生物序列模体的混合 Gibbs 抽样识别算法

刘立芳, 霍红卫, 王宝树
(西安电子科技大学计算机学院, 陕西西安 710071)

摘 要: 针对生物序列模体的识别问题, 提出了一个新的混合 Gibbs 抽样识别算法. 算法基于混合模体模型学习, 采用贪心策略, 通过似然度最大化, 逐次将新的模体加入到混合模型中. 算法中设计了位点抽样和模体抽样两种抽样方法, 这两种抽样方法交替进行. 为了加速搜索过程, 对输入数据集采用了基于 kd2trees 的分层划分策略. 实验结果表明, 该算法对序列家族大量模体特征的识别具有显著优势, 并且可建立更具统计特征的模体模型, 从而提高序列分类的准确性.

关键词: 生物信息学; 模体识别; Gibbs 抽样; 混合模体模型
中图分类号: Q81114, TP30116 文献标识码: A 文章编号: 037222112 (2008) 042750206

Multiple Motif Discovery in Biological Sequences by Mixture Gibbs Sampling

LIU Li2fang, HUO Hong2wei, WANG Bao2shu
(School of Computer Science and Technology, Xidian University, Xi. an, Shaanxi 710071, China)

Abstract: For the motif discovery problem of biological sequences, a mixture Gibbs sampling algorithm is presented. Based on mixture motifs model learning through likelihood maximization, a greedy strategy that adds sequentially new motif to a mixture model is employed. Two sampling methods are designed, site sampling and motif sampling, the two sampling methods are applied by turns. In order to speed up the searching procedure, a hierarchical partitioning scheme based on kd2trees is used for partitioning the input dataset. Experimental results indicate that the proposed algorithm is advantageous in identifying larger groups of motifs charac2teristic of biological families. In addition, it offers better diagnostic capabilities by building more powerful statistical motif models with improved classification accuracy.

Key words: bioinformatics; motif discovery; Gibbs sampling; mixture motifs model

1 引言

生物序列模体(motif)识别问题是生物序列分析的重要问题之一. 一组相关序列具有的模体特征与生物分子的功能位点, 三维结构特征等功能特征相关, 并且可根据序列模体特征判断一条未知序列是否属于某序列家族. 针对模体识别问题, 已有很多行之有效的方法, 例如: 基于贪心算法的 CONSENSUS[1]、基于 Gibbs 抽样的 Gibbs Motif Sampler[2~5], 基于期望最大化 EM(Expectation Maximization)算法的 MEME[6], 基于贪心混合学习算法的 Greedy EM[7] 等等. 描述序列模体通常采用确定性(如正则文法) 和概率模型. PROSITE 数据库使用正则文法来描述蛋白质家族模体. 一种简单的概率模型是位置权重矩阵 PWM(Position Weight Matrix), 用模体中每一位置上每一残基出现的概率来描述模体特征. 上述几种方法均采用 PWM 描述模体, 其中 MEME 基于两分量混合模体模型, Greedy EM 基于多分量混合模体模型, 文献[7]对这两种方法的性能做了比较. 根据文献[8]对高斯混合学习的描述, 本文基于多分量混合模体模型, 提出一个新的混合 Gibbs 抽样算法, 算法借鉴 Greedy EM 的思想, 采用贪心策略, 逐次将新的模体加入到混合模型中, 在模型训练过程中, 采用混合 Gibbs 抽样算法, 首先通过位点抽样确定新模体, 然后再通过模体抽样进行全局搜索, 与基于 EM 算法的 Greedy EM 相比性能有很大程度的提高. 实验结果表明, 该算法对序列家族大量模体特征的识别具有显著优势, 并且可建立更具统计特征的模体模型, 从而提高序列分类的准确性.

2 问题描述

给定一个有限字母表 2= {A1, A2, , , A8}, (8= | 2|), 则任一条序列 s= a1a2, aL(L \ 1, ai I 2)可称为长 L 的字符串. 设模体长度为 W, xi= ai1ai2, aiW(aik I

收稿日期: 2005208216; 修回日期: 2007210227
基金项目: 国家自然科学基金(No. 60705004); 陕西省自然科学基金(No. 2005F33)

2, k= 1, , , W) 为序列 s 的长度为 W 的子串. N 条序列 $S=\{s_1, s_2, , , s_N\}$, 其长度分别为 $L_1, L_2, , , L_N$, 则这 N 条序列长度为 W 的子串共有 n 条, $n=\sum_{s=1}^{N} m_s$, $m_s= L_s- W+ 1$. 通常 N 条序列内嵌不同模体具有不同长度, 本文假设 N 条序列内嵌不同模体的长度均为 W, 则这 n 条子串形成训练集 $X=\{x_1, , , x_n\}$. 若采用 PWM 来描述序列模体, 则 PWM 为一 8@W 的概率矩阵, 记为 H. 任一条长为 W 的子串 x_i 在 g 分量混合模体模型 f 下可表示为[7, 8]:

$$f(x_i;\ \theta_g)=\sum_{j=1}^{g} P_j\varphi_j(x_i;\ H_j) \tag{1}$$

其中 θ_g 是混合模型中所有的未知参数 $[P_1, , , P_{g-1}, H_1, , , H_g]$. 混合系数 $P_j(P_j\geq 0, \sum P_j= 1, , , g)$ 可认为是子串 x_i 由第 j 分量产生的先验概率, 有 $\sum_{j=1}^{g} P_j= 1$. H_1 为序列的背景概率向量(长为 8), 设 $H_1=[q_l]$, l= 1, , , 8, 其中 q_l 表示字母 A_l 在任意位置出现的概率. H_2~ H_g 为序列模体的 PWM, 设 $H_j=[p^j_{l,k}]$, l= 1, 2, , , 8, k= 1, 2, , W, 其中 $p^j_{l,k}$ 表示字母 A_l 在第 j 模体的第 k 位置出现的概率. $\varphi_j(x_i; H_j)$ 为子串 x_i 由第 j 分量产生的概率, 计算式如下[7, 8]:

$$\varphi_j(x_i; H_j)=\begin{cases}\prod_{k=1}^{W} p^j_{a_k,k}, & j \text{ is motif}\\ \prod_{k=1}^{W} q^j_{a_k}, & \text{otherwise}\end{cases} \tag{2}$$

在该混合模型下, 训练集 X 的对数似然度为[7, 8]:

$$L(\theta_g)=\sum_{i=1}^{n}\log f(x_i;\ \theta_g) \tag{3}$$

根据任一条序列的内嵌不同模体不重叠的特性, 即任一子串 x_i 只能由 g 个分量其中之一产生, 引入变量 z_{ij}, 当子串 x_i 由第 g 分量产生时, $z_{ij}= 1$, 当第 g 分量不产生子串 x_i 时, $z_{ij}= 0$, 则完备数据对数似然度为[7, 8]:

$$L^C(\theta_g)=\sum_{i=1}^{n}\sum_{j=1}^{g} z_{ij}(\log P_j+\log\varphi_j(x_i; H_j)) \tag{4}$$

将 z_{ij} 看作缺失数据, 针对该极大似然估计问题, 文献[7]提出了一个基于 EM 算法的求解方法 Greedy EM. EM 算法在计算过程中, 将所有的可能性进行相加, 其时间复杂度高. 另外 EM 算法为确定性算法, 易陷入局部最小. 本文提出了一个新的混合 Gibbs 抽样算法求解该问题. 在计算过程中, 不需将所有的可能性相加, 其时间复杂度降低, 并且随机的抽样使其不易陷入局部最小, 较之基于 EM 算法的 Greedy EM, 性能有很大程度的提高.

3　算法描述

311　确立初始参数候选集(candidate selection)

训练集 X 中每一条长 W 的子串 x_i(i= 1, , , n) 都与一个矩阵 $H_s=[p^S_{l,k}]$(S= 1, , , n) 对应,

$$p^S_{l,k}=\begin{cases}K, & \text{如果 } a_k= A_l\\ (1-K)/(8-1), & \text{其他}\end{cases} \tag{5}$$

其中 K 是(0, 1)之间的常数, 满足 $K\geq 1/8$.

若将这 n 个矩阵作为新分量参数初始化的候选集, 将导致搜索空间过大. 一种解决的办法是对 X 进行分层划分. 本文采用了文献[7]的基于 kd2trees 的分层划分策略, 将 n 条子串划分成 c(c n n)组, 每组按其调和序列生成矩阵 H_s(S= 1, , , c), 则参数初始化候选集的大小从 n 降到 c. 基于 kd2trees 的分层划分策略描述如下:

根结点包含所有 n 条子串, 计算 n 条子串每一位置 k(1~ W)上每一字母出现的频率, 则生成一 8@W 的矩阵 $F=[f_{l,k}]$, l= 1, , , 8, k= 1, , , W. 按式(6)计算每一位置的熵值 H_k:

$$H_k=-\sum_{\substack{A_l\\ f_{l,k}>0}} f_{l,k}\log f_{l,k} \tag{6}$$

然后选择位置 $q=\arg\max_{k=1,,,W}\{H_k\}$, 将所有子串中处于位置 q 的字母按其出现的频率 $f_{l,q}$ 从小到大排列并对其进行奇偶标注, 按照子串 q 位置字母的奇偶标注, 将根结点的 n 条子串划分成两个子集, 分别作为根结点的左孩子和右孩子. 分别对左孩子结点和右孩子结点进行上述操作, 直到所有的叶结点所包含的子串数目小于等于一个常数 L_n(如: $L_n= N$ 或 $L_n= N/2$).

312　位点抽样(site sampling)

位点抽样在原混合模型的基础上加入一个新分量. 设原混合模型为 $f(x_i;\ \theta_g)$, 现加入新分量 $\varphi_{g+1}(x_i; H_{g+1})$, 其中 H_{g+1} 为新模体的 PWM, 得到如下两分量混合模型, 如式(7)[7, 8].

$$f(x_i;\ \theta_{g+1})=(1-a)f(x_i;\ \theta_g)+a\varphi_{g+1}(x_i; H_{g+1}) \tag{7}$$

其中 a I (0, 1). X 在此模型下的对数似然度为式(8)[7, 8].

$$L(\theta_{g+1})=\sum_{i=1}^{n}\log f(x_i,\ \theta_{g+1}) \tag{8}$$

保持 θ_g 不变, 即 $f(x_i;\ \theta_g)$ 不变, 则待求解的问题为: 优化参数 H_{g+1} 和 a, 使得 $L(\theta_{g+1})$ 最大化.

位点抽样包括 4 步, 描述如下:

Step1　H_{g+1} 和 a 参数初始化. 位点抽样首先对 H_{g+1} 和 a 进行参数初始化. H_{g+1} 的值取自初始参数候选集 H_s(S= 1, , , c), 需对候选集空间进行全局搜索. 将式(8)在 $a= a_0$ 处进行二次 Taylor 展开, 二次 Taylor 展开式具有极大值, 如式(9)[7, 8].

$$\hat{L}(\theta_{g+1}|a_0)=L(\theta_{g+1}|a_0)-\frac{[\hat{L}(\theta_{g+1}|a_0)]^2}{2\ddot{L}(\theta_{g+1}|a_0)} \tag{9}$$

令 $a_0= 015$, $D(x_i, H_s)=\dfrac{f(x_i;\ \theta_g)-\varphi_{g+1}(x_i; H_s)}{f(x_i;\ \theta_g)+\varphi_{g+1}(x_i; H_s)}$

时[7,8],

$$\hat{L}(H_g) = \sum_{i=1}^{n} \log \frac{f(x_i;\ 7_g) + <_{g+1}(x_i;\ H_g)}{2} + \frac{1}{2} \frac{\left[\sum_{i=1}^{n} D(x_i,\ H_g)\right]^2}{\sum_{i=1}^{n} D^2(x_i,\ H_g)} \tag{10}$$

此时, $\hat{a} = \frac{1}{2} - \frac{1}{2} \frac{\sum_{i=1}^{n} D(x_i,\ H_g)}{\sum_{i=1}^{n} D^2(x_i,\ H_g)}$, 则 H_{g+1}和 a 初始化为式(11)、(12).

$$H_{g+1} = \arg\max_{H_g} \hat{L}(H_g) \tag{11}$$

$$a = \frac{1}{2} - \frac{1}{2} \frac{\sum_{i=1}^{n} D(x_i,\ H_{g+1})}{\sum_{i=1}^{n} D^2(x_i,\ H_{g+1})} \tag{12}$$

若 $a \mid (0, 1)$, 则当 g= 1 时, a= 015, 当 g \ 2 时, a= 1/(g+ 1)[7].

Step2 修改候选集. 设候选集是否为空标志 C_Set. 确定了 H_{g+1}和 a 的初值后, 删去候选集中与 H_{g+1}相应的项, 若候选集为空, C_Set= 0, 否则 C_Set= 1.

Step3 抽样. 每一条序列 s_i 其任一条长为 W 的子串x 均可能为模体, 根据 H_{g+1}和 H_1 的值, 给子串 x 赋权值 $A_x = <_{g+1}(x;\ H_{g+1}^{(t-1)}) / <_1(x;\ H_1)$, 其中 t 为迭代次数. 子串 x 以概率 $p = A_x / \sum_{j=1}^{L_i - W + 1} A_j$ 被选择, 并且保证不同的模体不重叠. 每条序列只任选一条子串作为模体, 则N 条序列找到N 条子串, 与之相应的变量 $Z_{x, g+1} = 1$.

Step4 参数更新. 根据新确定的模体, 含 N 条子串, 按式(13)、(14)重新计算 $H_{g+1}^{(t)}$和 $a^{(t)}$.

$$H_{g+1}^{(t)} = [p_{l,k}^{g+1}], \text{其中 } p_{l,k}^{g+1} = \frac{c_{l,k} + b_l}{N + B} \tag{13}$$

$$a^{(t)} = \frac{1}{n} \sum_{i=1}^{n} z_{i, g+1} \tag{14}$$

其中 $c_{l,k}$为新模体第 k 列上字母 A_l 的数目. 根据贝叶斯分析方法, 进行参数估计要使用/先验概率0, 本文采用了最简单的一种方法, /伪计数0法, $b_l = 0101 @ Qq_l$, $B = \sum_{l=1}^{8} b_l$, 其中 $Q = \sum_{s=1}^{N} L_s$, 为 N 条序列的字母总数, q_l 为字母 A_l 在N 条序列中出现的频率.

Step3 和 Step4 重复进行, 直到$| L^{(t)}(7_{g+1}) / L^{(t-1)}(7_{g+1}) - 1| < 10^{-6}$或达到最大迭代数 $t > T_{max}$.

313 模体抽样(motif sampling)

位点抽样在原混合模型参数不变的基础上, 初步确定新分量的参数, 并假设每一条序列仅包含一条新模体子串, 而事实上, 由于新分量的加入, 需对所有分量的参数进行调整, 并且一条序列可能包含多条或不包含模体子串. 进行位点抽样仅得到局部最优解, 模体抽样重新调整混合模型中的所有参数, 对 X 的每一子串 x_i 进行抽样, 从而得到全局最优解. 模体抽样描述如下:

Step1 抽样. X 的每一子串 x_i, 按式(15)计算其在第 j 分量下的权值 P_j, 则 j 分量在子串 x_i 上依概率 P_j 被选择.

$$P_j = \frac{P_j^{(t-1)} <_j(x_i;\ H_j^{(t-1)})}{(1 - P_j^{(t-1)}) <_1(x_i;\ H_1^{(t-1)}) + P_j^{(t-1)} <_j(x_i;\ H_j^{(t-1)})} \tag{15}$$

若 x_i 上 j 分量被选择, 则 $z_{ij}^{(t)} = 1$. 为了保证模体的不重叠性, 则下一要进行模体抽样的子串为 x_{i+W}(x_i与 x_{i+W}为同一序列的子串)或 x_{i+1}(x_i与 x_{i+1}为不同序列的子串).

Step2 参数更新. 根据抽样后的结果, 按式(16)、(17)重新计算混合模型的参数 7_g:

$$H_j^{(t)} = \begin{cases} \hat{p}_{l,k}^{j} = \dfrac{\hat{c}_{l,k}^{j} + b_l}{\sum_{l=1}^{8} \hat{c}_{l,k}^{j} + B}, & j \text{ is motif} \\ \hat{q}_l^{j} = \dfrac{\hat{c}_l^{j} + b_l}{\sum_{l=1}^{8} \hat{c}_l^{j} + B}, & \text{otherwise} \end{cases} \tag{16}$$

$$P_j^{(t)} = \frac{1}{n} \sum_{i=1}^{n} z_{ij}^{(t)} \tag{17}$$

其中 $\hat{c}_{l,k}^{j}$为模体j 第k 位置上字母 A_l 的数目, $\hat{c}_l^{j}$ 为处于背景下的字母 A_l 的数目.

Step1 和 Step2 重复进行, 直到$| L^{(t)}(7_g) / L^{(t-1)}(7_g) - 1| < 10^{-6}$或达到最大迭代数 $t > T_{max}$.

314 混合 Gibbs 抽样算法))) MSAM

混合 Gibbs 抽样算法包括确立初始参数候选集、位点抽样、模体抽样三部分, 描述如下:

```
Procedure MSAM
Begin
  G is the maximum number of motifs, g is the current number of motifs,
  C_Set is the symbol that the candidate set for initializing new model pa2
  rameters is not null.
  Initialization.
      11Candidate selection for initializing new model parameters,
      as described in section 3.1.
          21Initialize the model using one component ( g= 1) that
          represents the background with parameters H1 and P1= 1;
          31C_Set= 1;
  While((g< G+ 1) and (C_Set! = 0))
    11Perform site sampling as described in section 3.2;
      if (L(7g+1)> L(7g))
        Accept the new mixture model with g+ 1 components;
      else
        C_Set= 0;
    2. if(C_Set! = 0)
        Perform motif sampling as described in section 3.3;
  End While
End
```

图1 混合 Gibbs 抽样算法))) MSAM

3.5 算法复杂性分析

存储 N 条序列、n 条子串的中间计算结果和所识别模体的 8 @W 的概率矩阵是算法所需的主要存储空间, 因此算法空间复杂度为 O(NL), 其中 L 为N 条序列的平均长度.

算法由三部分组成, 分别对这三部分的时间复杂度加以分析. 第一部分确立初始参数候选集, 即构建 kd-tree 的过程. 构建 kd-tree 其主要运算为生成 8 @W 的矩阵 $F=[f_{l,k}]$, $l=1,\ ,\ ,\ 8$, $k=1,\ ,\ ,\ W$. 在树的每一层, 约进行 n @W 次加法运算, 而 kd-tree 的高度小于 logn, 因此确立初始参数候选集其时间复杂度为 O(nWlogn). 第二部分位点抽样, 包括参数初始化和抽样过程. 每计算一次 $\hat{L}(H_s)$ 需进行 O(n) 次运算, 引入变量 c_{cur}, 表示当前参数候选集大小, 则进行参数初始化其时间复杂度为 $O(nc_{cur})$. 抽样过程主要是计算 n 条子串的权值, 每进行一次位点抽样约进行 n @W 次乘法运算, 设迭代次数为 T_s, 则抽样过程其时间复杂度为 $O(T_snW)$. 第三部分模体抽样. 设 g_{cur} 为当前模体数, 其主要计算为计算 n 条子串在当前 g_{cur} 个分量下的权值, 约进行 2 @n @W @ g_{cur} 次乘法运算, 设迭代次数为 T_m, 则模体抽样时间复杂度为 $O(T_mnWg_{cur})$. 设算法识别模体的个数为 g_{max}, c 为 $\hat{L}(H_s)$ 的平均计算次数, T_s 为识别一个模体进行位点抽样的平均迭代次数, T_m 为识别一个模体进行模体抽样的平均迭代次数, 则算法总的时间复杂度为 $O(nWlogn+ncg_{max}+T_snWg_{max}+T_mnWg_{max}^2)$.

4 实验结果

为了验证算法的性能, 进行了两方面的实验. 选用真实蛋白质家族序列作为实验数据, 与文献[7]的选择相同. 首先从 PRINTS 数据库中选择了 6 组蛋白质家族序列, 验证算法能否有效识别蛋白质家族的指纹(fingerprints). 其次从 PROSITE 数据库选择了 4 组蛋白质家族序列, 根据算法识别的序列模体, 使用 MAST[9] 对 SWISS-PROT 数据库进行序列相似性搜索, 验证算法所识别模体的准确性. MAST 为一序列相似性搜索工具, 以序列模体作为输入量. MSAM 的实验结果与 Greedy EM 的实验结果进行了比较.

4.1.1 识别 PRINTS 蛋白质家族的指纹

选择的六组 PRINTS 蛋白质家族如表 1 所示. 对每一组序列, 设置不同的模体长度, 在每一模体长度下, 应用 MSAM 算法, 直到至多 15 个不同的模体被识别为止. 通过计算模体的信息量 IC(Information Content)[1,7] 来评价所识别模体的质量, 计算式如下:

$$IC_j=\sum_{k=1}^{W}\sum_{A\in\{A,C,G,T\}}p_{l,k}^{j}\log_2\frac{p_{l,k}^{j}}{f_{A_i}} \qquad (18)$$

其中 f_{A_i} 为N 条序列中字母 A_i 出现的频率. 实验结果如图 2 所示.

图 2 中柱状图顶端的数据为所识别模体的平均信息量. 实验结果表明, 在不同模体长度下, MSAM 所识别的模体数量多, 且与 Greedy EM 所识别模体的平均信息量相当, 即 MSAM 能够识别"又多又好"的序列模体, 可见 MSAM 对序列家族大量模体特征的识别具有明显优势.

表 1　6 组 PRINTS 蛋白质家族

PRINTS family	PR00058	PR00061	PR00810	PR01266	PR01267	PR01268
Number of sequences (average length)	16 (297)	24 (120)	6 (286)	24 (222)	22 (218)	19 (209)

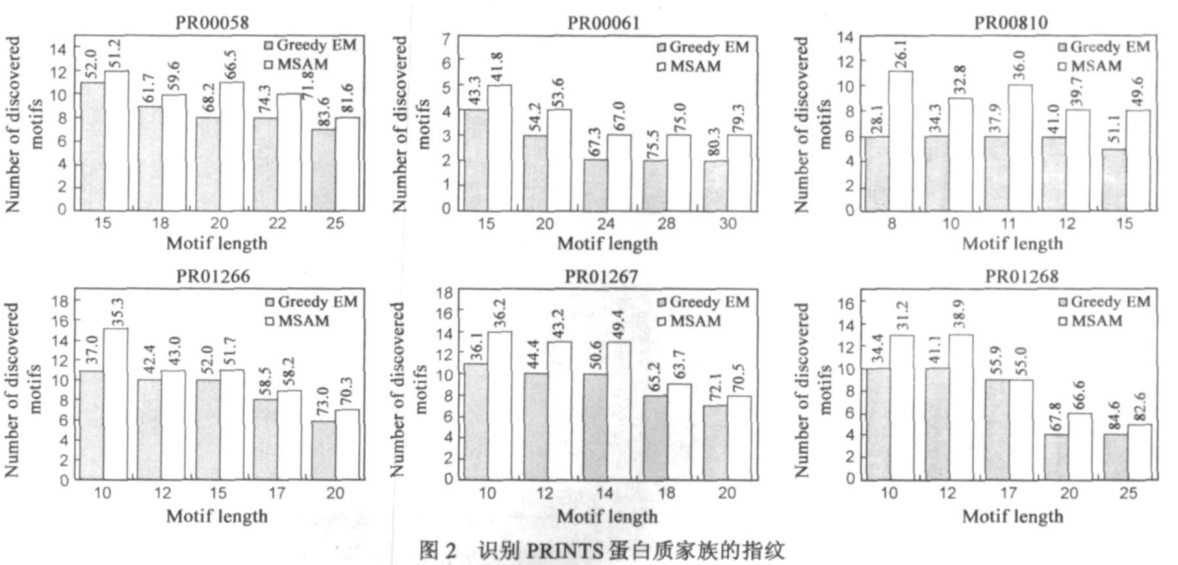

图 2　识别 PRINTS 蛋白质家族的指纹

4.1.2 蛋白质家族分类准确性验证

选择的四组 PROSITE 蛋白质家族序列和 SWISS-PROT 数据库如表 2 所示. 对每一组序列, 设置模体长度 W= 10, 最大模体数目 G= 10, 分别应用 MSAM 和 Greedy EM 算法, 将所识别模体作为 MAST 的输入, 对 SWISS-PROT 数据库进行搜索. MAST 通过计算 SWISS-PROT 数

据库中每一条序列的 E2value[9], 将 E2value 低于某一阈值的序列输出. 对 MAST 的输出, 计算其/ 真阳性 TP0 (True Positive) 序列的数目和/ 假阳性 FP0 (False Positive) 序列的数目, 画出其 ROC_{50} 曲线图[10], 如图 3 所示. ROC_{50}曲线以/ 真阳性0作为/ 假阳性0的函数, 直到发现 50 条/ 假阳性0 序列为止, 以此来评价 MSAM 和 Greedy EM 两种算法的灵敏度(sensitivity) 和特异度(specificity) , 验证算法所识别模体的准确性.

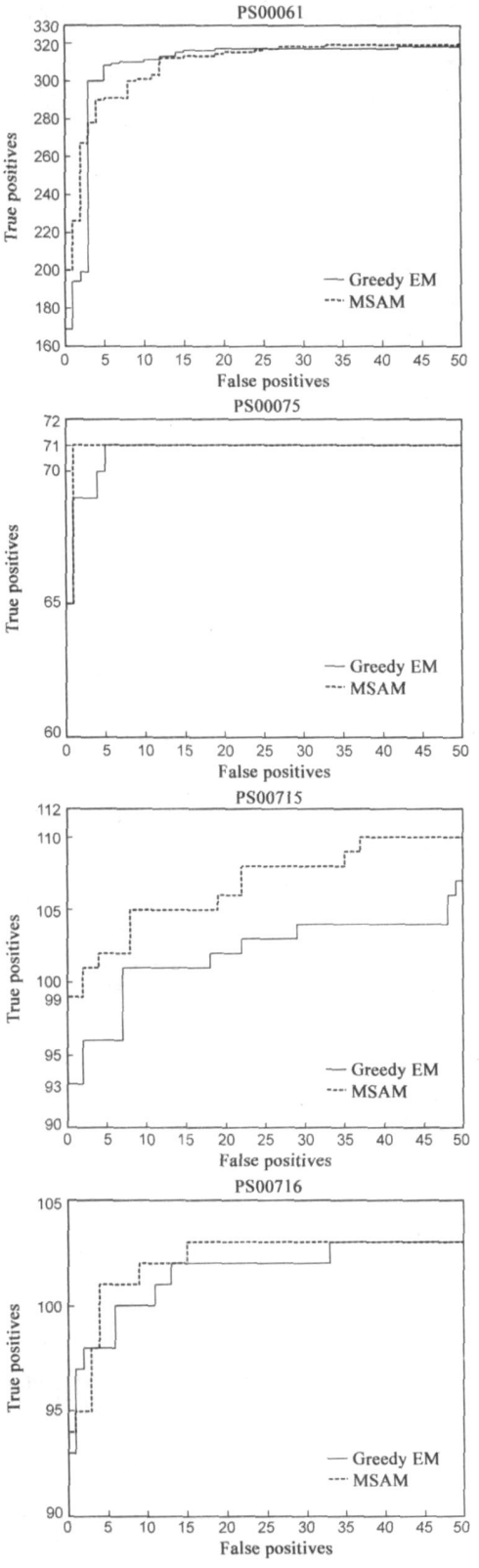

图 3 ROC_{50}曲线图

图 3 的结果表明, MSAM2MAST 的曲线高于 Greedy EM2MAST 的曲线, 特别是当序列 E2value 较小时, 说明 MSAM 算法所识别模体的准确性高于 Greedy EM.

表 2 4组 PROSITE 蛋白质家族和 SWISS PROT 数据库

PROSITE family	PS00061	PS00075	PS00715	PS00716
Number of positive data	346	74	125	120
Size of training set (average length of seqs)	30(308)	7(243)	10(362)	10(394)
SWISS PROT(test set)	173105 sequences			

5 结束语

本文在分析生物序列模体识别问题及已有算法的基础上, 提出了一个新的混合 Gibbs 抽样识别算法 MSAM. 与基于 EM 算法的识别算法相比, MSAM 通过位点抽样和模体抽样能够有效避免陷入局部最小. 通过两方面的实验, 验证了算法的性能. MSAM 对序列家族大量模体特征的识别具有显著优势, 并且所识别模体的准确性较高.

本文研究基于高斯混合模型并假定不同模体具有相同的长度, 进一步的工作将采用其它模型研究具有不同长度的序列模体识别问题, 并提出有效的识别方法.

参考文献:

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

[2] Lawrence C E, Altschul S F, Bogouski M S, Liu J S, Neuwald A F, Wooten J C. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment[J]. Science, 1993, 262(5131): 208- 214.

[3] Neuwald A F, Liu J S, Lawrence C E. Gibbs motif sampling: detection of bacterial outer membrane repeats[J]. Protein Science, 1995, 4(8): 1618- 1632.

[4] Liu J S, Neuwald A F, Lawrence C E. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies [J]. Journal of the American Statistical Association, 1995, 90(432): 1156- 1170.

[5] W Thompson, E C Rouchka, C E Lawrence. Gibbs recursive sampler: finding transcription factor binding sites[J]. Nucleic

Acids Research, 2003, 31(13): 3580- 3585.

[6] Timothy L Bailey, Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers [A]. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology [C]. Menlo Park, California: AAAI Press, 1994. 28- 36

[7] K Blekas, D Fotiadis, A Likas. Greedy mixture learning for multiple motif discovery in biological sequences[J]. Bioinformatics, 2003, 19(5): 607- 617.

[8] Vlassis N, Likas A. A greedy EM algorithm for Gaussian mixture learning[J]. Neural Processing Letters, 2002, 15(1): 77- 87.

[9] Timothy L Bailey, Michael Gribskov. Combining evidence using p-values: application to sequence homology searches[J]. Bioinformatics, 1998, 14(1): 48- 54.

[10] Gribskov M, Robinson N L. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching[J]. Computational Chemistry, 1996, 20(1): 25- 33.

作者简介:

刘立芳　女, 1972 年出生于甘肃兰州, 博士, 副教授, 2006 年毕业于西安电子科技大学. 主要研究方向为计算智能、智能信息处理、生物信息学. E-mail: lfliu@mail.xidian.edu.cn

霍红卫　女, 1963 年出生于陕西浦城, 博士, 教授. 主要研究方向为算法设计与分析、分布与并行计算、生物信息学算法.
E-mail: hwhuo@mail.xidian.edu.cn

王宝树　男, 1941 年出生于陕西西安, 教授, 博士生导师. 主要研究方向为: 计算机智能信息处理与控制、多传感器信息融合技术及应用. E-mail: bshwang@xidian.edu.cn