利用亚细胞位置特异的基因功能模块与表达调控网络识别疾病特征基因

张 敏[①] 朱 晶[①] 郭 政[①②*] 李 霞[①②] 杨 达[①] 王 磊[①] 饶绍奇[①③*]

(① 哈尔滨医科大学生物信息学系, 哈尔滨 150086; ② 同济大学生命科学与技术学院, 上海 200092; ③ Departments of Molecular Cardiology and Cardiovascular Medicine, the Cleveland Clinic Foundation, 9500 Euclid Avenue, Cleveland, Ohio 44195, USA. * 联系人, E-mail: guoz@ems.hrbmu.edu.cn, raos@ccf.org)

摘要 利用 Gene Ontology 中的生物过程(biological process)及细胞组分(cellular component)两种分类体系, 选择显著聚集差异表达基因的复合功能模块, 识别其中能够有效分类疾病样本的特征功能模块, 以特征功能模块中的差异表达基因作为特征并分析它们与疾病的相关性. 对前列腺癌、胃癌和白血病数据的分析结果表明, 基于特征功能模块的特征基因选择方法可以识别与疾病高度相关的功能一致的特征基因. 进一步的分析显示, 根据特征功能模块和基因表达调控信息构建基因表达调控网络, 可以从中挖掘可能的疾病关键特征基因, 并提示对复杂疾病同时应答的多功能模块间协同作用关系机理研究的重要线索. 同时, 本研究结合基因功能分类的疾病特征基因选择方法提示了一种高准确度的疾病分类方法, 分类结果有明确的生物学意义, 对复杂疾病的分子病理学研究亦有重要的意义.

关键词 基因表达谱 **Gene Ontology** 功能模块 特征基因 基因调控网络

根据高通量基因表达谱数据, 采用数据挖掘技术识别癌症等复杂疾病相关的特征基因与功能, 对研究疾病机理、预测疾病类型有重要的意义[1,2]. 基因的表达行为不是孤立的, 功能相关基因(如位于同一条代谢通路)的表达倾向于高度相关, 即基因有功能模块化协同表达的倾向[3,4]. 根据不同的研究目的, 已经有许多研究人员采用不同定义的基因模块(功能相同并且表达相似的基因集合、具有相同调控程序的基因集合等)对复杂疾病进行了有意义的研究[5~7]. 例如, Mootha等人[5]的研究发现, 利用疾病条件下功能相似并协同变化的一组基因(模块), 可以识别表达变化幅度较小且协同性很高的疾病相关基因. Segal等人[6]把模块定义为多种不同类型癌症样本中被激活或被抑制的基因集合, 并发现少数相关模块的结合可以描述特异肿瘤以及不同肿瘤形成的机制特征. 根据基因表达的功能模块化现象, 我们的研究也显示, 基于Gene Ontology (GO)[8]中聚集差异表达基因的功能模块, 可以对癌症样本进行高准确度的疾病分类[7,9]、疾病亚型发现(聚类分析)[10], 并可以有效地预测基因功能[11].

本研究进一步提出一种结合GO基因功能分类选择疾病特征的功能模块与特征基因, 从功能模块水平研究疾病机理与分类的方法: (ⅰ) 不同于已有的仅限于在GO的一种分类体系下选择差异表达功能模块的方法(如Onto-Express[12]与FatiGO[13]等), 提出结合GO中的生物过程(biological process, BP)和细胞组分(cellular component, CC)两个体系, 寻找显著聚集差异表达基因的亚细胞位置特异的复合功能结点作为差异表达功能模块. (ⅱ) 借鉴在传统的单基因水平上选择对疾病分类有鉴别意义的基因作为特征基因的原理, 通过评估差异表达功能模块中差异表达基因的疾病识别能力来分析模块与疾病的相关性, 选择具有高分类效能的差异表达功能模块作为特征功能模块. 本研究采用简单的k (k = 3)近邻分类器[1]作为评价特征功能模块(基因)分类能力的工具, 在选择疾病特征功能模块的基础上, 整合基因表达调控信息构建疾病相关的基因表达调控网络, 探讨进一步挖掘疾病相关的关键特征基因, 并分析对复杂疾病同时应答的多模块间协同作用的可能途径. 对前列腺癌、胃癌和白血病数据的初步分析结果显示, 通过本方法选择出的特征功能模块(特征基因)与疾病高度相关, 可以相当程度地在功能模块水平上揭示复杂疾病的功能机理. 另一方面, 对 3 套数据的分析结果也提示, 特征功能模块中的特征基因具有很强的疾病分类能力, 而且分类结果具有明确的生物学功能解释. 虽然传统的特征选择(过滤法和缠绕法等)可以有效地筛选掉大量对疾病缺乏分类能力的无关基因[14,15], 将高维特征(基因)空间压缩到低维特征空

间，但是只局限于采用模式识别方法直接筛选基因，还不能很好地发现集中于某几条代谢通路或功能类中的特征基因，将疾病分类结果与特征基因的功能明确地关联起来. 所以，本研究提出的结合基因功能分类知识的分析方法也是一种值得进一步探讨的特征基因选择与疾病分类方法，可以较好地解决根据基因表达谱对疾病进行分类时运用传统模式识别方法存在的维数灾难[16](几十例疾病样本而检测的基因数万个)及结果的生物学解释一致性差等问题. 同时，基于功能模块的分析结果受基因表达谱检测误差与检测缺失的影响较小，相对于从单基因角度分析具有较好的稳健性[9,17].

1 材料和方法

(ⅰ) 数据集. (1) 前列腺癌: Lapointe等人[18]用含 46000 个点样的cDNA芯片分析了 103 个样本，包括 62 例原发性前列腺癌和 41 例正常样本. 预处理后，共得到 11763 个基因(Gene ID)，含 1938 个差异表达基因. (2) 胃癌: Chen等人[19]用含 44500 多个点样的cDNA基因芯片分析了 132 个样本，包括 103 例原发性胃癌和 29 例正常样本. 预处理后，得到 13036 个基因，其中 3303 个基因差异表达. (3) 白血病: Yeoh等人[20]用含 12600 个点样的Oligo芯片分析了 327 例急性淋巴白血病样本，包括已知的 6 种亚型. 本研究选择其中样本量最大TEL-AML1 阳性急性淋巴细胞白血病(TEL-AML1，79 例)和高二倍体急性淋巴细胞白血病(超二倍体> 50，64 例)作为研究对象. 预处理后，共得到 8677 个基因，包含 1098 个差异表达基因.

(ⅱ) 预处理与差异表达基因筛选. 对cDNA芯片数据，采用类似于Dudoit等人[1]提出的芯片处理原则，删除缺失率大于 5%的基因，对剩余基因的表达值进行以 2 为底的对数转换，再对缺失值补 0；将得到的表达谱数据用BRB ArrayTools进行中值标准化(http://linus.nci.nih.gov/BRB-ArrayTools.html)，然后通过校正的t检验筛选差异表达基因($P \le 0.001$；错误发现率FDR ≤ 0.1). 对Oligo芯片数据，限定基因表达量检测值(intensity)为 100~16000，小于 100 以 100 代替；大于 16000，则以 16000 代替；对表达值进行以 2 为底的对数转换. 数据标准化方法和差异表达基因的选择标准同上.

(ⅲ) 基因功能注释与差异表达功能模块选择. GO 是跨越物种的基因功能分类体系，以严格定义的功能概念关系，从生物过程、细胞组分和分子功能(molecular function, MF) 3 个方面描述基因功能. 我们将数据集中的基因广义注释到 GO 功能概念结点，即如果基因被注释到 GO 的某个结点，则也被广义注释到该结点到根结点路径上的所有结点.

同一个生物过程可能发生于不同的亚细胞位置，因此参与的基因并不一定受到相同的调控机制调控[21,22]. 因此，将GO中生物过程和细胞组分两个体系中的所有结点组合，形成复合功能结点. 然后，利用超几何分布检验差异表达基因是否在复合功能结点中非随机地聚集，显著聚集差异表达基因的复合功能结点可能与疾病相关[12]. 设芯片中所有基因和差异表达基因的个数分别为 N和M，注释到某一复合功能结点中的基因和差异表达基因的个数分别是n和k，则此复合功能结点中至少注释有k个差异表达基因的概率为

$$P = 1 - \sum_{i=0}^{k-1} \frac{C_M^i C_{N-M}^{n-i}}{C_N^n}.$$

一个复合功能结点的 P 值越小，则反映该复合功能结点随机聚集差异表达基因的概率越小. 我们选择复合功能结点中$P \le 0.001$且功能相对具体细致(注释基因个数为 3~150，并且 BP 注释深度在 3 层以下)的结点作为候选的差异表达功能模块做进一步的分析. 这是由于目前的疾病机制研究多集中于信号、代谢通路等生物过程，如果同时限制 CC 深度，则可能因为产生的复合功能结点太小(包含的检测基因数小)，很具体的 BP 功能结点会被排除，故不再限制 CC. 显然，如果我们细致分析 CC，也可以选择限制 CC 深度.

我们采用随机重取样的方法来估计利用$P \le 0.001$ 筛选出的d个候选差异表达功能模块FDR[23]. 对每组数据集，我们在保证复合功能结点中注释基因不变的前提下，从芯片上所有的基因中随机抽取 1000 组与差异表达基因数目相等的基因，分别作为随机差异表达基因集合，重新计算各复合功能结点的P值，并统计在第i (i = 1, 2, …, 1000)次随机实验中 $P \le 0.001$ 的结点个数d_i. 设 $d_e = \sum_{i=1}^{1000} d_i/1000$，则利用$P \le 0.001$ 筛出d个选候选差异表达功能模块的FDR = d_e/d. 对前列腺癌、胃癌和白血病数据集，当按$P \le 0.001$ 选择候选差异表达功能模块时，已经可以获得足够小的FDR，分别为 6.9%, 5.1%和 3.7%.

根据一组差异表达基因，基于GO寻找差异表达功能模块是分析大量差异表达基因与疾病关系(生物

学解释)的主要途径. 这类分析的依据是功能相关的基因有模块化协同相关表达倾向的生物学假设[3,4]: 富集差异表达基因的“功能模块”有可能对疾病条件应答, 从而其中包含的基因有较大的与疾病相关联的可能性. 寻找富集差异表达基因的功能模块的过程, 实际上也可以理解为根据生物学假设, 利用先验的基因功能知识对“差异表达(特征)基因”再次进行(功能性)特征选择的过程, 使得后续分析集中于表达行为相对不“孤立”的基因(一般认为, “孤立”的基因差异表达不可靠或生物学意义不明确). 由于选择“差异表达基因”时利用了样本两类划分的信息, 这样选择出的特征一般具有一定的分类能力.

(iv) 样本分类与特征功能模块的选择. 我们进一步推测对疾病状态有较直接应答的功能结点中的基因表达模式可能有较强的疾病识别能力. 为此, 在设定如前选择的模块有特定生物学意义后, 采用分类器评价各模块中的差异表达基因的分类能力, 筛选包含有足够分类信息的模块做进一步的分析. 对每个差异表达功能模块, 分别以其中注释的差异表达基因为特征, 训练k (k = 3)近邻分类器[1]分类样本, 并采用留一法(Leave-One-Out Cross Validation, LOOCV)计算分类准确率. 在LOOCV证实过程中, 每次抽取一个样本为检验样本, 以剩余的样本构成训练集, 利用训练集训练 3 近邻分类器, 然后预测检验样本, 分别记录下每个 3 近邻分类器预测的真阳性数(TP)、真阴性数(TN)、假阳性数(FP)和假阴性数(FN). 采用如下的分类准确率(accuracy, acc)来评价分类器性能:

$$\mathrm{acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FN}+\mathrm{FP}}.$$

将分类准确率大于 90%的差异表达功能模块定义为特征功能模块, 其中注释的差异表达基因作为特征基因. 与根据缠绕法[15]选择特征的常规特征选择过程不同, 我们只限于评价按富集差异表达基因的“功能模块”组织起来有特定生物学意义的“特征子集”, 它们只是全部可能的特征基因组合的一部分, 而不是优化组合“产生”的新子集. 因此, 本研究主要利用模块的分类准确率(可理解为模块与疾病非线性相关程度的一种测度)排序选择有较强疾病识别能力的特征模块[24], 而没有采用LOOCV完全证实过程(在每次循环中重新筛选差异表达基因与模块等), 后者以试图得到分类准确率的无偏估计为主要目的[25].

(v) 构建基因表达调控网络选择疾病相关基因. 为了进一步分析各个特征功能模块中的基因与疾病机理的关系, 我们利用PathwayAssist[26]提供的基因表达调控信息, 构建特征功能模块中基因参与的表达调控网络. 将两个基因间的调控关系简化为调控和被调控两种, 而不区分正性调控和负性调控. 如果两个基因之间存在一条调控关系, 在表达调控网络中对应的结点(用蛋白质表示)间就存在一条有向边, 构成基因表达调控网络. 在蛋白质互相作用等生物网络中, 连接度很高的HUB结点(网络中的基因或者蛋白质)对网络结构有关键影响, 并且往往具有重要的生物学功能[27,28]. 同时, HUB结点可能在模块与模块间的协同作用中有关键性的枢纽作用[29]. 因此, HUB结点是分析生物网络的重点之一. 假设得到的网络中共包含t个基因参与的T对调控关系, 由于t个基因共可以形成$t(t-1)$个有向边(调控关系), 而且每个调控关系涉及一个调控基因与一个被调控基因, 所以一个基因调控另外一个基因的随机概率为$P_0 = T/t(t-1)$, 一个基因被调控的随机概率亦为P_0. 利用二项分布原理, 一个基因随机调控至少s个基因(或被至少s个基因调控)的概率为

$$P_r=1-\sum_{i=0}^{s-1}C_{t-1}^{i}\left(P_0\right)^{i}\left(1-P_0\right)^{t-i-1}.$$

值得注意的是, PathwayAssist提供的调控信息主要来自于文本挖掘, 而且两个基因间不同类型的调控关系可能发生于不同的时空条件下, 具有相当的不确定性. 所以本研究所描述的调控网络分析, 可以被保守地认为是一种有效的辅助综合利用文本信息寻找重要调控基因的途径. 作为初步探索, 选择显著多($P_r \le 0.001$)地调控其他基因或被其他基因调控的HUB基因作为可能的关键特征调控基因, 分析它们与疾病的相关关系.

2 结果

2.1 特征功能模块与特征基因

利用前列腺癌数据, 我们得到了 16 个候选差异表达功能模块, 分别用注释于各模块中的差异表达基因对样本进行分类后, 得到 12 个特征功能模块分类准确率大于 90%(表 1), 最高达到 97% (模块 PM1, GO 中的功能类描述为: BP, 器官发生; CC, 跨膜结构). 结果表明, 利用一些功能模块中的很少的基因就可以对疾病样本进行很好的分类. 例如, PM2 (BP, 离子运转;

表 1 前列腺癌相关特征功能模块

编号	生物过程	细胞组分	注释基因/差异基因	P值[a)]	分类准确率[b)]
PM1	器官发生	跨膜结构	147/43	7.11×10^{-5}	0.97
PM2	离子转运	胞外基质	42/17	1.81×10^{-4}	0.95
PM3	细胞黏附	细胞组分	85/26	8.74×10^{-4}	0.94
PM4	同种细胞黏附	膜	47/18	2.70×10^{-4}	0.94
PM5	细胞黏附	膜	75/24	6.67×10^{-4}	0.93
PM6	同种细胞黏附	跨膜结构	43/16	8.47×10^{-4}	0.93
PM7	细胞生长和/或维持	胞外基质	80/25	7.70×10^{-4}	0.93
PM8	转运	胞外基质	57/20	4.90×10^{-4}	0.93
PM9	磷酸盐转运	胶原蛋白	26/12	3.84×10^{-4}	0.93
PM10	器官发生	胞内非膜结合细胞器	67/23	2.76×10^{-4}	0.91
PM11	器官发生	细胞骨架	63/22	2.82×10^{-4}	0.91
PM12	肌肉发育	细胞组分	91/28	5.08×10^{-4}	0.90

a) 差异表达基因是否在复合功能结点中非随机地聚集的超几何分布检验的概率. b) 以模块中差异表达基因为特征对样本进行分类得到的准确率. 下同

CC, 胞外基质)和 PM9 (BP, 磷酸盐转运; CC, 胶原蛋白)分别包含 17 和 12 个差异表达基因, 分类准确率分别为 95%和 93%.

通过以上方法选择出的特征功能模块与疾病高度相关. 例如, 特征功能模块PM3, PM4, PM5 和PM6 代表发生在不同亚细胞位置的细胞黏附(cell-cell adhesion)生物过程, 具有很高的疾病分类能力(准确率都在 93%以上), 提示细胞黏附过程可能与前列腺癌相关. 相邻细胞间的连接决定细胞和组织的形态、调控细胞的运动、生长、分化和存活等. 细胞间黏附连接结构(adheren junction)[30]的形成是通过相邻细胞表面的钙依赖性钙黏着素(cadherin)受体的细胞外区域的相互作用实现的. 细胞的恶变通常表现为组织结构中细胞骨架的根本改变, 一方面通过下调钙黏素或连环蛋白(catenin)家族成员的表达实现, 另一方面可以通过激活一些可阻断细胞间黏附连接结构聚集的信号传导途径实现. 已有研究发现, 上皮型钙黏素(E-cad)表达的缺失可使腺瘤向腺癌转化[31]. 另有研究表明, 细胞黏附蛋白的共同异常表达对前列腺癌的预后诊断有重要意义[32].

根据胃癌和白血病数据, 其中分类准确率大于 90%的特征功能模块分别为 29 和 8 个. 由于不同模块的生物过程和细胞组分概念间存在包含关系, 我们仅分析同时在生物过程和细胞组分两个体系中注释最深(即概念最具体)的模块. 例如, 模块 1 (BP, 有丝分裂; CC, 细胞核)的生物过程和细胞组分都是模块 2 (BP, 有丝分裂调节; CC, 细胞核)的父结点, 我们只在表中显示模块 2 (表 2 和 3). 在胃癌数据中, 最小的特征功能模块GM6 (BP, 通过剪接体的核mRNA剪接, CC, snRNP复合物)包含 7 个差异表达基因, 以这 7 个基因作为特征的分类准确率达到 91%. 模块GM6 的生物过程是核mRNA剪切, 许多基因的选择性剪切与肿瘤相关联, 这些受影响的基因可归类为介导细胞与细胞、细胞与基质相互作用的受体, 激素、凋亡调节因子. Wang等人[33]通过对人类基因组的 26258 个选择性剪切进行筛选, 发现 845 个选择性剪切和人类癌症显著相关. Gunthert等人[34]发现, CD44 可以通过选择性剪切内部9个外显子赋予癌症不同的转移能力. 在白血病数据中, 最小的特征功能模块LM2 (BP, 损伤反应; CC, 胞内非膜结合细胞器)包含 4 个差异表达基因, 以这 4 个基因作为特征的分类准确率达到 95%. Yeoh等人[20]的研究发现, 其中CD27 的mRNA表达和TEL-AML相关, 而Vaskova等人[35]进一步从蛋白水平证实了CD27 特异地与TEL-AML相关.

2.2 基于特征功能模块的基因表达调控网络的构建

在选择了对疾病有高鉴别能力的特征功能模块后, 为了进一步挖掘可能与致病机理密切相关的关键特征基因, 我们选择由前列腺癌数据得到的 12 个特征功能模块, 结合基因表达调控信息构建基因表达调控网络, 共包含 58 个基因参与的 98 对基因表达调控关系. 利用二项分布原理, 发现特征模块中的蛋白质TNF, CD4, VEGF和SPP1 显著地($P_r\leq0.001$)调控其他蛋白, 而CD80 则显著地($P_r\leq0.001$)被其他蛋白质调控. 蛋白质CD4, CD80 和SPP1 注释到模块 PM1 (BP, 器官发生; CC, 跨膜结构), TNF来自模块 PM5 (BP, 细胞黏附; CC, 膜), VEGF与SPP1 来自模块PM7 (BP, 细胞生长和/或维持; CC, 胞外基质). 前

表 2 胃癌相关特征功能模块

编号	生物进程	细胞组分	注释基因/差异基因	*P* 值	分类准确率
GM1	细胞因子生物合成调节	细胞组分	13/9	5.56×10^{-4}	0.96
GM2	有丝分裂调节	细胞核	13/10	6.46×10^{-5}	0.96
GM3	细胞周期蛋白依赖的蛋白激酶活性调节	细胞组分	27/15	2.89×10^{-4}	0.94
GM4	DNA 依赖的转录调节	染色质	25/14	4.14×10^{-4}	0.94
GM5	细胞周期检查点	细胞核	15/11	5.61×10^{-5}	0.91
GM6	通过剪接体的核 mRNA 剪接	snRNP 复合物	9/7	8.39×10^{-4}	0.91
GM7	通过剪接体的核 mRNA 剪接	剪接体复合物	31/16	5.52×10^{-4}	0.90
GM8	rRNA 代谢	细胞组分	14/10	1.79×10^{-4}	0.90

表 3 白血病相关特征功能模块

编号	生物过程	细胞组分	注释基因/差异基因	*P* 值	分类准确率
LM1	防御反应	胞内非膜结合细胞器	10/6	5.40×10^{-4}	0.97
LM2	损伤反应	胞内非膜结合细胞器	4/4	2.55×10^{-4}	0.95
LM3	核糖核苷酸生物合成	细胞组分	54/19	1.74×10^{-5}	0.94
LM4	嘌呤核苷酸代谢	细胞内	24/10	3.70×10^{-4}	0.92
LM5	单价无机阳离子转运	细胞质	25/10	5.47×10^{-4}	0.90

列腺癌的 12 个特征功能模块分布于GO生物过程体系中的3条独立的路经, 以上HUB蛋白所涉及的模块PM1, PM5 和PM7 分别分布于这 3 路经, 所以不同HUB蛋白质间的表达调控关系可能体现了 3 类不同生物过程模块间的交互作用在HUB蛋白质连接的调控子网(图 1)中, PM5 中的TNF参与调控 22 个来自不同模块的蛋白表达, 很有可能在模块与模块间的协同作用中起主要作用. TNF是一种多功能细胞因子, Nakashima等人[36]发现血清的TNF水平与前列腺癌恶病质的临床表现高度相关; 而Oh等人[37]的研究表明, TNF-α +488, – 308 等位基因和基因型频率分布与前列腺癌的发病率显著相关. PM7 中的VEGF是主要的促血管生成因子, 对增加血管通透性及促进新生血管的形成有重要作用, 而新生血管的形成与肿瘤的发生、侵袭和转移有着密切的联系. 在调控网络中, 受到VEGF直接调控的蛋白质KDR, FLT1 和NRP1 (neuropilin-1)是VEGF受体蛋白, 都注释到模块PM1 (BP, 器官发生; CC, 跨膜结构). VEGF是肿瘤血管生成的重要调控因子, VEGF受体KDR/flk-1 的激活诱导上皮细胞的增殖, 能够引起肿瘤血管生成以及癌细胞的增殖, VEGF与新生血管的形成相关联[38,39]. 在前列腺癌细胞中, 血管内皮生长因子受体, 尤其是NRP-1 高度表达, 对癌细胞的生长和代谢、癌转移有重要的影响[40]. VEGF对细胞黏附分子ICAM1, VCAM1 也有直接调控关系. ICAM1, VCAM1 来自模块PM5 (BP, 细胞黏附; CC, 膜), 对癌细胞的转移性和侵袭性有重要的作用. 研究表明[41], E-selectin, ICAM-1 和VCAM-1 在前列腺癌细胞表面组成性表达(constitutive expression). 被调控HUB蛋白CD80 来自PM1, 显著地被其他蛋白调控. 与可能参与疾病发病机制的调控HUB蛋白相比, 被调控HUB蛋白可能是疾病阶段细胞内各种调控手段失常的集中体现. Gulley等人[42]用前列腺特异抗原(PSA)和CD80 的重组疫苗对前列腺癌病人进行免疫治疗的结果显示, 相对对照组, 显著多的病人产生了PSA-特异的细胞免疫.

3 讨论

已有的选择差异表达功能模块的方法(如Onto-Express[12], FatiGO[13]等), 都仅限于在GO的一种分类体系下, 分析差异表达基因在功能结点中的聚集程度的统计显著性. 本研究提出, 从显著聚集差异表达基因的功能类中进一步筛选有高分类效能的模块, 这样的功能模块与疾病有更强的相关. 值得注意的是, 对每组癌表达谱数据, 我们得到了对疾病有鉴别意义的多个特征功能模块, 除了由于GO层次化分类方式造成的冗余关系之外, 这也可能部分地反映了癌症这类复杂的多基因疾病的机理涉及较多的生物学通路的协同应答. 因此, 进一步结合基因特征功能模块和基因表达调控数据建立基因表达调控网络, 初步分析前列腺癌数据的结果提示, 大量参与特征模块中基因的调控关系的TNF和VEGF等HUB基因与前列腺癌高度相关, 可能在模块与模块间的协同作用过程中起到关键的连通作用. 分析模块与模块间的协同作用与疾病的关系是值得进一步深入研究的方向.

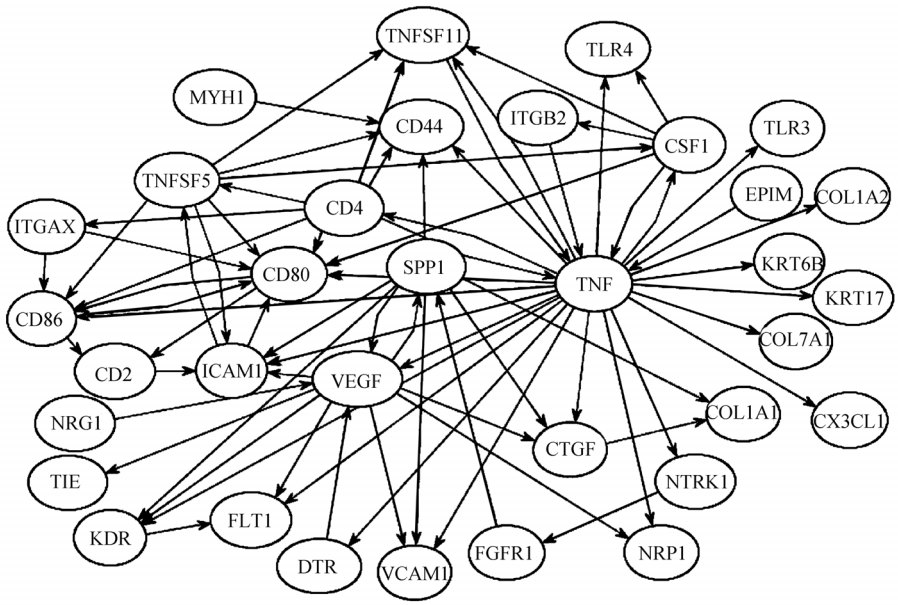


图 1 前列腺癌: HUB 蛋白质连接的表达调控子网

根据一组差异表达基因, 基于GO寻找差异表达功能模块是分析差异表达基因与疾病关系(生物学解释)的重要途径. 与已有的对GO的 3 个分类体系分开分析的方法不同, 本研究提出了将生物过程和细胞组分两个体系相结合的功能模块, 如此可以提供疾病相关的生物学过程所发生的亚细胞位置的信息, 使模块的生物学意义更加具体. 例如, 在前列腺癌相关特征功能模块中, PM4 和PM5 的细胞组分都是膜(membrane), 而已有研究表明一些膜蛋白在正常细胞和前列腺癌细胞中的表达有显著差异[43]; PM2, PM7 和PM8 的细胞组分是细胞外基质(extracellular matrix), 而细胞外基质胶原蛋白 I 参与调控前列腺癌细胞的增殖[44].

高通量基因表达谱数据具有高检测误差、高维等特点, 使得基因表达谱信息分析结果的稳定性与生物学解释的一致性较差. 虽然通过传统的特征选择方法, 也可以达到从高维特征(基因)中获取对样本具有高鉴别能力的特征基因的目的[14~16], 但是将本研究提出的方法应用于 3 套癌表达谱数据的结果表明, 新方法不仅可以降低分析维数, 而且更重要的是可以识别有很强的疾病分类能力、功能一致的特征功能模块与特征基因, 有助于进一步探索疾病机理. 同时, 模块分析受到基因表达谱数据中检测误差与检测缺失的影响较小, 相对于从单基因角度分析具有较好的稳健性[9,17]. 同理, 由于现有的人类基因表达调控数据尚不完整, 仍然存在较多误差[45], 利用一组基因的表达调控关系从基因网络中挖掘和识别可能的关键HUB基因的方法, 受单个基因调控信息的不可靠性的影响较小, 具有较高的统计稳健性, 值得进一步探讨.

致谢 本工作为国家自然科学基金(批准号: 30170515, 30370388, 30370798, 30570424 和 30571034)、国家重点基础研究发展计划(批准号: 2003AA2Z2051 和 2002AA2Z2052)、黑龙江科技攻关重点(批准号: GB03C602-4)、哈尔滨市科技攻关(批准号: 2003AA3CS113)、黑龙江自然科学基金(批准号: F0177)和黑龙江省教育厅杰出海外学人基金(批准号: 1055HG009)资助项目.

参 考 文 献

1 Dudoit S, Fridlyand J, Speed T P. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc, 2002, 97(457): 77—87[DOI]
2 Li J, Liu H, Downing J R, et al. Simple rules underlying gene expression profiles of more than six subtypes of acute lymphoblastic leukemia (ALL) patients. Bioinformatics, 2003, 19(1): 71—78[DOI]
3 Barabasi A L, Oltvai Z N. Network biology: Understanding the cell's functional organization. Nat Rev Genet, 2004, 5(2): 101—113[DOI]
4 Rives A W, Galitski T. Modular organization of cellular networks. Proc Natl Acad Sci USA, 2003, 100(3): 1128—1133[DOI]
5 Mootha V K, Lindgren C M, Eriksson K F, et al. PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet, 2003, 34(3): 267—273[DOI]
6 Segal E, Friedman N, Koller D, et al. A module map showing conditional activity of expression modules in cancer. Nat Genet, 2004, 36(10): 1090—1098

7 Guo Z, Zhang T, Li X, et al. Towards precise classification of cancers based on robust gene functional expression profiles. BMC Bioinformatics, 2005, 6(1): 58
8 Ashburner M, Ball C A, Blake J A, et al. Gene ontology: Tool for the unification of biology. Gene Ontol Consor Nat Genet, 2000, 25(1): 25—29
9 郭政, 张田文, 王琦, 等. 基于基因功能表达谱的疾病分类: 抗基因缺失稳健性. 高技术通讯, 2005, 15: 50—54
10 徐建震, 郭政, 李霞, 等. 结合基因功能分类体系筛选聚类特征基因. 生物物理学报, 2005, 21(3): 187—194
11 Tu K, Yu H, Guo Z, et al. Learnability-based further prediction of gene functions in Gene Ontology. Genomics, 2004, 84(6): 922—928[DOI]
12 Draghici S, Khatri P, Martins R P, et al. Global functional profiling of gene expression. Genomics, 2003, 81(2): 98—104[DOI]
13 Al-Shahrour F, Diaz-Uriarte R, Dopazo J. FatiGO: A web tool for finding significant associations of Gene Ontology terms with groups of genes. Bioinformatics, 2004, 20(4): 578—580[DOI]
14 Li L, Jiang W, Li X, et al. A robust hybrid between genetic algorithm and support vector machine for extracting an optimal feature gene subset. Genomics, 2005, 85(1): 16—23[DOI]
15 李霞, 饶绍奇, 张田文, 等. 应用 DNA 芯片数据挖掘复杂疾病相关基因的集成决策方法. 中国科学 C 辑: 生命科学, 2004, 34(2): 195—202
16 Xing E P, Jordan M I, Karp R M. Feature selection for high-dimensional genomic microarray data. In: Brodley C E, Danyluk A P, eds. Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA. San Fransisco: Morgan Kaufmann, 2001. 601— 608
17 Hosack D A, Dennis G Jr, Sherman B T, et al. Identifying biological themes within lists of genes with EASE. Genome Biol, 2003, 4(10): R70
18 Lapointe J, Li C, Higgins J P, et al. Gene expression profiling identifies clinically relevant subtypes of prostate cancer. Proc Natl Acad Sci USA, 2004, 101(3): 811—816[DOI]
19 Chen X, Leung S Y, Yuen S T, et al. Variation in gene expression patterns in human gastric cancers. Mol Biol Cell, 2003, 14(8): 3208—3215[DOI]
20 Yeoh E J, Ross M E, Shurtleff S A, et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. Cancer Cell, 2002, 1(2): 133—143[DOI]
21 Jimenez J L, Mitchell M P, Sgouros J G. Microarray analysis of orthologous genes: Conservation of the translational machinery across species at the sequence and expression level. Genome Biol, 2003, 4(1): R4[DOI]
22 Zhou X, Kao M C, Wong W H. Transitive functional annotation by shortest-path analysis of gene expression data. Proc Natl Acad Sci USA, 2002, 99(20): 12783—12788[DOI]
23 Osier M V, Zhao H, Cheung K H. Handling multiple testing while interpreting microarrays with the Gene Ontology Database. BMC Bioinformatics, 2004, 5(1): 124
24 Braga-Neto U M, Dougherty E R. Is cross-validation valid for small-sample microarray classification? Bioinformatics, 2004, 20(3): 374—380[DOI]
25 Simon R, Radmacher M D, Dobbin K, et al. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst, 2003, 95(1): 14—18
26 Nikitin A, Egorov S, Daraselia N, et al. Pathway studio: The analysis and navigation of molecular networks. Bioinformatics, 2003, 19(16): 2155—2157[DOI]
27 Jeong H, Mason S P, Barabasi A L, et al. Lethality and centrality in protein networks. Nature, 2001, 411: 41—42[DOI]
28 Carter S L, Brechbuhler C M, Griffin M, et al. Gene expression network topology provides a framework for molecular characterization of cellular state. Bioinformatics, 2004, 20(14): 2242—2250[DOI]
29 Han J D, Bertin N, Hao T, et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. Nature, 2004, 430(6995): 88—93[DOI]
30 Nagafuchi A. Molecular architecture of adherens junctions. Curr Opin Cell Biol, 2001, 13(5): 600—603[DOI]
31 Perl A K, Wilgenbus P, Dahl U, et al. A causal role for E-cadherin in the transition from adenoma to carcinoma. Nature, 1998, 392(6672): 190—193[DOI]
32 Kallakury B V, Sheehan C E, Ross J S. Co-downregulation of cell adhesion proteins α- and β-catenins, p120CTN, E-cadherin, and CD44 in prostatic adenocarcinomas. Hum Pathol, 2001, 32(8): 849—855[DOI]
33 Wang Z, Lo H S, Yang H, et al. Computational analysis and experimental validation of tumor-associated alternative RNA splicing in human cancer. Cancer Res, 2003, 63: 655—657
34 Gunthert U, Hofmann M, Rudy W, et al. A new variant of glycoprotein CD44 confers metastatic potential to rat carcinoma cells. Cell, 1991, 65: 13—24[DOI]
35 Vaskova M, Mejstrikova E, Kalina T, et al. Transfer of genomics information to flow cytometry: Expression of CD27 and CD44 discriminates subtypes of acute lymphoblastic leukemia. Leukemia, 2005, 19(5): 876—878[DOI]
36 Nakashima J, Tachibana M, Ueno M, et al. Association between tumor necrosis factor in serum and cachexia in patients with prostate cancer. Clin Cancer Res, 1998, 4(7): 1743—1748
37 Oh B R, Sasaki M, Perinchery G, et al. Frequent genotype changes at -308 and 488 regions of the tumor necrosis factor-α (TNF-α) gene in patients with prostate cancer. J Urol, 2000, 163: 1584—1587[DOI]
38 Sugamoto T, Tanji N, Sato K, et al. The expression of basic fibroblast growth factor and vascular endothelial growth factor in prostatic adenocarcinoma: Correlation with neovascularization. Anticancer Res, 2001, 21(1A): 77—88
39 Ortega N, Jonca F, Vincent S, et al. Modulation of the tumoral progression by anti-idiotypic antibodies of angiogenesis factors. C R Acad Sci Ⅲ, 1996, 319(5): 411—415
40 Qi L, Robinson W A, Brady B M, et al. Migration and invasion of human prostate cancer cells is related to expression of VEGF and its receptors. Anticancer Res, 2003, 23(5A): 3917—3922
41 Tozawa K. Activation of nuclera factor-kappa B and control of the expression of cell adhesion molecules in human prostate cancer cells. Nippon Hinyokika Gakkai Zasshi, 1996, 87: 1082—1091
42 Gulley J L, Arlen P M, Bastian A, et al. Combining a recombinant cancer vaccine with standard definitive radiotherapy in patients with localized prostate cancer. Clin Cancer Res, 2005, 11(9): 3353—3362[DOI]
43 Hastie C, Saxton M, Akpan A, et al. Combined affinity labelling and mass spectrometry analysis of differential cell surface protein expression in normal and prostate cancer cells. Oncogene, 2005, 24(38): 5905—5913[DOI]
44 Moro L, Arbini A A, Marra E, et al. Down-regulation of BRCA2 expression by collagen type Ⅰ promotes prostate cancer cell proliferation. J Biol Chem, 2005, 280(23): 22482—22491[DOI]
45 Rhodes D R, Chinnaiyan A M. Integrative analysis of the cancer transcriptome. Nat Genet, 2005, 37(Suppl): S31—S37[DOI]

(2005-12-18 收稿, 2006-06-14 接受)