第 58 卷 第 10 期 2009 年 10 月
1000-3290/2009/58(10)/6703-06

物 理 学 报
ACTA PHYSICA SINICA

Vol.58,No.10,October,2009

随机矩阵理论在肺癌基因网络识别中的应用*

李 蓉 颜平兰 陈 健 李 俊 李 金 张凯旺 钟建新†

(湘潭大学材料与光电物理学院,湘潭 411105)

(2008 年 8 月 1 日收到;2008 年 12 月 31 日收到修改稿)

利用随机矩阵理论(RMT)方法除去肺癌基因表达数据中的噪声,并将去噪后的数据分别用模块方法和等级聚类方法进行处理.比较两种方法处理后的结果,发现 RMT-等级聚类方法不仅能给出模块,还能给出模块间的关联强度.研究表明,RMT-等级聚类方法是一种有效的识别基因网络的新方法.

关键词:随机矩阵理论,等级聚类,基因网络,肺癌

PACC:0250,8710

1. 引 言

肺癌是目前世界上最常见的恶性肿瘤之一[1],它的发生发展与基因的改变有关,是一个多步骤多基因共同参与的过程[2—5].20 世纪 90 年代发展起来的基因芯片技术,能同时检测成千上万的基因,进行大批量的基因杂交处理,为发现基因间的相互关系提供了可能[6].如何从基因芯片技术得到的海量数据中,挖掘出隐含的生物信息,识别出基因网络,成为当前研究的热点.目前识别基因网络的方法有:布尔网络方法[7]、微分方程网络方法[8]、贝叶斯网络方法[9]、共表达网络方法[10]、聚类方法[11]等,但这些方法都存在着一定的不足.近年来,罗锋等[12,13]用随机矩阵理论(random matrix theory,简称 RMT)方法对生物网络进行去噪,结合模块方法,在不需要先验的生物知识的前提下成功地识别出基因网络模块、蛋白质网络模块和代谢网络模块.

我们利用 RMT 方法对肺癌基因表达数据进行去噪,通过标准误差方法确定体系的去噪参数,用模块方法[14]和等级聚类方法[15]处理去噪后的基因.研究表明,RMT-等级聚类方法不仅能得到基因模块,而且能进一步给出模块间的链接,是一种有效的识别基因网络的新方法.

2. 研究方法和研究对象

RMT 是 Wigner[16]为分析复杂核能谱而发展起来的一种统计理论.现已成功地应用于研究大原子光谱特性[17,18]、无序系统金属-绝缘体转变[19]、准周期体系光谱特性[20,21]、混沌体系[22,23]等领域.RMT 的一个重要的统计性质在于连续本征根的最近邻间隔分布(nearest-neighbour spacing distribution,简称 NNSD),对于一个代表时间反演不变性复杂体系的实对称随机矩阵,根据本征根关联性的强弱,它的分布有两个特点:代表富关联体系的随机矩阵,其本征根的 NNSD 遵循高斯分布;代表贫关联体系的随机矩阵,其本征根的 NNSD 遵循 Poisson 分布.当基因网络用随机矩阵理论方法进行研究时,根据上述性质,从基因网络变换而来的体系如遵循高斯分布,则体系中的元素存在着强弱关联的共同作用,含有因噪声引起的随机信息;如体系遵循 Poisson 分布,则体系间的元素仅有强相互作用.这样,我们通过移去小的关联系数来利用 RMT 方法除去基因网络中包含的噪声,得到真实基因网络.

高斯分布的 NNSD 近似服从 Wigner-Dyson 分布

$$p_{\mathrm{GOE}}(s) \approx \frac{1}{2}\pi s \exp\left(\frac{-\pi s}{4}\right), \tag{1}$$

Poisson 分布的 NNSD 服从

$$p_{\mathrm{Poisson}} = \exp(-s), \tag{2}$$

* 国家自然科学基金(批准号:30570432)资助的课题.

† 通讯联系人. E-mail:jxzhong@xtu.edu.cn

其中 Wigner-Dyson 分布和 Poisson 分布最显著的区别在于

$$p_{\mathrm{GOE}}(s \to 0) = 0,$$

$$p_{\mathrm{Poisson}}(s \to 0) = 1.$$

通过 Pearson 关联系数公式将肺癌基因表达矩阵转化成实对称肺癌基因关联矩阵.基因 g_i 和 g_j 之间的 Pearson 关联系数由如下公式计算得到：

$$c(g_i, g_j) = \frac{1}{n}\sum_{k=1,n}\left(\frac{g_{ik} - m_{g_i}}{\sigma_{g_i}}\right)\left(\frac{g_{jk} - m_{g_j}}{\sigma_{g_j}}\right), \quad (3)$$

其中 m_{g_i}，m_{g_j} 为基因 g_i 和 g_j 的平均值；σ_{g_i} 和 σ_{g_j} 为基因 g_i 和 g_j 的标准误差；n 为总基因个数.

肺癌基因数据下载的网址：http://www.camda.duke.edu/camda03/datasets/.所用的数据为该网址中斯坦福肺癌基因表达数据.

3. 结 果

首先用 RMT 方法对肺癌基因表达数据进行处理，除去噪声的干扰，然后分别用模块方法和等级聚类方法处理去噪后的基因，比较发现 RMT-等级聚类方法能得到真实模块且能给出模块间关联的强弱.

3.1. RMT 方法去噪

我们用 KNN 方法[24]对肺癌基因表达矩阵中所含的缺失值进行了处理；然后，通过 Pearson 关联系数公式将去噪处理后的表达矩阵转化成需要的关联矩阵；最后，设置去噪参数 $q(0 < q < 1)$，并计算每个去噪参数 q 值对应本征根的 NNSD.研究发现，随着 q 值的增大，在移去关联矩阵中小的关联系数后，关联矩阵本征根的 NNSD，从 Wigner-Dyson 分布逐渐过渡到 Poisson 分布，如图 1 所示.

我们分别计算肺癌基因关联矩阵本征根的 NNSD 对高斯分布和 Poisson 分布的标准误差，来确定体系由 Wigner-Dyson 分布过渡到 Poisson 分布的临界点及去噪参数.标准误差的公式定义如下：

$$SD_{\mathrm{GOE}}(q) = \sqrt{\frac{\sum_{i=1}^{m}(p(i) - P_{\mathrm{GOE}}(i))^2}{m-1}}, \quad (4)$$

$$SD_{\mathrm{Poisson}}(q) = \sqrt{\frac{\sum_{i=1}^{m}(p(i) - P_{\mathrm{Poisson}}(i))^2}{m-1}}, \quad (5)$$

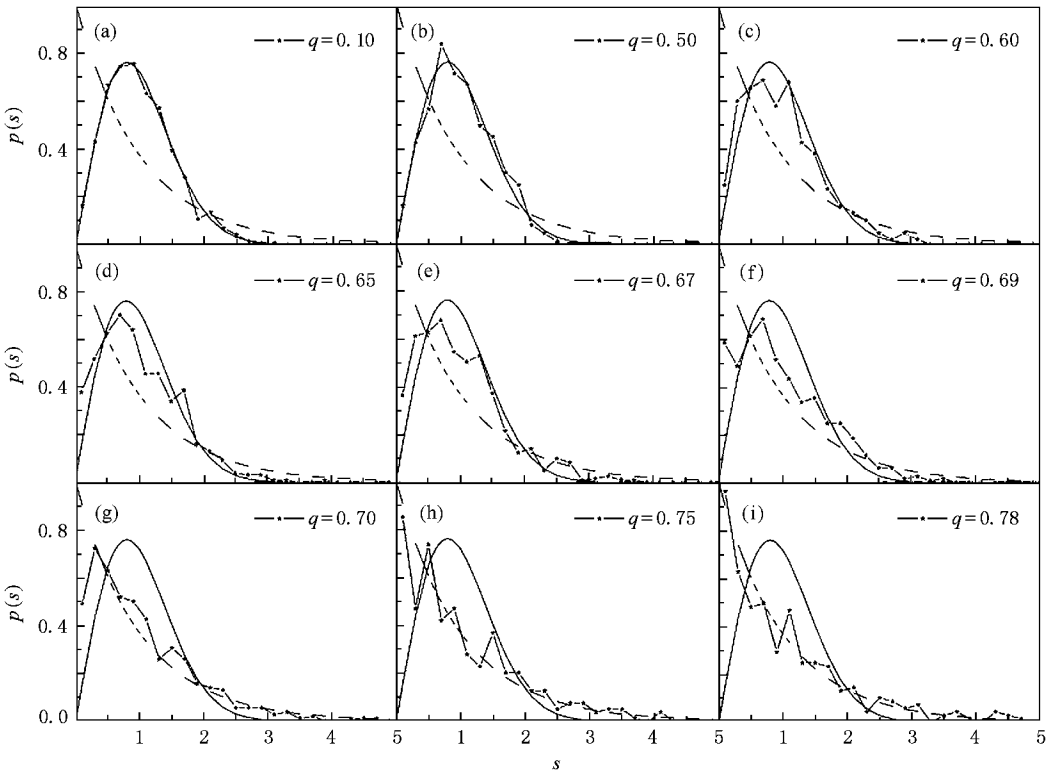

图 1 肺癌基因本征根的 NNSD 在不同 q 值下的分布 虚线表示 Poisson 分布，实线表示 Wigner-Dyson 分布，q 值为去噪参数.(a) $q = 0.10$，(b) $q = 0.50$，(c) $q = 0.60$，(d) $q = 0.65$，(e) $q = 0.67$，(f) $q = 0.69$，(g) $q = 0.70$，(h) $q = 0.75$，(i) $q = 0.78$

其中 $SD_{GOE}(q)$，$SD_{Poisson}(q)$ 分别代表本征根的 NNSD 对 Wigner-Dyson 分布和 Poisson 分布的标准误差．$p(i)$ 为第 i 个点对应本征根的 NNSD，$P_{GOE}(i)$，$P_{Poisson}(i)$ 分别为第 i 个点对应 Wigner-Dyson 分布和 Poisson 分布本征根的 NNSD．

当肺癌基因关联矩阵本征根的 NNSD 对高斯分布和 Poisson 分布的标准误差相等时，体系开始由 Wigner-Dyson 分布过渡到 Poisson 分布：此时体系中元素间的相互作用主要是强相互作用，元素间真实关联占主体地位，但还留有一些随机信息．当肺癌基因关联矩阵本征根的 NNSD 对高斯分布和 Poisson 分布的标准误差比值最大时，体系尽可能地偏离高斯系统而接近 Poisson 系统：此时体系保留了有用信息且充分地去掉了噪声．因此，我们选取肺癌基因关联矩阵本征根的 NNSD 对高斯分布和 Poisson 分布的标准误差比值最大时的点为去噪点．计算得到，在 $q = 0.69$ 处，肺癌基因关联矩阵本征根的 NNSD 开始由 Wigner-Dyson 分布过渡到 Poisson 分布（图 2(a)）．$q = 0.78$ 处，肺癌基因关联矩阵本征根的 NNSD 对高斯分布和 Poisson 分布的标准误差比值最大（图 2(b)），此时体系偏离高斯系统而接近 Poisson 系统，即 0.78 为 918 * 73 的肺癌基因体系的去噪点．

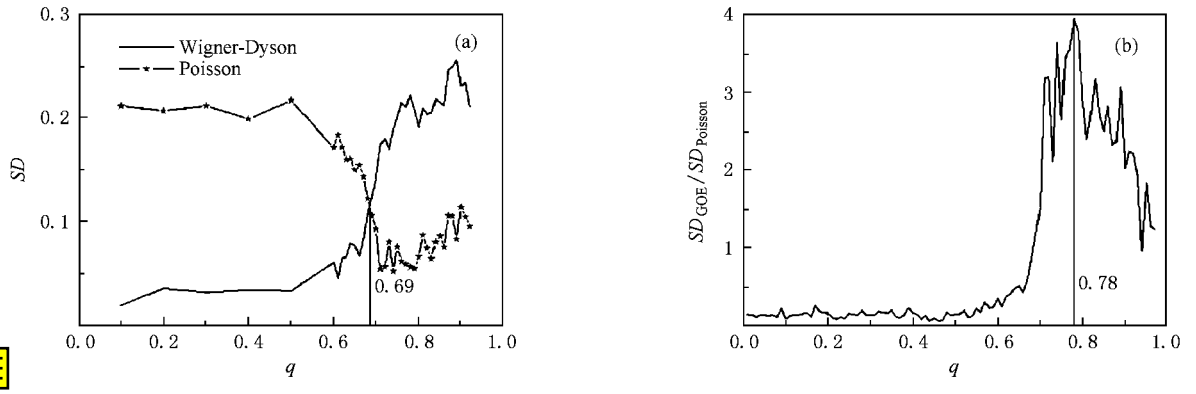

图 2　用标准误差方法得到的临界点和去噪点　(a)临界点，(b)去噪点

3.2. 采用 biolayout 软件观测肺癌基因模块

我们采用 biolayout 软件观测去噪前后基因模块结构的变化，结果如图 3 所示．图 3 中，每个节点代表一个基因，节点间的连线代表基因间不为 0 的关联，所有关联系数小于 q 值的都被视为 0．从图中可以看出，随着 q 值的增大，聚集在一起的基因逐渐分化出独立的模块，基因模块在 $q = 0.78$ 处明显呈现，即随着噪声的移除，真实的基因模块呈现出来．结合罗锋等[12,13]的工作，说明了 RMT 方法在生物模块识别中具有普适性．0.78 对应的基因网络模块是本文所研究基因对应的真实模块．

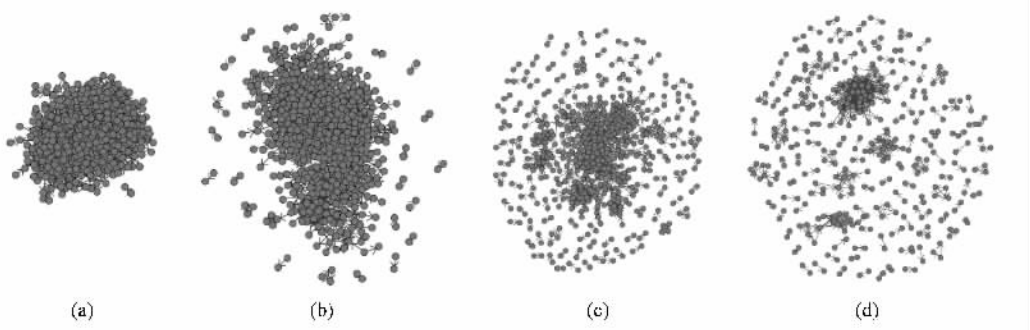

图 3　不同 q 值对应的基因模块　(a) $q = 0.50$，(b) $q = 0.60$，(c) $q = 0.69$，(d) $q = 0.78$

我们将 $q = 0.78$ 对应的网络模块放大并进行标记，如图 4 所示．去噪后系统保留的基因个数为 361 个，其中基因数小于 5 的模块有 101 个；大于或等于 5 的模块有 9 个；方框 1 到方框 9 为基因数目大于或等于 5 的 9 个模块；方框 10 中的模块为基因数目小于 5 的所有模块．从图 4 中我们可以清楚地看到真

实的基因网络模块.

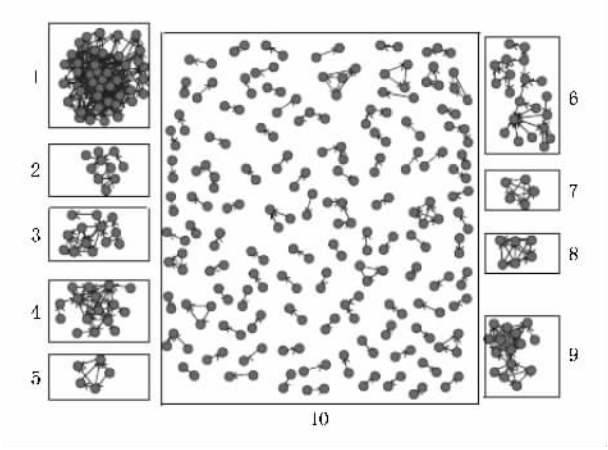

图 4 $q = 0.78$ 对应的 361 个基因构成的网络模块图

3.3. 采用等级聚类方法处理去噪后的肺癌基因

基因网络由众多的模块构成,且受模块间相互作用调控,模块内部基因间的关联强度很强,模块间基因的关联强度极弱.输入 biolayout 软件的数据,为大于或等于对应 q 值的关联系数及相应的基因,小于 q 值的关联系数就被去掉了.这样体系在去除噪声的同时也去掉了小于去噪参数的基因模块间的弱关联,因此无法表现基因模块间的联系,具有一定的缺陷.

为了解决这个问题,我们将与其他基因间关联系数大于或等于去噪参数 0.78 的基因挑选出来,构成一个新的肺癌基因表达矩阵.这样去噪过程中去掉的就仅仅是那些与其他基因存在弱关联的基因.而与其他基因存在强关联的基因及由这些强关联基因构成的模块间的弱关联被保留下来.在这个新的肺癌基因表达矩阵中保留了 361 个基因,我们没有去掉 361 个基因间任何的关联,因而新矩阵中的基因包含了基因间全部的信息.

我们用等级聚类方法处理新基因表达矩阵,聚类方法为凝聚法,相似性度量方法为 Pearson 关联,相似性指标为全连接方法,软件为 cluster-treeview,得到图 5(a)所示的树图.树图中的连线代表了基因间的关联,树枝的长度代表基因间关联的强弱.基因间的关联程度越强连线越短,关联程度越弱连线越长.

图 5(a)中标号 1 到标号 9 中的分支,分别对应

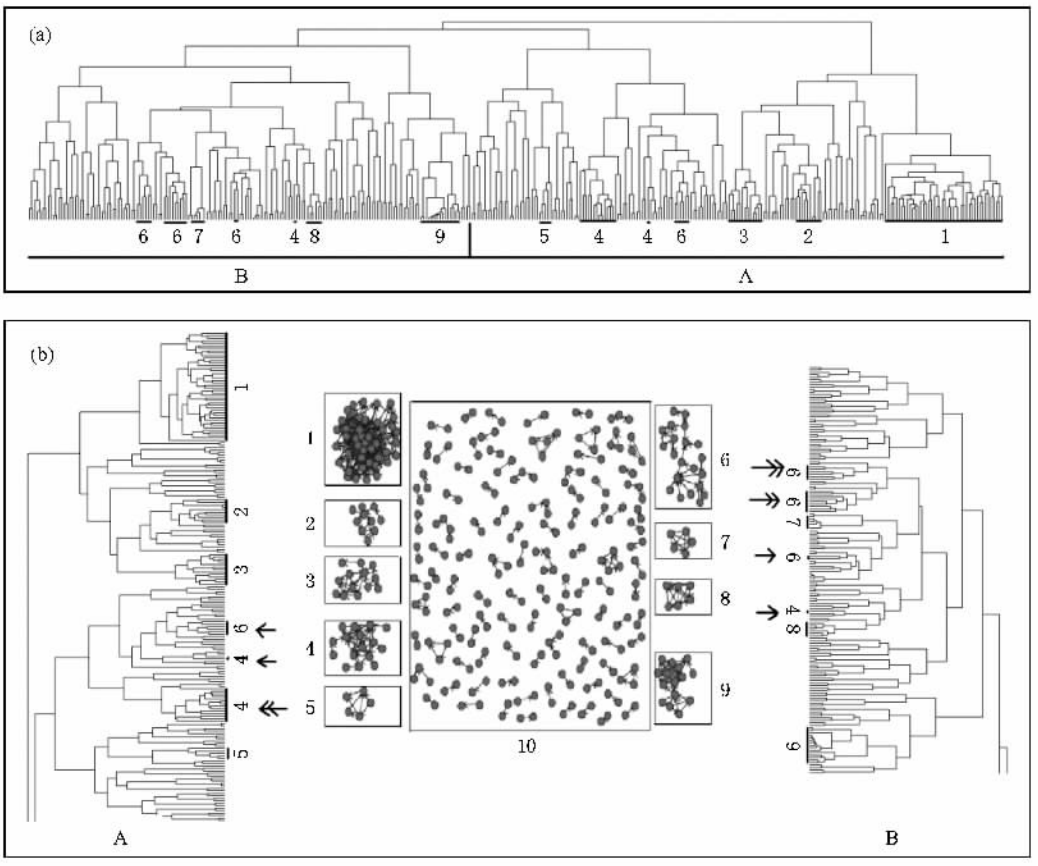

图 5 361 个基因构成的树图及模块图与树图之间的对应关系 (a)树图,(b)模块图和树图间的对应关系

图 4 中方框 1 到方框 9 中的模块；未标号的分支对应图 4 方框 10 中的模块．为了方便比较，将图 5(a)所示的树图分为 A，B 两部分，如图 5(b)所示．将树图结构与模块结构进行比较后发现：模块图中只有 10 个基因在树图中构成的分支偏离了各自所对应的主分支，这 10 个基因分别是：模块 4 中的 ARHGDIB，MNDA 基因，模块 6 中的 AQP3，ESTs，LDB2，EDG1，EDNRB，EDNRB，AGER，ADH2 基因(如图 5(b)中单箭头所示的分支里的基因)．这 10 个基因分别偏离了对应的主分支 4 和主分支 6(如图 5(b)中双箭头所示的分支)；除上述 10 个基因外，模块图中其他的属于同一模块中的基因在树图中属于同一个分支．统计表明，两种方法得到的结果相似度达到了 97%，即树图结构也可以给出基因的真实模块．相对于模块方法，从图 5(a)中不仅可以清楚观察到模块内部基因间的关联，而且可以有效地观察到模块与模块间的相互关联，并且根据树枝的长短得到基因间及模块间关联的强弱，识别出基因网络．RMT-等级聚类方法是一种有效的识别基因网络的新方法．

4. 结　论

采用 RMT 方法对肺癌基因表达数据进行去噪处理，通过标准误差的方法确定体系的去噪参数，用模块方法观测去噪后的基因，得到一系列肺癌基因模块，进一步证明了 RMT 方法在生物模块识别问题中的普适性．我们再用等级聚类方法处理去噪后的基因，并把两种方法得到的结果进行比较，发现两者所得到的模块相似程度达 97%．且 RMT-等级聚类方法相对于 RMT-模块方法在模块连接表示上更有效，它进一步给出了模块间的关联强度，能更真实地反映基因网络的信息，是一种有效识别基因网络的新方法．

[1] Greenlee R T, Hill-Harmon M B, Murray T, Thun M 2001 *CA-Cancer J. Clin.* **51** 15

[2] Wigle D A, Jurisica I, Radulovich N, Pintilie M, Rossant J, Liu N, Lu C, Woodgett J, Seiden I, Johnston M, Keshavjee S, Darling G, Winton T, Breitkreutz B, Jorgenson P, Tyers M, Shepherd F A, Tsao M S 2002 *Cancer Res.* **62** 3005

[3] Beer D G, Kardia S L R, Huang C, Giordano T J, Levin A M, Misek D E, Lin L, Chen G, Gharib T G, Thomas D G, Lizyness M L, Kuick R, Hayasaka S, Taylor J M G, Iannettoni M D, Orringer M B, Hanash S 2002 *Nat. Med.* **8** 8

[4] Bhattacharjee A, Richards W G, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark E J, Lander E S, Wong W, Johnson B E, Golub T R, Sugarbaker D J, Meyerson M 2001 *Proc. Natl. Acad. Sci.* **98** 13790

[5] Garber M E, Troyanskaya O G, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, Rijn M, Rosen G D, Perou C M, Whyte R I, Altman R B, Brown P O, Botstein D, Petersen I 2001 *Proc. Natl. Acad. Sci.* **98** 13784

[6] Li Y 2006 *Numerical Analysis and Processing of Gene Chip* (Beijing: Chemical Industry Press) (in Chinese) [李　瑶 2006 基因芯片数据分析与处理(北京：化学工业出版社)第 11 页]

[7] Akutsu T, Miyano S, Kuhara S 1999 *Pac. Symp. Biocomput.* **4** 17

[8] Chen T, He H, Church G 1999 *Pac. Symp. Biocomput.* **4** 29

[9] Wilkinson D J 2007 *Brief. Bioinform.* **8** 109

[10] Allocco D J, Kohane I S, Butte A J 2004 *Bioinformatics* **5** 18

[11] Törönen P, Kolehmainen M, Wong G, Castrén E 1999 *FEBS Lett.* **451** 142

[12] Luo F, Zhong J X 2006 *Phys. Lett.* A **357** 420

[13] Luo F, Zhong J X 2006 *Phys. Rev.* E **73** 031924

[14] Enright A J, Ouzounis C A 2001 *Bioinformatics* **17** 853

[15] Eisen M B, Spellman P T, Brown P O, Botstein D 1998 *Proc. Natl. Acad. Sci. USA* **95** 14863

[16] Wigner E P 1967 *SIAM Review* **9** 1

[17] Wigner E P 1951 *Proc. Cambridge Philos. Soc.* **299** 189

[18] Chen Z Q, Zheng R R, Chen H, Yao C Q 2000 *Acta Phys. Sin.* **49** 969 (in Chinese) [陈志谦、郑仁蓉、陈　洪、姚纯青 2000 物理学报 **49** 969]

[19] Hofstetter E, Schreiber M 1993 *Phys. Rev.* B **48** 16979

[20] Zhong J X, Grimm U, Romer R A, Schreiber M 1998 *Phys. Rev. Lett.* **80** 3996

[21] Zhong J X, Geisel T 1999 *Phys. Rev.* E **59** 4071

[22] Bohigas M J, Giannoni, Schmit C 1984 *Phys. Rev. Lett.* **52** 1

[23] Zhang F Z, Wang J, Gu Y 1999 *Acta Phys. Sin.* **48** 2169 (in Chinese) [张飞舟、王　矫、顾　雁 1999 物理学报 **48** 2169]

[24] Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman R B 2001 *Bioinformatics* **17** 520

Application of random matrix theory to identification of lung cancer gene networks*

Li Rong Yan Ping-Lan Chen Jian Li Jun Li Jin Zhang Kai-Wang Zhong Jian-Xin†

(*Faculty of Materials, Optoelectronics and Physics, Xiangtan University, Xiangtan* 411105, *China*)

(Received 1 August 2008; revised manuscript received 31 December 2008)

Abstract

We used random matrix theory (RMT) to remove the noises in lung cancer gene expression data and used the modules approach and the hierarchical clustering approach to construct the gene networks. Comparing the results given by the two methods, we found that RMT-hierarchical clustering method gives true modules as well as the correlations between the modules. The results indicate that RMT-hierarchical clustering method is an effective new method for identifying gene networks.

Keywords: random matrix theory, hierarchical clustering, gene network, lung cancer

PACC: 0250, 8710

* Project supported by the National Natural Science Foundation of China (Grant No. 30570432).

† Corresponding author. E-mail: jxzhong@xtu.edu.cn