PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2008, 35(12): 1451~1460
www.pibb.ac.cn

凋亡相关基因的整合分析及基因表达芯片的制备 *

黄茉莉[1)] 孙道春[2)] 娄雅欣[3)] 尹燕斌[1)] 李川昀[1)] 张 勇[1)]
高 歌[1)] 王升启[2)] 伯晓晨[2)] 魏丽萍[1)] 李松岗[1)**]

([1)]北京大学生物信息中心，北京大学生命科学学院，北京大学蛋白质工程和植物基因工程国家重点实验室，北京 100871；
[2)]北京放射医学研究所，北京 100850；[3)]中国医学科学院 肿瘤研究所 肿瘤医院，细胞和分子生物实验室，北京 100021)

摘要 利用 IPI 和 GenBank，收集了与凋亡相关的 5 333 条基因序列，用一定的标准，整合筛选得到 1 384 个凋亡相关的基因．通过亚细胞定位、组织表达差异显著性分析、自然正义 / 反义 RNA 对预测、基因簇在 pathway 上的定位、蛋白质 / 蛋白质相互作用等方面的分析发现，一些基因只在一个组织里显著差异表达，一些基因存在自然正义 / 反义 RNA 对现象，一些基因簇同时位于多条 pathway 等重要信息．同时，制作了一张凋亡相关基因的寡核苷酸芯片，并且对于一对 NAIF1 表达质粒转染前后的 HeLa 样品，通过该芯片的筛选，得到 24 个差异表达的基因．发现 NAIF1 过表达诱导的细胞凋亡，伴随了 PAX2、PDCD8、PDCD10、DFFA、CASP7 等基因表达的显著变化，同时还发现，U58668，该 mRNA 没有任何基因或蛋白质的注释信息，当 camptothecin 诱导 U937 细胞系凋亡时上调，在这里的 NAIF1 过表达诱导的 HeLa 细胞系中也上调(上述数据结果见 http://gpcrome.cbi.pku.edu.cn:2005/chip).

关键词 凋亡，生物信息，NAIF1，生物芯片
学科分类号 Q81

随着人类基因组测序工作的完成，数据资源的急剧膨胀，使得数据之间的交叉融合分析开始兴起．各种致力于某一领域方向的数据整合分析正成为研究的一个热点.

细胞凋亡是细胞生命周期中的重要组成部分，与胚胎发育[1]、组织发生[2, 3]、组织分化[2, 3]、组织修复[4]、内环境的稳定[4, 5]、细胞癌变[6]等过程有紧密的联系[7]．2003 年 Kutbuddin 等从 domain 入手，作了凋亡蛋白的二级数据库 (http://www.apoptosis-db.org/)[8]．Thomson Scientific 搜集了与凋亡相关的文献(http://www.esi-topics.com/apoptosis/)，IHC 对凋亡方面的文献作了分门别类工作 (http://www.ihcworld.com/apoptosis.htm)．本文不仅收集了较全的凋亡数据而且对这批数据进行了较系统的分析说明.

自从 20 世纪 90 年代中期芯片技术开始引起广泛注意以来，有成千上万个点的表达谱芯片，也有百来个点的功能分类基因芯片．就凋亡而言，目前至少已经有细胞凋亡和细胞周期基因芯片、DNA 损伤信号通路基因芯片、细胞凋亡基因芯片、p53 信号通路基因芯片、应激和毒性通路发现者基因芯片、细胞周期基因芯片等．如 SuperArray 公司制作了包含 128 个凋亡相关基因的 OHS-012 寡核苷酸芯片和 96 个凋亡基因的 HS-002 寡核苷酸芯片；DualChip 制作了包含 136 个凋亡基因的 cDNA 芯片；上海生物芯片有限公司制作了含 458 个凋亡相关基因的 cDNA 芯片(V1.0)；此外 GEArray，BD clontech，R&D Systems，Sigma-Genosys，晶美生物工程有限公司等都相继推出了凋亡芯片．目前，我们将所搜集到的 1 384 个凋亡相关的基因设计制作成 40 bp 的寡核苷酸芯片，属于一张信息相对更全面的芯片.

最近发现的 NAIF1 基因，在 HeLa 细胞中超表达可以引起细胞凋亡[9]，但其作用的分子机制尚无报道．为此，我们应用自己设计制作的人类凋亡基

* 国家高技术研究发展计划(863)资助项目(2002AA231051).
** 通讯联系人. Tel: 010-62751862, Fax: 010-62751861
E-mail: lsg@pku.edu.cn
收稿日期：2008-04-16，接受日期：2008-06-30

因芯片，从基因组的角度系统地检测 NAIF1 在 HeLa 细胞内的超表达，会引起哪些凋亡相关基因 mRNA 的表达变化，为进一步的分子机制研究和可能的信号转导途径提供线索.

1 凋亡相关基因的筛选和分析

1.1 材料与方法

1.1.1 数据库和工具. NCBI(http://www.ncbi.nlm.nih.gov/)的 mRNA(Jun 5th, 2004)和 Protein(Jun 5th, 2004)，EBI(http://www.ebi.ac.uk/IPI/IPIhelp.html)的 IPI (human IPI 2.31 (April 1st, 2004)), GoldenPath (http://genome.ucsc.edu)的基因组序列(human build hg36.1) , SWISSPROT 和 Trembl(http://www.expasy.org/sprot/), Ensembl (http://www.ebi.ac.uk/ensembl/), NATS(http://nats.cbi.pku.edu.cn/nats_list.php), BLAT[10], OligoArray[11].

1.1.2 方法.

a. 数据的收集. 所有凋亡数据的整合筛选：参考公开的文献[8, 12~14]收集描述细胞凋亡相关的结构域关键词. 用这些关键词搜索 InterPro7.0 数据库，得到 29 个 InterPro 条目. 通过 perl 程序解析 IPI 文件[15]，提取出被索引到上述 29 个 InterPro 结构域 ID 的 401 个凋亡相关的蛋白质. 同时用(“apoptotic” [All Fields] OR “apoptosis” [All Fields]) AND “human” [Organism] Limits:mRNA, excluding ESTs”)搜索 NCBI Entrez，得到 2 312 条 mRNA 和 2 620 条蛋白质序列.

b. 数据的整合.

去除冗余：通过 BLAT 将凋亡相关的 mRNA 和蛋白质定位到基因组上. 根据 length coverage>= 0.95 和 alignment identity>=0.99 筛选数据. 如果某些序列在基因组上重叠，认为这些序列是冗余的或者是可变剪切的产物，取这些序列定位位置的最左端和最右端作为基因的边界. 结果有 277 条 GenBank 的 mRNA，117 条 GenBank 的 protein 和 9 条 IPI 的 protein 无法比对到基因组上(表 1). 最后得到 1 697(1294+277+117+9)条非冗余的凋亡相关的蛋白质和 mRNA(表 1).

Table 1 The BLAT result of apoptosis sequences mapped in human genome

Data sources	Number of human mRNA/protein entries	Cutoff of BLAT mapping	Number of mapped gene location	Number of mapped entries	Total genes
Entrez nucleitide DB	2 312	0.99/0.95	1 084	2 035	
Entrez protein DB	2 620	0.99/0.95	1 134	2 503	1 294
Human IPI	401	0.99/0.95	237	392	

LocusLink 注释：解析 LocusLink ID, loc2acc, ipi.HUMAN.xref, kgXref.txt, ipi.HUMAN.xref, ensTranscript.txt 等文件，将凋亡相关的 mRNA 和蛋白质映射到对应的 LocusLink. 结果共有 5 136 条序列对应 1 234 个已知的基因，还有 163 条序列没有相应的 LocusLinkID(表 2). 如此，我们共得到 1 397(1 234+163)个与凋亡相关的基因.

Table 2 Apoptosis mRNA or protein mapped corresponding GeneID link

Data sources	Number of human mRNA/protein entries	Number of entries mapped to public gene/LocusLink IDs	Number of genes/ LocusLink IDs	Number of entries not mapped to public gene/LocusLink IDs	Total genes
Entrez nucleitide DB	2 312	1 551+594	1 040	167	
Entrez protein DB	2 620	1 578+824	1 115	218	1 213
Human IPI	401	281+39+40+17	230	24	

1.2 结果

1.2.1 亚细胞定位. 从 Swissprot 中解析出这 1 397 个基因对应的蛋白质亚细胞定位条目，根据我们自己定义的亚细胞定位词汇表，得到 849 个蛋白质有不同的亚细胞定位结果. 其中 259 个蛋白质至少有 2 个以上的不同亚细胞定位，45 个蛋白质位于线粒

体，大多数蛋白质定位于细胞膜和细胞质(图 1)，这与细胞凋亡发生的关键环节不在细胞核而在细胞质一致．为了验证该结果是否可靠，我们解析了

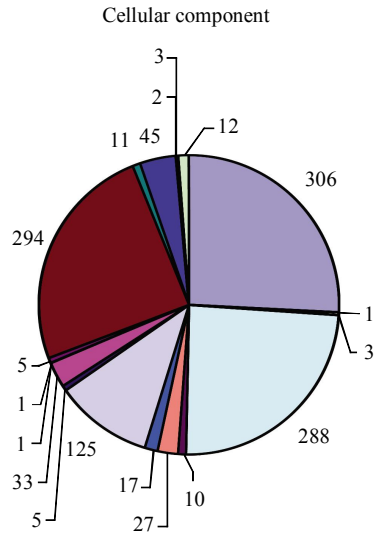

Fig. 1　Distribution in various subcellular location

■: Cell membrane; ■: Synaptosome; □: Synapse; □: Cytoplasm; ■: Lysosome; ■: Golgi apparatus; ■: Extracellular matrix; □: Extracellular; ■: Endosome; ■: Endoplasmic reticulum; □: Vesicle; □: Vacuole; ■: Peroxisome; ■: Nucleus; ■: Nucleolar; ■: Mitochondrion; □: Midbody; □: Microsome.

GeneCards 的亚细胞定位结果，发现只有 229 个蛋白质在 GeneCards 中有亚细胞定位注释，且与我们的结果近 100%的一致．此外，我们随机挑选了几个蛋白质，发现结果与文献一致：如 Green 和他的研究小组发现，p53 蛋白可从细胞核转移到细胞质，一旦在细胞质中稳定下来，就会诱导细胞凋亡．总的说来，我们注释得到的亚细胞定位信息与 GeneCards 和文献的结果相当一致，且对于某些蛋白质，我们发现了更多的亚细胞定位，如 MAPK8IP2 不仅仅位于细胞质内，而且也在细胞外基质和细胞膜．

1.2.2　组织显著性差异表达．解析 CGAP 的数据，对于每个基因，用超几何分布检验对应的 EST 丰度显著性($P < 0.01$)，发现 7 个基因只在一个组织里检测到 EST 片段(表 3)．其中 GAST 与 Dockray 等[16]所说的一致，是促进胃酸的分泌和上皮细胞的增殖，且在胃癌组织里高表达[17]．IL9、CCL1、KRTP9-2 等 6 个基因没有在目前的文献中指出只在一个组织里表达．另外，还有 91 个基因虽然在多个组织里表达，但是只在一个组织里差异显著表达．这些基因呈现的组织特异性过表达和这样的过表达方式如何调控的，有待于我们进一步研究发现．

Table 3　Genes of over/under representations in only one tissue

GeneID	Symbol	Gene name	Tissue	The observed number of EST sequences	The expected number of EST sequences[1)]	Over/under representation[2)]
3578	IL9	Interleukin 9	Kidney	2	0.08	1.29E-03
6346	CCL1	Chemokine (C-C motif) ligand 1	Lung	2	0.12	4.12E-03
83899	KRTAP9-2	Keratin associated protein 9-2	Heart	3	0.001	3.98E-06
3906	LALBA	Lactalbumin, alpha	Germ cell	2	0.001	1.05E-04
1237	CCR8	Chemokine (C-C motif) receptor 8	Germ cell	2	0.001	1.05E-04
5657	PRTN3	Proteinase 3 (serine proteinase, neutrophil, Wegener granulomatosis autoantigen)	Bone marrow	2	0.001	6.80E-05
2520	GAST	Gastrin	Stomach	22	0.5	3.69E-34

[1)]Expected number of EST sequences for a specific gene in a specific tissue = Total number of EST sequences for the gene in all tissues× Total number of EST sequences for all genes in the tissue/Total number of EST sequences for all genes in all tissues.

[2)]The *P* value is calculated by hypergeometric test.

1.2.3　自然正义 / 反义 RNA 对(NATS)．我们搜索来自 UniGene 数据库中反向重叠至少 20 个碱基的转录本对(mRNA/EST)，其染色体位置来自存储于 GoldenPath 数据库中的 BLAT 比对信息．Ensembl 人类转录本中内含子的最大长度[18](200 kb)作为内含子长度的极大估计．为确定 EST 方向，使用 polyA 信号、polyA 尾和标准剪切位点一致性序列 GT-AG 作为三个证据．发现有 415 条基因可能存

在NATS，该比例(～34%＝415/1234)远远高于预测的人类所有可能NATS水平(～20%)(http://nats.cbi.pku.edu.cn/nats_list.php)．这也许说明凋亡基因有更丰富的转录后调控．表4为挑选出的在overlap区保守性＞90%的NATS．其中至少有2对(MKRN2、RHOBTB2)NATS已经被报道．如MKRN2[19]仅当选择性多聚腺苷酸化不同的时候，转录体不同．较长的转录体，与RAF1的3′UTR重叠．较短的转录体则与RAF1没有重叠区．用RNA印迹分析发现，较长的转录体在所有的正常组织和大多数肿瘤细胞系中能检测到．而较短的转录体在直肠和外围血白细胞中没有监测到．又如在肺癌细胞系中，RHOBTB2与泛素连接酶骨架蛋白CUL3结合降解．如果第一个BTB domain区的第284位点Y-＞D，则无法与CUL3结合，从而逃避被泛素降解，导致体内的RHOBTB2积累．与此[20]对应的正义RNA--TNFRSF10B在第1 065位点ACAC插入2 bp，转录就提前终止．当这种突变的cDNA转染进直肠癌细胞系和卵巢细胞系时，会抑制集落的形成．

Table 4 Genes seem to be involved in sense/antisense pairs，which demostrate high conservation in overlapping regions

Gene1	Gene name	Gene2	Gene name	Overlap
ILK	Integrin-linked kinase	TAF10	TAF10 RNA polymerase Ⅱ, TATA box binding protein (TBP)-associated factor	0.986
RAF1	v-raf-1 Murine leukemia viral oncogene homolog 1	MKRN2	Makorin, ring finger protein, 2	0.973
STAT6	Signal transducer and activator of transcription 6, interleukin-4 Induced	NAB2	NGFI-A binding protein 2 (EGR1 binding protein 2)	0.983
BAT3	HLA-B Associated transcript 3	APOM	Apolipoprotein M	0.971
SNN	Stannin	TXNDC11	Thioredoxin domain containing 11	0.971
TNFRSF10B	Tumor necrosis factor receptor superfamily, member 10b	RHOBTB2	Rho-related BTB domain containing 2	0.959
DEDD	Death effector domain containing	NIT1	Nitrilase 1	0.979
AKT3	v-akt Murine thymoma viral oncogene homolog 3 (protein kinase B, gamma)	SDCCAG8	Serologically defined colon cancer antigen 8	0.984
WDR3	WD repeat domain 3	PF6	Projection protein	0.943
APPL	Adaptor protein containing pH domain, PTB domain and leucine zipper motif 1	ASB14	Ankyrin repeat and SOCS box-containing 14	0.912
CYFIP2	Cytoplasmic FMR1 interacting protein 2	ADAM19	A disintegrin and metalloproteinase domain 19 (meltrin beta)	0.936
PCBP4	Poly(rC) binding protein 4	GPR62	G protein-coupled receptor 62	0.987
SYVN1	Synovial apoptosis inhibitor 1, synoviolin	MRPL49	Mitochondrial ribosomal protein L49	0.933
UNC5A	Unc-5 homolog A (*C. elegans*)	HK3	Hexokinase 3 (white cell), nuclear gene encoding mitochondrial protein	0.991

1.2.4 基因簇．我们发现，这批基因在染色体上有很明显的基因簇现象存在(图2)，于是收集NCBI MapView中所有人类基因在染色体上的位置信息和个数，并以此为背景，通过修正泊松分布，调节P值计算出这批基因在每条染色体上成簇的最大距离．具体方法如下：

取零假设(H0)为：基因在染色体上的位置是均匀分布的．设此染色体长度为L，其上共有N个基因．在一段长为l的区域中，出现k个基因的分布服从二项分布：

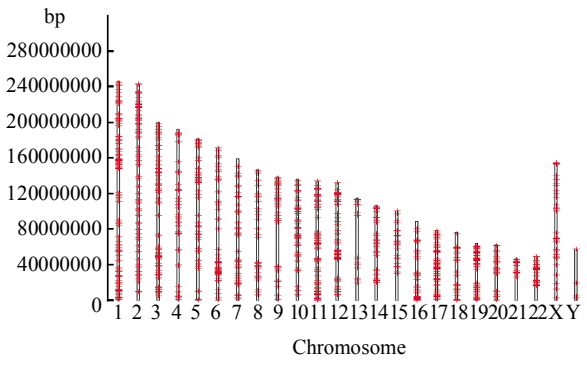

Fig. 2 Distribution of apoptosis genes in chromsome

$$P(n=k)=\left(\begin{array}{c}N\\k\end{array}\right)\left(\frac{l}{L}\right)^{K}\left(1-\frac{l}{L}\right)^{N-k}$$

N/L 较小时，二项分布可以用 Possion distribution 来近似表示. 所以在长为 l 的区域中，出现一个及一个以上的基因的概率：$P(n>=1)\approx 1-e^{-d_0 \cdot l}$

设给定 cluster 中共有 *N* 个基因，定义两个相邻基因间的距离分别为 l_1，l_2……l_{n-1}，依照上式计算出 $n-1$ 个 *P*-value，为 P_1，P_2……P_{n-1}. 此时：

$$P(n>=N)=\left(\prod_{i=1}^{N-1}P_i\right)$$

我们来检验一条染色体上基因的整体聚集现象是否存在，具体过程如下：

a. 对给定染色体进行随机化，将所有基因随机分配位置.

b. 针对随机化后的染色体，采用现有参数重新计算并寻找基因簇(在实现中，当按照全局 D 已经找到一个可能的基因簇后，我们将以这个新找到的基因簇为中心，取一个 4M 的窗口，统计其中的基因密度，重新计算一个 Local D，对于基因密集区域来说，Local D 会小于 Global D. 因而，作为一个保守的估计，我们将根据 min (local D，global D)重新对这个基因簇进行调整)，而后计算在基因簇中的基因数与总基因数的比例(基因簇密度，gene cluster density)，作为统计量.

$$\text{gcd}=\frac{\text{the number of genes in cluster}}{\text{the number of all genes}}$$

c. 重复上述过程若干次，根据结果计算 $P(\text{gcd} >= \text{gcd}_0)$，即基因簇密度大于原始值的次数.

由此，我们计算得到 251 个凋亡相关的基因成簇分布在染色体的不同部位，并且有 31 簇基因分布在 8 条 pathway 上(map 到 KEGG Pathway). 从基因组的角度来说，同功能(或类似功能)的基因往往倾向于在不同时刻、以不同的水平表达，从而完成特定的生物学功能，以基因簇形式组织对它们正常发挥作用并没有优势. 基因簇主要是一种表达调控单位而非功能单位，参与同一代谢途径的基因需要同时表达，因而倾向于以基因簇的方式聚集在同一簇中. 例如 IL1A\IL1B\CHCHD5、IL3\CSF2、TNFRSF10B\TNFRSF10C\TNFRSF10D\TNFRSF10A、IFNA2\IFNA1 至少在 3 条通路上同时出现(表 5). 有意思的是，有些基因簇中的基因是新近发现的基因，但是在目前的 pathway 上没有出现. 我们认为这些新基因很可能位于同一 pathway 的上下游. 例如 IL1A\IL1B\CHCHD5，IL1A\IL1B 为白介素家族成员，参与免疫反应、炎症和造血，与风湿性关节炎、神经退行性疾病相关，同时在多条 pathway 出现，而 CHCHD5 为一个新的基因，在目前的数据库里没有任何功能性注释信息，很可能 CHCHD5 也应该位于这些 pathway. 还有 CASP4/CASP5/CASP1/COP1/LOC648470，COP1 与 CASP1 前体

Table 5 Distribution of 31 gene cluster in 8 pathways

Cytokine-cytokine receptor interaction:			
IL10	L19	IL24	FAIM3
IL3	CSF2		
IFNA2	IFNA1		
IL1A	IL1B	CHCHD5	
CXCL9	CXCL10	CXCL11	
LTA	TNF		
LTBR	TNFRSF7	GAPDH	
MFN2	TNFRSF8	TNFRSF1B	
TNFRSF18	TNFRSF4		
TNFRSF10B	TNFRSF10C	TNFRSF10D	TNFRSF10A
IL18R1	IL18RAP		
Apoptosis:			
BIRC3	BIRC2		
IL1A	IL1B	CHCHD5	
CFLAR	CASP10	CASP8	
TNFRSF10B	TNFRSF10C	TNFRSF10D	TNFRSF10A
MAPK signaling pathway:			
MAP3K10	AKT2		
IL1A	IL1B	CHCHD5	
LOC648470	CASP4	CASP5	CASP1
CFLAR	CASP10	CASP8	
Focal adhesion:			
BIRC3	BIRC2		
CAV2	CAV1		
Toll-like receptor signaling pathway:			
TLR10	TLR1	TLR6	
IFNA2	IFNA1		
CXCL9	CXCL10	CXCL11	
Jak-STAT signaling pathway:			
IL10	IL19	IL24	FAIM3
IL3	CSF2		
IFNA2	IFNA1		
STAT5B	STAT5A	STAT3	
Natural killer cell mediated cytotoxicity:			
IFNA2	IFNA1		
TNFRSF10B	TNFRSF10C	TNFRSF10D	TNFRSF10A
Hematopoietic cell lineage:			
IL3	CSF2		
IL1A	IL1B	CHCHD5	

相似，LOC648470 为新基因，与 CASP4 前体相似，这些基因很可能与 CASP4、CASP5、CASP1 共表达，位于相同的通路，只是目前没有实验验证或者被发现而已.

1.2.5 Pathway 分析[21]. 以人的所有基因为背景，用超几何分布检验基因落在各条 pathway 上的显著性水平，取 $P < 0.001$，发现有 463 个基因差异显著的分布在 33 条 pathway 上(图 3). 观测基因分布最多、差异最显著的这些 pathway，发现它们都与细胞凋亡相关. 例如几乎覆盖了 Cytokine-cytokine receptor interaction 通路上的 IL2RG shared 和 TNF family. 在 MAPK signaling pathway 中，几乎覆盖了 Wnt signaling pathway 和 apoptosis 子通路上的所有基因，也说明凋亡是一个复杂的相互作用过程.

1.2.6 蛋白质相互作用. 本文对人类凋亡相关蛋白质相互作用网络进行了分析，证实了蛋白质相互作用网络具有无尺度性质的存在. 我们从 BIND[22]中抽取这 1 384 个蛋白质相互作用数据，用 Medusa 软件勾画出蛋白质 / 蛋白质相互作用网络，数据基本吻合 power-law 分布(图 4). MYC(95)、JUN (85)、MAX (79)、TP53 (54)、E2F1 (44)、RBL2 (36)、RB1(36)为该网络的 hub 节点(表 6). 这些 hub 节点与已有文献报道相当一致. 例如 MYC[23] 基因的突变与许多癌症(乳腺癌、肠癌、黑色素瘤等)有关，它所编码的蛋白质结合人类基因的 15%，又如 JUN，广泛的分布在 44 个 pathway.

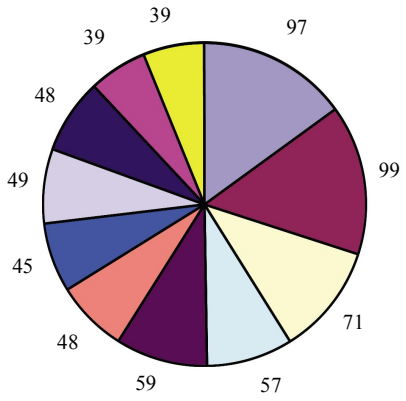

Fig. 3　Numbers of genes distributed in main pathways

■: MAPK signaling pathway; ■: Cytokine-cytokine receptor interaction; ■: Focal adhesion ; ■: Apoptosis ; ■: Toll-like receptor signaling pathway ; ■: Regulation of actin cytoskeleton ; ■: T cell receptor signaling pathway ; ■: Natural killer cell mediated cytotoxicity ; ■: Jak-STAT signaling pathway; ■: Cell cycle; ■: Wnt signaling pathway.

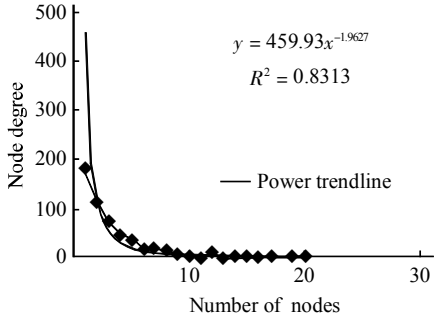

Fig. 4　Distribution of node degrees

Table 6　Hub proteins and their corressponding interacting proteins Hub interaction protein

MYC	BRCA1///CDKN2A///CHUK///CSNK2A1///E2F1///JUN///BCL2L11///PARP2///KIF20A///CDK2///CDK4///CDKN1A///CDKN1B///MAEA///PAK4///HBXIP///NFAT5///WDR4///PDCD10///TUSC2 ///CISH///CLN3///CREB1///DAXX///DDB2///ERCC6///ATF6///DKK1///CORO1C///FRAP1///XRCC6///PRDX5/////TNFRSF21///PYCARD///GPR132///HIP2///IGFBP1///IL15///ILK///ING1///IRF3///MAPT///MAX///MDM4///MAP3K5///MT2A///PPP1R12A///NCL///ATM///NME1///PAK2///LSR/////PML///GNB1L///PPIA///PPP2R1B///PRKAB1///C20orf24///PRKRIR///DIABLO///PTGER3///BIRC6///BARD1///RAC2///SAV1///RPS6KB1///SOD1///SREBF2///TP53///DAP3///CALR///TRRAP///PLA2G6///PIAS1///MADD/////CRADD///RIPK2///SGPL1///NMI///LATS1///PDCD8///GSTO1///RNF7///CD79B///CDC6///THOC1///POLR2A///MAPK3///RB1///RBL2///XRCC5///PARP10
JUN	BRCA1///MAPK14///FOS///BCL2L11///CDK4///CDKN1B///TADA3L///GADD45G///CLU///CREB1///CREM///PARP1///CSNK2A1///CTSD///CYP1A1///DAPK1///JAG1///EGR1///EGR2///EP300///AKT1///F7///FGF2///FGFR1///FOXO3A///G6PD///GJA1///BBC3///GLRX///HIPK2///GSTP1///HIF1A///HLA-A///HMOX1///BIRC5///FASLG///IL3///IL4///ING1///AR///JUNB///JUND///LGALS8///MCL1///MFGE8///MKI67///MME///MMP2///MYBL2///MYC///ATF3 ///ATM///NFKB2///SERPINB5///PIN1///PKD1///PMS2///MAPK8///PSEN2///BAX///BCL2///CCL2///BID///BIK///SHC1///BNIP1///SREBF1///TGFB1///TGFBR2///TIMP3///TNF///TNFAIP3///TRAF3///TRRAP///CAV1///HDAC3///CCNA2///CCNB1///GDF15///CDC2///MAPK11///MAPK9///MAPK10///RB1///RELA
MAX	PARP2///KIF20A///CDKN1B///MAEA///RTN3///HBXIP///NFAT5///WDR4///PDCD10///TUSC2///CISH///CLN3///CREB1///DAXX///ERCC6///ATF6///CORO1C///FRAP1///XRCC6///PRDX5/////TNFRSF21///PYCARD///GPR132///HIP2///ILK///ING1///IRF3///LGALS1///MAP3K5///MT2A///GADD45B///PPP1R12A///NCL///ATM///NME1///LSR/////TLR9///GNB1L///PPIA///PPP2R1B///PRKAB1///PKN1///C20orf24///PRKRIR///DIABLO///BIRC6///BARD1///RAC2///SAV1///RPS6KB1///MOAP1///SOD1///SREBF1///SREBF2///TP53///DAP3///FXR1///CALR///PLA2G6///IKBKAP///PIAS1///MADD/////CRADD///RIPK2///SGPL1///LATS1///PDCD8///STK17B///GSTO1///BAG5///RNF7///CD79B///CDC6///THOC1///CASP8AP2///MYC

Continued

TP53	ATF3///ATM///ATR///BCL2L1///BCL6///BLM///BRCA1///DHCR24///EP300///FOXO3A///HD///HIF1A///MAX///MDM2///MUC1///MYC///MAPK8///PTGS2///CDKN1A///CREBBP///E2F1///XRCC6///GAPDH///SFN///NR3C1///GSK3B///APEX1///ING1///ING4///PIAS4///PTEN///BAK1///BARD1///BAX/////BNIP3L///STK6///TXN///USP7///ING5///TNFRSF10C///TNFRSF10B///PTTG1///TP53I3///TP53BP1///TP53BP2///WRN///PIAS1///MTA2///PPP1R13L///CHEK2/////SIRT1///HIPK2
E2F1	BRCA1///CDKN2A///PARP2///MAEA///RTN3///HTATIP2///DCC///DDB2///PPP1R13B///XRCC6///AATF///BIRC5///KCNA3///MDM4///MYBL2///MYC///PPP1R12A///NCL///ATM///PIN1///PPIA///DIABLO///BARD1///RB1///RBL2///BMP4///ZFP36L1///TFDP1///TIAM1///TOP2A///TP53BP2///XRCC5///YWHAE///PLA2G6///PIAS1///CCNA2///LATS1///TP53INP1///GSTO1///CD79B///CDC2///CDC6///CASP8AP2///TP53
RBL2	E2F1///MCM7///RAF1///KIF20A///ZNF443///TRAIP///PAK4///CHEK2///AK2///SIRT1///GAPDH///AATF///GZMA///APEX1///BIRC5///IRF3///MDM2///MYBL2///MYC///ATM///GTSE1///PLK1///PLSCR1///BLM///BRCA1///MAP3K7///TERF1///TOP2A///ANP32E///CCNA2///CCNB1///PTTG1///CDC2///CDC6///THOC1///CASP8AP2
RB1	BRCA1///CASP2///CASP3///CASP6///CASP7///CASP8///CASP9///CASP10///E2F1///MCM7///RAF1///CDK2///CDK4///CDKN1A///EPO///NCOA6///CRI1///ABL1///HDAC1///HIF1A///BIRC5///AR///JUN///MDM2///MDM4///MYC///SERPINB2///PPIA///MAPK1///MAPK9///CCND1///BCL2///RFP///VEGF///YWHAE///AATF

2 芯片的制作和应用

2.1 材料与方法

2.1.1 探针的设计．用 OligoArray2_1 对这 1 397 个基因设计探针，探针长度 38～42nt，GC 含量 35%～50%，去除简单重复片段，TM82～88，自身无二级结构，选取编码区近 3′端探针进行 BLAST 同源性分析，本地的数据库为从 NCBI 中筛选出去除 EST 的人所有的 220 402 条 mRNA．最后得到 1 384 条有效的探针序列．

2.1.2 探针的布局．每张芯片共有 1 548 个探针(包括质控点)，其中 mRNA 1 492 个点 (118 个重复的 mRNA)．质控方法采用阳性对照、看家基因、阴性对照和点样液，以及随机选择布置重复点．具体如下：阳性对照荧光素酶基因探针 13 个点，分布于四周与中间，看家基因共 29 个点，阴性对照共 11 个点，空白点样液 3 个点．每张芯片上随机选取 118 条探针重复地随机分散在芯片各个部位．

2.1.3 细胞的培养和转染．人宫颈癌细胞系 HeLa 由北京大学生物信息中心传代培养，细胞正常培养在含 10%(体积分数)胎牛血清(fetal bovine serum，FBS)、100 U/ml 青霉素、100 mg/L 链霉素的 DMEM 培养基中．细胞培养过程中，维持细胞的密度不超过 80%融合．细胞转染采用电穿孔法，即消化细胞，用无血清 DMEM 培养基洗细胞 3 次，调整细胞浓度为 2×10^6～5×10^6/ml，取细胞 300 μl/ 样品，加 NAIF1表达质粒或 pcDNA3 对照质粒 10 μg，混匀后移入 2 mm 电转杯，电击条件 120 V，20 ms．转染后室温放置 10 min，吸出并重悬于完全 DMEM 培养基中，培养 6～8 h 细胞贴壁后更换新鲜的培养基，于转染后 48 h 收获细胞提取细胞总 RNA．

2.1.4 芯片的杂交．按照 Trizol RNA 提取试剂盒说明书提取总 RNA，紫外分光光度计定量，A_{260}/A_{280}在 1.8～2.0 之间，RNA 甲醛变性电泳显示无降解．然后进行逆转录荧光标记，总 RNA 分别用不同的荧光(Cy3 或 Cy5)标记．按 1∶3 的比例将逆转录产物与杂交液(50%甲酰胺，5×Denhardt 液，6×SSC，0.5%SDS)混合，94℃ 水浴变性 2 min，42℃ 杂交 3 h，室温下分别用洗液 A(1×SSC，0.2% SDS)，洗液 B(0.2×SSC)，洗液 C(0.1×SSC)清洗，室温晾干．使用 Genepix 4000 扫描仪对杂交结果扫描，调节 PMT 值使阳性对照的 Cy5/Cy3 信号值比值接近 1．使用 GenePix4.0 软件对杂交结果进行数据分析和归一化处理，用看家基因和阳性对照对 Cy3 和 Cy5 扫描结果进行校正[24].

2.2 芯片分析结果

总共做了两次重复的染色交换实验，对于扫描校正的结果，去除强度值 < 80 的点．采用成组设计的 t 检验，$P < 0.01$ 且比值大于 2 或小于 0.5 倍的，挑选出 24 个差异表达的基因(5 个下调，19 个上调，表 7)．随机选取 PAX2、FADD、TNFSF10、PDCD10、PDCD8、DFFA 等 6 个差异表达的基因做 RT-PCR 分析(图 5)，结果和芯片基因一致．

Table 7 List of representative differentially expressed genes in the HeLa-NAIF1 cells and HeLa-pcDNA3 cells

ProbeID	Symbol	Gene name	Mean ratio	*P*-value
NM_004205	USP2	Ubiquitin specific peptidase 2	0.21	3.02E-10
NM_004672	MAP3K6	Mitogen-activated protein kinase kinase kinase 6	0.40	2.19E-06
NM_003988	PAX2	Paired box gene 2	0.37	2.50E-06
NM_001375	DNASE2	Deoxyribonuclease Ⅱ, lysosomal	4.76	2.52E-06
NM_002117	HLA-C	Major histocompatibility complex, class Ⅰ, C	4.26	2.35E-05
NM_001558	IL10RA	Interleukin 10 receptor, alpha	0.16	2.40E-05
NM_005356	LCK	Lymphocyte-specific protein tyrosine kinase	0.17	7.16E-05
U58668			4.69	8.69E-05
NM_213566	DFFA	DNA fragmentation factor, 45 ku, alpha polypeptide	4.11	0.000 165
NM_003824	FADD	Fas (TNFRSF6)-associated via death domain	4.87	0.000 176
XM_371008	LOC388326	Similar to death effector filament-forming Ced-4-like apoptosis protein	0.29	0.000 339
NM_001748	CAPN2	Calpain 2, (m/Ⅱ) large subunit	4.27	0.000 361
NM_020532	RTN4	Reticulon 4	4.96	0.000 41
NM_016166	PIAS1	Protein inhibitor of activated STAT, 1	4.99	0.000 477
NM_002661	PLCG2	Phospholipase C, gamma 2 (phosphatidylinositol-specific)	4.67	0.000 52
NM_000660	TGFB1	Transforming growth factor, beta 1 (Camurati-Engelmann disease)	3.70	0.000 763
NM_003810	TNFSF10	Tumor necrosis factor (ligand) superfamily, member 10	4.88	0.000 864
NM_007217	PDCD10	Programmed cell death 10	3.96	0.000 945
NM_004208	PDCD8	Programmed cell death 8 (apoptosis-inducing factor)	4.52	0.001 409
NM_002117	HLA-C	Major histocompatibility complex, class Ⅰ, C	4.41	0.002 673
NM_000885	ITGA4	Integrin, alpha 4 (antigen CD49D, alpha 4 subunit of VLA-4 receptor)	4.51	0.003 302
NM_014326	DAPK2	Death-associated protein kinase 2	4.22	0.003 318
NM_033339	CASP7	Caspase 7, apoptosis-related cysteine peptidase	3.73	0.004 014
NM_005729	PPIF	Peptidylprolyl isomerase F (cyclophilin F)	4.52	0.004 946

Shown here are the means of the intensity ratios of 4 separate experiments.

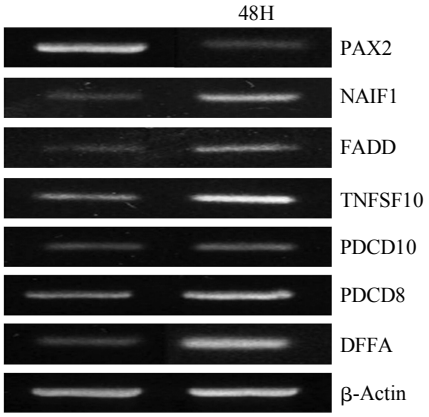

Fig. 5 Validation of oligonucleotid microarray results of 6 selected genes by semiquantitative RT-PCR

3 讨 论

通过结合公开的数据和统计方法，对 1 384 个凋亡相关的基因作了亚细胞定位、组织差异表达显著性分析、自然正义 / 反义 RNA 对预测、基因簇与 pathway 分析、在 pathway 上的显著性分布分析、蛋白质 - 蛋白质相互作用等方面的分析，在某些方面发现了一些有趣的现象.

例如，一个蛋白质对应多个亚细胞定位，很可能预示着某种功能的变化，在我们的结果里，近 31%的蛋白质具有 2 个以上的亚细胞定位，45 个蛋白质位于线粒体. 此外，我们利用 CGAP 的数据，计算各类组织的 cDNA 文库中代表该基因的 EST 数量，发现 7 个基因只在一个组织里表达，98 个基因只在一个组织里显著差异表达.

此外，Fukuoka 等[25]指出，共表达现象在真核生物中是普遍存在的，而且共表达的基因有成簇[10 k 距离]存在的趋势. 在这次试验中，我们也发现了多个基因簇不仅位于同一条 pathway，而且同时在多条 pathway 出现. 有些基因簇中有某个基因几乎没有功能注释，例如 CHCHD5，在 pathway 中也没有出现，很可能是没有被实验发现. 又如，NATs 在真核生物基因调节的多个层次上起作用：包括 X- 染色体失活、印记基因(imprinted gene)的产生、RNA 的稳定性和转运、可变剪切和翻译调节[26, 27]. 另外，反义 RNA 的改变会导致某些疾病，

包括癌症[28]和神经退行性疾病[29, 30]．我们通过本实验室的预测 NATS 流程，发现 415 对可能的凋亡相关基因存在自然正义 / 反义 RNA 对．尽管我们还没用实验验证这种准确性有多高，为什么与凋亡的基因中有如此多的自然反义 RNA 存在，但是至少给了我们下一步检测反义 RNA 提供了强有力的证据．

对于这批数据，我们制作了寡核苷酸芯片，应用于检测 NAIF1 在 HeLa 细胞内的超表达会引起哪些凋亡相关基因 mRNA 的表达变化．结果发现了 24 个差异表达的基因，并随机地选择 6 个做 RT-PCR，与芯片的结果基本吻合，初步认为主要是通过促凋亡基因而诱导细胞凋亡，为进一步的 NA 诱导细胞凋亡的分子机制研究和可能的信号转导途径提供线索．

由于整张芯片，是包含了所有凋亡可能的数据，所以，从系统的角度，宏观地分析问题，应该是很占优势的．这也是芯片的一个极其重要的特征．我们相信，通过这张芯片，可以对多种样品进行系统的检测与凋亡相关的生物学问题，为进一步研究细胞的发育和疾病提供基础．

致谢 本文得到了赵敏博士、吴健民博士的指点和帮助，在此一并表示感谢．

参 考 文 献

1 Huppertz B, Kadyrov M, Kingdom J C. Apoptosis and its role in the trophoblast. Am J Obstet Gynecol, 2006, **195**(1): 29～39

2 O′Connor L, Huang D C, O′Reilly L A, *et al*. Apoptosis and cell division. Curr Opin Cell Biol, 2000, **12**(2): 257～263

3 Blagosklonny M V. Apoptosis, proliferation, differentiation: in search of the order. Semin Cancer Biol, 2003, **13**(2): 97～105

4 Xavier P A. Apoptosis and human reproduction. Acta Med Port, 2002, **15**(4): 287～291

5 Altieri D C. Survivin and apoptosis control. Adv Cancer Res, 2003, **88**: 31～52

6 Schuchmann M, Galle P R, Kanzler S. Apoptosis in disease. Med Klin (Munich), 2002, **97**(12): 738～746

7 赵卫红, 寿好长, 福 岭, 等. 细胞凋亡. 郑州：河南医科大学出版社, 1997. 5～270
Zhao W H, Shou H Z, Fu L, *et al*. Cell Apoptosis. Zhengzhou: Henan medical University Press, 1997. 5～270

8 Doctor K S, Reed J C, Godzik A, *et al*. The apoptosis database. Cell Death Differ, 2003, **10**(6): 621～633

9 Lv B, Shi T, Wang X, *et al*. Overexpression of the novel human gene, nuclear apoptosis-inducing factor 1, induces apoptosis. Int J Biochem Cell Biol, 2006, **38**(4): 671～683

10 Kent W J. BLAT--the BLAST-like alignment tool. Genome Res, 2002, **12**(4): 656～664

11 Rouillard J M, Zuker M, Gulari E. OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res, 2003, **31**(12): 3057～3062

12 Aravind L, Dixit V M, Koonin E V. The domains of death: evolution of the apoptosis machinery. Trends Biochem Sci, 1999, **24**(2): 47～53

13 Aravind L, Dixit V M, Koonin E V. Apoptotic molecular machinery: vastly increased complexity in vertebrates revealed by genome comparisons. Science, 2001, **291**(5507): 1279～128

14 Yin Y B, Zhang Y, Yu P, *et al*. Comparative study of apoptosis-related gene loci in human, mouse and rat genomes. Acta Biochim Biophys Sin, 2005, **37**(5): 341～348

15 Kersey P J, Duarte J, Williams A, *et al*. The international protein index: an integrated database for proteomics experiments. Proteomics, 2004, **4**(7): 1985～1988

16 Dockray G J. Clinical endocrinology and metabolism. Gastrin. Best Pract Res Clin Endocrinol Metab, 2004, **18**(4): 555～568

17 Mukawa K, Fujii S. Analysis of K-ras mutations and expression of cyclooxygenase-2 and gastrin protein in laterally spreading tumors. J Gastroenterol Hepatol, 2005, **20**(10): 1584～1590

18 Birney E, Andrews T D, Bevan P, *et al*. An overview of Ensembl. Genome Res, 2004, **14**(5): 925～928

19 Gray T A, Azama K, Whitmore K, *et al*. Phylogenetic conservation of the makorin-2 gene, encoding a multiple zinc-finger protein, antisense to the RAF1 proto-oncogene. Genomics, 2001, **77** (3): 119～126

20 Wilkins A, Ping Q, Carpenter C L, *et al*. RhoBTB2 is a substrate of the mammalian Cul3 ubiquitin ligase complex. Genes Dev, 2004, **18** (8): 856～861

21 Kanehisa M. The KEGG database. Novartis Found Symp, 2002, **247**: 91～101; discussion 101～103, 119～128, 244～252

22 Bader G D, Betel D, Hogue C W. BIND: the biomolecular interaction network database. Nucleic Acids Res, 2003, **31** (1): 248～250

23 Knoepfler P S, Zhang X Y, Cheng P F, *et al*. Myc influences global chromatin structure. EMBO J, 2006, **25**(12): 2723～2734

24 Guo W F, Lin R X, Huang J, *et al*. Identification of differentially expressed genes contributing to radioresistance in lung cancer cells using microarray analysis. Radiat Res, 2005, **164**(1): 27～35

25 Fukuoka Y, Inaoka I, Kohane I S. Inter-species differences of co-expression of neighboring genes in eukaryotic. Genomes, 2004, **5** (1): 4

26 Borsani O, Zhu J, Verslues P E, *et al*. Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in Arabidopsis. Cell, 2005, **123**(7): 1279～1291

27 Vanhee-Brossollet C, Vaquero C. Do natural antisense transcripts make sense in eukaryotes?. Gene, 1998, **211**(1): 1～9

28 Lavorgna G, Dahary D, Lehner B, *et al*. In search of antisense. Trends Biochem Sci, 2004, **29**(2): 88～94

29 Korneev S, O′Shea M. Natural antisense RNAs in the nervous system. Rev Neurosci, 2005, **16**(3): 213～222

30 Nagase T, Seki N, Tanaka A, *et al*. Prediction of the coding sequences of unidentified human genes. Ⅳ. The coding sequences of 40 new genes (KIAA0121-KIAA0160) deduced by analysis of cDNA clones from human cell line KG-1. DNA Res, 1995, **2** (4): 167～174, 199～210

Analysis and Preparation of Oligonucletide Microarray of Apoptosis-related Genes*

HUANG Mo-Li[1)], SUN Dao-Chun[2)], Lou Ya-Xin[3)], YIN Yan-Bin[1)], LI Chuan-Yun[1)], ZHANG Yong[1)], GAO Ge[1)], WANG Sheng-Qi[2)], BO Xiao-Chen[2)], WEI Li-Ping[1)], LI Song-Gang[1)]**

([1)]*Center of Bioinformatics, National Laboratory of Genetic Engineering and Protein Engineering,College of Life Sciences, Peking University, Beijing* 100871, China; [2)]*Beijing Institute of Radiation Medicine, Beijing* 100850, China; [3)]*Laboratory of Cell and Molecular Biology, Cancer Institute & Cancer Hospital, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing* 100121, China)

Abstract With some conditions, there are 1 384 genes selected from 5 333 possible apoptosis sequences which were mined from IPI and GenBank. After the analysis of their subcellular location, tissue expression remarkably, natural antisense transcripts predict, gene clusters distributed in pathways, protein protein network, some interested things were mined. Some genes are differentially expressed in a tissue type; Some genes are NATS and some gene clusters are in more than one pathway. Meanwhile, one 40-bp oligonucleotide microarray which includes most apoptosis genes was made. With the microarray and samples of HeLaT-NAIF1 and HeLaT-pcDNA3, there were 24 genes differentialy expressed. Perhaps, when the Naif1 was over expressed. PAX2, PDCD8, PACD10, DFFA, CASP7 were also expressed differentially. And there is one mRNA, U58668, neither any gene nor protein information annotation, upregulated during camptothecin-induced apoptosis of U937 cells , also upregulated when the NAIF1 induced HeLaT cell apoptosis(all data is public in http://gpcrome.cbi.pku.edu.cn:2005/chip).

Key words apoptosis, bioinformatics, NAIF1, microarray

---

*This work was supported by a grant from Hi-Tech Research and Development Program of China (2002AA231051).

**Corresponding author.
Tel: 86-10-62751862, Fax: 86-10-62751861, E-mail: lsg@pku.edu.cn

Received: April 16, 2008 Accepted: June 30, 2008