doi:10.3969/j.issn.1672-5565.2014.01.11

基于转录调控模体的人不同组织基因差异性的统计分析

杨 敏,张 静*
(云南大学数学与统计学院统计系,昆明 650091)

摘 要:转录调控是基因表达调控的主要过程,而转录调控模体使用的差异性可能是导致基因组织特异性的因素之一。本文提出一种不同组织基因调控差异性的统计分析方法,首先结合泊松分布和主成分分析提取基因启动子中过表达模体作为潜在的转录因子结合位点。基于这些位点通过Wilcoxon秩和检验获得不同组织基因结构的差异性。再用超几何分布确定出现次数显著的模体作为组织基因的特有模体,并分析特有模体的碱基特征及在启动子序列中的位置分布。将特有模体与TRANSFAC数据库进行对照,得到潜在的调控组织特异性基因的转录因子结合位点。以人管家基因及30个组织特异性基因为分析对象,得到不同组织调控模体使用的差异性信息。

关键词:人组织基因;转录调控模体;泊松分布;主成分分析;秩和检验
中图分类号:Q753 **文献标志码**:A **文章编号**:1672-5565(2014)-01-065-07

Statistical analysis on differences of human specific tissue genes based on transcriptional

YANG Min, ZHANG Jing*
(*Department of Statistics, School of Mathematics and Statistics, Yunnan university, Kunming* 650091, *China*)

Abstract: Transcriptional regulation is the main regulatory process by gene expression. The differences in the use of transcriptional regulatory motifs may be one of the factors leading to gene tissue specificity. This paper presents a statistical method for analysis of the regulatory differences between different tissue genes. Firstly, over-represented motifs in gene promoters were extracted as potential transcription factor binding sites based on Poisson distribution and principal components analysis. Based on these sites, differences of gene structures in different tissues were obtained based on Wilcoxon rank sum test. Then, over-represented motifs were determined as specific motifs for certain tissue genes based on hypergeometric distribution, and the distribution and characteristics of these specific motifs in the promoter sequences were analyzed. By comparing these specific motifs with TRANSFAC database, the potential transcription factor binding sites in tissue-specific genes were selected. Finally, housekeeping genes and 30 tissue-specific genes were analyzed and the use differences of regulatory motifs in different tissues were found out.

Keywords: Tissue genes of human; Transcriptional regulatory motif; Poisson distribution; Principal component analysis; Wilcoxon rank sum test

基因表达是指基因在生物体内的转录、剪接、翻译以及转变成具有生物活性的蛋白质分子之前的所有加工过程,基因的转录调控是基因表达调控中最重要的过程,正确的转录调控使得整个生物体内的能量和资源得以正确的分配[1]。由于调控元件(或称模体)是一些具有保守性功能的片段,在基因的长期进化中具有不变的趋势,因此可以从调控元件的角度来分析基因表达的差异性。不同组织所使用

收稿日期:2013-10-18;**修回日期**:2013-12-23.
基金项目:国家自然科学基金资助项目(11261066)。
作者简介:杨敏,硕士研究生,研究方向:生物统计学;E-mail:yangminggh@163.com.
*通信作者:张静,教授,研究方向:生物信息学;E-mail:zhangjing@ynu.edu.cn.

的调控元件既有相同的,也有不同的。可能几个组织共同使用一些调控元件,也可能某个组织单独使用一些调控元件。Yu [2]等人基于转录因子之间的相互作用,识别了人类 30 个组织特异性基因的调控模块,开发并制作了 TiGER 数据库[3]。用这个数据库可以方便地查询各个组织的调控模块,但是不同组织可能共同使用了一些调控元件,因此,还缺少各个组织所特有的调控元件信息,即对组织基因调控的差异性缺乏研究。本文利用传统识别调控元件的方法[4],并应用泊松分布和主成分分析提取出各组织的调控元件,再基于这些调控元件,用统计方法统计出现次数具有显著性特征的特有模体,基于这些特有模体,寻找控制各组织基因调控的元件。

1 方 法

1.1 提取过表达模体的方法

提取过表达模体的方法分为两步:第一步是采用泊松分布计算出各个模体出现的概率;第二步是把模体作为变量,每个组织中的每条基因启动子序列作为实验条件,通过主成分分析,确定一组"主要模体"作为过表达模体。

1.1.1 泊松分布

van Helden 提出泊松分布来比较两条序列的相似性[5]。他认为模体在某条基因序列上出现是服从泊松分布的,可用泊松分布来度量每个模体出现的概率。由于模体是一些具有保守性的功能片段,在基因的长期进化中具有不变的趋势,因此通过泊松分布度量的每个模体出现的概率大小可以作为模体功能大小的估计。

为不同模体的出现打分,需要定义被考虑的模体在基因组中出现的概率。把每个组织的全部基因作为背景序列来计算背景频率。频率 f_i^j 表示给定模体 i 在组织 j 中出现的频率,m_i^j 表示模体 i 在组织 j 中出现次数的期望,L 表示启动子序列的长度。则有

$$f_i^j=\frac{n_i^g}{N_i}, \tag{1}$$

$$m_i^j=f_i^jT=f_i^j(L-w+1)。 \tag{2}$$

其中 n_i^g 表示模体 i 在组织 j 中第 g 条基因启动子序列上出现的次数,N_i 表示模体 i 在组织 j 中出现的次数,w 是模体 i 的长度,T 表示模体 i 在组织 j 中可能出现位置的数目。

模体 i 在组织 j 的基因启动子序列上出现服从期望为 m_i^j 的泊松分布,因此,采用泊松分布来计算模体 i 在组织 j 中出现的概率 p_i^j。p_i^j 只依赖于模体出现的次数和期望概率,没有要求额外的信息,而且各个模体出现的概率值是相互独立的。

模体 i 在第 k 条启动子序列上的出现次数服从泊松分布,泊松分布的分布函数 $F(x,m_i^j)$ 表示期望值为 m_i^j 时,观察到出现次数$\leq x$ 的概率。定义第 k 条基因启动子序列上模体 i 至少出现 n_i^g 次的概率为

$$p(x\geq n_i^g)=\begin{cases}1-F(n_i^g-1,m_i^j),(\text{当 } n_i^g>0 \text{ 时});\\ 1\qquad\qquad,\quad(\text{其它})。\end{cases} \tag{3}$$

采用 MATLAB 软件编程即可得到每个模体在每条序列中出现的概率。

1.1.2 主成分分析 PCA

主成分分析的主要思想是以某些线性组合(主成分)来表示原始数据,再从这些线性组合中尽快提取原始数据的信息[6]。给定 n 维空间中的 m 个点,寻求一个 $n\times n$ 维的矩阵 W,使得 $Y=[y_1,y_2,\cdots,y_m]=W^TX$,同时满足新坐标系下各维之间数据的相关性最小[7]。

主成分分析的一般步骤为[8]:

假设数据为 $X=[x_1,x_2,\cdots,x_i,\cdots,x_m]$,维数为 n,在下列所有运算中均有 $i\in[1,n]$,$j\in[1,m]$。

(1)计算每维(行)数据的平均值$\overline{x_i}=\sum_{j=1}^{m}x_{ij}/m$,得到矩阵 $\bar{X}$。

(2)中心平移每个数据得到矩阵 X,即 $x_{ij}=x_{ij}-\overline{x_i}$。

(3)计算协方差矩阵 $S_{n\times m}$,当 $P_{\text{occ}}*P_{\text{d}}$ 时,$S[a,b]=\text{cov}(x_a,x_b)=\sum_{k=1}^{m}x_{ak}\cdot x_{bk}/(m-1)$。

(4)对协方差矩阵进行特征分析,并将特征值由大到小排列:$\lambda_1>\lambda_2>\cdots>\lambda_d$,对应的特征向量也作相应排列。

(5)取前 d 个特征值 $\Lambda_d=\text{diag}[\lambda_1,\lambda_2,\cdots\lambda_d]$ 和特征向量 $W_d=[w_1,w_2,\cdots,w_d]$,主成分分析可以由 X 在 W_d 上投影得到,即:$Y=W_d^TX$,原始数据重建为:$X=WY+\bar{X}$。

(6)进一步分析每个主成分对信息的贡献,确定 d。令 λ_i 表示第 $i(i=1,2,\cdots,d)$ 个特征值,定义第 i 个主成分的贡献率为 $\lambda_i/\sum_{i=1}^{d}\lambda_i$。则有前 r 个主成分的累计贡献率为 $\sum_{i=1}^{r}\lambda_i/\sum_{i=1}^{d}\lambda_i$。一般要求累计贡献率达到 70%以上。

n 个变量的 m 个观察值,形成一个 $n\times m$ 的数据矩阵,n 通常比较大。由于每个新变量是原有变量的线性组合,体现原有变量的综合效果,具有一定的实际意义。采用 PCA 的方法提取过表达模体,把模

体 i 作为变量,每个组织中的每条基因启动子序列作为实验条件,通过主成分分析,确定一组“主要模体”作为过表达模体。

1.2　不同组织基因调控模体使用的差异性分析

1.2.1　Wilcoxon 秩和检验

Wilcoxon 秩和检验[9]并不要求数据满足某种分布假设且这种方法也非常适合小样本数据集。根据基因表达数据的大小排序,然后得到数据的秩,再利用数据的秩而不是数据本身计算基因的秩和统计量。Wilcoxon 秩和检验是一种建立在二项分布理论基础上的总体分布位置差异非参数检验方法。

设 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_n$ 是分别来自具有 $F(x-\mu_1)$ 和 $F(x-\mu_2)$ 连续分布形式的独立总体的两个随机样本,假设

$$H_0:\mu_1=\mu_2(\mu=\mu_1-\mu_2=0),$$

$$H_1:\mu_1\neq\mu_2(\mu=\mu_1-\mu_2\neq0)。$$

将样本 $X_1,\cdots,X_m$ 和 $Y_1,\cdots,Y_m$ 混合在一起,并按从小到大的顺序排列起来。每一个观测值在混合排列中都有自己的秩。令 R_i 为 Y_i 在这 $N=m+n$ 个数中的秩(即 Y_i 是第 R_i 小的)。令 I_m 和 I_n 分别表示两样本的指标集,则

$$W_Y=\sum_{i=1}^{n}\{\#(X_j<Y_i,j\in I_m)+\#(Y_k\leq Y_i,k\in I_n)\}。\quad(4)$$

同样,对于 X 样本也可以得到 W_X:

$$W_X=\sum_{j=1}^{m}\{\#(Y_i<X_j,i\in I_n)+\#(X_k\leq X_j,k\in I_m)\}。\quad(5)$$

称 W_Y 或 W_X 为 Wilcoxon 秩和统计量。令

$$W_{XY}=\sum_{i=1}^{n}\{\#(X_j<Y_i,j\in I_m,i\in I_n)\},\quad(6)$$

$$W_{YX}=\sum_{j=1}^{m}\{\#(Y_i<X_j,j\in I_m,i\in I_n)\}。\quad(7)$$

W_{XY}表示在混合样本中所有 X 的观测值小于 Y 的观测值的个数。则有

$$W_Y=W_{XY}+\frac{n(n+1)}{2},\quad(8)$$

$$W_X=W_{YX}+\frac{m(m+1)}{2},\quad(9)$$

而 $W_X+W_Y=\frac{(m+n)(m+n+1)}{2}$,于是有 $W_{XY}+W_{YX}=mn$。因此 W_{XY}(或 W_{YX})也可作为上述检验问题的检验统计量,它们称为 Manm-Whitney 统计量。当 W_{XY}很小时可拒绝零假设。

当 m,n 均较大时(大于 10),在 H_0 假设下,可用正态分布近似。此时 W_{XY}渐近服从均值为$\frac{mn}{2}$,方差为$\frac{mn(m+n+1)}{12}$的正态分布。因此可通过标准正态分布作检验,其检验统计量为:

$$Z=\frac{W_{XY}-\frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}}\sim N(0,1)。\quad(10)$$

在许多情况下,数据中有相同的数字,称为结(tie)。结中数字的秩为它们按升幂排列后位置的平均值。对于打结的情况,此时大样本近似用的 Z 值应修正为:

$$Z=\frac{W_{XY}-\frac{mn}{2}}{\sqrt{\frac{mn(m+n+1)}{12}-\frac{mn(\sum_{i=1}^{t}\tau_i^3-\sum_{i=1}^{t}\tau_i)}{12(m+n)(m+n-1)}}}\sim N(0,1)。\quad(11)$$

这里 τ_i 为结统计量,而 t 为结的个数。

对于显著性水平 α,当 p 值<α 时,拒绝 H_0;否则不能拒绝。

1.2.2　基因表达的差异性分析

本文主要考虑不同组织所使用的调控元件的差异情况。具体做法是:把某个组织的过表达模体与其他与之有显著性差异的所有组织的过表达模体相比,提取出那些在这个组织中出现,而在其他与之有显著性差异的所有组织中不出现的模体作为测试集 S。并统计出各个模体出现的次数。用超几何分布(Hypergeometric distribution)计算模体在 S 中出现的概率:

$$p_t=1-\sum_{k=0}^{t-1}\frac{\binom{n}{k}\binom{N-n}{T-k}}{\binom{N}{T}}。\quad(12)$$

其中 N 表示与这个组织有显著性差异的组织个数,T 表示某个模体 m_i 在测试集 S 中出现的次数,把与这个组织有显著性差异的组织分为两类,一类是具有模体 m_i 的组织,另一类是不具有模体 m_i 的组织,前者组织个数记为 $N-n$,后者记为 n。p_t 值越小,说明模体 m_i 是某个组织特有的模体的可能性越大。设定阈值 b,当 $p_t<b$ 时,认为模体 m_i 在测试集 S 中出现的频率显著高于其他模体的出现频率,称模体 m_i 是这个组织的特有模体。然后分析这些特有模体的特征,并分析它们在启动子序列中出现的位置,以此为基础分析各组织使用的模体参与情况。

2　样本和启动子序列的选取

2.1　样本

本文样本包括人的 HK(管家)基因以及 30 组 TSP(组织特异性)基因。人 HK 基因序列数据从 HuGEIndex 数据库中获得。该数据库给出了 451 条

HK 基因的 id,剔除无内含子和线粒体中的基因,共得到 412 条 HK 基因启动子序列。30 组 TSP 基因,来自 Yu 等[2]文献中采用聚类方法所得。根据文献中分析得到的 30 个组织特异性基因的 id 号,从 UCSC 数据库提取人 30 个组织特异性基因的基因序列(见表 1)。

表 1 各组织的基因条数

Table 1 Number of genes in each tissues

组织	条数	组织	条数
膀胱	174	喉	258
血液	352	肝脏	290
骨	100	肺	24
骨髓	242	淋巴	321
大脑	224	乳腺	108
宫颈	201	肌肉	231
结肠	175	卵巢	127
眼睛	207	胰腺	160
心脏	211	次级神经系统	33
肾脏	314	胎盘	45
前列腺	133	小肠	103
皮肤	136	脾脏	132
软组织	128	胃	185
睾丸	611	胸腺	80
舌头	324	子宫	45

2.2 启动子序列的选取

研究表明,基因上游的转录调控位点一般位于翻译起始位点(ATG)上游 1 000 bp 区域内,目前大部分工作主要集中在基因上游区域。因此采取基因上游 1 000 bp 区域作为启动子序列。在下载的 7 261条基因序列中剔除启动子序列相同的基因序列,共得到 4 672 条启动子序列。本文主要考察 6 核苷酸。

3 潜在的转录调控模体

3.1 提取过表达模体

根据 $P=F(n_i^g-1,m_i^j)$ 计算各个组织中 4 096 个模体在每条基因启动子序列中出现的概率,其数据形式如下:

$$\begin{array}{c} \\ m_1 \\ m_2 \\ \vdots \\ m_{4\,096} \end{array} \begin{array}{c} \begin{array}{cccc} g_1 & g_2 & \cdots & g_t \end{array} \\ \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1t} \\ p_{21} & p_{22} & \cdots & p_{2t} \\ \vdots & \vdots & \cdots & \vdots \\ p_{4\,096,1} & p_{4\,096,2} & \cdots & p_{4\,096,t} \end{pmatrix} \end{array} 。$$

其中 $m_i(i=1,2,\cdots,4\,096)$ 表示模体,$g_j(j=1,2,\cdots,t)$ 表示组织中的基因启动子序列。为了消除概率很小甚至为零的数据的不利影响,对 4 096 个模体的出现概率进行加权平均,即对于模体 m_i,在组织中出现的概率为:

$$p'_i=\frac{n_{i1}}{N_i}p_{i1}+\frac{n_{i2}}{N_i}p_{i2}+\cdots+\frac{n_{it}}{N_i}p_{it}。$$

其中 $n_{ij}(j=1,2,\cdots,t)$ 表示模体 i 在第 j 条基因序列上出现的数目,N_i 表示模体 i 在这个组织中出现的数目,且有 $N_i=n_{i1}+n_{i2}+\cdots+n_{it}$。

本文提取出现率 $p_{occ}=1-p'_i<0.05$ 的模体进行主成分分析,取累计贡献率在 70%以上的模体,作为过表达模体,各个组织中提取的过表达模体数目见表 2。

表 2 各组织中应用 PCA 提取的过表达模体数目

Table 2 Number of over-represented motifs extracted by principal components analysis

组织	过表达模体数目	组织	过表达模体数目
管家基因	137	眼睛	144
膀胱	82	心脏	145
血液	117	肾脏	144
骨	132	喉	101
骨髓	92	肝脏	188
大脑	119	肺	224
宫颈	119	淋巴	116
结肠	149	乳腺	129
次级神经系统	190	脾脏	136
卵巢	126	软组织	147
胰腺	151	胃	126
肌肉	131	睾丸	103
胎盘	144	胸腺	145
前列腺	122	舌头	83
小肠	119	子宫	186
皮肤	135		

提取出的过表达模体的准确率,可以通过与 TRANSFAC 数据库[10]进行比对。模体识别准确率的计算为识别正确的模体的个数除以提取的过表达模体的个数。通过搜集整理得到 TRANSFAC 数据库中人的 213 个转录因子。与 TRANSFAC 数据库比对,准确率最低是 87.50%,最高可达到 95.98%(表 3),说明提取出的过表达模体具有生物学上的意义。

表 3　各组织的过表达模体与 TRANSFAC 数据库比对的情况

Table 3　Comparison of over-represented motifs with TRANSFAC database

组织	过表达模体数目	正确数目	准确率(%)	组织	过表达模体数目	正确数目	准确率(%)
管家基因	137	126	91.97	次级神经系统	190	180	94.74
膀胱	82	74	90.24	卵巢	126	115	91.27
血液	117	106	90.60	胰腺	151	139	92.05
骨	132	124	93.94	肌肉	131	119	90.84
骨髓	92	84	91.30	胎盘	144	136	94.44
大脑	119	110	92.44	前列腺	122	111	90.98
宫颈	119	111	93.28	小肠	119	106	89.08
结肠	149	136	91.28	皮肤	135	123	91.11
眼睛	144	132	91.67	脾脏	136	126	92.65
心脏	145	134	92.41	软组织	145	135	93.10
肾脏	144	126	87.50	胃	126	119	94.44
喉	101	93	92.08	睾丸	103	92	89.32
肝脏	188	171	90.96	胸腺	147	136	92.52
肺	224	215	95.98	舌头	83	77	92.77
淋巴	116	111	95.69	子宫	186	172	92.47
乳腺	129	117	90.70				

3.2　过表达模体的特征分析

针对过表达模体中碱基的出现情况,将模体进行适当分类[11]。如果 6 核苷酸中有 4 个或 4 个以上的碱基是 A 或 T,认为该 6 核苷酸富含 A、T,称为 AT_rich 模体;同样如果 6 核苷酸中有 4 个或 4 个以上的碱基是 C 或 G,认为该 6 核苷酸富含 C、G,称为 CG_rich 模体;其它既非 AT_rich 模体又非 CG_rich 模体,称为 CG/AT_lack 模体。

HK 基因启动子中碱基 A+T 含量和 C+G 含量分别为 52% 和 48%,即在 HK 基因启动子序列中,A+T含量较高。提取出的过表达模体中富含 A、T 的模体比率高于富含 C、G 模体的比率。各组织的过表达模体特征见表 4。

表 4　各组织中模体的碱基使用情况

Table 4　Base usage of motifs in each tissues

组织	AT_rich(%)	CG_rich(%)	CG/AT_lack(%)	组织	AT_rich(%)	CG_rich(%)	CG/AT_lack(%)
管家基因	27.74	16.06	56.20	次级神经系统	22.63	35.79	41.58
膀胱	17.07	30.49	52.44	卵巢	28.57	19.84	51.59
血液	19.66	21.37	58.97	胰腺	19.20	22.52	58.28
骨	25.00	22.73	52.27	肌肉	18.32	26.72	54.96
骨髓	16.30	26.09	57.61	胎盘	17.36	25.00	57.64
大脑	21.85	15.13	63.02	前列腺	14.75	31.97	53.28
宫颈	26.05	15.13	58.82	小肠	19.33	26.89	53.78
结肠	18.12	29.53	52.35	皮肤	19.26	21.48	59.26
眼睛	19.45	27.08	53.47	脾脏	18.38	27.94	53.68
心脏	15.17	29.66	55.17	软组织	19.31	31.72	48.97
肾脏	15.28	31.25	53.47	胃	23.81	18.25	57.94
喉	18.81	23.76	57.43	睾丸	19.42	24.27	56.31
肝脏	11.70	41.49	46.81	胸腺	24.10	16.87	59.03
肺	24.55	39.73	35.71	舌头	23.81	20.41	55.78
淋巴	18.96	19.83	61.21	子宫	17.20	37.10	45.70
乳腺	17.83	27.13	55.04				

把上表做成图形以便于更直观的观察(见图 1)。

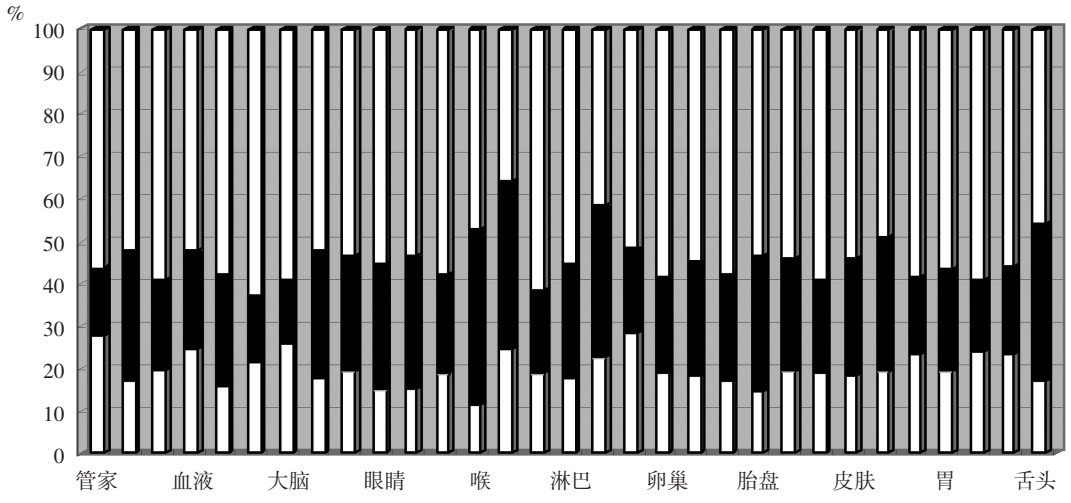

图 1 各组织中各富含模体的比率

Fig.1 The rate of rich motifs in each tissues

注:上:CG/AT_lack 模体; 中:CG_rich 模体; 下:AT_rich 模体。

Notes:top:CG/AT_lack motifs; middle:CG_rich motifs; below:AT_rich motifs.

从上图知,各个组织的调控元件中 CG/AT_lack 模体比率大致相同,而 AT_rich 模体、CG_rich 模体比率变化大,说明其主要差异在于 AT_rich 模体、CG_rich模体。虽然 CG/AT_lack 模体在各个组织中的比重都远远高于 AT_rich 模体、CG_rich 模体,但在各个组织中基本都出现,而在各个组织中 AT_rich 模体、CG_rich 模体出现不尽相同。由此可以推测控制基因表达差异性的元件是 AT_rich 模体或CG_rich 模体。从以上分析可以看出,HK 基因的调控模体中 AT_rich 模体的比率高于 CG_rich 模体比率,而大部分 TSP 基因(除骨、大脑、宫颈、卵巢、胃、胸腺、舌头外)的调控模体中 CG_rich 模体的比率高于AT_rich 模体的比率。由此可以推测 HK 基因和 TSP 基因的调控元件特征不同。即 HK 基因的调控元件偏向 AT_rich 模体,而 TSP 基因的调控元件偏向 CG_rich 模体。

4 各组织调控元件使用的差异性分析

一个组织的表现形式不同于另一个组织,虽然是由众多的因素造成,但是否与每个组织自身特有的调控元件有关呢? 为此我们考查了不同组织组织调控元件使用的差异性情况。

4.1 组织间调控元件的显著性差异结果

计算每两个组织调控元件的 Z 统计量,并根据此统计量检验每两个组织的调控元件是否具有显著性差异。结果显示,除组织次级神经系统、睾丸的调控元件与 HK 基因的调控元件不能判断具有显著性差异外,其他 TSP 基因的调控元件都与 HK 基因的调控元件具有显著性差异。

4.2 调控元件的差异性分析

以 HK 基因为例,HK 基因与除次级神经系统、睾丸这两个组织的其他 28 个组织都具有显著性差异,而通过统计发现 ATTTCG、GGCATA、GGTATA、GTATAG、TATAGT 和 TATGAC 这 6 个模体在测试集 S 中都出现了 28 次,说明这 6 个模体与在其他 28 个组织中都不出现的,而只出现在 HK 基因中,因此,我们可以推测这 6 个模体很可能是 HK 基因的特有调控元件。这 6 个模体所对应的转录因子是 HTF。再分析这 6 个模体,发现它们中有 5 个是AT_rich模体。若把启动子序列分为五个区域:0~200 bp、200~400 bp、400~600 bp、600~800 bp 和 800~1 000 bp。统计发现,这六个模体在启动子序列上的分布情况如下:

表 5 HK 基因中特有模体在启动子序列上的位置分布

Table 5 Location of specific motifs in the promoter sequences of HK genes

启动子区域	模体个数
0~200 bp	5
200~400 bp	1
400~600 bp	4
600~800 bp	0
800~1 000 bp	1

由上表知,特有模体的位置分布也具有一定的偏向性,在 0~200 bp 上分布最密集,其次在 400~600 bp上也有分布,而其他序列区域里则分布得少,甚至没有。

对每个组织仿效 HK 基因,对得到的测试集进行分析并确定这个组织的特有模体。结果显示各个组织的特有模体大部分都不相同,其中不乏共同使用特有模体的。如肾脏与小肠共同使用特有模体 CGCATG,这也能从一个侧面得到证实:肾脏和小肠[12]都具有消化功能。膀胱与子宫共同使用特有模体 CCGACG,有临床试验表明:子宫切除术后,由于支配膀胱的神经损伤和膀胱解剖位置的改变,常常引起膀胱功能的障碍[13]。再如肌肉、骨髓与肾脏共同使用特有模体 GCGCTA,这是由于体内肌肉组织代谢产物肌酐,经血循环到达肾脏,从肾小球滤过后从尿中排泄,因此可以通过测定血清肌酐浓度判断肾小球滤过功能。同时,肾脏分泌促进红细胞生成素,促进骨髓造血,生成红细胞。各组织特有模体的模体特征都偏向于 AT_rich 模体或 CG_rich 模体。

5　结　语

本文用简单易行的方法提取出各个组织的调控元件及特有模体。结果显示,在管家基因和 30 个组织特异性基因中,它们的调控元件都具有一定的偏向性,偏向于 AT_rich 模体或 CG_rich 模体,而具有 CG/AT_lack 模体特征的调控元件是极少的。由调控元件的这一偏向规律,在今后的调控元件识别中可以首先忽略 CG/AT_lack 模体,从而为识别过程缩小了范围。分析发现各组织之间有共同使用的调控元件,例如管家基因和眼睛,其调控元件个数分别为 126 和 132,他们共同使用的调控元件个数为 55。然而各个组织也有不少自己单独使用的调控元件即特有模体,这些特有模体专职调控这些组织的基因表达,模体特征都偏向于 AT_rich 模体或 CG_rich 模体,与过表达模体的偏向是一致的。由此可推测,特有模体控制着各组织基因的表达。本文的工作对于掌握人基因的转录调控机制和调控模体的作用具有一定的指导意义。

参考文献(References)

[1] 孙啸,陆祖宏,谢建明.生物信息学基础[M].北京:清华大学出版社,2005.
SUN Xiao, LU Zuhong, XIE Jianming. Foundation of Bioinformatics [M]. Beijing: Tsinghua University press, 2005.

[2] YU X, LIN J, ZACK D J, et al. Computational analysis of tissue-specific combinatorial gene regulation: predicting interaction between transcription factors in human tissues [J]. Nucleic Acids Research, 2006,34(17):4925-4936.

[3] LIU Xiong, YU Xueping, DONALD Z, et al. TiGER: A database for tissue-specific gene expression and regulation [J]. BMCBioinformatics,2008,9:271.

[4] 李婷婷,蒋博,汪小我,等. 转录因子结合位点的计算分析方法[J]. 生物物理学报,2008,24(5).
LI Tingting, JIANG Bo, WANG Xiaowo, et al. The method of calculation and analysis for transcription factor binding sites [J]. Acta Biophysica Sinica, 2008, 24 (5): 334-347.

[5] VAN HELDEN J. Metrics for comparing regulatory sequences on the basis of pattern counts [J].Bioinformatics, 2004,20(3):399-406.

[6] 朱建平.应用多元统计分析[M].北京:科学出版社,2006.
ZHU Jianping. Application of multivariate statistical analysis [M]. Beijing: Science Press, 2006.

[7] 李刚,高政. 人脸自动识别方法综述[J]. 计算机应用研究,2003,(8):4-9,40.
LI Gang, GAO Zheng. A survey of automatic human face recognition [J]. Application Research of Computers, 2003,(8):4-9,40.

[8] 侯咏佳,方东博,袁生光,等.主成分分析算法的 FPGA 实现[J].机电工程,2008,25(9):37-40.
HOU Yongjia, FANG Dongbo, YUAN Shengguang, et al. The principal component analysis algorithm realized by FPGA [J]. Mechanical & Electrical Engineering, 2008, 25 (9): 37-40.

[9] 吴喜之.非参数统计[M].北京:中国统计出版社,1999.
WU Xizhi. Non-parametric statistics [M]. Beijing: China Statistical Press, 1999.

[10] WINGENDER E, CHEN X, HEHL R, et al. TRANSFA: an integrated system for gene expression regulation [J]. Nucleic Acids Res., 2000, 28, 316-319.

[11] LI Huimin, CHEN Dan, ZHANG Jing. Statistical analysis of combinatorial transcriptional regulatory motifs in human intron-containing promoter sequences [J]. Computational Biology and Chemistry, 2013, 43(2):35-45.

[12] 庞智.小肠消化吸收营养基因的区段性表达和调节[J]. 国外医学(卫生学分册),1998,25(4):230-234.
PANG Zhi. The digestion and absorption section expression and regulation of nutrient gene in small intestinal [J]. Foreign Medical Sciences (Health Sciences), 1998, 25 (4): 230-234.

[13] 张文淼,施铮铮,吴雪清,等.子宫切除术后膀胱功能障碍患者的尿动力学分析[J].中华妇产科杂志,2005, 40 (11): 778-779.
ZHANG Wenmiao, SHI Zhengzheng, WU Xueqing, et al. Analysis of hysterectomy urinary dynamics in patients with bladder dysfunction [J]. Chinese Journal of Obstetrics and Gynecology, 2005, 40 (11): 778-779.