DOI:10.19405/j.cnki.issn1000-1492.2014.10.003

乳腺癌组织学分级下目标基因提取及转录调控网络构建

孔 薇[1],李海燕[1],牟晓阳[2],杨 旸[1]

摘要 **目的** 乳腺癌类型和分级多样性导致其预后差别显著,探寻乳腺癌不同分级情况下的基因表达差异及调控关系能够为乳腺癌致病机制的发现提供重要依据。**方法** 对不同分级下的乳腺癌基因表达数据利用快速独立成分分析(FastICA)方法提取特征基因,并结合人类蛋白质相互作用(PPI)数据选取目标基因。在此基础上,结合转录因子对靶基因调控的先验信息,利用网络成分分析(NCA)方法对与乳腺癌发病有密切关系的转录因子及其靶基因构建转录调控网络。**结果** 筛选出的基因经过数据库验证与乳腺癌相关的占48.15%,构建的调控网络发现了多个转录因子及靶基因在不同分级情况下的活性变化趋势。**结论** FastICA算法结合PPI数据提取目标基因的方法较为有效,通过NCA算法构建的转录调控网络为研究乳腺癌发生发展机制提供了新的方法。

关键词 乳腺癌;基因表达数据;快速独立成分分析;蛋白质相互作用数据;网络成分分析

中图分类号 Q 343.1

文献标志码 A **文章编号** 1000-1492(2014)10-1365-06

乳腺癌是女性最常见的恶性肿瘤之一,国内外发病率呈逐年上升趋势,已经高居女性恶性肿瘤发病率的首位[1]。目前医学上对乳腺癌的发病机制尚不清楚。乳腺癌的组织学分级是重要的预后指标[2],作为一个单独的预后指标能有效预测乳腺癌的预后。乳腺癌组织学分级与DNA增值指数、癌基因产物的表达和肿瘤细胞的分化程度等有关。通常,分级高的乳腺癌DNA增值指数较高,肿瘤细胞的分化程度较低,预后较差。研究不同分级情况下乳腺癌样本的显著基因、探寻其传导通路及调控网络,对乳腺癌的诊断、治疗和发生发展机制的研究具有重大意义。

2014-04-10 接收

基金项目:国家自然科学基金(编号:61271446、61003093);上海市科委青年科技启明星计划(A类)(编号:11QA1402900)

作者单位:[1]上海海事大学信息工程学院,上海 201306

[2]美国罗文大学生物化学系,新泽西 08028

作者简介:孔 薇,女,副教授,硕士生导师,责任作者,E-mail:weikong@shmtu.edu.cn

1 材料与方法

1.1 数据来源 选用的乳腺癌基因表达数据来自美国国立生物技术信息中心(national center for biotechnology information,NCBI)基因表达数据库中的数据集GSE42568。该数据集收集104例乳腺癌患者样本,年龄31~89岁,平均年龄58岁,包括11个乳腺癌Ⅰ级(G1,高分化)样本、40个乳腺癌Ⅱ级(G2,中分化)样本和53个乳腺癌Ⅲ级(G3,高分化)样本,另取了17个正常的乳腺组织作为对照样本。人类蛋白质-蛋白质相互作用(protein-protein interaction,PPI)数据来自BioGRID(Biological General Repository for Interaction Datasets,http://thebiogrid.org/),该数据库存放有酵母菌、人类等蛋白质的相互作用数据,本文采用的发布版本为3.2.104,包含人类18 170个蛋白节点和139 539对蛋白相互作用关系。转录因子对基因的调控关系库选用ITFP(Integrated Transcription Factor Platform,http://itfp.biosino.org/itfp/),该平台提供了人类转录因子4 105个,靶基因1 974个,以及转录因子-靶基因69 496个调控对的信息。

1.2 基因表达数据的独立成分分析(independent component analysis,ICA)模型及其快速算法 ICA[3]最初用于从线性混合信号中恢复出统计独立的源信号。基于在基因表达谱是由一些具有特定生物意义的相互独立的基因表达模式线性组合而成的假设,将ICA算法应用于基因数据中。设基因表达数据表示为矩阵$X_{m\times n}$,其中:m为样本个数,n为基因个数,通常$n>>m$。基因表达数据的ICA模型如下:

$$X=AS \tag{1}$$

ICA算法将基因表达数据矩阵$X_{m\times n}$分解为矩阵$X_{m\times m}$和$S_{m\times n}$,其中:$A_{m\times m}$是混合矩阵;$S_{m\times n}$是基因标记矩阵,其行向量是相互统计独立的,称为基因标记向量。为了得到矩阵$A_{m\times m}$和$S_{m\times n}$,可以通过找到线性变换矩阵W,使得:

$$Y=WX=WAS=\hat{S} \tag{2}$$

在当前的各种ICA应用算法中,快速独立成分

分析(Fast independent component analysis, FastICA)算法性能较好,收敛速度快。文中采用的是以负熵极大化作为搜寻方向的 FastICA 算法。

1.3 PPI 网络匹配 由于基因表达数据本身高噪声、信息量缺失的特性,近年来,很多学者[4-5]在不同程度上提出了整合基因表达数据集以及融合其他数据源的方法。其中,PPI 数据能为基因功能和网络的研究提供丰富的相互作用信息,尤其是对应用序列相似性不能注释的基因。因此,文章将 FastICA 所提取的表达显著基因与 PPI 数据进行匹配,保留具有相关关系的蛋白质所对应的基因,去除孤立的蛋白节点及其对应的基因,这样可以更好地筛选和弥补基因表达数据的不足。利用 R 软件包 Bionet[6],步骤为:① 用人类 PPI 数据构建出原始的蛋白质网络;② 将差异基因蛋白与原始网络匹配,得到只包含这些蛋白节点的子网;③ 将子网中度为 0 的节点删除(孤立点,即子网中不与其他蛋白节点相互作用的节点);④ 将最终得到的子网中的节点保存,即目标节点,其对应的基因即目标基因。

1.4 网络成分分析算法原理 网络成分分析(network component analysis, NCA)[7]是一种从基因表达数据和转录因子对基因调控关系的连通网络出发,推导转录因子活性以及对靶基因调控强度的算法。其数学模型为:

$$E = AP \tag{3}$$

其中矩阵 E 表示 $N \times M$ 维的基因表达数据,N 表示基因个数,M 表示样本个数;矩阵 A 表示 $N \times L$ 维的转录因子调控矩阵,L 表示转录因子个数;矩阵 P 表示 $L \times M$ 维的转录因子在不同时间点或者样本下的活性矩阵。NCA 算法的输入为基因表达数据矩阵 E 和根据转录因子对基因的调控关系构建的初始调控矩阵 A_0。最终的调控矩阵 A 和转录因子活性矩阵 P 的最优解可以通过下式求得:

$$\min_{A,P} \| E - AP \|^2 \tag{4}$$

$$s.t.\ A \in A(Z_0) \quad a_{i,j}^{(l)} \leqslant a_{i,j} \leqslant a_{i,j}^{(u)}, P_{i,j}^{(l)} \leqslant P_{ij} \leqslant P_{i,j}^{(u)}$$

其中 $a_{i,j}^{(l)}$,$a_{i,j}^{(u)}$,$P_{i,j}^{(l)}$ 和 $P_{i,j}^{(u)}$ 是边界约束条件,以连接模式矩阵 Z_0 来定义双凸优化问题。对给定矩阵 A 或者矩阵 P,总存在唯一的最小二乘解,以此来确定其它矩阵且同时满足约束条件。使用迭代优化算法,对矩阵和矩阵进行修正。步骤如下:

(1) 通过连接模式矩阵 Z_0 对矩阵 A_0 初始化。

(2) 修正矩阵 P。对给定矩阵 A_{k-1},通过用最小二乘法求解下式得出新的估计值 P_k:

$$\min_{P_k} \| E - A_{k-1}P_k \|^2 \tag{5}$$

$$s.t.\ P_{i,j}^{(l)} \leqslant P_{i,j}^{(u)}$$

将矩阵和矩阵写成如下形式:

$$E = [e_{c,1}\ e_{c,2}\ \mathrm{L}\ e_{c,M}], P_k = [P_{c,1}^{(k)}\ P_{c,2}^{(k)}\ \mathrm{L}\ P_{c,M}^{(k)}] \tag{6}$$

其中 $e_{c,i}$ 是 E 的第 i 列,$P_{c,i}^{(k)}$ 是 P_k 的第 i 列。再将矩阵 E 和矩阵 P 分别写成为 $NM \times 1$ 和 $LM \times 1$ 的列向量:

$$e_c = \begin{bmatrix} e_{c,1} \\ e_{c,2} \\ \mathrm{M} \\ e_{c,M} \end{bmatrix}, P_c^{(k)} = \begin{bmatrix} P_{c,1}^{(k)} \\ P_{c,2}^{(k)} \\ \mathrm{M} \\ P_{c,M}^{(k)} \end{bmatrix} \tag{7}$$

如此,式(5)定义的优化问题可写成下面的标准形式:

$$\min_{P_c^{(k)}} \| e_c - A_{k-1}P_c^{(k)} \|^2 \tag{8}$$

$$\mathrm{s.t.}\ P_{i,j}^{(k)} \leqslant P_{i,j} \leqslant P_{i,j}^{(u)}$$

其中 $\begin{bmatrix} \mathrm{A_{k-1}} & & \mathrm{L} & \mathrm{0} \\ \mathrm{M} & \mathrm{A_{k-1}} & & \\ & & \mathrm{O} & \mathrm{M} \\ & & \mathrm{L} & \mathrm{A_{k-1}} \end{bmatrix}$ 是 $NM \times LM$ 的块对角矩阵。文中对标准凸优化问题选用基于内点法的 SBLS 算法。

(3) 修正矩阵 A。对给定的 P_k,用下式更新 A_k:

$$\min_{A_k} \| E - A_kP_k \|^2 \tag{9}$$

$$\mathrm{s.t.}\ A_k \in \mathrm{A}(\mathrm{Z_0})\ \mathrm{a_{i,j}^{(l)}} \leqslant \mathrm{a_{i,j}} \leqslant \mathrm{a_{i,j}^{(u)}}$$

将矩阵 E 和矩阵 A_k 写成如下形式:

$$E = \begin{bmatrix} e_{r,1} \\ e_{r,2} \\ \mathrm{M} \\ e_{r,N} \end{bmatrix}, A_k = \begin{bmatrix} a_{r,1}^{(k)} \\ a_{r,2}^{(k)} \\ \mathrm{M} \\ a_{r,N}^{(k)} \end{bmatrix} \tag{10}$$

其中 $e_{r,1}$ 是 E 的第 i 行,$a_{r,i}^{(k)}$ 是 A_k 的第 i 行。则式(9)等价于:

$$\min_{\mathrm{a_{r,i}^{(k)}}} \| \mathrm{e_{r,i}} - \mathrm{a_{r,i}^{(k)}P_k} \|^2, \mathrm{i} = 1,2,K,\mathrm{N} \tag{11}$$

$$\mathrm{s.t.}\ \mathrm{A_k} \in A(Z_0),\ a_{i,j}^{(l)} \leqslant a_{i,j} \leqslant a_{i,j}^{(u)}$$

通过去除 $\mathrm{a_{r,i}^{(k)}}$ 中元素为 0 的值处理矩阵 $\mathrm{A_k}$ 的关联性约束,得到最终的行向量 $\mathrm{a_{r,i}^{(k)}}$。因此式(11)等价为:

$$\min_{a_{\beta,i}^{(k)}} \| e_{r,i} - a_{r,i}^{(k)}P_k \|^2, i = 1,2,\mathrm{K},N \tag{12}$$

$$\mathrm{s.t.}\ a_{\mathrm{i,j}}^{(\mathrm{l})} \leqslant \mathrm{a_{i,j}} \leqslant \mathrm{a_{i,j}^{(u)}}$$

其中矩阵 P_k 是通过删除矩阵 P_k 中对应 $a_{r,j}$ 中为 0 的行所得的矩阵。同理，可用相同的方法处理式(8)。

2 结果

2.1 数据预处理 为了去除基因表达数据中大量的冗余数据，在 FastICA 特征提取前首先用 T 统计对数据进行预处理。由于原始的 54 675 个基因探针对应的基因有数值缺失、对应重复的现象，经处理最终得到 21 026 个基因的表达数据。将 17 个正常组织样本分别与 G1 组、G2 组和 G3 组对照进行 T 统计分析，分别得到 5 681、6 167 和 5 216 个差异基因。

2.2 特征提取和 PPI 匹配 对以上 3 组预处理筛选出的基因表达数据用 FastICA 算法进行特征基因提取。FastICA 算法运行时随机生成的初始化矩阵会导致每次运行结果有所不同。为了保证结果的稳健性，本文对每组数据集重复运行 20 次 FastICA，统计显著基因出现的频次作为特征基因的选取依据。

图 1 为正常和乳腺癌 G2 样本集某次 FastICA 运行后混合矩阵 A 的热图。图中的行表示样本，列对应于独立分量；白色和黑色分别代表正值和负值，数值的绝对值越大对应正方形的面积越大。以第 25 列为例，对应前 17 个正常样本基本上为正值，而对应后 40 个乳腺癌 G2 样本的值则为负。这一列 A 向量可以较好的表征原数据矩阵的分类情况，且根据矩阵相乘的对应关系，此列 A 向量均与第 25 行 S 向量相乘，因此第 25 行 S 向量中对应表达值显著的基因即为对分类具有关键作用的基因。如图 2 给出了 IC25 的基因签名表达值取绝对值的结果。图中的横坐标代表基因，纵坐标代表 6 167 个基因签名表达值的绝对值。此行基因表达值的绝对值较高的基因即为显著基因。图 1 中虚线框分别给出了所选取的第 25、33 和 48 列。

完成 FastICA 算法后，正常与 G1 样本集共选择了 89 个 IC，对预处理后的 5 681 个差异基因，统计显著基因出现的频次，将频次不小于 14 的基因选出，共得到 1 200 个差异表达基因。同理，正常与 G2 样本集共选择了 1 201 个差异表达基因(共选择 83 个 IC，频数不小于 11)，正常与 G3 样本集共选择了 1 172 个差异表达基因(共选择 51 个 IC，频数不小于 9)。

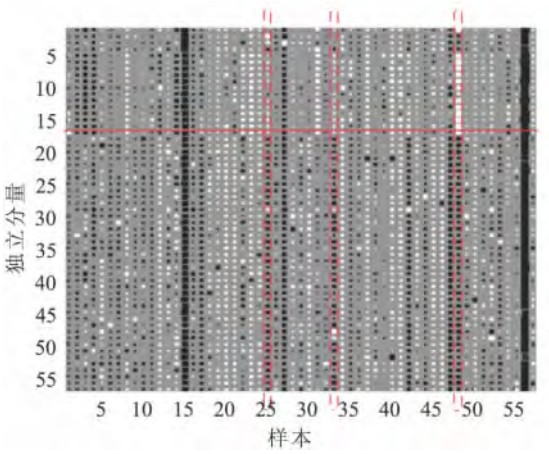

图 1 正常和乳腺癌 G2 样本集 FastICA 分解后 A 矩阵热图

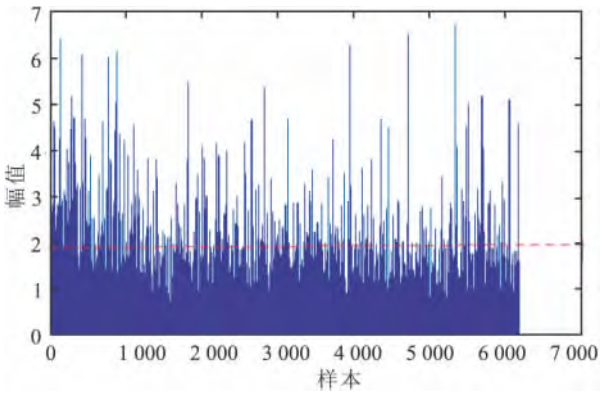

图 2 IC25 表达绝对值的柱状图

为了统一考察显著基因在不同分级样本中的表达情况，将以上 3 组所得的差异表达基因取交集，得到 666 个在 3 个分级样本中均具有差异表达的显著基因。

将这 666 个差异表达基因对应的蛋白与 PPI 网络进行匹配，共得到 162 个基因及 209 条相互关系信息，作为构建调控网络的目标基因。对这 162 个乳腺癌特征基因进行 GO[8] 功能注释，分析表明，这些特征基因主要集中在细胞程序性死亡、细胞增殖、磷酸代谢过程、类固醇激素刺激响应等多个生物功能，见表 1。接着进行 KEGG[9] 通路分析，结果表明，占主导地位的通路为癌症通路、丝裂原活化蛋白激酶信号通路、黄体酮卵母细胞成熟等相关生物过程，见表 2。为了分析提取的目标基因与乳腺癌的相关性，将得到的这 162 个目标基因导入乳腺癌数据库 G2SBC[10]，其中 78 个基因与乳腺癌相关，约占 48.15%，见表 3。

表 1 162 个目标基因部分 GO 注释结果

GO 功能注释	基因数	GO 功能注释	基因数
雌性激素刺激响应	7	激酶活性调控	12
细胞程序性死亡调控	25	细胞增殖调节	23
细胞坏死调控	25	类固醇激素刺激响应	11
联结受体蛋白信号通路酶	15	内源性刺激响应	15
激素刺激响应	14	蛋白激酶级联反应	13
脂肪细胞分化	6	转移酶活性调控	13
磷酸化作用调节	15	磷酸代谢过程调控	15

表 2 162 个目标基因参与的 KEGG 通路

KEGG 通路	基因数	KEGG 通路	基因数
癌症通路	13	黄体酮卵母细胞成熟	5
过氧化物酶体增生物激活受体信号通路	6	糖质新生	4
丝裂原活化蛋白激酶信号通路	10	Toll 样受体信号通路	5
造血细胞谱系	5	脂肪细胞因子信号通路	4

表 3 162 个目标基因与乳腺癌相关性的统计

与乳腺癌相关的 78 个显著差异表达基因	与乳腺癌没有明显相关性但是显著差异表达的 84 个基因
VIM、XBP1、BCL11A、SOD2、CSTB、CTBP2、BGN、SDC1、PRNP、GPX3、SLC9A3R1、NQO1、DKC1、APOD、SORD、KRT18、KRT19、TPD52、CD14、LTF、SERPINA3、TOB1、TRAF4、BIN1、PHOBTB3、MYO6、CDH3、F13A1、TF、MFAP2、HEBP2、MME、LPL、IGF1R、MID1、MAPT、CA12、FABP4、CDKN2C、BCAS1、FGFR3、PC、EER1A2、TFAP2A、MAPK10、CSTA、TFF1、AUH、ESR1、LRP2、HSD17B1、CLGN、PRLR、CD36、HP、MUC1、PPARG、FLNB、CRIP2、CRYAB、SIAH2、GATA3、ADH1B、CES1、VLDLR、NCF2、MAPK13、SDC2、FASN、SPDEF、ELOVL2、FANCL、VAV3、ECT2、HSPB8、GJB2、RASD1、CREB3L4	CSNK1A1、RASAL2、MDFIC、DSP、SPTBN1、TRIM28、CD9、TXNIP、TOP2A、ALDH2、FHL1、SH3BP5、EPCAM、ACSL1、ALDOC、MATN2、C14orf1、PLOD2、SOX9、PYGL、FBLN5、PAFAH1B3、BAMBI、RPS6KA1、NUCB2、CEACAM6、PCYOX1、DNAJB4、FBLN2、SPRY2、CEBPA、PRELP、S100P、FGF2、VCAN、MYB、BMPR1A、XRCC4、FGF13、PEX11A、NR3C2、ISG15、HSPB2、ZBTB16、MSX1、ASPA、ADRB2、LY96、EPB41L3、SMAD6、ASS1、DLX2、PCK1、MAGED2、PLIN2、FOS、RERMT2、GPC3、NDN、GNAI1、NKX3-1、VEGFC、CAMK2B、SPAG1、KCNB1、ALDH1A1、ZNF3、CCL8、ITGA1、PTPLAD1、LUC7L2、MICAL1、URGCP、HSPB7、C14orf139、FAM173A、SLC19A3、CIDEA、MPHOSPH8、PTGFRN、SUDS3、KCTD1、LNX2、SHISA2

2.3 转录调控网络的构建 将 162 个目标基因作为转录因子调控的靶基因，根据转录因子库中转录因子对基因的调控关系，得到其中的 52 个目标基因和调控 321 个转录因子，共 434 对调控关系。在这些调控对中，有单个转录因子调控多个目标基因，也有同一个目标基因受多个转录因子调控的情况。为了方便研究和分析，本文选取调控目标基因个数较多的转录因子及受调控的目标基因，通过 NCA 算法构建由 10 个转录因子和 15 个目标基因组成的转录调控网络。在仿真实验中，这 15 个目标基因的基因表达数据作为 NCA 的输入矩阵，10 个转录因子与 15 个目标基因的调控关系构成初始调控矩阵（转录因子调控基因为 1，否则为 0）。转录因子与目标基因的调控关系如表 4 所示。对部分的转录因子和基因的功能说明见表 5。

表 4 选取的 10 个转录因子与其对应的目标基因

转录因子 TF	受转录因子调控的靶基因
CCNB1	ECT2，TOP2A，PAFAH1B3
MSH6	C14orf1，ECT2，TOP2A
EZH2	ECT2，TOP2A，HSPB2
CCNA2	ECT2，CD14，TOP2A
DBF4	ECT2，CRYAB，HSPB2
RAD18	ECT2，CAMK2B，XRCC4
ZNF771	HSPB7，PC，RHOBTB3
BRCA2	CRYAB，CAMK2B
KRT8	DSP，KRT18
NCOR1	BCAS1，KRT18

表 5 10 个转录因子与 15 个目标基因中部分的功能说明

名称	功能描述
ECT2	细胞形态发生
CD14	细胞凋亡相关
CRYAB	葡萄糖代谢过程
CAMK2B	ATP 结合，蛋白质丝氨酸活性相关
TOP2A	正调控 RNA 聚合酶 II 启动子转录
XRCC4	正调控神经形成，子宫内胚胎发育
RHOBTB3	小 GTP 酶介导信号传导相关
KRT18	细胞凋亡过程
MSH6	错配修复
EZH2	调控细胞增殖
CCNA2	细胞有丝分裂，Ras 蛋白信号传导
DBF4	DNA 复制，细胞周期有丝分裂
BRCA2	雌性性腺发育，细胞衰老
KRT8	细胞凋亡过程，胚胎胎盘发育中的细胞分化
NCOR1	胆固醇平衡，生长调控

为了更好地观测这 10 个转录因子及其 15 个靶基因在正常组样本和乳腺癌不同病程下的表达活性和调控强度，经过 NCA 运算后，进一步用 Cytoscape（http://www.cytoscape.org/）软件绘制出在正常、G1、G2 和 G3 样本组下的转录调控网络图，见图 3。图中圆形节点代表靶基因，菱形节点代表转录因子；节点为红色代表表达水平上调，绿色代表表达水平下调，颜色越深表示表达水平越高；连线为红色代表转录因子对靶基因是正调控，绿色代表负调控。

3 讨论

本文对正常样本和经典的诺丁汉分级体系下的

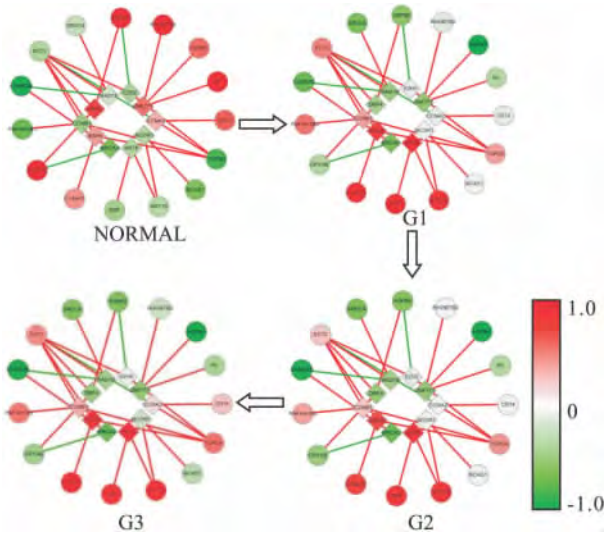

图3 10个转录因子及其靶基因在正常、G1、G2和G3样本下的转录调控图

乳腺癌Ⅰ级、Ⅱ级、Ⅲ级样本基因表达数据进行分析。首先利用T统计对数据进行降噪预处理，然后运用FastICA算法提取显著基因，接着结合人类蛋白质相互作用数据，得到在同一调控网络中的蛋白所对应的目标基因。最后根据转录因子对基因的调控关系，运用NCA算法构建乳腺癌不同分级下的基因转录调控网络。经过乳腺癌数据库验证，本文提取的这些目标基因中与乳腺癌相关的基因占较大的比重（提出显著基因经数据库验证约占比30%），本文将FastICA算法结合PPI数据提取目标基因的方法是有效的，相比仅对基因表达数据进行特征提取的方法所提取的乳腺癌相关基因有所增加。

同时，在分子生物学分析方面，由图3构建的正常样本和乳腺癌不同病程下的动态转录调控网络图可知，转录因子DBF4和BRCA2共同调控了靶基因CRYAB。虽然BRCA2是乳腺癌易感基因[11]，但在本文的转录调控网络中，在正常样本和乳腺癌各级样本中的表达水平变化不大，对靶基因CRYAB的表达几乎不起作用。而转录因子DBF4随着乳腺癌病级的加重表达活性逐渐下降，通过正调控靶基因CRYAB，使得CRYAB在正常样本中的高表达变为在乳腺癌各级样本中的低表达，且随着病程加重，表达水平有逐渐降低的趋势（颜色从红变绿，且绿色程度加深）。而转录因子DBF4的表达与乳腺癌易感基因p53的表达高度相关[12]，可以推断，p53影响DBF4的表达，DBF4调控CRYAB的转录表达，进而推动乳腺癌病级向前发展。在构建的转录调控网络中，基因ECT2受到较多转录因子调控，故ECT2在该转录调控网络中地位更为显著。在调控ECT2的转录因子中，CCNA2与乳腺癌雌激素分泌相关[13]，MSH6的少数变体与家族性乳腺癌的发生有关[14]，EZH2是乳腺癌易感基因[15]。上述的这些转录因子与其他转录因子CCNB1、RAD18加上上文提及的DBF4共同调控靶基因ECT2的表达。在这些转录因子的作用下，ECT2的表达水平最终表现为在正常样本中较低，在乳腺癌各级样本中较高，且随着病级的加重表达水平逐渐升高（颜色由绿色至红色，且红色变深）。由此推断，ECT2与乳腺癌的发生密切相关，且可能参与乳腺癌相关的多个生物通路，有待进一步验证。BCAS1受到转录因子NCOR1的调控，在正常样本和乳腺癌各级样本中的表达水平逐渐下降。而NCOR1可作为乳腺癌诊断的因子[16]，在本文的调控网络中表达水平逐渐下降，可推断BCAS1可能在NCOR1参与的乳腺癌生物通路中发挥重要作用。此外，基因TOP2A在CCNA2、MSH6和EZH2等乳腺癌相关转录因子的调控下，在正常样本和乳腺癌各级样本中的表达水平逐渐降低。这提示生物学家或许可以从CCNA2、MSH6和EZH2相关的生物通路着手，探究TOP2A与乳腺癌发生之间的联系。

本文通过对正常和乳腺癌不同分级样本转录调控网络的分析，发现表达水平受p53影响的DBF4能调控CRYAB的表达以推动乳腺癌病级发展、ECT2可能参与乳腺癌相关的多个信号通路、TOP2A可能与乳腺癌的发生相关等结论。进一步利用本文方法可以构建相关的转录调控网络，以对乳腺癌病理上各种子类进行研究，同样地，文章中的方法也可以应用到其他疾病上。总之，该研究为乳腺癌相关基因的提取、乳腺癌发生发展机制的探究与临床的诊断、预后提供新的方法和参考。

参考文献

[1] 韩晓雪. 乳腺癌病因病机及证治的文献研究[D]. 北京中医药大学，2012.

[2] Contesso G，Mouriesse H，Friedman S，et al. The importance of histologic grade in long-term prognosis of breast cancer：a study of 1，010 patients，uniformly treated at the Institut Gustave-Roussy[J]. J Clin Oncol，1987，5(9)：1378－86.

[3] Hyvärinen A，Oja E. Independent component analysis：algorithms and applications[J]. Neural Networks，2000，13(4－5)：411－30.

[4] 江丽华，李亦学，刘 琪. 综合ChIP-chip数据、基因敲除数据

和表达谱数据构建基因调控网络[J]. 生物化学与生物物理进展,2010,37(9):996 - 1005.

[5] Engin H B, Guney E, Keskin O, et al. Integrating structure to protein-protein interaction networks that drive metastasis to brain and lung in breast cancer [J]. PLoS One, 2013, 8(11): e81035.

[6] Beisser D, Klau G W, Dandekar T, et al. BioNet: an R-Package for the functional analysis of biological networks [J]. Bioinformatics, 2010, 26(8): 1129 - 30.

[7] Seok J, Xiao W, Moldawer L L,et al. A dynamic network of transcription in LPS-treated human subjects [J]. BMC Syst Biol, 2009, 3: 78.

[8] Harris M A, Clark J,Ireland A, et al. The Gene Ontology (GO) database and informatics resource [J]. Nucleic Acids Res, 2004, 32(Database issue): D258 - 61.

[9] Ogata H, Goto S, Sato K, et al. KEGG: Kyoto encyclopedia of genes and genomes [J]. Nucleic Acids Res, 1999, 27(1): 29 - 34.

[10] Mosca E, Alfieri R, Merelli I, et al. A multilevel data integration resource for breast cancer study [J]. BMC Syst Biol, 2010, 4: 76.

[11] 叶 云. 乳腺癌组织学分级特征基因提取及基因集富集分析[D]. 南方医科大学, 2010.

[12] Bonte D, Lindvall C, Liu H, et al. Cdc7-Dbf4 kinase overexpression in multiple cancers and tumor cell lines is correlated with p53 inactivation [J]. Neoplasia, 2008, 10(9): 920 - 31.

[13] Vendrell J A, Magnino F, Danis E, et al. Estrogen regulation in human breast cancer cells of new downstream gene targets involved in estrogen metabolism, cell proliferation and cell transformation [J]. J Mol Endocrinol, 2004, 32(2): 397 - 414.

[14] Wasielewski M, Riaz M, Vermeulen J, et al. Association of rare MSH6 variants with familial breast cancer [J]. Breast Cancer Res Treat, 2010, 123(2): 315 - 20.

[15] Panousis D, Xepapadakis G, Lagoudianakis E, et al. Prognostic value of EZH2, paxillin expression and DNA ploidy of breast adenocarcinoma: correlation to pathologic predictors [J]. J BUON, 2013, 18(4):879 - 85.

[16] Zhang Z, Yamashita H, Toyama T, et al. NCOR1 mRNA is an independent prognostic factor for breast cancer [J]. Cancer Lett, 2006, 237(1): 123 - 9.

Aimed genes' extraction and construction of transcription regulatory network under different grading levels of breast cancer

Kong Wei[1], Li Haiyan[1], Mou Xiaoyang[2], et al

([1]*Information Engineering College, Shanghai Maritime Univ, Shanghai* 201306;
[2]*DNJ Pharma, Rowan University, NJ* 08028)

Abstract ***Objective*** The diversities of breast cancer types and grading levels lead to distinct difference for breast cancer prognosis. Studying the gene difference expression and regulatory relationship among genes under different grading levels of breast cancer could provide an important basis for finding breast cancer pathogenesis. ***Methods*** Using fast independent component analysis (FastICA) method to extract feature genes of gene expression data of breast cancer, and then selected the aimed genes by combining with human protein-protein interaction data (PPI). On this basis, introducing prior information which described regulatory relationships about how transcription factors regulated their target genes, we continued to analize transcription factors and their target genes, which were closely associated with the incidence of breast cancer, by using network components analysis method (NCA), and then constructed a transcriptional regulatory network. ***Results*** Selected aimed gene which was closely associated with breast cancer is about 48.15%, that had been validated by breast cancer database. And from the built regulatory network, found out the activity change trend of multiple transcription factors and their target genes under different grading levels. ***Conclusion*** FastICA algorithm combined with PPI data for extracting aimed gene is a relatively effective method. Simultaneously, constructing transcription regulatory network with NCA method provides a novel way for studying progression mechanism of breast cancer.

Key words breast cancer; gene expression data; fast independent component analysis; protein-protein interaction data; network component analysis