转录因子结合位点的计算预测方法研究进展

杜耀华,王正志

(国防科技大学 机电工程与自动化学院,中国湖南 长沙 410073)

摘 要 转录因子结合位点的计算预测是研究基因转录调控的重要环节,但现有算法的预测特异性偏低. 在深入分析转录因子结合位点生物特征的基础上,对当前基于保守模体和基于比较基因组学的两类计算预测方法进行了综述,指出了方法各自的优点和不足,并探讨了可能的改进方向.

关键词 转录因子结合位点;计算预测;模体;比较基因组学
中图分类号:Q527 文献标识码:A 文章编号:1007-7847(2006)S0-0024-08

Review on Computational Prediction of Transcription Factor Binding Sites

DU Yao-hua, WANG Zheng-zhi

(*College of Mechatronics Engineering and Automation, National University of Defense Technology, Changsha* 410073, *Hunan, China*)

Abstract: Computational prediction of transcription factor binding sites is an essential task in the research of transcription regulation, but the specificity of recognition results achieved by the current algorithms is quite low. According to the thorough analysis about biologic features of transcription factor binding sites, two classes of predicting methods, based on motif signals and comparative genomics, are reviewed respectively. For each method, besides the merits and defects, the possible directions for improvement are also discussed.

Key words: transcription factor binding site (TFBS); computational prediction; motif; comparative genomics

(*Life Science Research*, 2006, 10(2): 024~031)

揭示基因表达调控的复杂机制是后基因组时代所面临的重大挑战之一. 根据中心法则可知,转录是基因表达的第一步,对转录过程的调控则是表达调控的重要形式. 转录过程的激活、抑制和调节主要通过转录因子(transcription factor, TF)蛋白与其在基因组序列中对应的结合位点(transcription factor binding site, TFBS)之间的交互作用来实现. 当前,对大多数转录因子及其结合位点的认识还很有限. 因此,对转录因子结合位点的计算预测将有助于分析其与转录因子间的相互作用机理,为构建转录调控网络奠定基础,从而推动对基因表达调控的研究.

对经过生物实验验证的已知位点进行分析可知,转录因子结合位点往往以保守短序列片段(亦称作模体(motif))的形式出现. 对于原核基因组,模体的长度一般为 10~30 bp,而对于真核基因组,其长度更短,通常为 5~15 bp. 与其它常见的序列模体信号相比,转录因子结合位点模体除了

收稿日期:2006-03-28;修回日期:2006-05-08
基金项目:国家自然科学基金资助项目(60471003)
作者简介:杜耀华(1978-),男,河北唐山人,博士研究生,从事生物信息学研究,Tel:0731-4574991,E-mail:qsyaohua@nudt.edu.cn;
王正志(1945-),男,上海人,国防科技大学教授,博士生导师,主要从事生物信息学、模式识别与信息处理等方面的研究.

长度较短以外，其碱基组成也更加灵活，容许较多的错配．这些特征造成位点信号的保守性偏弱，特异性不强，很容易与长序列中随机出现的类似信号混淆在一起．另外，转录因子结合位点在基因组中的分布范围比较广，虽然大多数集中位于转录单元或基因上游的启动子区域内，但也有一些分布在转录单元或基因的下游，甚至在内含子或编码区内．即便只考虑启动子区域，对于真核基因组，其范围也常常能达到数千碱基对．上述的种种因素使得转录因子结合位点的计算预测成为一项困难的任务[1,2].

随着基因组序列数据的积累和计算技术的发展，针对转录因子结合位点计算预测的算法和工具也越来越多[3]．几乎所有的方法都以结合位点信号的特异保守性作为预测的出发点[4,5]．根据识别策略和搜索对象的不同，已有的预测方法可大致分成两类：基于保守模体的方法和基于比较基因组学的方法．前者主要在同一物种基因组的协同调控基因（co-regulated gene）调控区域内通过发现或搜索过显现的保守模体（statistically over-represented motif）来预测可能的结合位点．而后者则利用比较基因组学方法，例如系统发生足迹法（phylogenetic footprinting），通过比对多个相关物种基因组的对应区域来发现具有公共保守特性的模体位点．本文将对这两类方法进行简要介绍，指出它们的优点以及存在的问题，再探讨其可能的组合和改进方向．

1　基于保守模体的方法

在某种模式生物基因组内，受同一种转录因子调控的一组基因称为协同调控基因．这些基因调控区域中的过显现模体（在多个区域内出现频率均高于随机水平的短序列片段）很可能具有一定的调控功能，可作为此种转录因子的备选结合位点．基于保守模体方法的主要目标就是在一组协同调控序列区域中识别出符合阈值条件的过显现模体．此类方法的一般步骤是：

1）在输入序列中检测出一组或多组保守的短序列片段（组内的片断具有足够的相似性）作为备选的过显现模体信号；

2）根据所含片断的数目和保守性，选取适当的评价度量，估算每一组模体信号的统计显著性；

3）将各组模体信号按统计显著性排序，显著性最高的一组或多组即为预测结果．

在执行上述步骤时，需要重点解决两个问题：模体信号的描述和模体信号的检测．

1.1　模体的描述模型

如何合理的描述转录因子结合位点的模体信号将对后续的识别过程产生较大影响．理想的描述模型应该在适当的参数规模下，尽最大可能的表征结合位点的生物特征信息．当前最常见的模体描述模型是一致序列（consensus）和比对谱（alignment profile）[6].

1.1.1　一致序列模型

一致序列由结合位点实例片断中每个位置上出现频率最高的碱基组成，直观地反映了模体组成的偏好．每一种结合位点对应一段等长的一致序列，形式简单，便于搜索．通过引入碱基通配符（IUPAC-IUB 编码），一致序列的表达可以更加灵活，每个位置上可同时表示多种优势碱基．但即使如此，一致序列对每个位置上的碱基分布描述依然不够完整，在保守性偏弱的位置丢失了较多信息，只能对模体进行粗略的描述．

1.1.2　比对谱模型

比对谱则通过对结合位点实例片断进行比对来得到模体每个位置上的碱基分布，比一致序列更加灵活．比对谱有多种定义形式，实际中应用最多的则是位置权重矩阵（position weight matrix, PWM）．PWM 是一个 $4 \times l$ 的矩阵，l 为所描述模体的长度．矩阵的每一列给出对应模体位置上 4 种碱基的出现频率．设 $m_{i,j}$ 为 PWM 第 i 行第 j 列的元素，则有：

$$m_{i,j} = \frac{n_{i,j}}{\sum_i n_{i,j}} \tag{1}$$

其中 $n_{i,j}$ 为实例片断中第 j 个位置上碱基 i 的出现次数，$i \in \{A, C, G, T\}$.

由（1）式可知，PWM 对模体内各位置的碱基偏好进行了量化，因此可作为计算实例片断与模体之间相似性的依据．PWM 模型参数较少，便于计算，训练速度快．因此当前许多保守模体预测方法都以 PWM 模型为基础．

1.1.3　改进的 PWM 模型

鉴于转录因子结合位点模体的特点，在较长的基因组序列中单纯利用 PWM 进行预测往往会得到大量假阳性结果，使得真实位点被随机噪声所湮没[3~5]．为了提高模型的特异性，增强信噪比，研究者陆续提出了多种改进 PWM 模型，例如

位置相关 PWM、混合 PWM 等等.

由定义可知，一致序列和 PWM 的计算均基于独立性假设，即认为模体中的各个位置是相互独立的. 但相关研究表明，对于转录因子结合位点，其模体各个位置上的碱基存在着一定的相关性，独立性假设与实际情况并不完全相符[7]. 虽然如此，基于独立性假设的模型在多数情况下仍不失为一种良好的近似[8]. 然而，为了使描述更加准确，就需要摒弃独立性假设，在 PWM 模型中引入位置相关性. 由于 PWM 本质上等同于 0 阶马尔可夫链，所以一种很自然的思路就是将 PWM 扩展为 k 阶（$k>0$）PWM（相当于 k 阶马尔可夫链），用以描述模体内相邻 $k+1$ 个连续位置上碱基的概率转移关系[9,10]. 对于 DNA 序列，k 阶 PWM 的参数数目为 4^{k+1}. 所以随着阶次的升高，估计参数所需的训练数据也将急剧增加. 当前各种调控元件数据库中可利用的转录因子结合位点数据并不多，这就从客观上限制了高阶 PWM 的广泛应用. 应该看到，k 阶 PWM 只能描述模体连续位置上的相关性，而对不相邻位置之间的相关性则无能为力. 在转录因子与结合位点的实际作用过程中，这种远程相关往往是很重要的. 一种改进思路是利用 χ^2 统计量计算模体任意两个位置之间的相关性，再根据此相关性对所有位置进行重新排列，使得新顺序下原有的相关性较强但不相邻的位置均变为连续相邻[11]. 调整位置之后，即可直接利用 k 阶 PWM. 除了上面两种方案，还有一种可直接描述模体内位置所有类型相关性的广义 PWM 模型（GWM）[12].

同一类转录因子的结合偏好在不同的条件下可能会发生分化，使得其结合位点具有多种保守模体. 因此可以考虑为每种模体分别建立 PWM，用一组 PWM 来代替原有的单一 PWM. 这样得到的混合 PWM 模型对于多子类模体具有更好的适应性[13]. 同样，应用混合 PWM 也需要每一子类都具有充足的训练数据. 基于这种‘分而治之’思想的还有核估计方法[14]、家族结合谱方法[15]以及将单一 PWM 和混合 PWM 通过权重系数相组合的非参数方法[16].

相比原始模型，改进的 PWM 更符合生物实际，对结合位点模体的描述也更加准确，但代价是模型复杂度的增加，同时在结构设计以及参数估计方法等方面还需要进一步的优化[17].

1.1.4 PWM 的相似性度量

随着基因组数据的不断积累，许多专用数据库收集了经过实验证实的序列调控元件，其中最具代表性的有 RegulonDB[18]、TRANSFAC[19] 和 JASPAR[20] 等. 调控元件预测模型的训练数据集大多取自这些数据库. 对于各类转录因子结合位点，数据库在列出原始序列数据的同时，还常常提供其对应的 PWM 信息. 这些 PWM 可作为在未知序列中预测结合位点的依据. 由于种种原因，数据库中的某些 PWM 具有一定的相似性. 合理的估算 PWM 相似性的大小，可以将结合偏好相近的 PWM 进行聚类，以减少模型集的冗余，并对模型的比较和精确模型的选取都有重要意义. 常用的 PWM 相似性度量有：皮尔逊 χ^2 距离[21]、得分谱相关系数[21]、平均对数似然比[22]以及列比对似然度[23]等等，其具体的定义和计算方法请参阅原始文献.

除了一致序列和比对谱，字典模型[24]、隐马尔可夫模型（HMM）[25]、贝叶斯网络[26]以及序列特异结合能模型[27]均可用于转录因子结合位点保守模体的描述和预测. 在实际的应用中，往往要在计算准确度和复杂度之间寻求折衷，并根据具体的需求来选择适当的描述模型.

1.2 模体的检测方法

在一组协同调控序列中检测转录因子结合位点模体的方法可大致分为两类：模体发现方法和模体搜索方法. 前者适用于在序列中发现未知的结合位点，主要根据保守序列片断的显现度差异来确定可能的模体. 而后者则对应结合位点模体已知的情况，用于在序列中搜索与已知位点同类的可能位点. 由于结合位点模体比较短，利用这两类方法在长基因组序列中进行初步检测，一般会得到大量的备选位点. 因此，通常还需要一个后处理过程，对这些备选结果进行评价，从中筛选出符合特定准则的最优位点作为真实位点的预测.

1.2.1 模体发现方法

模体发现方法又可细分为基于一致序列的方法和基于比对的方法两类.

1.2.1.1 基于一致序列的方法

此类方法将满足特定条件的序列片断均当作模体可能的一致序列，计算其实例在协同调控序列中的显现度. 方法的主要步骤如下：

1）设待寻模体长度为 l，列出所有满足特定条件的长度为的序列片断，将其作为模体的候选一致序列集合；

2）统计候选集合中每条片断在协同调控序

列中的出现次数（允许少量碱基错配），从中筛选出在全部（或大多数）的协同调控序列中均有出现的片段；

3）选取适当的统计性度量，计算所选片断的显现度，将显现度最高的一致序列作为模体预测的结果．

如果待寻模体长度 l 未知，则需根据先验信息确定其大致取值范围，然后进行逐一尝试．在选取候选序列集合时，最直接的方法是考虑所有长度为 l 的序列片断．这种穷举策略的优点是如果模体存在，就一定能将其找到．但其计算量约为 4^l，当 l 较大时，将变得不可实现．但由于转录因子结合位点模体一般比较短，所以穷举的方法还是得到了一定的应用[28~30]．当前基于片断穷举思想的转录因子结合位点预测工具有 ITB[31]、Weeder[32]、YMF[33, 34]等等．另一种方法是考虑利用启发式策略，例如只将输入协同调控序列中出现过的寻列片断作为候选集．启发式的方法可以降低计算量，但有可能丢失保守模体．利用启发式思想的预测工具有 MobyDick[35]、POBO[36]等等．

1.2.1.2　基于比对的方法

此类方法的基本思想是在每条协同调控序列中选取长度相同的序列片断进行多序列比对，再根据特定准则下的最优比对来获得模体的一致序列或 PWM．对于 k 条长度为 n 的协同调控序列，大约需要进行 n^k 次多序列比对．为了加快计算速度，通常利用启发式策略寻找最优比对．通过比对来发现保守模体的典型方法有 CONSENSUS[37]、MEME[38]、GibbsDNA[39]和 AlignACE[40]．另外，比对的方法还可以与其它模型相结合，例如 ANN-Spec[41]就采用了 Gibbs 采样与人工神经网络组合的方案．

1.2.1.3　模体的统计性度量

转录因子结合位点模体在协同调控序列中是通常都是过显现的．用模体发现方法得到初步预测结果后，可通过计算特定的统计性度量来评价备选模体的显现度．对于一致序列，最常用的度量是 z-score[29]：

$$z(p) = \frac{\mathrm{Obs}(p) - \mathrm{Exp}(p)}{\sqrt{\mathrm{Var}(p)}} \tag{2}$$

其中 $\mathrm{Obs}(p)$ 为一致序列 p 在输入序列中的出现次数，$\mathrm{Exp}(p)$ 和 $\mathrm{Var}(p)$ 分别是 $\mathrm{Obs}(p)$ 的期望和方差．$z(p)$ 越大，一致序列对应的模体显现度越高．

对于 $4 \times l$ 的 PWM，常用的度量有信息量（information content, IC）和极大后验概率（maximum a posteriori, MAP）[42]：

$$\mathrm{IC}(M) = \sum_{i=1}^{4}\sum_{j=1}^{l} m_{i,j}\log\frac{m_{i,j}}{b_i} \tag{3}$$

$$\mathrm{MAP}(M) = \sum_{i=1}^{4}\sum_{j=1}^{l} n_{i,j}\log\frac{m_{i,j}}{b_i} \tag{4}$$

其中 $m_{i,j}$ 为 PWM 第 i 行第 j 列的元素，b_i 为碱基 i 的期望频率，$n_{i,j} = n \times m_{i,j}$，$n$ 为用于统计的片断数目．信息量和后验概率越大，PWM 对应的模体显现度越高．

1.2.2　模体搜索方法

模体搜索方法根据已知的转录因子结合位点序列建立模体得分模型，在基因组序列中对同类位点进行预测，其基本步骤为：

1）根据已知的转录因子结合位点序列数据或 PWM，建立模体得分模型；

2）利用得分模型在目标序列中进行扫描，依次计算各个位置对应序列片断的模体得分；

3）得分超过特定阈值的序列片断即为位点预测结果．

最常用的模体得分模型是根据 PWM 建立的位置特异得分矩阵（position-specific scoring matrix, PSSM）．对于 $4 \times l$ 的 PWM，设 $s_{i,j}$ 为相应 PSSM 第 i 行第 j 列的元素，则有：

$$s_{i,j} = \log\frac{m_{i,j}}{b_i} \tag{5}$$

其中 $m_{i,j}$ 为 PWM 对应位置的元素，b_i 为碱基 i 的期望频率．基于可加性假设，对于长度为 l 的序列片断，其 PSSM 模体得分为各位置碱基对应的 $s_{i,j}(j = 1, \cdots, l)$ 之和．

在计算 PSSM 时，碱基 i 的期望频率 b_i 通常根据背景序列集来获得．背景序列理论上是指不含转录因子结合位点信号的序列．但在实际中，尚没有充足的证据表明哪些序列肯定不含结合位点信号．因此，需要根据先验知识来合理选取背景序列，尽量避开结合位点集中分布的区域[3, 28]．

模体得分阈值的设定也是一个很关键的问题．阈值过高，对真实信号位点的预测正确率将会降低；阈值过低，预测结果中又会充斥大量

的假阳性信号. 因此必须根据实际的要求在准确性和特异性之间寻求折衷. 通常的做法是利用得分模型计算已知位点的得分分布，在保证对已知位点的预测正确率达到一定水平的前提下，选取使假阳性结果最少的阈值. 显然，这种设定方法的主观性比较强，如何尽量客观的选取阈值仍然需要深入的研究[28, 43].

多数情况下，扫描预测的结果还要结合一个后处理过程，以进一步减少假阳性位点的数目. 通常的做法是利用备选位点的统计显著性进行筛选[44]. p-value 是用得最多的一种统计显著性度量[45, 46]. 对于一个得分为 S 的位点片断，其 p-value 表示在背景序列中出现得分不低于 S 的位点片断的概率. p-value 越小，统计显著性越高. 当前，寻找快速有效的 p-value 计算方法仍是研究的热点[47].

以 PSSM 为基础的得分模型可通过自身的优化或与其它模型相结合来改进其模体预测性能. 对于前者，可将信息量 IC 引入，作为 PSSM 各位置的得分权重[9, 48]；也可根据“核”模体中保守性较强的连续区域）的相似性得分对备选片断进行预过滤[48]；还可以把预测结果反馈给已知模体集，通过反复迭代训练来校正模型[49]. 对于后者，最常用的是将一致序列的方法与 PSSM 相结合[31, 50, 51].

2 基于比较基因组学的方法

随着大规模测序技术的应用，越来越多的模式生物基因组被测序完毕，在基因组层次上对序列进行大规模比较分析已经成为可能. 比较基因组学方法可以通过比对来发现不同物种基因组序列中的公共保守区域，而这些在进化过程中显示保守特性的区域应该具有重要的功能. 相关研究已经表明，相关物种基因组的非编码公共保守序列区域富含调控功能元件[52]. 因此，可以通过比对不同物种的同源基因的非编码序列区域来发现公共保守模体，并将其作为可能的转录因子结合位点. 这种比较基因组学方法又称为系统发生足迹法.

比对的方法是系统发生足迹法的一个重要问题. 无论是全局比对还是局部比对，当前都有很多成熟的算法和工具，可以根据实际需求来确定. 至于阈值的选取和结果的统计显著性计算等问题，与基于保守模体的方法十分类似，此处不再详述. 另外，用于比对的物种也需精心选择，进化距离太近或太远都可能导致方法失效. 对于转录因子结合位点预测，当前常用的是多种细菌/古细菌基因组之间[53, 54]、多种酵母基因组之间[55, 56]或人 - 鼠基因组之间[55, 57]的比对.

3 讨论与展望

对于转录因子结合位点的计算预测，基于保守模体的方法适用于模体较短的位点，预测结果的位置精度较高；基于比较基因组学的方法则适用于模体较长的位点，预测结果的位置精度偏低. 因此，可以将两类方法结合起来，首先用比较基因组学的方法确定保守区域的大致范围，再通过保守模体的方法对结合位点进行精确定位. 相关研究证实，组合的方法利用了物种间和物种内的特征信息，预测性能比单一方法有大幅度提高[22, 45, 54, 58, 59].

在实际的转录因子结合位点预测中，现有的方法平均每 500 ~ 5 000 bp 产生一个预测位点. 对于较长的真核基因组序列，预测位点的数目往往比真实位点数目高出几个数量级[2]. 然而，假阳性结果过多并不代表计算预测方法的完全无效，却客观地反映了生物过程的实际情况. 转录因子与靶位点的结合是多种因素综合作用的结果，除了特异的序列保守模体，染色质的空间构象、位点之间的竞争以及外界环境等因素也起着重要作用[60]. 因此，合理量化这些外源特征，实现多种信息的融合，无疑将有助于提高预测的特异性. 对于前者，目前已有利用结构信息（序列的局部构象特征[61]或 DNA- 蛋白复合物中的氨基酸 - 碱基结合偏好[62, 63]）和序列物理化学性质[64, 65]进行转录因子结合位点预测的尝试. 对于后者，有研究者提出了基于贝叶斯网络的信息融合框架，可以不断容纳新的结合位点特征信息[66].

由于转录因子结合位点在基因组中的分布并不均匀，主要集中分布于基因或转录单元的上游，因此可以根据先验信息来合理确定预测结合位点的序列范围，以减少假阳性结果[1]. 对于原核基因组，可取基因翻译起始位点上游一定范围的非编码序列[67]；对真核基因组，则可取基因 5′端非翻译区（UTR）上游一定范围的非编码序列[31, 56].

转录因子与 DNA 序列的结合通常需要多个相关结合位点的协同作用. 这些功能相近或相关的结合位点往往成簇分布，形成所谓的“模块”（module）. 对相似的保守模体进行聚类来预测结

合位点模块，不但更符合生物实际，还能进一步减少假阳性结果[68, 69]. 随着研究的深入，计算预测方法的识别对象将逐渐由单一转录因子结合位点转向结合位点对或结合位点模块，其最终的目标是建立基因组水平下的转录调控网络. 反之，通过分析调控网络通路中的局部保守拓扑结构，在网络的层次上总结功能位点间的约束关系也将有助于结合位点的预测[70].

另外，随着以基因芯片为代表的高通量实验技术和基因本体论（gene ontology, GO）技术的发展，芯片表达数据和 GO 数据也都可以用于转录因子结合位点的预测，以提高预测的特异性[71~73]. 总之，多种特征信息的有机融合和高通量实验数据的合理利用将是计算预测方法的总体发展方向.

参考文献（References）：

[1] PENNACCHIO L, RUBIN E. Genomic strategies to identify mammalian regulatory sequences [J]. Nat Rev Genet, 2001, 2(2): 100-109.

[2] WASSERMAN W, SANDELIN A. Applied bioinformatics for the identification of regulatory elements [J]. Nat Rev Genet, 2004, 5(4): 276-287.

[3] TOMPA M, LI N, BAILEY T, *et al*. Assessing computational tools for the discovery of transcription factor binding sites [J]. Nature Biotechnology, 2005, 23(1): 137-144.

[4] BULYK M. Computational prediction of transcription-factor binding site locations[J]. Genome Biology, 2003, 5(1): 201.

[5] PAVESI G, MAURI G, PESOLE G. *In silico* representation and discovery of transcription factor binding sites [J]. Brief Bioinform, 2004, 5(3): 217-236.

[6] STORMO G. DNA binding sites: representation and discovery [J]. Bioinformatics, 2000, 16(1): 16-23.

[7] BULYK M, JOHNSON P, CHURCH G. Nucleotides of transcription factor binding sites exert interdependent effects on the binding affinities of transcription factors[J]. Nucleic Acids Research, 2002, 30(5): 1255-1261.

[8] BENOS P, BULYK M, STORMO G. Additivity in protein-DNA interactions: how good an approximation is it? [J]. Nucleic Acids Research, 2002, 30(20): 4442-4451.

[9] OSADA R, ZASLAVSKY E, SINGH M. Comparative analysis of methods for representing and searching for transcription factor binding sites[J]. Bioinformatics, 2004, 20(18): 3516-3525.

[10] PONOMARENKO P, PONOMARENKO J, FROLOW A, *et al*. Oligonucleotide frequency matrices addressed to recognizing functional DNA sites[J]. Bioinformatics, 1999, 15(7-8): 631-643.

[11] ELLROTT K, YANG C, SLADEK F, *et al*. Identifying transcription factor binding sites through Markov chain optimization [J]. Bioinformatics, 2002, 18(S2): S100-S109.

[12] ZHOU Q, LIU J S. Modeling within-motif dependence for transcription factor binding site predictions[J]. Bioinformatics, 2004, 20(6): 909-916.

[13] HANNENHALLI S, WANG L S. Enhanced position weight matrices using mixture models [J]. Bioinformatics, 2005, 21(S1): i204-i212.

[14] KEL A, TIKUNOV Y, VOSS N, *et al*. Recognition of multiple patterns in unaligned sets of sequences: comparison of kernel clustering method with other methods[J]. Bioinformatics, 2004, 20(10): 1512-1516.

[15] SANDELIN A, WASSERMAN W. Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics [J]. Journal of Molecular Biology, 2004, 338(2): 207-215.

[16] KING O, ROTH F. A non-parametric model for transcription factor binding sites[J]. Nucleic Acids Research, 2003, 31(19): e116.

[17] HARRISON R, DELISI C. Condition specific transcription factor binding site characterization in *Saccharomyces cerevisiae* [J]. Bioinformatics, 2002, 18(10): 1289-1296.

[18] SALGADO H, GAMA-CASTRO S, PERALTA-GIL M, *et al*. RegulonDB(version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization and growth conditions [J]. Nucleic Acids Research, 2006, 34(Database issue): D394-D397.

[19] MATYS V, KEL-MARGOULIS O, FRICKE E, *et al*. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes[J]. Nucleic Acids Research, 2006, 34(Database issue): D108-D110.

[20] VLIEGHE D, SANDELIN A, DE BLESER P, *et al*. A new generation of JASPAR, the open-access repository for transcription factor binding site profiles[J]. Nucleic Acids Research, 2006, 34(Database issue): D95-D97.

[21] KIELBASA S, GONZE D, HERZEL H. Measuring similarities between transcription factor binding sites[J]. BMC Bioinformatics, 2005, 6: 237.

[22] WANG T, STORMO G. Combining phylogenetic data with co-regulated genes to identify regulatory motifs[J]. Bioinformatics, 2003, 19(18): 2369-2380.

[23] SCHONES D, SUMAZIN P, ZHANG M Q. Similarity of position frequency matrices for transcription factor binding sites [J]. Bioinformatics, 2005, 21(3): 307-313.

[24] SABATTI C, ROHLIN L, LANGE K, *et al*. Vocabulon: a dictionary model approach for reconstruction and localization of transcription factor binding sites[J]. Bioinformatics, 2005, 21(7): 922-931.

[25] MARINESCU V, KOHANE I, RIVA A. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes[J]. BMC Bioinformatics, 2005, 6: 79.

[26] BARASH Y, ELIDAN G, FRIEDMAN N, *et al*. Modeling dependencies in protein-DNA binding sites [C] // VINGRON M, ISTRAIL S, PEVZNER P, *et al*. Proceedings of the Seventh International Conference on Research in Computational Molecular Biology(RECOMB). Berlin, Germany: ACM Press, 2003, 28-37.

[27] DJORDJEVIC M, SENGUPTA A, SHRAIMAN B. A bio-

physical approach to transcription factor binding site discovery [J]. Genome Research, 2003, 13(11): 2381-2390.

[28] VAN HELDEN J, ANDRE B, COLLADO-VIDES J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies[J]. Journal of Molecular Biology, 1998, 281(5): 827-842.

[29] SINHA S, TOMPA M. A statistical method for finding transcription factor binding sites [C] // BOURNE P, GRIBSKOV M, ALTMAN R, *et al*. Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology. Menlo Park, CA, USA: AAAI Press, 2000, 344-354.

[30] CHAN B, KIBLER D. Using hexamers to predict cis-regulatory motifs in Drosophila [J]. BMC Bioinformatics, 2005, 6: 262.

[31] KIELBASA S, KORBEL J, BEULE D, *et al*. Combining frequency and positional information to predict transcription factor binding sites[J]. Bioinformatics, 2001, 17(11): 1019-1026.

[32] 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[J]. Nucleic Acids Research, 2004, 32(Web Server issue): W199-W203.

[33] SINHA S, TOMPA M. Discovery of novel transcription factor binding sites by statistical overrepresentation[J]. Nucleic Acids Research, 2002, 30(24): 5549-5560.

[34] SINHA S, TOMPA M. YMF: a program for discovery of novel transcription factor binding sites by statistical overrepresentation[J]. Nucleic Acids Research, 2003, 31(13): 3586-3588.

[35] BUSSEMAKER H, LI H, SIGGIA E. Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis [J]. Proc Natl Acad Sci USA, 2000, 97(18): 10096-10100.

[36] KANKAINEN M, HOLM L. POBO, transcription factor binding site verification with bootstrapping[J]. Nucleic Acids Research, 2004, 32(Web Server issue): W222-W229.

[37] HERTZ G, STORMO G. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences [J]. Bioinformatics, 1999, 15(7-8): 563-577.

[38] BAILEY T, ELKAN C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization[J]. Machine Learning, 1995, 21(1-2): 51-80.

[39] LAWRENCE C, ALTSCHUL S, BOGUSKI M, *et* al. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment[J]. Science, 1993, 262(5131): 208-214.

[40] HUGHES J, ESTEP P, TAVAZOIE S, *et al*. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae* [J]. Journal of Molecular Biology, 2000, 296(5): 1205-1214.

[41] WORKMAN C, STORMO G. ANN-Spec: a method for discovering transcription factor binding sites with improved specificity [C] // ALTMAN R, DUNKER A, HUNTER L, *et al*. Pac. Symp. Biocomp. Hawaii, USA: World Scientific Press, 2000. 467-478.

[42] FRIBERG M, VON ROHR P, GONNET G. Scoring functions for transcription factor binding site prediction[J]. BMC Bioinformatics, 2005, 6: 84.

[43] ZHENG J, WU J, SUN Z. An approach to identify over-represented cis-elements in related sequences[J]. Nucleic Acids Research, 2003, 31(7): 1995-2005.

[44] BLANCHETTE M, SINHA S. Separating real motifs from their artifacts [J]. Bioinformatics, 2001, 17(S1): S30-S38.

[45] LEVY S, HANNENHALLI S. Identification of transcription factor binding sites in the human genome sequence[J]. Mammalian Genome, 2002, 13(9): 510-514.

[46] FRITH M, FU Y T, YU L Q, *et al*. Detection of functional DNA motifs via statistical over-representation[J]. Nucleic Acids Research, 2004, 32(4): 1372-1381.

[47] BARASH Y, ELIDAN G, KAPLAN T, *et al*. CIS: compound importance sampling method for protein-DNA binding site p-value estimation[J]. Bioinformatics, 2005, 21(5): 596-600.

[48] QUANDT K, FRECH K, KARAS H, *et al*. MatInd and MatInspector-new fast and versatile tools for detection of consensus matches in nucleotide sequence data[J]. Nucleic Acids Research, 1995, 23(23): 4878-4884.

[49] GERSHENZON N, STORMO G, IOSHIKHES I. Computational technique for improvement of the position-weight matrices for the DNA/protein binding sites[J]. Nucleic Acids Research, 2005, 33(7): 2290-2301.

[50] THIEFFRY D, SALGADO H, HUERTA A, *et al*. Prediction of transcriptional regulatory sites in the complete genome sequence of *Escherichia coli* K-12[J]. Bioinformatics, 1998, 14(5): 391-400.

[51] CHEKMENEV D, HAID C, KEL A. P-Match: transcription factor binding site search by combining patterns and weight matrices[J]. Nucleic Acids Research, 2005, 33(Web Server issue): W432-W437.

[52] LEVY S, HANNENHALLI S, WORKMAN C. Enrichment of regulatory signals in conserved non-coding genomic sequence [J]. Bioinformatics, 2001, 17(10): 871-877.

[53] GELFAND M, NOVICHKOV P, NOVICHKOVA E, *et al*. Comparative analysis of regulatory patterns in bacterial genomes [J]. Brief Bioinform, 2000, 1(4): 357-371.

[54] GELFAND M, KOONIN E, MIRONOV A. Prediction of transcription regulatory sites in *Archaea* by a comparative genomic approach[J]. Nucleic Acids Research, 2000, 28(3): 695-705.

[55] LIU Y, LIU X, WEI L, *et al*. Eukaryotic regulatory element conservation analysis and identification using comparative genomics [J]. Genome Research, 2004, 14(3): 451-458.

[56] CLIFTEN P, HILLIER L, FULTON L, *et al*. Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis[J]. Genome Research, 2001, 11(7): 1175-1186.

[57] LOOTS G, OVCHARENKO I, PACHTER L, *et al*. rVista for comparative sequence-based discovery of functional transcription factor binding sites[J]. Genome Research, 2002, 12(5): 832-839.

[58] SINHA S, BLANCHETTE M, TOMPA M. PhyME: A probabilistic algorithm for finding motifs in sets of orthologous sequences [J]. BMC Bioinformatics, 2004, 5: 170.

[59] JOLLY E, CHIN C S, HERSKOWITZ I, *et al*. Genome-wide identification of the regulatory targets of a transcription factor using biochemical characterization and computational genomic analysis [J]. BMC Bioinformatics, 2005, 6: 275.

[60] HALFORD S, MARKO J. How do site-specific DNA-binding

proteins find their targets? [J]. Nucleic Acids Research, 2004, 32(10): 3040-3052.

[61] LIU R X, BLACKWELL T, STATES D. Conformational model for binding site recognition by the *E. coli MetJ* transcription factor [J]. Bioinformatics, 2001, 17(7): 622-633.

[62] KAPLAN T, FRIEDMAN N, MARGALIT H. *Ab initio* prediction of transcription factor targets using structural knowledge [J]. PLoS Computational Biology, 2005, 1(1): 5-13.

[63] EISEN M. All motifs are not created equal: structural properties of transcription factor-dna interactions and the inference of sequence specificity[J]. Genome Biology, 2005, 6(5): P7.

[64] PONOMARENKO J, PONOMARENKO M, FROLOV A, *et al*. Conformational and physicochemical DNA features specific for transcription factor binding sites[J]. Bioinformatics, 1999, 15(7-8): 654-668.

[65] OSHCHEPKOV D, VITYAEV E, GRIGOROVICH D, *et al*. SITECON: a tool for detecting conservative conformational and physicochemical properties in transcription factor binding site alignments and for site recognition[J]. Nucleic Acids Research, 2004, 32(Web Server issue): W208-W212.

[66] PUDIMAT R, SCHUKAT-TALAMAZZINI E, BACKOFEN R. A multiple-feature framework for modeling and predicting transcription factor binding sites[J]. Bioinformatics, 2005, 21(14): 3082-3088.

[67] MCGUIRE A, HUGHES J, CHURCH G. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes [J]. Genome Research, 2000, 10(6): 744-757.

[68] GUHATHAKURTA D, STORMO G. Identifying target sites for cooperatively binding factors[J]. Bioinformatics, 2001, 17(7): 608-621.

[69] HANNENHALLI S, LEVY S. Predicting transcription factor synergism[J]. Nucleic Acids Research, 2002, 30(19): 4278-4284.

[70] PRITSKER M, LIU Y C, BEER M, *et al*. Whole-genome discovery of transcription factor binding sites by network-level conservation [J]. Genome Research, 2004, 14(1): 99-108.

[71] TADESSE M, VANNUCCI M, LIO P. Identification of DNA regulatory motifs using Bayesian variable selection[J]. Bioinformatics, 2004, 20(16): 2553-2561.

[72] CORA D, HERRMANN C, DIETERICH C, *et al*. *Ab initio* identification of putative human transcription factor binding sites by comparative genomics[J]. BMC Bioinformatics, 2005, 6: 110.

[73] CORA D, CUNTO F, PROVERO P, *et al*. Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing over-represented upstream motifs [J]. BMC Bioinformatics, 2004, 5: 57.

(上接第 23 页)

参考文献(References):

[1] SILMAN I, KATCHALSKI E. Water-insoluble derivatives of enzymes, antigens, and antibodies[J]. Annu Rev Biochem 1966, 35: 873-908.

[2] ENGVALL E P. Enzyme-linked immunosorbent assay(ELISA). Quantitative assay of immunoglobulin G[J]. Immunochemistry, 1971, 8: 871-874.

[3] BOUTELL J M, HART D J, GODBER B L, *et al*. Functional protein microarrays for parallel characterisation of p53 mutants [J]. Proteomics, 2004, 4: 1950-1958.

[4] SNAPYAN M，LECOCG M, GUEVEL L, *et al*. Dissecting DNA-protein and protein-protein interactions involved in bacterial transcriptional regulation bya sensitive protein array method combining a near-infrared fluorescence detection[J]. Proteomics, 2003, 3: 647-657.

[5] KERSTENB, POSSLING A, BLAESING F, *et al*. Protein microarray technology and ultravioletcrosslinking combined with mass spectrometry for theanalysis of protein-DNA interactions[J]. Anal Biochem, 2004, 331: 303-313.

[6] HALL D A, ZHU H, ZHU X *et al*. Regulation of gene expression by a metabolic enzyme[J]. Science, 2004, 306: 482-484.

[7] NEWMAN J R, KEATING A E. Comprehensive identification of human bZIP interactions with coiled-coil arrays[J]. Science, 2003, 300: 2097-2101.

[8] RAMACHANDRAN N, HAINSWORTH E, WALTER J. Self-assembling proteinmicroarrays[J]. Science, 2004, 305: 86-90.

[9] NAKAR D, HANDELSMANA T, BAYEREA. Pinpoint mapping of recognitionresidues on the cohesin surface by progressive homologueswapping[J]. Biol Chem, 2004, 279: 42881-42888.

[10] ALCOCER M J, MURTAGH G J, WILSON P B, *et al*. The major human structural IgE epitope of the Brazil nutallergen Ber e 1: a chimaeric and proteinmicroarray approach[J]. Mol Biol, 2004, 343: 759-769.

[11] FANGY, FRUTOS A G, LAHIRI J. Membrane protein microarrays[J]. Am Chem Soc, 2002, 124: 2394-2395.

[12] HUANG J, ZHU H, HAGGARTYS J, *et al*. Finding new components of the target of rapamycin (TOR) signaling network through chemicalgenetics and proteome chips[J]. Proc Natl Acad Sci USA, 2004, 101: 16594-16599.

[13] CHEN G Y, YAO S Q. Developinga strategy for activity-based detection of enzymes in a proteinmicroarray[J]. Chem Bio Chem, 2003, 4: 336-339.

[14] OUYANG Z, TAKATS Z, BLAKE T A. Preparing protein microarrays by soft-landing of mass-selected ions[J]. Science, 2003, 301: 1351-1354.

[15] JUNG G Y, STEPHANOPOULOS G. A functional protein chip for pathway optimization and *in vitro* metabolic engineering [J]. Science, 2004, 304: 428-431.

[16] LEE M Y, PARK C B, DORDICK J S, *et al*. Metabolizing enzymetoxicology assay chip(MetaChip) for high-throughput microscale toxicity analyses[J]. Proc Natl Acad Sci USA, 2005, 102: 983-987.