生物化学与生物物理进展 *Progress in Biochemistry and Biophysics*, 2006, 33(3): 254~261
www.pibb.ac.cn

用生物信息学方法寻找肝癌特异性表达基因转录调控模式 *

王士雷　黄　波　周　云　孙之荣 **
(清华大学生物信息与系统生物学研究所，教育部生物信息学重点实验室，生物膜和膜生物技术国家重点实验室，北京 100084)

摘要　为了对肝癌 (hepatocellular carcinoma，HCC) 的分子发病机理进行研究，首先对肝癌基因表达谱数据用 *t*- 检验算法进行了分析，找到了肝癌中特异性表达基因(characteristic genes). 然后把这些基因结合已知的肝 HNF 家族转录因子染色质免疫共沉淀结合 DNA 启动子芯片 (ChIP-chip) 实验数据用 SAEM 算法进行分析，得到了肝癌特异性表达基因的转录调控关系，并寻找到了多个 HNF 家族转录因子调控单基因的转录调控模式. 结果表明 HNF 家族转录因子对大量具有重要功能的肝癌特异性表达基因进行了转录调控，并且多个 HNF 家族转录因子调控单基因可以形成前馈环和多输入调控等模式.

关键词　转录因子，转录调控，基因表达谱，ChIP-chip 实验，模式
学科分类号　Q811.4

细胞必须通过选择基因表达的模式来适应环境的不断改变，而基因表达模式改变的一个主要因素就是通过转录因子(transcription factor)对基因的转录调控[1]. 最近，新的高通量数据和计算方法或算法为研究基因转录表达调控的网络提供了可能.

利用基因表达谱对肿瘤细胞的分析，可以找到它的表达特异性基因，但这些特异性表达基因的转录调控关系却不能从基因表达谱中分析到. 对基因转录调控的研究，最近几年来发展了许多算法[2~5]，染色质免疫共沉淀结合 DNA 启动子芯片实验 (chromatin immunoprecipitation with promoter DNA array，ChIP-chip)是第一个用来研究酵母并随后用于哺乳动物细胞和果蝇研究的[3]. 近几年来，国际上利用 ChIP-chip 实验数据构建酵母等模式生物转录因子和基因之间转录调控网络(transcriptional regulatory network)的研究不断有文献报道[6~8]. 本文利用 ChIP-chip 数据结合基因表达谱数据，研究转录因子和癌症中表达特异性基因之间的转录调控关系，目标是更进一步弄清楚癌症发生、发展的分子机理[1].

肝癌(hepatocellular carcinoma，HCC)是一种常见的恶性肿瘤，在全世界范围内属于 5 种最主要致死癌症之一[9]，在中国因肝癌死亡的病人居所有癌症死亡人数的第 2 位. 肝癌的发生、发展是一个复杂的过程，涉及到多阶段、多因素、多基因的参与. 现在已经知道，绝大部分的肝癌都和肝炎 B 病毒 (hepatitis B vivus，HBV) 或者肝炎 C 病毒 (hepatitis C vivus，HCV)的慢性感染有关[10,11]，但这些感染导致肝癌发生的具体分子机理到现在还不很清楚. 因此，肝癌的治疗手段也是很有限的，外科手术切除(surgical resection)是现在唯一比较有效的方法[12]，即使这样，在做了肝癌组织外科切除的病人中，仍有 50%以上的病人会复发[13]. 因此，对肝癌发生、发展分子机理的研究是很重要的. 本文利用肝癌的基因表达谱数据和肝癌中主要转录因子的 ChIP-chip 数据，试图系统地分析肝癌的特异性表达基因和这些基因受转录表达调控的关系，从而寻找肝癌特异性表达基因的转录调控模式，进一步弄清楚肝癌发生、发展的分子机理. 这对肝癌临床诊断和治疗药物的设计都具有非常重要的应用 价值.

*国家自然科学基金(90303017, 90408019), 国家高技术研究发展计划 (863) (2002AA234041) 和国家重点基础研究发展计划 (973) (2003CB715900)资助项目.
** 通讯联系人.
Tel/Fax: 010-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn
收稿日期：2005-09-28，接受日期：2005-11-29

1 材料和方法

1.1 材料

1.1.1 肝癌基因表达谱数据. 数据来自 Chen 等[9]在 2002 年的肝癌基因表达谱，该表达谱共做了 23 075 个 cDNA 克隆，约 17 400 个人类基因，样本共 157 个，其中 82 个为原发肝癌(primary hepatocellular carcinoma)样本，正常组织样本为 75 个.

1.1.2 肝癌转录因子 ChIP-chip 数据. 来自 Young 等[14]2004 年做的肝脏中 HNF 家族转录因子(HNF family transcription factors)对肝细胞基因转录调控的 ChIP-chip 数据. 该数据主要包含 HNF1α，HNF4α，HNF6 转录因子和 RNA 聚合酶Ⅱ(RNA polymerase Ⅱ) 对 13 047 个肝细胞基因启动子区域的结合关系.

1.2 方法

1.2.1 肝癌基因表达谱分析. 对肝癌基因表达谱数据的分析，目的是找到在表达谱中那些基因表达量和正常肝组织细胞中基因表达量差异比较大 (过表达或者低表达) 的基因，也就是表达特异性基因. 寻找基因表达谱中表达特异性基因的算法有很多种，如 *t*- 检验 (*t*-test)、层次聚类 (hierarchical clustering)、K- 均值聚类(K-means clustering)、自组织图谱 SOM (self organizing map)、PCA (principle component analysis) 等统计算法. 本文对肝癌基因表达谱用 *t*- 检验 ($P < 0.01$) 算法.

1.2.2 ChIP-chip 实验数据分析. 从肝癌基因表达谱中找到特异表达性基因，这些基因理论上应该和肝癌的发生、发展有着直接或间接的关系. 本文对 ChIP-chip 实验数据进行第一步处理，就是把 ChIP-chip 实验数据中有肝癌特异性表达基因的数据留下来，其他的数据去除掉. 得到这些数据后，第二步就是利用 SAEM (single array error model)算法[15] ($P < 0.001$) 来对做过预处理的 ChIP-chip 实验数据进行分析，找到 HNF 家族转录因子和肝癌特异性表达基因的转录调控关系. 第三步处理是结合第二步的结果，寻找 HNF 家族转录因子和被它们调控基因之间的转录调控模式.

算法描述

SAEM 算法：SAEM 为统计学算法，在数据处理中，任何一个实验对象的重要程度用 X 表示：

$$X=\frac{a_2-a_1}{(\sigma_1^2+\sigma_2^2+f^2\times(a_1^2+a_2^2))^{1/2}} \tag{1}$$

其中 a_1、a_2 表示芯片中实验对象对照和测试实验获得的亮度值，σ_1、σ_2 表示由于数据采集背景导致的不确定性(uncertainty)，f 为相对乘法误差(fractional multiplicative error)，造成误差的原因主要是试验中杂交不均匀，染料结合效率有波动以及数据扫描器自身误差等. 整个芯片所有点的 X 值近似正态分布，参数 σ 和 f 基于对照杂交(control hybridizations)来选择，这样 X 就具有统一的变量. 那么，所算结果 X 的重要性为：

$$P=1-\mathrm{Erf}(X) \tag{2}$$

$$\mathrm{Erf}(x)=\frac{2}{\sqrt{\pi}}\int_0^X e^{-t^2}\mathrm{d}t \tag{3}$$

对任一对象，都做 3 次独立实验，并且对每次实验数据都独立地用(1)和(2)公式计算. 这 3 次实验的转录因子，基因平均结合率和相关 P 值用加权平均分析算法得到. 对每个实验对象，转录因子和基因结合的能力都是用 $\lg(a_2/a_1)$表示，那么这个比值的不确定性因素可以用以下公式表示：

$$\sigma_{\lg(a_1/a_2)}=\lg(a_2/a_1)/X \tag{4}$$

多次重复试验就用最小方差加权平均(minimum-variance weighted average) 的方法计算每个实验对象的 $\lg(a_2/a_1)$ 均值：

$$w_i=1/\sigma_i^2 \tag{5}$$

$$\overline{X}=\sum_{i=1,n} w_i x_i / \sum x_i \tag{6}$$

σ_i 表示第 i 次试验 $\lg(a_2/a_1)$的误差，见公式(4)，X_i 表示第 i 次试验的 X 值，n 表示实验重复次数. $\overline{X}$ 的误差表示为：

$$\sigma_p^2=1/\sum w_i \tag{7}$$

对于多次重复实验的数据，最终 X 值为：

$$X=\overline{X}/\sigma_p \tag{8}$$

所得结果的重要程度用公式(2) 求得.

用 SAEM 对做过预处理的 ChIP-chip 实验数据进行分析，主要就是要计算得到公式(2)中 P 的值，当 $P < 0.001$ 时，我们就认为 HNF 家族转录因子和肝癌特异性表达基因存在转录调控关系.

SAEM 分析 ChIP-chip 实验数据的假阳性(false positives)率在严格的阈值控制下($P < 0.001$)一般不超过 16%[14].

寻找转录因子调控模式的算法[6,8]：

第一步，用转录因子和基因是否结合的数据构建一个二元全矩阵 $\boldsymbol{D}_{ij}$，矩阵中用 1 表示转录因子 j 和基因 i 有结合，用 0 表示它们没有结合.

第二步，构建矩阵 ***D*** 的子矩阵 ***R***，矩阵 ***R*** 中只包含矩阵 ***D*** 中那些基因对应转录因子赋值的行，和转录因子对应基因赋值的列.

第三步，构建完以上两个矩阵后，用 matlab (版本 6.5.0.180913a)对矩阵进行分析，得到转录调控模式[1,6,8]：

a. 自调控模式(autoregulatory motif)：***R*** 矩阵对角线非 0 值对应的调控关系.

b. 前馈环(feedforward loop)：对矩阵 ***R*** 中每一个主转录因子(矩阵 ***R*** 的每一列)分析，找列中的非 0 值，从中找到与主转录因子结合的转录因子，构成转录因子对. 对矩阵 ***D*** 所有行分析，找与矩阵 ***R*** 中得到的转录因子对都结合的基因. 得到的调控关系即为前馈环.

c. 多元素环(muti-componet loop)：从矩阵 ***R*** 中的每个转录因子(矩阵 ***R*** 的每一列)中，找到与它结合的转录因子，构成转录因子对. 然后对它们进行分析，如果它们是相互调控转录，并不受其他转录因子调控，那么就构成含两个转录因子的两元素环. 依此类推，可以找到三个转录因子和多转录因子的多元素环.

d. 单输入模式(single input module)：从矩阵中找到只结合当转录因子的基因. 即对矩阵 ***D*** 的每行取总和得 1 的行的基因. 然后对每列寻找非 0 值区域，它们的交集对应的转录模式就为单输入模式.

e. 多输入模式(multi-input module)：寻找矩阵 ***D*** 中每行数值和大于 1 的行，即结合多个转录因子的行. 然后仍对矩阵 ***D*** 的每一行分析，找到与上面行结合同样转录因子的行. 最后得到的这些行的基因就同时受同样几个转录因子的调控，即多输入模式. 然后去除这些行再做同样的分析，就可以找到更多的多输入模式.

f. 调控链(regulator cascade)：对每一个转录因子(矩阵 ***R*** 的每一列)，用递归算法找到最长的链. 即对每个转录因子，寻找它的启动子是否和第二个转录因子其他结合，然后寻找与第二个转录因子的启动子结合的转录因子，直到寻找不到为止. 这样得到的转录结果就是调控链模式.

2 结　果

2.1 肝癌基因表达谱的表达特异性基因

通过 *t*- 检验算法，在 $P < 0.01$ 的情况下，我们从肝癌基因表达谱数据中共获得 1 648 个表达特异性基因，占整个表达谱基因总数的 9.5%，假阳性率为 10.5% ($P < 0.01$) 左右. 对这些提取出来的基因进一步分析，去调 ESTs 和 Homo 的数值，因为它们的基因具体名称不确定. 另外对同一基因做的多次实验数据，就按一次计算，最终得到 1 100 个肝癌特异性表达基因.

2.2 肝癌特异性表达基因转录调控分析

把得到的肝癌特异性表达的基因和已知的肝 ChIP-chip 实验数据做比较，在 ChIP-chip 实验数据得到肝癌特异性表达基因的数据. 在做完数据的初步处理后，用 SAEM 算法分析 ChIP-chip 实验数据，得到 HNF 家族转录因子和肝癌特异性表达基因的转录调控关系.

2.2.1 单个转录因子对肝癌特异性表达基因的调控关系. 用 SAEM 算法分析 ChIP-chip 数据中的肝癌特异性表达基因和转录因子的关系，首先找到 HNF 家族单个转录因子和肝癌特异性表达基因的转录调控关系，如表 1 所示.

以上为 3 个转录因子总共调控的基因，去掉相互重叠的基因，共有 256 个基因受这 3 个转录因子的调控，占肝癌特异性表达基因的 23.3%. 对以上结果，用基因功能分类软件 ProtoGo(www.protogo.cs.huji.ac.il)做进一步分析，得到这些转录因子调控的肝癌特异性基因表达的蛋白质产物的功能类. 那些没有在 GO 数据库中自动分类的基因，就利用 NCBI 的分类信息进行功能分类，现在还未知功能的肝癌特异性表达基因不做分类. 本文以 HNF1α 调控的肝癌特异性表达基因为例，如表 2 所示.

从 HNF1α 调控的基因功能分类来看，它调控的癌症特异性表达基因都是那些具有重要功能的基因，进一步对 HNF1α 和整个肝细胞基因调控关系分析，发现它对整个肝细胞调控的基因中，功能类为分子伴侣的基因居然全都是肝癌中特异性的表达基因，这说明 HNF1α 在肝癌的发生中的确有重要的作用.

如图 1 所示，HNF1α 调控的肝癌特异性表达基因中，功能类为酶的最多，其次为分子伴侣，然后是具有转运功能的蛋白质和脂质及小分子，最后是配体结合蛋白和信号转导蛋白. 从 HNF1α 调控的基因功能类，我们也能看出 HNF1α 转录因子的重要性，在它调控基因中，占第二位的为分子伴侣，这类基因表达的蛋白质具有辅助其他蛋白质折叠的功能，如果这些分子伴侣表达失调，自然会影响它们辅助折叠蛋白质的功能，从而导致整个机体的紊乱.

Table 1 List of HNF family transcriptional factors regulating HCC's characteristic genes

TF	HCC's characteristic genes regulated by TF
HNF1α	HSPC216 SERPINA1 GOT1 AADAC TSC22 DUSP6 C21ORF50 SERPINE1 AQP3 LY6E IGFBP1 SLC17A2 C4BPA TNFRSF6 FHL1 KIAA0022 SERPING1 SLC23A1 C21ORF4 ASGR2 ASGR1 C1S HSD11B1 PCK1 CPB2 F11 GRHPR PLGL CYB5 MTHFD1 MT1H MT1L GCKR UGT2B15 APCS MTP LOC54518 KNG SERPINA6 HAL FHR-3 VTN G6PT1 ALB TNFRSF11B NPC1L1 AGTR1 RAMP1 HSPC210 C22ORF3 M96 STRAIT11499 HSPC224 FLJ20080 APOH PAX8 CDC25B CKS1 C20ORF1 CDC20 CDC2 AKR1C3 TXNRD1
HNF6	STAT4 OAT HSPC216 TNFSF10 EIF2B2 CRADD TSC22 RNF9 EIF4A1 C21ORF50 NFKBIA IFITM1 LIF SREBF1 PON1 GRO1 TNFRSF6 SSI-3 ADAMTS1 SERPING1 APOH SLC23A1 C21ORF4 C1S PCK1 F11 PLGL UGT2B15 APCS IF SAS10 TNFRSF11B AGTR1 HSPC210 EIF4B Homo RPLP1 PPP2R5A HSPC224 TIF1 SIAHBP1 C22ORF3 TIF1B CDC25B TAF2E E2F1 CKS1 EIF4A2 TCFL1 C20ORF1 CDC20 CDC2 CDKN3 IFIT1 EIF3S6 FABP5 SLC29A1 AF038169 H2BFS H2BFL H2BFQ DHFR
HNF4α	STAT4 SERPINA5 HMOX2 HSPC216 SLC31A2 TF CSF2RA NCF1 EPB72 TNFSF10 NAPA CSF2 TFAP2A CST3 PIR121 LOC55862 FLJ20010 SLC4A4 KIAA0914 SERPINA1 GOT1 CRADD C4A IL1RAP TSC22 RNF9 DUSP6 JUND JUN C21ORF50 FOSL2 CLECSF2 MAFF SERPINE1 NTN4 IL15RA TFPI2 MAPK7 SAA1 IFITM1 FLJ11286 FGF12B AQP3 CST6 FNTA EHM2 HMGCS2 SREBF1 PON1 IGF1 CRSP3 GCHFR PCBD CBS GCSH IGFBP1 SLC17A2 HADH2 C4BPA TNFRSF6 TF ENC1 ICSBP1 FHL1 JUNB jSSI-3 ADAMTS1 EPHA2 PHLDA1 MBLL SGK PLSCR1 PBEF PROZ TFG HSPA5 SERPING1 SCYA14 SLC23A1 C21ORF4 ASGR2 ASGR1 MST1 APOA1 C1S CYP2C8 HAAO PIPOX HSD11B1 ANG PCK1 ACAA2 VLCS-H2 PCK2 CPB2 EHHADH GRHPR ACADSB PLGL PEMT CYB5 MUT GYS2 MTHFD1 VLCS-H1 MT1H MT1L GCKR ACY1 UGT2B15 APCS MTP LOC54518 ORM1 KNG SULT2A1 CYP2J2 EHD3 FTHFD SOD1 SLC10A1 PDK4 CSF1R FETUB SERPINA6 COX7A2 HAL VTN G6PT1 ATF5 FNTA SAS10 FLJ20037 TNFRSF11B KIAA1017 NPC1L1 AGTR1 GSTM4 RAMP1 HSPC210 ABCC6 C22ORF3 DKFZp762L0311 M96 CETN2 SCD RAB11A IK3C3 STRAIT11499 ORC3L RPLP1 AF093680 PYCS LOC51107 HMG17 CSNK2B LOC51596 MCP ABCB10 TCOF1 DXS1357E HSPC224 RPS16 RPS19 MAFG PRKAB2 PRPSAP1 FLJ20080 AD022 APOH CS KIAA0205 SIAHBP1 PAX8 UBE2M CLPTM1 TAF2E VARS2 CDC25B FLJ10604 GPC3 CKS1 PRCC USP21 C20ORF1 CDC20 CDC2 CDKN3 USP1 TYMS RDBP GNPAT ACLY LOC51606 POLR2K FLJ10511 PTK2 XPR1 PSMD10 NFYA NME1 PPGB GSTA4 BCAT2 DCK AKR1C3 TXNRD1 PIR SNRPA1 TACC3 SLC29A1 NCOR1 DUSP12 NDRG1 TAGLN2 NRCAM YKT6 H2BFS H2BFL H2BFQ

Table 2 List of genes' function about HNF1α regulating HCC's characteristic genes

Functional category	Gene	Functional category	Gene
Chaperone	C4BPA APCS F11 VTN C1S	Enzyme-regulator	SERPINA1 SERPING1
Transferase	GOT1 UGT2 B15	Hydrolase	AADAC CPB2
Lyase	PCK1 HAL	Oxidoreductase	CYB5
Transporter-channel/pore	AQP3 SLC17 A2	Transporter-lipids and small molecules	APOH G6PT1 ALB
Ligand bingding	IGFBP1	Signal transduction-receptor	ASGR2

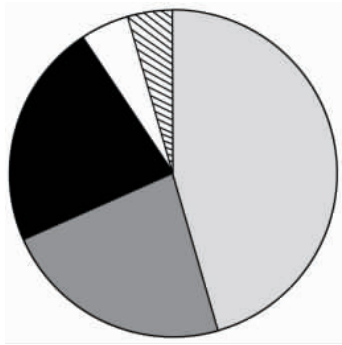

Fig. 1 Functional map about HNF1α regulating HCC's characteristic genes

□: Enzyme; ▩: Chaperone; ■: Transporter function; □: Ligand binding; ▨: Signal transduction-receptor.

2.2.2 转录因子对肝癌特异性表达基因组合调控关系.

所谓组合调控，就是在基因转录的过程中，多个转录因子相互协作，对某一基因进行转录调控. 利用 ChIP-chip 实验数据，用 SAEM 算法分析，同样可以找到这些转录调控关系，如表 3 所示.

而对于这些转录因子调控基因的关系，由 Lee 等[6]最先提出，并由 Yeger-Lotem 等[8]做了进一步的研究，主要可以分为以下几类，如图 2 所示.

根据图 2 关系，本文也对 HNF 家族转录因子和肝癌特异性表达基因的调控模式 (motif) 进行了研究，并从肝癌特异性表达基因的 ChIP-chip 实验数据中找到了前馈环 (图 2b) 和多输入调控模式 (图 2f).

Table 3 Relations of combined regulation about HNF family transcriptional factors and HCC's characteristic genes

TFs	HCC's characteristic genes regulated by TFs
HNF4α、HNF6	C1S SAS10 CDK2 TNFRSF6 PON1 CASP2 SERPING1 PCK1 VTN PLGL APCS APOH
HNF4α、HNF1α	C1S PCK1 TNFRSF6 VTN PLGL SERPING1 CPB2 C4BPA FLJ20080 LOC54518 DUSP6 AQP3 MTP AGT ASGR1 PAX8 IGFBP1 RAMP SERPINA1 UGT2B15 HAL MT1H GRHPR HSD11B1 SLC17A2 APCS SERPINE1 MTHFD1 APOH GOT1 ASGR2 STRAIT11499 MT1L M96
HNF1α、HNF6	F11 C1S PLGL SERPING1 TNFRSF6 PCK1 VTN UGT2B15 APCS APOH SLC23A1
HNF4α、HNF1α、HNF6	C1S PLGL TNFRSF6 VTN PCK1 SERPING1 UGT2B15 APCS APOH

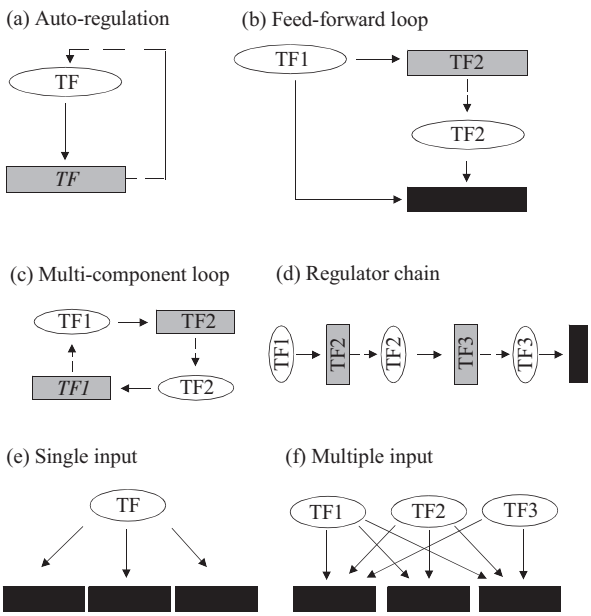

Fig. 2 Relations of Combined regulation about transcriptional factors and genes

→: Translation; →: TF binding to DNA; ■: Target gene; TF: TF-encoding gene; (TF): TF protein.

从表 4 中可以看出，转录因子 HNF6 和转录因子 HNF4α 的基因启动子区结合，调控 HNF4α 的表达，然后 HNF4α 和 HNF6 在共同调控肝癌的某些特异表达基因(表 4). 而 HNF4α 和 HNF1α 转录因子可以相互结合在对方基因的启动子区域调控对方的表达，然后再共同调控肝癌的某些特异性表达基因(表 4). 而转录因子 HNF4α 和 HNF1α 可以相互结合对方基因的启动子区域，也和早期的文献[16,17]相符. 前馈环为基因调控提供了多方面的优势：前馈环可能为基因持续性表达提供一个敏感性开关，来控制基因表达量[18]，最终靶点基因(ultimate target gene)的表达可能依赖于主调控子(master regulator)和第二调控子(second regulator)量的集聚，前馈环可能是最终靶点基因表达的多级敏感性调控[19] (multistep ultrasensitivity). 在前馈环中，主调控子的一个微小的量或者活性的改变，都有可能使最终靶点基因的表达量有很大的改变，因为主调控子和第二调控子的结合活性是在主调控子调控控制下的.

受到多个转录因子调控的基因，往往是比较重

要的基因，从表 5 中我们可以看出，C1S、APCS、F11 和 VTN 是分子伴侣，PCK1 为裂解酶，SERPING1 是调控酶，UGT2B15 是转移酶，APOH 具有蛋白质转运功能，可见这些基因都是很重要的. 多输入调控模式，为基因在各种生长环境中都能得到转录调控表达提供了可能. 在不同的环境中，基因可以和不同的调控子结合，从而能够使它们在不同的环境中都能得到转录调控[6].

Table 4 Feedforward loop about HNF family transcriptional factors regulating HCC's characteristic genes

TF1	HNF6	HNF4α/ HNF1α
TF2	HNF4α	HNF1α/ HNF4α
Gene	C1S SAS10 CDK2 PON1 CASP2 PCK1 PLGL TNFRSF6 SERPING1 VTN	C1S PCK1 TNFRSF6 VTN PLGL SERPING1 CPB2 C4BPA FLJ20080 LOC54518 DUSP6 AQP3 MTP AGT ASGR1 PAX8 IGFBP1 RAMP SERPINA1 UGT2B15 HAL MT1H GRHPR HSD11B1 SLC17A2 APCS SERPINE1 MTHFD1 APOH GOT1 ASGR2 STRAIT11499 MT1L

Table 5 Multi-input motifs about HNF family transcriptional factors regulating HCC's characteristic genes

TF1	HNF6	HNF6
TF2	HNF4α	HNF1α
TF3	HNF1α	
Gene	C1S PLGL TNFRSF6 APOH VTN PCK1 SERPING1 UGT2B15 APCS	APCS C1S PLGL SERPING1 TNFRSF6 PCK1 VTN UGT2B15 APOH F11

3 讨 论

肝癌(HCC)分子发病机理目前仍不清楚，因此对于肝癌的治疗，现在仍没有取得突破. 对于人类癌症全基因表达谱的系统分析可以为癌症发病机理的研究提供新的视角[20]，利用基因表达谱数据，可以把基因进行功能聚类，从而预测未知基因的功能，可以找到表达谱中那些癌症样本(tumor sample)中基因表达量和非癌症样本(nontumor sample)中基因表达量差异比较大的基因(即特异性表达基因)，也可以为把不同样本根据基因表达关系对疾病进行分型等.

但是，在只有基因表达谱的情况下，无法找到疾病特异性表达基因的转录调控关系. 而只有知道这些基因的转录调控关系，才能进一步弄明白疾病发生、发展的分子机理. 本文就是基于肝癌基因表达谱数据和肝 HNF 转录因子家族的 ChIP-chip 实验数据对肝癌的特异性表达基因进行转录调控的研究，并对多转录因子调控单基因的转录调控关系进行了研究. HNF 家族转录因子对很多肝癌特异性表达基因具有转录调控作用，在从肝癌中选择得到的特异表达基因中，受 HNF 转录因子调控的就占 23.3% (表 1). 并且被 HNF 家族转录因子调控的基因多数具有重要的功能 (表 2)，进一步对 HNF1α 调控的肝癌特异性表达基因的功能类进行划分，已知功能的基因中酶功能基因最多，占 43%，其次是分子伴侣和具有转运功能的蛋白质、脂质及小分子，各占 24%，最后是配体结合蛋白和信号转导受体，各占 4.5%左右. 以上各类功能蛋白，对整个机体代谢过程都有至关重要的作用，无论哪种蛋白质在体内失调，都会引起疾病. 并且 HNF1α 对整个肝细胞调控的基因中，功能类为分子伴侣的基因居然全都是肝癌中特异性表达的基因(C4BPA、C1S、F11、APCS、VTN)，这说明 HNF1α 在肝癌的发生中的确有重要的作用，并且可以推测在肝癌的发生、发展过程中，肝细胞中功能为分子伴侣的基因可能起到至关重要的作用，也许由于它的表达失调，导致其他一些蛋白质不能折叠或者过快折叠最终导致机体病变. 因此，对于某些肝癌的特异表达基因，我们至少从理论上可以得出它们的表达失调是因为肝癌 HNF 家族转录因子的缘故.

本文也研究了多个 HNF 家族转录因子协同调控单基因的关系. 首先，我们利用 SAEM 算法，得到了多个转录因子调控单基因的结果 (表 3). 根据 Lee 等[6]和 Yeger-Lotem 等[8]的研究，多转录因子调控单基因的过程存在不同的模式(图 2). 利用他们给出的算法，本文对 HNF 家族转录因子协同调控单

基因的模式进行了研究，得到了前馈环模式(表 4)和多输入调控模式(表 5). 从表 4 中可以看出，转录因子 HNF6 和转录因子 HNF4α 的基因启动子区结合，调控 HNF4α 的表达，然后 HNF4α 和 HNF6 在共同调控肝癌的某些特异表达基因，而 HNF4α 和 HNF1α 转录因子可以相互结合在对方基因的启动子区域调控对方的表达，然后再共同调控肝癌的某些特异性表达基因. 这些转录调控模式的存在，为肝癌的分子发病机理的研究提供了新的突破点：也许可能正因为这些前馈环模式或者多输入调控模式的存在，使某些基因的表达调控变得更为精细，但是如果受到环境的影响，使这些模式调控发生紊乱，就有可能导致它们调控的基因表达失调，从而导致机体的整个代谢系统紊乱，最终导致疾病的发生.

根据多输入调控模式，我们也可以大致了解某些基因的功能或者是它们的重要性程度，对表 5 中调控基因的功能进行分析，可以看出 C1S、APCS、F11 和 VTN 是分子伴侣，PCK1 为裂解酶，SERPING1 是调控酶 UGT2B15 是转移酶，APOH 具有蛋白质转运功能，可见这些基因都是很重要的. 因此，我们也可以得出这样的结论：受到多个转录因子调控的基因往往是比较重要的基因.

致谢 本文得到郑家顺博士和李立博士等多方面的建议和指导，在此一并表示感谢.

参 考 文 献

1 Blais A, Dynlacht B D. Constructing transcriptional regulatory networks. Genes & Dev, 2005, **19** (13): 1499～1511

2 Weinmann A S, Farnham P J. Isolating human transcription factor tar- gets by coupling chromatin immunoprecipi- tation and CpG island microarray analysis. Genes & Dev, 2002, **16** (2): 235～244

3 Ren B, Hannett N, Kanin E, *et al*. Genome wide location and function of DNA binding proteins. Science, 2000, **290** (5500): 2306～2309

4 Cawley S, Williams A J, Gingeras T R, *et al*. Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. Cell, 2004, **116**(4):499～509

5 Orian A,Yost C, Eisenman R N, *et al*. Genomic binding by the *Drosophila* Myc, Max,Mad/Mnt transc- ription factor network. Genes & Dev, 2003, **17** (9): 1101～1114

6 Lee T I, Simon I, Young R A, *et al*. Transcriptional regulatory networks in Saccharomyces cerevisiae. Science, 2002, **298** (5594): 799～804

7 Harbison C T, Yoo J, Young R A, *et al*. Transcriptional regulatory code of a eukaryotic genome. Nature, 2004, **431** (7004): 99～104

8 Yeger-Lotem E, Sattath S, Margalit H, *et al*. Network motifs in integrated cellular networks of transcription regulation and protein-protein interaction. Proc Natl Acad Sci USA, 2004, **101** (16): 5934～5939

9 Chen X, Cheung S T, Brown P O, *et al*. Gene expression patterns in human liver cancers. Mol Biol Cell, 2002, **13** (6): 1929～1939

10 Beasley R P. Hepatitis B virus.The major etiology of hepatocellular carcinoma. Cancer, 1988, **61** (10): 1942～1956

11 Hasan F, Jeffers L J, Kuo G, *et al*. Hepatitis Cassociated hepatocellular carcinoma. Hepatology, 1990, **12** (3 Pt 1): 589～591

12 Lin T Y, Lee C S, Chen K M, *et al*. Role of surgery in the treatment of primary carcinoma of the liver: a 31-year experience. Br J Surg, 1987, **74** (9): 839～842

13 Okuda K, Obata H, Ohnishi K, *et al*. Prognosis of primary hepatocellular carcinoma. Hepatology, 1984, **4** (1 Suppl): 3S～6S

14 Odom Duncan T, Zizlsperger N, Young Richard A, *et al*. Control of pancreas and liver gene expression by HNF transcription factors. Science, 2004, **303** (5662): 1378～1381

15 Hughes T R, Marton M J, Friend S H, *et al*. Functional discovery via a compendium of expression profiles. Cell, 2000, **102** (1): 109～126

16 Sylvia F B, Parrizas M, Ferrer J, *et al*. From the cover: A transcription factor regulatory circuit in differentiated pancreatic cells. Proc Natl Acad Sci USA, 2001,**98** (25): 14481～14486

17 Milo R, Shen-Drr S, Alon U, *et al*. Network motifs: simple building blocks of complex networks. Science, 2002, **298** (5594): 824～827

18 Guelzim D, Bottani S, Képès F, *et al*. Topological and causal structure of the yeast transcriptional regulatory network. Nature genet, 2002, **31**(1):60～63

19 Goldbeter A, Koshland Jr D E , *et al*. Ultrasensitivity in biochemical systems controlled by covalent modification. Interplay between zero-order and multistep effects. J Biol Chem, 1984, **259** (23): 14441～14447

20 Golub T R, Slonim D K, Lander E S, *et al*. Molecular classifica-ion of cancer: class discovery and class prediction by gene expression monitoring. Science, 1999,**286** (5439): 531～537

Finding Transcriptional Regulatory Motifs of Hepatocellular Carcinoma's Characteristic Genes Using Bioinformatics Methods*

WANG Shi-Lei, HUANG Bo, ZHOU Yun, SUN Zhi-Rong**

(*Institute of Bioinformatics and System Biology, Ministry of Education Key Laboratory of Bioinformatics, State Key Laboratory of Biomembrane and Membrane Biotechnology, Department of Biological Science and Biotechnology, Tsinghua University, Beijing* 100084, China)

Abstract In order to research the molecular pathogenesis of hepatocellular carcinoma (HCC), firstly, gene expression profile data of HCC, which dealt with *t*-test method, was used and HCC's characteristic genes were found, secondly, SAEM method was used to analyze these data combining HCC's characteristic genes and human liver's HNF family transcription factors ChIP-chip data, finally, transcriptional regulatory relations of HCC's characteristic genes and HNF family transcription factors were got, furthermore the transcriptional regulatory motifs of HNF family transcription factors regulating HCC's characteristic genes were found. All the results indicate that HNF family transcription factors regulate many HCC's characteristic genes, a great number of those have very important functions, and multi-HNF family transcription factors regulating HCC's characteristic genes can form feedforward loop motif and multi-input regulatory motifs.

Key words transcription factor, transcriptional regulation, gene expression profile, motif, ChIP-chip experiment

---

*This work was supported by grants from The National Natural Science Foundation of China (90303017, 90408019), Hi-Tech Research and Development Program of China (2002AA234041) and National Basic Research Program of China (2003CB715900).

**Corresponding author . Tel/Fax: 86-10-62772237, E-mail: sunzhr@mail.tsinghua.edu.cn

Received: September 28, 2005 Accepted: November 29, 2005