生物化学与生物物理进展 *Progress in Biochemistry and Biophysics*, 2006, 33(11): 1044~1050
www.pibb.ac.cn

PIBB

用非联配方法预测人类转录调节模体 *

吕 军[1,2] 罗辽复[1)**] 张 颖[1,2] 赵巨东[1,2]
([1)]内蒙古大学物理系，呼和浩特 010021；[2)]内蒙古工业大学物理系，呼和浩特 010051)

摘要 通过对 TRANSFAC 数据库中转录因子结合位点 (TFBS) 所包含核苷 k 联体 (k-mer) 在人类和小鼠基因组启动子区中分布的比较分析，提出一种在人类全基因组启动子区搜索转录调节 k-mer 模体 (transcription regulatory k-mer motifs，TRKMs) 的非联配快速算法——基于距离的保守 k-mer 搜索算法 (distance-based conservative k-mer searching algorithm，DCKS algorithm). 应用该算法，对人 7-mer 转录调节模体进行预测，预测结果敏感性为 90%，特异性为 78%，相关系数为 0.65.

关键词 转录调节模体，非联配途径，基于距离的保守 k-mer 搜索算法，二次判别分析
学科分类号 Q61

高等真核生物转录调节元件及其相互作用网络的研究已成为分子生物学的研究热点，随着实验技术及生物信息学方法的发展[1,2]，在全基因组范围高通量地预测转录因子调节元件成为可能[3~15].从实验角度看，微阵列基因表达图谱分析[3]、染色质免疫沉淀法[4]和 Dam 甲基化酶鉴定法[5]是通用的方法.文献[1]对这些技术作了较详细的综述.就生物信息学方法而言，有统计显著性分析[6,7]，系统发育足迹法 (phylogenic footprint)[8~11]，从 Motif 下游的基因表达水平来发现调节元件[12]，以及以上方法的综合运用[13].文献[14]用 4 个物种的标准数据集测试比较了现有 13 种转录因子结合位点 (transcription factor binding sites，TFBS) 的计算工具.

现有 TFBS 的预测算法，大多数是基于序列联配的.序列联配尤其是多序列联配，其时间和空间复杂度很高，而且由于调节元件一般都是 10 bp 左右的短信号，存在大量统计噪声，进化中的基因组重排已使 TFBS 的位置和序列发生了很大变化，对于特异性的或低保守性的调节元件极难准确发现，同时，基因组间的基础保守性 (basal conservation) 还产生了大量假阳性，所以现有算法只能给出少量的与 TRANSFAC[16]数据库匹配的 TFBS.文献[14]指出，所有 13 种调节元件预测工具的预测能力都比较低，位点水平敏感性指标最大为 0.22，核酸水平相关系数最大为 0.20. 针对 TFBS 预测中的问题，探索新的算法极有必要.最近有工作[15]提出一种保守调节元件的非联配算法，试图绕过序列联配，但该算法要用 BLAST 搜索同源开放阅读框(ORF)对，所以是一种“准非联配算法”.

本文通过对 TRANSFAC[16]数据库中人类基因组 TFBS 所包含的 k-mers 在人类和小鼠的基因组 PolⅡ启动子中分布的比较分析，提出一种在人类全基因组启动子区，搜索转录调节 k-mer 模体 (transcription regulatory k-mer motifs，TRKMs)的非联配快速算法——基于距离的保守 k-mer 搜索算法 (distance-based conservative k-mer searching algorithm，DCKS algorithm). 应用该算法，对转录调节模体进行预测，得到较好的结果.

这个算法的核心是用一对 k-mer 的距离来描述该对 k-mer 的保守性. 更具体地说， 是用人鼠基因组中一对 k-mer 的距离来描述该对 k-mer 在人鼠基因组中的保守性.不是着眼于一段序列的相似性，而是着眼于一个 k-mer 对是否在两基因组中同时在近距离出现的保守性，因此不需要序列比对，而且这种保守性也不需要以基因组中严格定位为前提. 这种保守性称为 k-mer 对保守性. 在本文中我们取 $k = 7$.

*国家自然科学基金资助项目(90403010).
** 通讯联系人. Tel: 0471-4992676, E-mail: lfluo@mail.imu.edu.cn
收稿日期：2006-04-30，接受日期：2006-06-02

1 数据和方法

1.1 数据集

本文使用的基本数据集，分为人类基因组转录因子结合位点数据集与人类和小鼠基因组 PolⅡ启动子序列数据集两部分.

1.1.1 人类基因组 TFBS 数据集. 人类基因组 TFBS 数据集从 TRANSFAC7.0 [16] 数据库 (http://www.gene-regulation.com/pub/databases.html)的 Site 子库下载.由 TRANSFAC 数据库下载其登录的人类全部 1 388 个 TFBS 数据，去掉一些只有转录因子记录而没有结合位点 DNA 序列记录的数据. 我们共得到 1 342 条结合位点 DNA 序列，核苷酸数为 25 129 bp.其中有 1 045 条序列在 TRANSFAC 数据库中登录的位置约在−700～+300 bp 间 (以转录起始位点——transcription start site(TSS)为参考+1 位点)，核苷酸数为 19 008 bp. 另外，还有 297 条序列位置约在−700～+300 bp 之外，核苷酸数为 6 121 bp.

1.1.2 人类和小鼠基因组 PolⅡ启动子序列数据集. 人类和小鼠基因组 PolⅡ启动子序列数据集取自 ZHANG 的实验室 (ftp://cshl.org/pub/science/mzhanglab/PromoterSet/Refseq)[17]. 其中人类的启动子数据集共登录 16 175 条序列，小鼠的启动子数据集共登录 12 801 条序列. 每条启动子序列的长度为 1 000 bp (−700～+300 bp，以 TSS 参考+1 位点). 我们手工剔除掉序列中含有未知碱基的序列，整理得到人类的启动子序列 16 163 条，小鼠的启动子序列 11 968 条.

1.1.3 7-mer 数据整理. 搜索 1 045 条位置在−700～+300 bp 间的 TFBS 序列所包含的 7-mer，共得到 6 772 个 7-mer，这 6 772 个 7-mer 作为我们预测的训练正集，称为 Ps 集.另外，搜索 297 条位置在−700～+300 bp 之外的 TFBS 序列所包含的 7-mer，得到不同于训练正集的 1 480 个 7-mer，作为对预测结果的一个核对集，称为 Pc 集. 其余 8 132 (16 384 − 6 772 − 1 480 = 8 132)个 7-mer 组成的集合，称为 Ntf 集，在 Ntf 集中有 680 个被选出作为预测的训练负集，称为 Nt 集， Nt 集是统计性质明显异于 Ps 集的 7-mer 集合，剩余的 7 452 个 7-mer 作为有待发现新的 TRKMs 的集合，称为 Nf 集.关于预测负集筛选的方法将在 **1.2.1** 节中介绍.

1.2 方法

本文提出一种在人类全基因组启动子区搜索 TRKMs 的非联配快速算法——基于距离的保守 k-mer 搜索算法，我们称之为 DCKS 算法.该算法包含 2 个子算法，其一是 k-mer 到转录起始位点的距离分布算法，称为 DCKS1 算法，用于对预测负集筛选.其二是一对 k-mer 的距离分布算法，称为 DCKS2 算法，用来与二次判别分析相结合，预测 TRKMs. 下面我们对 DCKS 算法进行详细描述.

1.2.1 DCKS1 算法.

为了从 Ntf 集的 8132 个 7-mer 中筛选出 Nt 集，我们提出 DCKS1 算法来实现这一目的.该算法的基本思想为 Nt 集元素的保守性较 Ps 集明显低，算法的详细步骤如下.

x 集 (x = Ps 或 Ntf) 中的第 i 个 7-mer，第 j 次出现在物种 z (z = Hs 代表人或 Mm 代表鼠) 的启动子序列集中时，与 TSS 的距离记为 $L_{ij}^{z}(x)$，距离分布的均值和方差分别记为 $D_i^z(x)$和 $S_i^z(x)$. 无论 7-mer 出现在 TSS 上游还是下游，其与 TSS 的距离一概取绝对值，用 7-mer 的第一位碱基位置与 TSS 之间间隔的碱基数目来度量.

$$D_i^z(x)=\frac{1}{N_i^z(x)}\sum_{j=1}^{N_i^z(x)}L_{ij}^z(x) \tag{1}$$

$$S_i^z(x)=\frac{1}{N_i^z(x)}\sum_{j=1}^{N_i^z(x)}[L_{ij}^z(x)-D_i^z(x)]^2 \tag{2}$$

这里，$N_i^z(x)$表示 x 集中第 i 个 7-mer 出现在物种 z 的启动子序列集中的总次数.

研究 Ps 集元素与 TSS 的距离方差在两个物种中的差别. 统计 $|S_i^{Hs}(Ps)-S_i^{Mm}(Ps)|$ 的分布，其均值和偏差分别记为 A 和 σ，即令

$$A=\frac{1}{P}\sum_{i=1}^{P}|S_i^{Hs}(Ps)-S_i^{Mm}(Ps)| \tag{3}$$

$$\sigma=\sqrt{\frac{1}{P}\sum_{i=1}^{P}[|S_i^{Hs}(Ps)-S_i^{Mm}(Ps)|-A]^2} \tag{4}$$

这里，P 为 Ps 集中的 7-mer 总数，本文中 P = 6 772. 考虑到 Ntf 集元素一般比 Ps 集元素的保守性弱 (见讨论 **2.2.1**)，为了在 Ntf 集中明确分出那些保守性特别低的 7-mer 元素，作为 Nt 集，我们定义：

当 Ntf 集中的任一个 7-mer 满足下式时，就将此 7-mer 判别为 Nt 集元素.

$$|S_i^{Hs}(Ntf)-S_i^{Mm}(Ntf)|>A+2\sigma \tag{5}$$

此式表示，当用 7-mer 与 TSS 距离分布的方差来

度量时，在 95%的置信水平下，Nt 集元素在人鼠间的保守性低于 Ps 集元素. 依据这一算法，我们从 Ntf 集的 8 132 个 7-mer 中筛选出了 680 个作为 Nt 集元素.

1.2.2 DCKS2 算法.

DCKS2 算法是 DCKS 算法的主要部分. 这个算法的核心是用一对 k-mer 的距离来描述该对 k-mer 的保守性. 更具体地说， 是用人鼠启动子中一对 k-mer 的距离来描述该对 k-mer 在人鼠基因组中的保守性. 这种保守性称为 k-mer 对保守性. 下面我们将详细描述该算法.

x 集中的第 i 个 7-mer 与 y 集中的第 j 个 7-mer (x，y = Ps，Pc，Nt 或 Nf) 第 k 次同时出现在物种 z (z = Hs 或 Mm) 的某一条启动子序列上时，ij 之间的距离绝对值 (用 7-mer 对的第一位碱基的位置之间间隔的碱基数目来度量) 记为 $L^z_{ijk}(xy)$. 令 ij 之间距离分布的均值和方差分别记为 $D^z_{ij}(xy)$和 $S^z_{ij}(xy)$，

$$D^z_{ij}(xy)=\frac{1}{N^z_{ij}(xy)}\sum_{k=1}^{N^z_{ij}(xy)}L^z_{ijk}(xy) \tag{6}$$

$$S^z_{ij}(xy)=\frac{1}{N^z_{ij}(xy)}\sum_{k=1}^{N^z_{ij}(xy)}[L^z_{ijk}(xy)-D^z_{ijk}(xy)]^2 \tag{7}$$

这里，$N^z_{ij}(xy)$为 x 集中的第 i 个 7-mer 与 y 集中的第 j 个 7-mer 在物种 z 的启动子序列集中出现的总次数.

研究$D^z_{ij}(xy)$和$S^z_{ij}(xy)$在 2 个物种中的差别. 指定 x 集中的第 i 个 7-mer 与 y 集中各个 7-mer 构成 7-mer 对，每一 7-mer 对的平均距离及距离方差在 2 物种间的绝对差值在所有 7-mer 对中的平均分别记为 $\overline{\Delta D_i}(xy)$和 $\overline{\Delta S_i}(xy)$，即

$$\overline{\Delta D_i}(xy)=\frac{1}{W_i(xy)}\sum_{j=1}^{W_i(xy)}|D^{\mathrm{Hs}}_{ij}(xy)-D^{\mathrm{Mm}}_{ij}(xy)| \tag{8}$$

$$\overline{\Delta S_i}(xy)=\frac{1}{W_i(xy)}\sum_{j=1}^{W_i(xy)}|S^{\mathrm{Hs}}_{ij}(xy)-S^{\mathrm{Mm}}_{ij}(xy)| \tag{9}$$

这里，$W_i(xy)$为指定 x 集中的第 i 个 7-mer 与 y 集中的所有 7-mer 构成的 7-mer 对总数，由于可能有些 7-mer 对在人或鼠的启动子序列集中不出现，所以，总有 $W_i(xy)$小于等于 y 集包含的 7-mer 数.

至此，我们构建了衡量某一 7-mer 对在 2 个物种间保守性的参数——平均距离之差及方差之差. 如何整合这些参数，从而对任一 7-mer 作出判别，给出该 7-mer 为 TRKMs 可能性的一个评价? 我们使用的参数整合方法是二次判别分析.下面将对此作简要叙述，详细步骤可以参见文献[18～21].

1.2.3 信息的二次判别函数 (quadratic discriminant analysis， QD) 整合.

对 x 集中的第 i 个 7-mer，当 y 集分别取为 Ps 集和 Nt 集时，我们可以由 DCKS2 算法得到下面 4 个参数，由这 4 个参数构成一个 4 维的二次判别向量，记为 $\boldsymbol{R}_i(x)$，即有

$$\boldsymbol{R}_i(x)=[\overline{\Delta D_i}(x\mathrm{Ps}),\overline{\Delta S_i}(x\mathrm{Ps}),\overline{\Delta D_i}(x\mathrm{Nt}),\overline{\Delta S_i}(x\mathrm{Nt})] \tag{10}$$

若当 x 集又分别取为 Ps 集和 Nt 集时，我们将得到训练集向量 $\boldsymbol{R}_i(\mathrm{Ps})$和 $\boldsymbol{R}_i(\mathrm{Nt})$. 记 P 和 N 分别为 Ps 集和 Nt 集中的 7-mer 总数. 训练集向量的平均向量分别记为 $\boldsymbol{\mu}(\mathrm{Ps})$和 $\boldsymbol{\mu}(\mathrm{Nt})$，协方差矩阵分别记为 $\Sigma(\mathrm{Ps})$和 $\Sigma(\mathrm{Nt})$.

则对任意 x 集中的第 i 个 7-mer 其二次判别值 ξ由下式给出.

$$\xi=\ln\frac{P}{N}-\frac{\delta_i(x\mathrm{Ps})-\delta_i(x\mathrm{Nt})}{2}-\frac{1}{2}\ln\frac{|\Sigma(\mathrm{Ps})|}{|\Sigma(\mathrm{Nt})|} \tag{11}$$

其中，$\delta_i(xy)=[\boldsymbol{R}_i(x)-\boldsymbol{\mu}(y)]^T\Sigma^{-1}(y)[\boldsymbol{R}_i(x)-\boldsymbol{\mu}(y)]$，是 $\boldsymbol{R}_i(x)$与$\boldsymbol{\mu}(y)$之间的马氏距离，$|\Sigma(y)|$为 $\Sigma(y)$的行列式值，这里 y = Ps, Nt.

(11)式由 Bayes 理论导出，这里 ξ 是正负集后验概率比的自然对数. 如果特征(信息参数)选的合适， 正负集可在 ξ 空间的 0 附近分得很开.本文中 ξ 的分类 threshold 值选为 ξ_0= 0.

2 结果和讨论

2.1 结果

2.1.1 Ps 和 Nt 集中的检验结果.

首先以 Ps 集和 Nt 集为对象.将 Ps 集的 6 772 个 7-mer 随机分为 10 组，每组元素为 678 或 677 个，和 Nt 集的 680 个元素合在一起进行检验.对每一组进行检验时，在 678 或 677 个 Ps 集 7-mer 中随机抽出 100 个作为检验集，其余为训练集. 680 个 Nt 集元素中也随机抽出 100 个作为检验集，其余为训练集.这样给出独立的 10 组检验结果见表 1.

对结果的评价指标我们采用常规的敏感性指标 Sn，特异性指标 Sp 及相关系数 CC，

$$Sn=[TP/(TP+FN)]\times 100\%$$

$$Sp=[TP/(TP+FP)]\times 100\%$$

$$CC=\frac{(TP\times TN)-(FN\times FP)}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \tag{12}$$

由表 1 可以看出，我们的算法对 TRKMs 具有很高的识别精度，10 组的平均结果为 Sn = 90.0%，Sp = 77.6%，CC = 0.65. 对每一组也进行了自洽检验，10 组平均结果为 Sn = 89.6%，Sp = 75.4%，CC = 0.62. 这个算法的成功说明，用 k-mer 对的距离来描述物种间保守性是可行的.

Table 1 Test on distance-based conservative 7-mer searching algorithm in Pc and Nt set

	Sn%	*Sp*%	*CC*
1	84.0	76.4	0.58
2	89.0	80.9	0.68
3	91.0	77.8	0.66
4	95.0	75.4	0.66
5	89.0	74.2	0.59
6	89.0	80.2	0.67
7	92.0	75.4	0.64
8	86.0	76.8	0.60
9	96.0	81.4	0.75
10	89.0	78.1	0.65
Average	90.0	77.6	0.65

2.1.2 Pc 集和 Nf 集中的预测结果.

进一步，我们将 Ps 集随机分为 6 组作为训练正集，每组的训练负集为 Nt 集，用 DCKS2 算法对 Pc 集和 Nf 集进行了预测. 然后将 6 组判别结果的 ξ 值进行平均，给出 Pc 集和 Nf 集中的 7-mer 按 ξ 值的一个排序.

令 $\xi_0 = 0$ 为判别 threshold 值，Pc 集中预测为真的 7-mer 有 1 324 个，预测精度为 (1 324/1 480)× 100% = 89.5%，这与以上对 Ps 集的检验结果是吻合的. 这可以看作是一个对我们的预测方法准确性的评估， 准确性约为 90%.

Nf 集中预测为真的 7-mer 有 5 500 个，预测为假的 7-mer 有 1 952 个. 这 1 952 个 7-mer 与 Nt 集的 680 个 7-mer 合在一起，总数为 2 632 个，这些 7-mer 在本文方法中被判定为非 TRKMs，约占总 7-mer 数的 16%. 本文判定的参与转录调节的 7-mer 模体 (TRKMs) 约为 13 000 个左右，而目前 TRANSFAC 登录的 TFBS 所包含的 7-mer 数约为 8 000 个，相差约 5 000 个.

在 Nf 集中高 ξ (>5) 值的 320 个 7-mer，可以作为 TRKMs 的可靠候选者，列在表 2 中.

2.2 讨论

2.2.1 DCKS 算法的依据是 k-mer 对保守性.

本文非联配的 DCKS 算法的核心是 k-mer 对保守性，它提出的依据是什么？在 16 384 个 7-mer (65 536 个 8-mer)中，TRANSFAC 库中列出作为人类 TRKM 的约占一半(1/6)，它们在 Promoter 区的频次分布和随机情形相比，并未显示出向高频端的偏移，相反，有略向低频端偏移的特征，因此从 7-mer(或 8-mer)的出现频次来发现 TRKM 是不可能的. 事实上，单从人类基因组的序列资料，而不利用和其他基因组的进化比较，很难准确找出 TRKM[22]. 通过人鼠基因组的比较，我们发现：a. 7-mer 与 TSS 的距离分布，b. 7-mer 对的距离分布，具有较高的保守性. 由 7-mer 与 TSS 距离均值 $D_i^z(x)$ 构成 $D_p = \langle D_i^{Hs}(\mathrm{Ps}) - D_i^{Mm}(\mathrm{Ps})\rangle_i$ 和 $D_N = \langle D_i^{Hs}(\mathrm{Ntf}) - D_i^{Mm}(\mathrm{Ntf})\rangle_i$ ($\langle\rangle_i$ 代表对 i 平均)， 由 7-mer 与 TSS 距离方差 $S_i^2(\mathrm{x})$ 构成 $S_p = \langle S_i^{Hs}(\mathrm{Ps}) - S_i^{Mm}(\mathrm{Ps})\rangle_i$ 和 $S_N = \langle S_i^{Hs}(\mathrm{Ntf}) - S_i^{Mm}(\mathrm{Ntf})\rangle_i$. 统计结果给出：$D_P$ 和 D_N 几乎无差别，但 S_P 和 S_N 差别显著， S_N/S_P >1.3. 因此， 7-mer 与 TSS 的距离方差具有保守性，可以用做寻找 TRKM 的依据. 进一步， 由一对 7-mer 的距离均值 $D_y^2(\mathrm{xy})$ 构成 $E_p = \langle D_{ij}^{Hs}(\mathrm{PsPs}) - D_{ij}^{Mm}(\mathrm{PsPs})\rangle_{ij}$ 和 $E_N = \langle D_{ij}^{Hs}(\mathrm{NtfNtf}) - D_{ij}^{Mm}(\mathrm{NtfNtf})\rangle_{ij}$ ($\langle\rangle_{ij}$ 代表对 ij 平均)， 由 7-mer 对的距离方差 $S_{ij}^z(xy)$ 构成 $T_p = \langle S_{ij}^{Hs}(\mathrm{PsPs}) - S_{ij}^{Mm}(\mathrm{PsPs})\rangle_{ij}$ 和 $T_N = \langle S_{ij}^{Hs}(\mathrm{NtfNtf}) - S_{ij}^{Mm}(\mathrm{NtfNtf})\rangle_{ij}$. 统计结果给出： E_P 和 E_N 的差别及 T_P 和 T_N 的差别都很显著， $E_N/E_P \approx 1.69$， $T_N/T_P \approx 1.57$. 因此， 7-mer 对的距离均值和方差都具有强保守性，都可以用来作为寻找 TRKM 的依据.又由于这类保守性不依赖序列联配，它们具有新的广泛的应用价值. 以上分析是我们提出 k-mer 对保守性以及建立在此保守性基础上的 DCKS 算法的依据.

2.2.2 k-mer 对保守性机制的分析.

k-mer 对保守性有两种可能机制. 机制 1： 人和鼠基因组中共有某些长的 (长度 > 7 bp) TFBS；机制 2：由于转录因子的协同作用，TFBS 组织为模块(module).

第一种机制是显然的，人和鼠基因组间的长的保守 TFBS 将直接被体现为 k-mer 对保守性.关于第二种机制的作用，可作如下的统计.

为了考虑 2 种机制的相对贡献大小，我们研究正集 k-mer 对在人和鼠的启动子序列集中的距离分布： 统计人和鼠基因组中每一 PsPs 对的距离分布的峰值和均值，求得峰值和均值对各个 PsPs 对的平均. 统计的结果见表 3.由表 3 可见，PsPs 对的峰值和均值一般都较大. 这说明 k-mer 对保守性的主要机制是第二种，即顺式调节模块(*cis*-regulatory modules，CRMs) 结构.

Table 2 Top 7-mers in Nf set with threshold $\xi > 5$ (320 7-mers in ξ order)

Rank	7-mer	Rank	7-mer	Rank	7-mer	Rank	7-mer	Rank	7-mer	Rank	7-mer
1	cagctcc	56	tctcttt	111	cgcagag	166	agccggc	221	gctgcgc	276	agggtcc
2	gaggaag	57	gggcaga	112	tggcttc	167	gggagac	222	ggccggc	277	agcaagg
3	ccagctc	58	tgcctgg	113	ttaaaaa	168	cgccctg	223	gtctccc	278	gccaccc
4	gggaaag	59	ctgagct	114	gagaccc	169	cctcacc	224	agagagc	279	aaggtgg
5	ggctgcc	60	cagcggc	115	ccccttt	170	tggcagg	225	ctgttcc	280	tgcagct
6	ctctggg	61	agccagc	116	gagcaga	171	aggcctg	226	ccatctc	281	ggctttg
7	ccggaag	62	ggagctc	117	gcgcctc	172	tggccct	227	tcaggaa	282	ggaaccc
8	gcagctg	63	gcagaga	118	cagggaa	173	gccagcg	228	cagttcc	283	gagtgag
9	cctctgc	64	tgcaggg	119	ggcaggt	174	cggagag	229	ctctgcg	284	tcagcag
10	aggaggc	65	cccacag	120	cccaggt	175	aggctct	230	tggaggt	285	ttcagga
11	agagagg	66	agccccg	121	agctccg	176	ccccaag	231	cggggaa	286	agtccca
12	cctttcc	67	agtgggg	122	gctgctg	177	tccttct	232	caagatg	287	ccagagt
13	cagcagc	68	aagatgg	123	ggagcgg	178	gccagca	233	gagctct	288	tctgcgc
14	cagaggc	69	ctctttc	124	gaagccc	179	ggcttct	234	actgcag	289	gccggaa
15	ttttctt	70	cggaagt	125	tccacct	180	ggactgg	235	caggact	290	ccagcgg
16	agcagcc	71	cagccgc	126	ggctgtg	181	tgtcccc	236	gccaggt	291	gaaggtg
17	aggaaga	72	cggctcc	127	gcggaga	182	aggactg	237	cagcggg	292	aggagcg
18	gagctgc	73	ggaggca	128	ggcggct	183	ttctctg	238	cggctgg	293	cgcaggg
19	ttcctgg	74	agctccc	129	tcctgag	184	agaagca	239	cgggtcc	294	agagcaa
20	cctggga	75	caggcgg	130	tggagct	185	aaggggg	240	ccgcgga	295	catctcc
21	gctctgc	76	ctggaag	131	acctggg	186	aagagag	241	tgtcttt	296	cgcgctg
22	aggctgc	77	cagaaag	132	aggccgg	187	aggacag	242	ttcagcc	297	gaggtgc
23	gagagaa	78	gtcccag	133	tgtttct	188	tttctca	243	agcggag	298	gacagga
24	caggagc	79	gcaggct	134	cggcagc	189	ctgtgcc	244	tccctca	299	ctggtga
25	tctgggc	80	gtccctg	135	cctgaag	190	ctgcgcg	245	gcggctc	300	ggtgagc
26	gctgccc	81	gagctcc	136	cagccgg	191	cactccc	246	ggcgccg	301	ctgaagg
27	ggaagcc	82	aggagca	137	cagtccc	192	ctccagt	247	ggcaagg	302	gagagca
28	ctcccac	83	ggtcccc	138	aggggct	193	tcctcac	248	ctgaagc	303	ggccaca
29	gcggagc	84	agccgcc	139	tttctgc	194	gccggct	249	gcagaaa	304	agacagc
30	ccctggg	85	ccctttc	140	gccgagc	195	gggactc	250	aaaaccc	305	gaagagc
31	tctctgc	86	tctgctc	141	gctgtgg	196	tggtggg	251	ctccgag	306	gagcaag
32	cagagga	87	gccacag	142	ggtgctg	197	ccacttc	252	ttaaaat	307	ggacctg
33	ctggagc	88	ggcttgg	143	tgctggc	198	ccacaga	253	agaaaca	308	ggcagaa
34	ggaagaa	89	ccaggtg	144	cggaagc	199	tgctctc	254	tgaggga	309	attttct
35	gggagct	90	ctggtgg	145	ggcagca	200	gggaacc	255	ctcaggt	310	gcactgg
36	agcagag	91	cctgtcc	146	acagctg	201	ggtctcc	256	cttcagc	311	atggcgg
37	ctctctg	92	tgcgcag	147	tgaggag	202	cttgggc	257	ggtgaga	312	agccact
38	ctctgct	93	agcagaa	148	cgcggag	203	gaggagt	258	ggttccc	313	gagagac
39	cggcccc	94	ctgggtt	149	cccggga	204	gcagctt	259	agagcgc	314	caggaca
40	gcttctc	95	tccccgg	150	gtctctg	205	accccca	260	tgggacc	315	gagcgga
41	tctgcag	96	cagcaga	151	cgccgcg	206	agaaaat	261	ctttaaa	316	tgccagc
42	ctgaggg	97	gggaagc	152	cactgcc	207	ctgaaga	262	gctctct	317	tttcttg
43	cagcttc	98	cagcacc	153	gtcccca	208	cttggag	263	actccca	318	caagaaa
44	cctgggt	99	ctctctt	154	ccggccg	209	gaaaaca	264	gaaggcc	319	gggctgt
45	tttaaaa	100	gaaagag	155	acttccg	210	gcagccg	265	ccgggtc	320	tgggaac
46	gctccca	101	ggagacc	156	ctgccgg	211	ctcaccc	266	ccctttt		
47	tccctgc	102	cgctggg	157	aggtgga	212	aagaagc	267	agccctc		
48	ttccagg	103	cagcccg	158	cttcagg	213	tgttttc	268	gaagatg		
49	ttctctt	104	agcggcg	159	gagcgcg	214	gcgggct	269	gggccac		
50	ggagtgg	105	caccctg	160	cacctgc	215	tgaaaaa	270	tctctgt		
51	agcccca	106	gctgaag	161	agcagcg	216	ttctcag	271	ccatttt		
52	ccgcagc	107	agcaggc	162	gaggcct	217	gggccct	272	agctgga		
53	gctccct	108	tttgggg	163	tgcgcgc	218	cctggtg	273	tgtctct		
54	gaagcag	109	ggaagct	164	ccagccg	219	tgtgttt	274	cagaccc		
55	gccgcag	110	aacccag	165	cacagct	220	gcgctgg	275	cctgccg		

Table 3 Distribution of 7-mer pair distances in human and mouse Ps set

Species	Peak value of 7-mer pair distances (average)	Mean value of 7-mer pair distances (average)
Human	190	337
Mouse	195	337

高等真核生物由于转录因子的协同作用，TFBS 常常会成簇出现，组织为模块结构.最近的研究工作[23]估算，人类基因组中 58%的 CRMs 长度小于 500 bp，CRMs 的平均长度为 635 bp，此计算值和实验值很接近.如果模块长度为 L，一个模块包含 6 个等长 TFBS，则可算得 2 个 TFBS 的平均距离为 0.4L；一个模块包含 5 个等长 TFBS，则可算得 2 个 TFBS 的平均距离为 0.5L. 因此，模块长度大约为 TFBS 平均距离的 2 倍. 由 CRMs 的平均长度可得 TFBS 的平均距离约为 300 bp. 这和表 3 给出的 7-mer 对距离平均值一致.故可以认为， 7-mer 对保守性的主要机制是 TFBS 模块结构.

2.2.3 Motif 搜索问题.

直接对单个的 TFBS 进行预测是一个极为困难的问题.一方面，是由于 TFBS 本身只是一些短的(5～15 bp) DNA 序列，这样的短片段本身所包含的信息是极为有限的；另一方面，TFBS 的碱基组成是高度退化的，这一点对 TFBS 的精确预测也设置了极大的障碍.

我们把 TFBS 预测问题分成两步，先对全部的 k-mer 进行 TRKMs 预测，对每一个 k-mer 都给出一个 TRKMs 的倾向性评价，然后再进行 TRKM 定位研究. 本文所做的是第一步，这一做法可能有助于处理进化中基因组重排引起的 TFBS 变化的复杂性.

本文提出并应用 DCKS 算法将所有 7-mer 进行预测. 从结果看：a. 我们估计出参与转录调节的 7-mer 模体约 13 000 个左右，其中有 8 000 个左右已经在 TRANSFAC7.0 数据库中登录，还有 5 000 个可以作为实验检验的对象，尤其是表 2 中给出的顶端的 320 个 7-mer，这是我们重点预测的新的 TRKMs. b. 我们预测非 TRKMs 的 7-mer 有约 3 000 左右，仅占 7-mer 总数的不到 20%. TRKMs 的数量这么大，这一点也与大部分 TFBS 是高度退化的 DNA 序列这一事实相一致.

曾有研究工作对人类启动子中的 8-mer 频数分布进行研究和分类[24]，但很少见到对 k-mer 的直接预测. 有工作[15]提出了保守调节元件的非联配算法，在得到两物种间的同源 ORF 对的基础上，统计任一 7-mer 同时出现在同源 ORF 对上游序列的频次，从而给出该 7-mer 的保守得分，以此为据评价该 7-mer 是否可认为是保守调节元件的候选者. 此工作没有考虑 k-mer 对的信息 (我们认为 k-mer 对信息对于 TFBS 预测至关重要)，也未能给出一个 7-mer 是否为 TRKM 的明确界定. 从结果看，本文预测的顶端 320 个 7-mer 和文献[15]中保守得分在 72 nat 以上的比较，仅有 24.1%是重合的，但也有 75.9%是存在重要差别的. 从 TRANSFAC 库的进一步扩大将有可能对两种方法的差别作出评价. 然而更重要的是 TRKM 的定位研究， 我们希望这个工作能为进一步的全面解决 TFBS 预测问题提供直接的可依据的线索.

参 考 文 献

1 Taverner N V, Smith J C, Wardle F C. Identifying transcriptional targets. Genome Biology, 2004, **5** (3): 210.1～210.7

2 Wasserman W W, Sandelin A. Applied bioinformatics for the identification of regulatory elements. Nature Reviews Genetics, 2004, **5** (4): 276～287

3 Schena M, Shalon D, Davis R W, *et al*. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science, 1995, **270** (5235): 467～470

4 Shannon M F, Rao S. Transcription:Of chips and ChIPs. Science, 2002, **296** (5568): 666～669

5 van Steensel B, Henikoff S. Identification of *in vivo* DNA targets of chromatin proteins using tethered Dam methyltransferase. Nature Biotechnology, 2000, **18** (4): 424～428

6 Frith M C, Hansen U, Spouge J L, *et al*. Finding functional sequence elements by multiple local alignment. Nucleic Acids Research, 2004, **32** (1): 189～200

7 Pavesi G, Mereghetti P, Mauri G, *et al*. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Research, 2004, **32**(Web Server issue): W199～203

8 Blanchette M, Tompa M. Discovery of regulatory elements by a computational method for phylogenetic footprinting. Genome Research, 2002, **12** (5): 739～748

9 Cliften P, Sudarsanam P, Desikan A, *et al*. Finding functional features in *Saccharomyces* genomes by phylogenetic footprinting. Science, 2003, **301** (5629): 71～76

10 Moses A M, Chiang D Y, Pollard D A, *et al*. MONKEY: identifying conserved transcription factor binding sites in multiple alignments using a binding site-specific evolutionary model. Genome Biology, 2004, **5** (12): R98.1～15

11 Chin C S, Chuang J H, Li H. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral

sequence. Genome Research, 2005, **15** (2): 205～213

12 Bussemaker H J, Li H, Siggia E D. Regulatory element detection using correlation with expression. Nature Genetics, 2001, **27** (2): 167～171

13 Xie X, Lu J, Kulbokas E J, *et al*. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. Nature, 2005, **434** (7031): 338～345

14 Tompa M, Li N, Bailey T L, *et al*. Assessing computational tools for the discovery of transcription factor binding sites. Nature Biotechnology, 2005, **23** (1): 137～144

15 Elemento O, Tavazoie S. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. Genome Biology, 2005, **6** (2): R18.1～27

16 Matys V, Kel-Margoulis O V, Fricke E, *et al*. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Research, 2006, **34** (Database issue): D108～110

17 Xuan Z, Zhao F, Wang J, *et al*. Genome-wide promoter extraction and analysis in human, mouse and rat. Genome Biology, 2005, **6** (8): R72.1～12

18 McLachlan G J. Discriminant analysis and statistical pattern recognition. New York :Wiley, 1992. 1～526

19 Zhang M Q.Identification of protein coding regions in the human genome by quadratic discriminant analysis. Proc Natl Acad Sci, 1997, **94** (2): 565～568

20 Zhang L R, Luo L F. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucleic Acids Research, 2003, **31** (21): 6214～6220

21 吕 军, 罗辽复. 人类 polⅡ启动子的识别. 生物化学与生物物理进展, 2005, **32** (12): 1185～1191
Lu J, Luo L F. Prog Biochem Biophys, 2005, **32** (12): 1185～1191

22 Doniger S W, Huh J, Fay J C. Identification of functional transcription factor binding sites using closely related *Saccharomyces* Species. Genome Research, 2005, **15** (5): 701～709

23 Blanchette M, Bataille A R, Chen X, *et al*. Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. Genome Research, 2006, **16** (5): 656～668

24 FitzGerald P C, Shlyakhtenko A, Mir A A, *et al*. Clustering of DNA sequences in human promoters. Genome Research, 2004, **14** (8): 1562～1574

Prediction of Human Transcription Regulatory Motifs by Using Non-alignment Based Method*

LÜ Jun[1,2], LUO Liao-Fu[1)**], ZHANG Ying[1,2], ZHAO Ju-Dong[1,2]

([1)]*Department of Physics, Inner Mongolia University, Hohhot* 010021, China;

[2)] *Department of Physics, Inner Mongolia University of Technology, Hohhot* 010051, China)

Abstract The comparative studies of k-mer distribution in human and mouse TFBS sequences listed in TRANSFAC database are given. A non-alignment based approach for fast genome-wide discovery of transcription regulatory k-mer motifs (TRKMs) is proposed. The method is called distance-based conservative k-mer searching algorithm (DCKS) which is based on the conservation of k-mer pair distance. By use of DCKS the prediction accuracy of human transcription regulatory 7-mer motifs is: sensitivity 90%, specificity 78%, and correlation coefficient 0.65.

Key words transcription regulatory motifs, non-alignment based approach, distance-based conservative k-mer searching algorithm, quadratic discriminant analysis

---

*This work was supported by a grant from The National Natural Science Foundation of China (90403010).

**Corresponding author . Tel: 86-471-4992676, E-mail: lfluo@mail.imu.edu.cn

Received: April 30, 2006 Accepted: June 2, 2006