2000) Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry, and Engineering. Perseus Books Publishing, New York.

[3] Liu, X., Xiao, C., Rui, L., *et al.* (2017) Quantifying Critical States of Complex Diseases Using Single-Samples Dynamic Network Biomarkers. *PLoS Computational Biology*, **13**, e1005633. https://doi.org/10.1371/journal.pcbi.1005633

[4] 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

[5] 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*: *An Official Journal of the Society of Toxicology*, **131**, 375. https://doi.org/10.1093/toxsci/kfs303

[6] Nesbit, C.E., Tersak, J.M. and Prochownik, E.V. (1999) MYC Oncogenes and Human Neoplastic Disease. *Oncogene*, **18**, 3004-3016. https://doi.org/10.1038/sj.onc.1202746

[7] Jenkins, R.B., Qian, J., Lieber, M.M., *et al.* (1997) Detection of C-MYC Oncogene Amplification and Chromosomal

Anomalies in Metastatic Prostatic Carcinoma by Fluorescence *in Situ* Hybridization. *Cancer Research*, **57**, 524-531.

[8] Sun, C., Dobi, A., Mohamed, A., *et al.* (2008) TMPRSS2-ERG Fusion, a Common Genomic Alteration in Prostate Cancer Activates C-MYC and Abrogates Prostate Epithelial Differentiation. *Oncogene*, **27**, 5348-5353. https://doi.org/10.1038/onc.2008.183

[9] Chi, S.G., White, R.W.D., Meyers, F.J., *et al.* (1994) p53 in Prostate Cancer: Frequent Expressed Transition Mutations. *Journal of the National Cancer Institute*, **86**, 926-933. https://doi.org/10.1093/jnci/86.12.926

[10] Battaglia, S., Maguire, O., Thorne, J.L., *et al.* (2010) Elevated NCOR1 Disrupts PPARα/γ Signaling in Prostate Cancer and Forms a Targetable Epigenetic Lesion. *Carcinogenesis*, **31**, 1650. https://doi.org/10.1093/carcin/bgq086

[11] Lai, K.K., Shang, S., Lohia, N., *et al.* (2011) Extracellular Matrix Dynamics in Hepatocarcinogenesis: A Comparative Proteomics Study of PDGFC Transgenic and Pten Null Mouse Models. *PLoS Genetics*, **7**, e1002147. https://doi.org/10.1371/journal.pgen.1002147

[12] Clegg, N., Ferguson, C., True, L.D., *et al.* (2003) Molecular Characterization of Prostatic Small-Cell Neuroendocrine Carcinoma. *Prostate*, **55**, 55-64. https://doi.org/10.1002/pros.10217

[13] Tesori, V., Piscaglia, A.C., Samengo, D., *et al.* (2015) The Multikinase Inhibitor Sorafenib Enhances Glycolysis and Synergizes with Glycolysis Blockade for Cancer Cell Killing. *Scientific Reports*, **5**, 9149. https://doi.org/10.1038/srep09149

[14] Judson, R.S., Houck, K.A., Kavlock, R.J., *et al.* (2010) *In Vitro* Screening of Environmental Chemicals for Targeted Testing Prioritization: The ToxCast Project. *Environmental Health Perspectives*, **118**, 485. https://doi.org/10.1289/ehp.0901392

[15] Camarero, N., Mascaró, C., Mayordomo, C., *et al.* (2006) Ketogenic HMGCS2 Is a C-MYC Target Gene Expressed in Differentiated Cells of Human Colonic Epithelium and Down-Regulated in Colon Cancer. *Molecular Cancer Research*, **4**, 645. https://doi.org/10.1158/1541-7786.MCR-05-0267

[16] Eferl, R., Ricci, R., Kenner, L., *et al.* (2003) Liver Tumor Development. C-JUN Antagonizes the Proapoptotic Activity of p53. *Cell*, **112**, 181-192. https://doi.org/10.1016/S0092-8674(03)00042-4

附　录

1. 基于多样本动态网络生物标志物法

基于以上定理和结论，对于复杂生物系统的多样本数据或高维时间序列数据，利用基于多样本动态网络生物标志物法我们可以检测生物系统复杂疾病的动态网络生物标志物(DNB)和早期预警信号，为疾病的临床诊断和治疗做贡献。DNB 可以仅依据数据得到，数据具体处理过程如下：

Step 1：标准化

$$N=\frac{D_{\text{normal}}-\text{mean}\left(D_{\text{normal}}\right)}{SD\left(D_{\text{normal}}\right)} \tag{1}$$

$$D=\frac{D_{disease}-mean\left(D_{normal}\right)}{SD\left(D_{normal}\right)} \tag{2}$$

D_{normal} 表示正常样本数据；D_{disease} 表示疾病样本数据；N,D 分别表示正常样本数据和疾病样本数据标准化后的数据；$\text{mean}\left(D_{\text{normal}}\right)$，$SD\left(D_{\text{normal}}\right)$ 分别表示正常样本数据的均值和标准差。

Step 2：*T*-检验

在每个时间点，设置显著性水平 $p<0.05$ 或者 $p<0.01$ 对数据进行 *T*-检验，挑选出在正常样本和疾病样本间具有显著差异的基因。

Step 3：计算伪发现率 *FDR*

伪发现率(*FDR*, *false discovery rate*)是统计学名词，其意义是拒绝真的原假设的个数占所有被拒绝的原假设个数的比例的期望值，可以作为筛选出的差异变量的评价指标。*FDR* 的取值可以灵活调整，通常定为 0.05。设定一个显著性水平阈值 $q\left(0\le q\le 1\right)$，使得 *FDR* 被限制在某一固定水平，分两步进行：首先将所有检验的 p 值按升序排序，即 $p_1\le p_2\le\cdots\le p_n$；然后逐步后退比较 $p_i\le q\left(i=n,n-1,\cdots,1\right)$，取第一个满足条件的 p_i，理论上可以证明在此情况下可以将 *FDR* 控制在水平 q 之下。

Step 4：差异表达 *Fold-change*

利用标准差进一步筛选显著性基因。

Step 5：聚类

利用皮尔逊相关系数对基因聚类，每个类中基因间的相关性高。由以上理论知识可知，若一个时间点接近或是临界点，那么这个时间点的类中满足条件的基因即为备选 *DNB*，生物系统在这个时间点达到临界状态，疾病恶化。

Step 6：显著性分析

根据 *DNB* 的三个指标，确定 *DNB*。

计算综合指标值 *CI*，

$$CI=SD\frac{PCC_{in}}{PCC_{out}+\varepsilon} \tag{3}$$

其中，*SD* 是 *DNB* 中所有基因的标准差的均值，PCC_{in} 是 *DNB* 中基因间皮尔逊相关系数的绝对值的均值，PCC_{out} 是 *DNB* 中的基因与其他基因间皮尔逊相关系数的绝对值的均值，ε 是一个充分小的数，以保证整个分式有意义。

最后观察 SD,PCC_{in},PCC_{out},CI 的动态变化确定疾病恶化的早期预警信号和动态网络生物标志物。

2. 基于单样本动态网络生物标志物法

样本数据少甚至只有一个样本数据时，无法计算皮尔逊相关系数。基于单样本动态网络生物标志物

法(*single-sample DNB*, *sDNB*)解决了这个问题。类似动态网络生物标志物法，此方法要求挑选出的 *DNB* 满足正文 2.1 中的三个条件具体过程如下：

Step 1：计算表达偏差 ΔED 筛选基因

$$\Delta ED = \left| g - \bar{g} \right| \tag{4}$$

计算每个基因的表达偏差值，设置合适的阈值筛选基因。

Step 2：聚类

利用皮尔逊相关系数差值的绝对值对基因聚类，距离函数设为 $\left|2-\Delta PCC\right|$。

Step 3：计算 ΔPCC

假设有 n 个对照组，再加入一个测试样本后，计算每类基因组中的 ΔPCC，根据其显著性挑选显著表达的边。用 *Z*-检验计算边的显著性值 *p value*，设定阈值以筛选边，

$$\Delta PCC_{in}\left(g_1,g_2\right) = \left| PCC_{n+1}\left(g_1,g_2\right) - PCC_n\left(g_1,g_2\right) \right| \tag{5}$$

$$Z\left(g_1,g_2\right) = \frac{\Delta PCC\left(g_1,g_2\right)}{\left(1-PCC_n^2\left(g_1,g_2\right)\right)\Big/\left(n-1\right)} \tag{6}$$

Step 4：计算指标值 ΔCI：

$$\Delta CI = \frac{\Delta ED \times \Delta PCC_{in}}{\Delta PCC_{out} + \varepsilon} \tag{7}$$

依据条件筛选基因和边，确定疾病恶化的临界点和 DNB。

3. 基于多样本动态网络生物标志物方法的应用

从 *NCBI* 数据库下载两个个真实数据(http://www.ncbi.nlm.nih.gov/geo/)，找到每个指针对应的基因名，没有对应基因名的指针数据忽略不计，指向同一基因名的指针对应的数据取均值，数据具体处理过程如下。

3.1. 前列腺癌

从 *NCBI* 数据库下载前列腺癌数据 *GSE*5345。该实验数据是基于前人的实验工作得到的。简单地说，用合成雄激素 *R*1881(实验组)和酒精(对照组)刺激人类前列腺癌细胞系 *LNCap*，研究合成雄激素 *R*1881 对前列腺癌细胞系 *LNCap* 的基因表达的影响。每个时间点 4 个样本，研究人员对实验组和控制组分别在 0，6，9，12，18，24，48 小时(共 7 个时间点，表 S1)时分离 *RNA*，收集基因表达数据，一共 58 个样本数据。*NCBI* 没有提供实验组在 0 小时处的数据，这里用对照组在 0 小时处的数据代替。利用微阵列数据显著性分析发现至少有 3 个连续时间点对雄激素刺激反应有统计学意义的变化，实验的具体过程请详见原文。合成雄激素 *R1881* 是一种口服活性蛋白同化雄性类固醇，是被广泛应用于科学研究中的雄激素受体。

基于多样本动态网路生物标志物法，该数据的具体处理过程及结果如下：

Step 1：数据预处理

此数据具有 4224 个原始指针，忽略没有对应基因名的指针并对指向同一基因名的指针数据取均值，最后剩下 2674 个基因数据。根据 *T*-检验(显著性水平 $p<0.05$)挑选出 1004 个显著表达基因映射到 *STRING* 上得到 976 个基因和 31,496 条边，用以画动态网络生物图。按照公式(1)、(2)对实验组和对照组数据标准化处理。

Table A1. Descriptions of the two datasets
表 A1. 前列腺癌数据和肝癌数据

数据	描述
前列腺癌	合成雄激素 *R1881* 和酒精分别刺激人类前列腺癌细胞系 LNCap
样本点	0，6，9，12，18，24，48 小时
观测基因	2674 个
对照组	4 个样本
实验组	4 个样本
肝癌	给老鼠分别喂食含 0.05% *wt/vol PB* 的饮用水和不含 *PB* 的饮用水
样本点	1，3，7，14，28，57，91 天
观测基因	21495 个
对照组	5 个样本
实验组	5 个样本

Step 2：筛选显著性基因

在每个时间点对数据进行 *T*-检验(显著性水平 $p<0.05$)和 FDR 处理($k_i < controlsize(1)$)，在每个时间点筛选出的基因数都是 2674 个；设置差异表达倍数值为 4，在每个时间点得到的基因数为[0, 153, 139, 178, 301, 336, 74]。

Step 3：聚类

在每个时间点，利用皮尔逊相关系数对基因聚类，聚类数是 40。

Step 4：计算综合指标值 *CI*

依据公式(3)计算前列腺的综合指标值，确定临界点。

综合指标 *CI* 的值在第 24 小时显著增加并达到峰值，动态网络生物标志物是第 164 类，共 264 个基因，动态网路生物标志物中的基因在 12 小时处综合指标值 *CI* 开始骤增并在 24 小时处达到最大值，系统在 24 h 处发生突变，出现早期预警信号(图 1(a))，结果与实验观测吻合。

对两类数据所得的动态生物标志物用 *KEGG* 富集分析和生存分析，发现动态生物标志物中的基因与疾病发展密切相关(表 1，图 S1)。

3.2. 肝癌

从 NCBI 数据库下载肝癌数据 *GSE*80018。对公鼠 *B*6*C*3*F*1 分别喂食含 0.05% [*wt*/*vol*]苯巴比妥(*Phenobarbital*, *PB*)饮用水和不添加 *PB* 的正常饮用水，分别作为实验组和对照组。在实验时间第 1，3，7，14，28，57，91 天收集基因表达数据，每个时间点 5 个样本(表 S1)，研究 *PB* 对小鼠肝脏的致癌作用。

根据基于多样本动态网路生物标志物法的理论知识，该数据的具体处理过程及结果如下：

Step 1：数据预处理

此数据具有 39653 个原始指针，忽略没有对应基因名的指针并对指向同一基因名的指针数据取均值，最后剩下 21495 个基因数据。再根据(1)、(2)对实验组和对照组数据标准化处理。

Step 2：筛选显著性基因

在每个时间点对数据进行 *T*-检验(显著性水平 $p<0.05$)和 FDR 处理

($p_i(k_i) < \left(k_i / controlsize(1)\right) \times q, q = 0.05$)，筛选出的基因数为[6789, 3436, 1678, 1756, 3090, 973, 1770]；设

置差异表达倍数值为 2.5，在每个时间点得到的基因数为[788, 204, 96, 141, 110, 80, 100]。

Step 3：聚类

对以上的到的基因根据皮尔逊相关系数聚类,每个时间点基因分为 40 类。

Step 4：计算指标值 *CI*

根据(3)计算综合指标值，观察它的变化情况，确定临界点。

得到该数据的动态网络生物标志物是第 3 天处的第 4 类，共 139 个基因，包括 12 个上游转录因子，动态网络生物标志物中的基因在第 1 天的综合指标值开始骤增并在第 3 天处达到最大值，出现早期预警信号，即第 3 天是该肝癌数据的临界点(图 1(c))，结果与实验观测吻合。

对两类数据所得的动态生物标志物用 *KEGG* 富集分析和生存分析，发现动态生物标志物中的基因与疾病发展密切相关(表 1，图 S2)。

4. 基于单样本动态网络生物标志物方法的应用

基于充足样本数据，我们可以根据正文 2.1 的三个条件检测复杂疾病的早期预警信息，确定疾病恶化的临界点，为临床诊断和治疗做贡献。但是出于现实原因，我们无法得到充足完整的样本数据。可基于单样本数据，根据正常样本和病人在某个时间点的数据构建网络。数据具体处理过程如下：

4.1. 前列腺癌

根据公式(4)~(7)计算各个指标值，基于前列腺癌数据，以对照组的第 1~9 列为对照组，对于实验组中的 4 个测试样本设定表达偏差 ΔED 的阈值分别为 3.3，2.5，2.0，1.7；层次聚类数分别是 50，20，20，10；在每个时间点下的聚类数是[12, 5, 2, 14, 40, 270, 18]，[63, 243, 20, 42, 150, 304, 21]，[12, 5, 2, 14, 40, 270, 18]，[12, 5, 2, 14, 40, 270, 18]，得到动态生物网络标志物中的基因分别有 262，327，168，180 个(含上游转录因子)，它们两两间公共基因分别有 103，38，37，37，84，45 个。这里把每一列数据都看做是一个单样本，是不同个体在不同时间下测得的数据，于是将得到的综合指标差值 ΔCI 取均值，确定病人疾病恶性突变的临界信号，根据综合指标差值变化得知该数据的临界点是第 24 小时(图 1(b))。对这些公共基因(183 个)做生存分析知基因 *AIM*1*L*，*ANXA*6，*CCNA*2，*C-MYC*，*E*2*F*-2，*ECT*2，*POLD*1，*RAD*51*AP*1，*TMEM*19，*TP*53 表达过高或过低时会使得病人的存活时间变短，存活率下降(图 S3)。

4.2. 肝癌

基于肝癌数据，以对照组的第 27~35 列为基于单样本动态网络生物标志物方法中的对照组，对于实验组中的 5 个测试样本设定表达偏差 ΔED 的阈值分别为 3.4，3.6，2.8，3.0，3.5；层次聚类数都是 30；在每个时间点下的聚类数是[169, 210, 21, 101, 118, 5, 47]，[52, 162, 99, 13, 7, 2, 2]，[973, 207, 1146, 346, 41, 156, 430]，[523, 147, 189, 97, 117, 186, 229]，[72, 174, 13, 31, 8, 88, 34]。动态生物网络标志物中的基因分别有 190，138，186，131，170 个，并且其中分别包括 9，5，8，13，25 个上游转录因子，它们两两间公共基因分别有 6，9，14，19，15，9，5，28，14，17 个。这里把每一列数据都看做是一个单样本，是不同个体在不同时间下测得的数据，于是将得到的综合指标差值 ΔCI 取均值，确定病人疾病恶性突变的临界信号，根据综合指标差值变化得知该数据的临界点是第 3 天(图 S1(d))。对这些公共基因(82 个)做生存分析可知基因 *OLIG*1，*RFK*，*HMGCS*2，*UAP*1*L*1，*COL*15*A*1，*APOE*，*JUN* 表达过高或过低时会使得病人的存活时间变短，存活率下降(图 S4)。

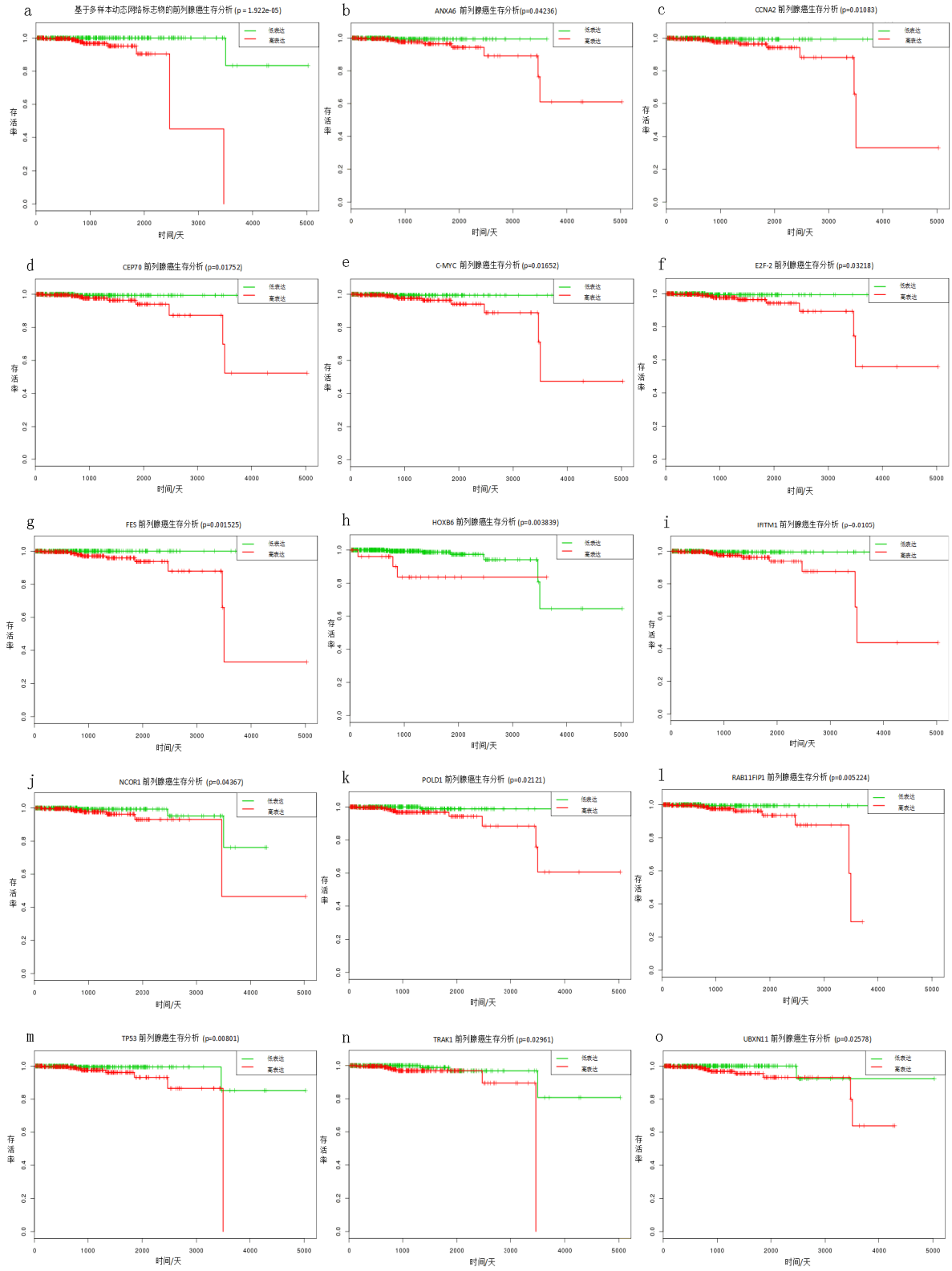

Figure A1. Survival analysis based on multi-samples of prostate cancer
图 A1. 基于多样本动态网络生物标志物法的前列腺癌生存分析

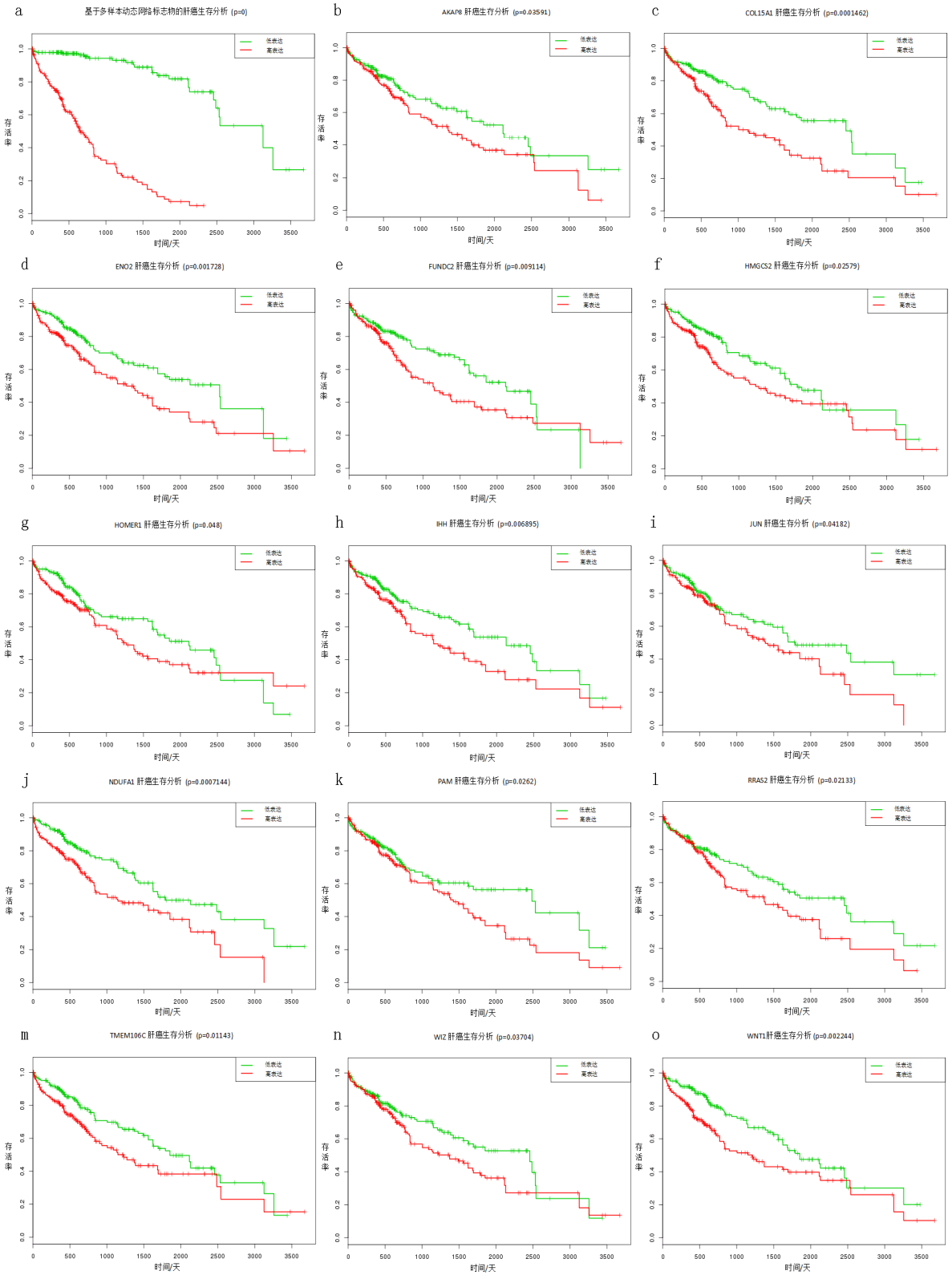

Figure A2. Survival analysis based on multi-samples of liver cancer
图 A2. 基于多样本动态网络生物标志物法的肝癌生存分析

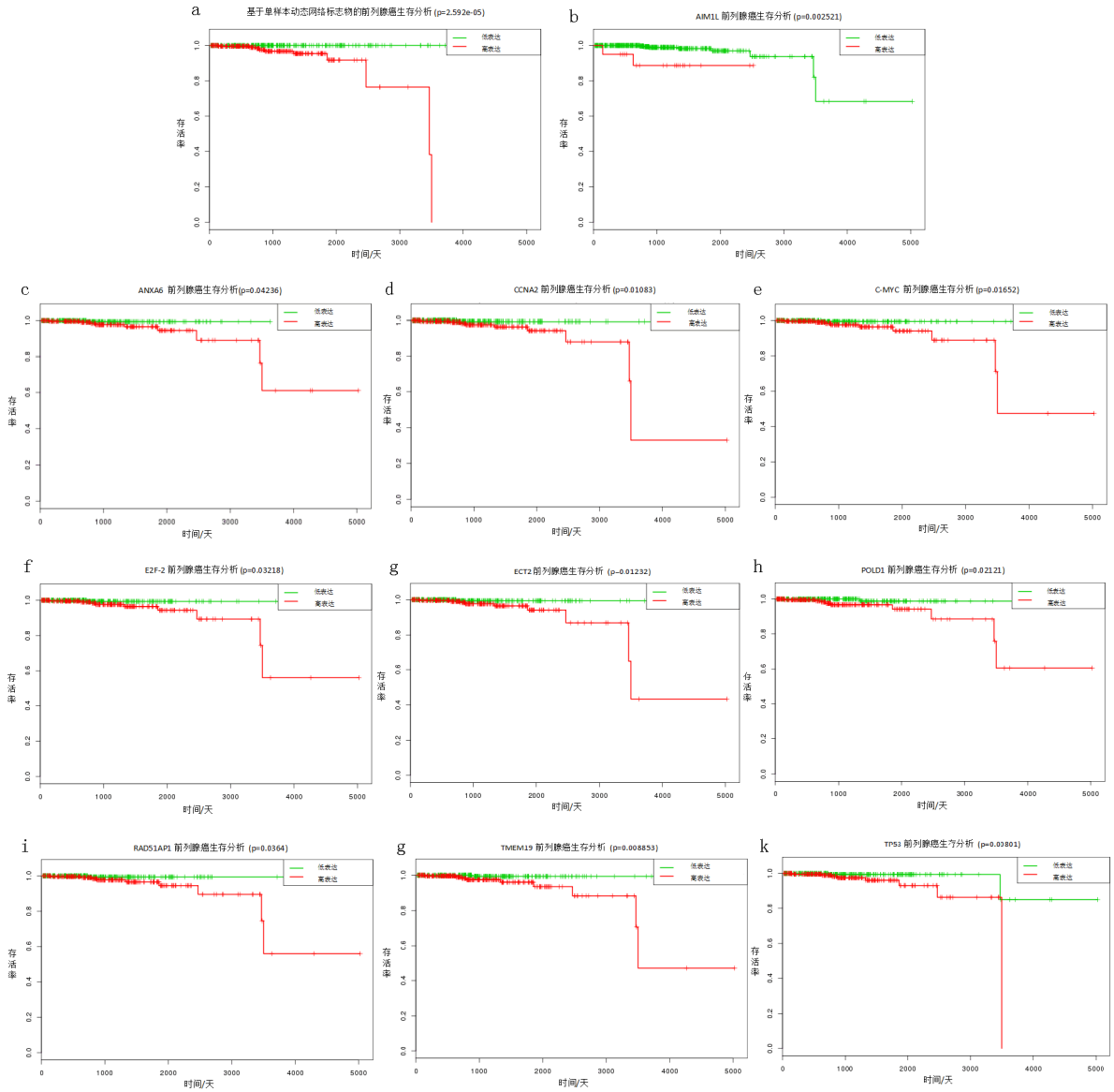

Figure A3. Survival analysis based on single-samples of prostate cancer
图 A3. 基于单样本动态网络生物标志物法的前列腺癌生存分析

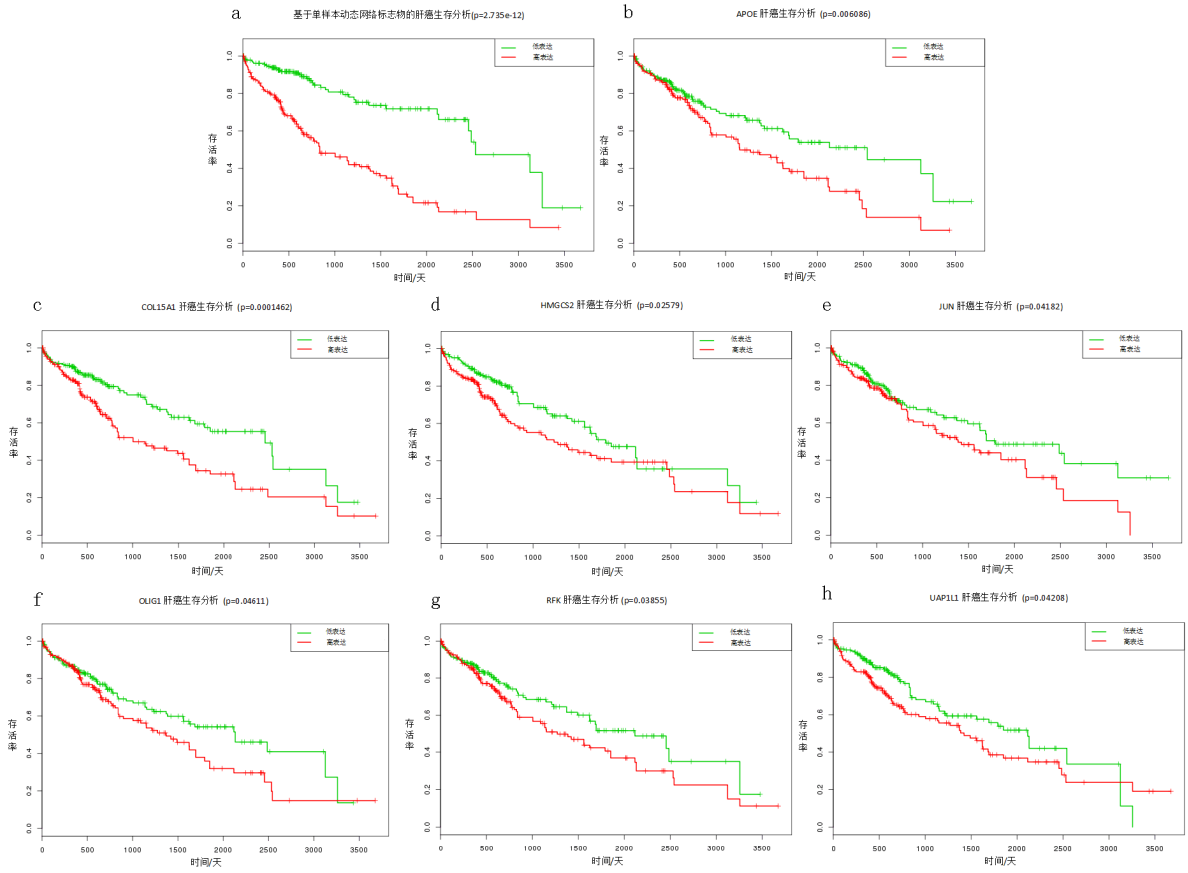

Figure A4. Survival analysis based on single-samples of liver cancer
图 A4. 基于单样本动态网络生物标志物法的肝癌生存分析

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2324-7991，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：aam@hanspub.org