基于隐马尔可夫模型的多重序列分析*

罗泽举[1]，朱思铭[1]，何 淼[2]
（1. 中山大学数学与计算机科学学院，广东 广州 510275；
2. 中山大学生命科学学院，广东 广州 510275）

摘 要：隐马尔可夫模型是最近几年在许多机器学习领域都得到成功应用的关于序列分析的重要统计模型，特别是在蛋白质家族的识别方面。这主要是由于生物数据的急剧增长导致 2 个领域（计算科学和生物学）走向结合引起的。探讨了多重序列比对和序列谱隐马尔可夫模型，讨论了隐马尔可夫模型的基本算法以及如何建立 HMMs。根据 E 值和训练分数进行蛋白质家族的识别和分类。

关键词：隐马尔可夫模型；多重序列比对；蛋白质家族的分类
中图分类号：O211.62 文献标识码：A 文章编号：0529-6579（2005）02-0009-05

1 研究背景

隐马尔可夫模型是最近几年在许多机器学习领域都得到成功应用的关于序列分析的重要统计模型，特别是在蛋白质家族的识别方面。多重序列分析是寻找整个基因家族的共同特征的重要方法，通过多重序列比对，发现它们特征序列，某个家族成员进化可以看成是由特征序列经过若干年代一系列的插入、替代、删除的结果。多重序列分析是一个非常困难的问题，涉及许多模型的选择，Carillo 和 Lipman 引入了基于两两最优化比对分数和的多重序列比对方法，并得到了广泛应用；但是这种方法对于计算时间和空间的耗费极大，被证明是 NP 难题[1]。许多研究者利用启发式和近似算法改进了比对分数算法[2]，包括 Feng 和 Doolittle 的 Clustal[3]，Thompson 等[4] 著名的多重序列数据库分析的 CLUSTALW 和新近的基于隐马尔可夫的算法。

隐马尔可夫理论最初是由 Baum 及他的同事[5] 于 20 世纪 60 年代末 70 年代初提出，并最早用于语音识别[6—8]。隐马尔可夫模型最早用于计算生物学是于 80 年代末 90 年代初，目前已经用于 DNA 模型构建，蛋白质二级结构预测，基因预测，横跨膜蛋白识别，其中最为普遍的应用是 Krogh 等[9] 基于 profile 家族共同特征提取的蛋白质序列分析。

隐马尔可夫之所以在生物序列分析中得到普遍应用是因为它正好模拟了生物基因的突变、插入、缺失、匹配过程。

2 多重序列比对

2.1 多重序列比对的描述

一个多重序列比对可以看成是三元组 $\Omega=(\Sigma, S, A)$，其中 Σ 是字母表的集合，若对 DNA 或 RNA，$\Sigma=\{A, T, G, C, —\}$ 或 $\Sigma=\{A, U, G, C, —\}$（其中“—”表示空位或删除态）；若是针对蛋白质，Σ 是 20 种氨基酸字母和“—”的集合，即 $\Sigma=\{G, A, L, M, F, W, K, S, N, D, P, V, I, C, Y, H, R, T, Q, E, —\}$；$S=\{S_1, S_2, \cdots, S_k\}$ 是比对序列的集合，其中 S_i（$i=1, 2, \cdots, k$）是以集合的形式代表一条序列，例如 $S_1=\{A, A, G, G, C, T, T, A\}$，代表序列 AAGGCTTA，比对时，一般取每条序列长度相等，但也可以不等；$\boldsymbol{A}=(a_{ij})$ 是一个比对矩阵，其元素是 Σ 中的元素；如图 1 是有 3 个序列的比对，图中每条序列的长度相等。

S_1：Y E G V A — — T
S_2：Y E G — A T — A
S_3：F E G — C — V A

图 1 一个有 3 条序列的多重序列比对
Fig. 1 A multiple alignment of three strings

由于基于比对和分数的多重序列计算是 NP 难题，用线性罚分的优化比对和分数计算方法，对 k 个序列，每个序列的长度长为 n，则计算时间和空间耗费将分别是 $O(2^k \cdot n^k)$ 和 $O(n^k)$，若 k 和 n 较

* 收稿日期：2004—06—01
基金项目：国家自然科学基金资助项目（10371135）
作者简介：罗泽举（1965 年生），男，博士生；通讯联系人：朱思铭；E-mail：stszsm@zsu.edu.cn

大，将超出计算机容量。因此必须改进比对的计算方法。

算法的改进要考虑到 2 个问题：①采用什么标准和用什么样的计分函数来计算多重序列比对？②如何计算其最优化分数？Feng 和 Doolittle 的 Clustal，Thompson 等利用启发式和近似算法改进了比对分数算法，著名多重序列数据库分析工具 ClustalW 也是这类方法的典型代表；另一个重要的问题是一个多重序列比对首先考虑的是一个家族的进化关系，但上述算法却忽略了这个重要事实，故若能将进行多重序列比对的各序列具有进化上的相关关系引入比对分数计算，是不是可以大大改进计算时间和空间的耗费呢？隐马尔可夫方法正是利用了这个思想，它利用特征序列（或叫一致序列）的概念，将多重序列比对建立在进化关系这一思想下，使算法得到大大改进，计算时间和空间都大为减少，且算法收敛速度快。

2.2 特征序列

一个多重序列的特征序列是最能描绘这个多重序列的共同本质的序列，虽然目前还没有关于特征序列的统一定义，但可以用子序列（Subsequence）方法，从多重序列比对中找出每列元素中出现字符最多的元素来定义，例如图 1 的 S_1，S_2，S_3 的特征序列是 YEGAA。定义特征序列的意义至少有 3 点：①可以对一个序列进行数据库搜索，以寻找它的所在家族；②可以比较不同家族的进化关系；③它是构建隐马尔可夫模型等的理论基础。

3 隐马尔可夫模型

3.1 隐马尔可夫模型的生物学背景

现代某类生物的基因组可以认为是某个祖先基因组经过若干代的进化而来的，和我们研究基因组序列的相似性是为了讨论生物的同源性一样，生物的特征序列正代表着某个祖先基因，这个祖先基因经过插入、删除和匹配而不断进化，最终衍变为一个基因家族，这个过程正好可以用如下模型来描述。

模型用 3 个状态来描述，分别称为删除态、插入态、匹配态，图中分别用圆形、菱形及正方形表示。

基因的进化就可以认为是这 3 个状态之间的随机转移的结果。删除态代表基因序列中的空位和缺失，插入态代表基因的突变，匹配态代表某个特征序列。为了简化起见，假设原始祖先序列是 CC，开始以某种转移概率插入了一个碱基 A，再以某随机概率转移到匹配态 C，再随机转移到匹配态 C，

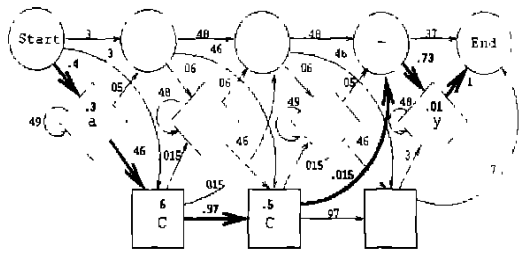

图 2 隐马尔可夫的描述

Fig. 2 The description of a profile hidden Markov model

圆形为删除态，菱形为插入态，正方形为匹配态

再进入一个删除态，最后转入插入态，插入碱基 Y，从而由特征序列 CC 最终形成了序列 ACCY。当然这只是进化的一种途径，由模型还可以形成其它许多序列，理论上讲，形成的路径可以有无数多条，因为有无穷多种插入的可能。

3.2 隐马尔可夫模型的定义

定义 一个模型 $\lambda=(S, \Sigma, A, B, \pi)$ 称为隐马尔可夫模型，其中：

（1）$S = \{S_1, S_2, \cdots, S_N\}$ 为状态集合，$N=|S|$ 是状态个数；

（2）$\Sigma=\{O_1, O_2, \cdots, O_M\}$ 是观察符号或观察向量的集合，$M=|\Sigma|$ 是观察符号或观察向量的个数；

（3）$\boldsymbol{A}=(a_{ij})$ 为状态转移概率矩阵，其元素 a_{ij} 表示从状态 S_i 转移到状态 S_j 的转移概率，有

$$a_{ij} = P(q_{t+1} = S_j \mid q_t = S_i), 1 \leqslant i, j \leqslant N \quad (1)$$

满足

$$a_{ij} \geqslant 0, \sum_{j=1}^{N} a_{ij} = 1; 1 \leqslant i, j \leqslant N \quad (2)$$

（4）$\boldsymbol{B}=(b_j(k))$ 表示在状态 S_j 时产生观察符号 $v_k \in O$ 的离散概率值（v_k 为离散符号）或连续概率密度（v_k 是连续的观察矢量）矩阵：

$$b_j(k) = P(v_k \mid q_t = S_j),$$
$$1 \leqslant j \leqslant N, 1 \leqslant k \leqslant M \quad (3)$$

当 v_k 是连续的观察矢量时，$b_j(k)$ 一般取正态概率密度函数，即 $b_j(k) = \sum_{m=1}^{M} C_{jm}\pi(v_k, \mu_{jm}, U_{jm})$，其中 C_{jm} 是混合系数，满足：

$$C_{jm} \geqslant 0, \sum_{m=1}^{M} C_{jm} = 1, 1 \leqslant j \leqslant N, 1 \leqslant m \leqslant M$$

μ_{jm}, U_{jm} 是正态概率密度中第 m 个分量的均值和协方差矩阵

$$\pi(v_k, \mu_{jm}, U_{jm}) = \frac{1}{\sqrt{2\pi}\ |U_{jm}|^{\frac{1}{2}}}。$$

$$\exp\left[-\frac{1}{2}(v_k-\mu_{jm})^{\mathrm{T}}U_{jm}^{-1}(v_k-\mu_{jm})\right] \quad (4)$$

(5) $\pi=(\pi_j)$ 是初始状态分布矩阵，其中：

$$\pi_j=P(q_1=S_j),\ 1\leqslant j\leqslant N$$

满足条件：

$$\pi_j\geqslant 0,\ \sum_{j=1}^{N}\pi_j=1$$

上述定义中当观察符号 v_k 是离散符号时，叫离散马尔可夫模型；当 v_k 是连续矢量时，叫连续马尔可夫模型；其中关键的参数是 A，B，π，从而模型可以简记为 $\lambda=(A,\ B,\ \pi)$。

3.3　向前向后算法的改进[6]

由模型 λ 产生序列 $O_1^*O_2^*\cdots O_k^*$ 的概率是：

$$P(O_1^*O_2^*\cdots O_k^*\mid\lambda)=\sum_{\text{allpath}}\pi_1 b_1(O_1^*)a_{12}b_2(O_2^*)\cdots a_{k-1k}b_k(O_k^*) \quad (5)$$

产生序列 $O_1^*O_2^*\cdots O_k^*$ 所需计算量是 $O(k\cdot N^k)$，若 $N=10$，观察序列长度是 $k=100$，则 10^{100} 级的计算量计算机是根本吃不消的！为此必须对算法进行改进，定义向前向后变量 $\alpha_t(i)$ 及 $\beta_t(i)$ 分别如下：

$$\alpha_t(i)=P(O_1^*O_2^*\cdots O_t^*,\ q_t=S_i^*\mid\lambda) \quad (6)$$

$$\beta_t(i)=P(O_{t+1}^*O_{t+2}^*\cdots O_k^*\mid q_t=S_i^*,\ \lambda) \quad (7)$$

故关于评估问题 $P(O^*\mid\lambda)$ 算法的可以改进为：

①初始化：

$$\alpha_1(i)=\pi_i b_i(O_1^*),\ 1\leqslant i\leqslant N \quad (8)$$

②迭代向前：

$$\alpha_{t+1}(j)=\left(\sum_{i=1}^{N}\alpha_t(i)a_{ij}\right)b_j(O_{t+1}^*)$$
$$1\leqslant t\leqslant k-1,\ 1\leqslant j\leqslant N \quad (9)$$

③终止：

$$P(O_1^*O_2^*\cdots O_k^*\mid\lambda)=\sum_{i=1}^{N}\alpha_k(i) \quad (10)$$

由此可知，改进后的算法，其运算量减少为 $O(k\cdot N^2)$，比起改进前的 $O(k\cdot N^k)$，其减少的量级是指数级的。

4　Pfam 数据库简介

Pfam 数据库[10] 是以序列谱为基础，用 profile 隐马尔可夫模型方法组建的蛋白质家族多重序列比对数据库，由 Sanger Centre 开发和维护(Sonnhammer 等，1998)，目前的版本是 Pfam version 13.0 (April 2004)，它是基于 Swissprot 42.12 和 SP-TrEMBL 25.12 蛋白质序列数据库，至 2004 年 4 月共有 7426 个蛋白质家族被收录。

Pfam 数据库可用于识别一未知序列是否属于某个已知的蛋白质家族，甚至其同源性非常微弱也能识别。不同于标准的双序列比对数据库搜索（如 BLAST，FASTA），Pfam 数据库注重于多重蛋白质域的搜索。

Pfam 数据库共分为 2 个子库，PfamA 和 PfamB，PfamA 是基于一组人工比对得到的种子序列，并对结果进行编辑，其结果较准确；PfamB 则是用计算机程序对 Swissprot 数据库进行多序列比对自动生成的非冗余蛋白质数据库。

5　实验结果和讨论

5.1　建立隐马尔可夫模型

从与 Pfam 数据库相连的 SCOP 蛋白质数据库[11] 中选择了类为 β，折叠为前清蛋白，超族和族为淀粉黏合物的已知结构的蛋白质序列 100 条作为训练序列，先进行多重序列比对，后建立多重序列的 profile HMMS 模型，采用 http://pfam.wustl.edu/ 为训练平台，经过 100 个循环的训练，得到如下的 HMMs 模型（表 1）。

表 1　HMMs 模型参数

Tab. 1　Parameters of HMMs

模型长度			训练分值 (Score)			E 值		
插入节点	删除节点	匹配节点	最高	最低	平均	最高	最低	平均
102	102	102	199.9	131.3	153	2.2×10^{-36}	7.6×10^{-57}	2.3×10^{-43}

表 1 中的训练分数越高表明序列的保守性越强，E 值是判断比对结果的期望值（置信度），期望值越接近 0，表明置信度越高，本例中最大和最小 E 值都非常接近 0，说明由训练集产生的模型比对置信度是非常高的。图 3 是模型的训练熵，是直接由训练分数根据训练的时间得到的。计算公式是

$$\mathrm{Score}_{\mathrm{entropy}}(t)=\mathrm{Score}+(t-t_i),$$
$$1\leqslant i\leqslant 100 \quad (11)$$

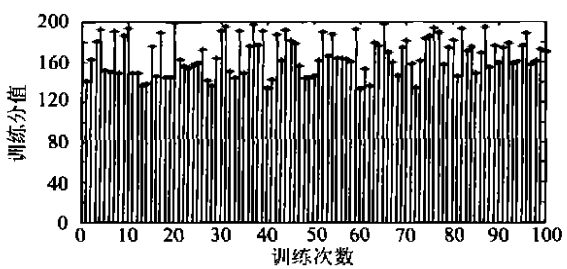

图 3　模型的训练熵

Fig. 3　Entropy of different training by HMMs

5.2 用 HMMs 进行蛋白质家族的模式分类

5.2.1 两类问题的模式分类问题 以训练好的模型为基础，再在 SWISSPORT 数据库中选取已知结构的非簇类（非淀粉黏合物家族类）38 条，以上述模型计算训练分值，以 153 分和平均阈值 $2.3\times^{-43}$ 为分界值（有的识别只考虑 HMMs 的分数值而不考虑 E 值的保守性，笔者认为这是不全面的），得到表 2 所示的分类结果。

表 2 HMMs 用于两类蛋白家族分类

Tab. 2 Two kinds of protein families classified by HMMs

家族	训练序列	测试序列	正确识别	识别率 %	真阴性错误
淀粉黏合物	19	19	18	95	1
非淀粉黏合物	19	19	18	95	1

从分类结果来看，在识别本类非本类方面都有较高的区分水平。

5.2.2 建立多类 HMMs 分类模型 而对多个蛋白质家族，选取了 3 个蛋白质家族作为样本，分别是球蛋白、流感红血球凝聚素和糖苷酶。并为每一类建立一个多重序列的 HMMs 模型，如表 3 所示。

从表 3 可以看出这 3 个蛋白质家族各有不同的隐马尔可夫分值和 E 值，从而它们可以代表不同的家族类。

表 3 HMMs 用于多类蛋白家族识别

Tab. 3 Many kinds of protein families recognized by HMMs

家族	序列	训练分值	E 值
Globins	Q21978/165—314	55.5	5×10^{-13}
	Q20638/74—216	62.6	1.1×10^{-15}
	Q19601/54—189	65.0	2×10^{-16}
	Q18311/32—175	56.9	5.7×10^{-14}
	Q18209/233—375	35.2	1.9×10^{-7}
Influenza hemagglutinin	HEMA-INCJH/19—421	573.5	1.6×10^{-169}
	HEMA-INCCA/19—420	573.1	2.2×10^{-169}
	HEMA-CVMJH/9—395	561.8	5.6×10^{-166}
	HEMA-CVBF/4—379	565.5	4.4×10^{-167}
	HEMA-CVMDV/8—387	571.7	6.1×10^{-169}
family 19 glycoidase	CHIX-PEA/83—314	588.3	5.8×10^{-174}
	CHID-LYCES/8—239	586.1	2.8×10^{-173}
	CHI5-PHAVU/78—308	594.1	1.1×10^{-175}
	CHI2-BRANA/74—305	613.5	5×10^{-181}
	CHI1-TOBAC/83—314	613.5	1.5×10^{-181}

HMMs 模型对于各类蛋白质家族的这些值是非常不同的，有的 E 值相当保守。

5.3 关于模型的局限性讨论

虽然我们可以为不同的问题建立 HMM 模型，但是和其它任何模型一样，隐马尔可夫模型也有它的局限性，主要体现在以下几个方面：

（1）它假设连续的观察符号是相互独立的。而事实上它们可能并非如此，这些信号之间会相互依赖；

（2）t 时刻的状态只依赖于 $t-1$ 时刻，然而对于像语音识别问题，t 时刻的状态依赖于前面多个时刻，因此这种模型的一个局限性正是它的马尔可夫假设；

（3）隐马尔可夫模型只是线性模型，没有描绘蛋白质各因素之间如非相邻氨基酸之间和链与链之间的氢键关系，以及蛋白质的折叠关系等高阶相关性，这也是所有线性模型的局限性；

（4）对于小样本训练的偏差性。由于数据库中能取样的序列总是有限的，因此训练的模型总存在偏差；

（5）域值选取的界限问题。过紧和过松的域值选取，会导致样本的少分或错分，因此要结合序列的同源性（结合其它比对方式）或其它性质来取域值可能会更好；

（6）隐马尔可夫模型训练和长度相关。对于有些非常短的序列（比如病毒序列），则无法和其它序列的模型比较；

除了上述缺点外，隐马尔可夫模型仍然是一个非常好的概率模型，因为它已经成功进行了语音识别，手写数字识别，蛋白质分类，基因预测等重大问题，和其它机器学习方法一样，结合计算机计算能力的不断提高，其算法必将继续得到改进，其应用也将越来越广泛。

参考文献：

[1] WANG L, JIANG T. On the complexity of multiple sequence alignment[J]. Journal of Computational Biology, 1994, 1: 337—348.

[2] GUSELD D. E-cient methods for multiple sequence alignment with guaranteed error bounds[J]. Bulletin of Mathematical Biology, 1993, 55: 141—154.

[3] FENG D, DOOLITTLE R. Progressive sequence alignment as prerequisite to correct phylogenetic trees[J]. Journal of Molecular Evolution, 1987, 25: 351—360.

[4] THOMPSON J, HIGGINS D, GIBSON T. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties, and weight matrix choice[J]. Nucl Acids Res, 1994, 22: 4673—4680.

[5] BAUM L E, EGON J A. An inequality with applications to statistical estimation for probabilistic functions of a Markov process and to a model for ecology[J]. Bull Amer Meteoral,

1967, 73: 360—363.
[6] BAUM L E, PETRIE T. Statistical inference for probabilitic functions of finite state Markov chains[J]. Annmath Stat, 1996, 37: 1554—1563.
[7] BAKER J K. The dragon system —— an overview[J]. IEEE Trans Acount Speech Signal Prpcessing, 1975, 23(1): 24—29.
[8] JELINEK F. A fast sequential decoding algorithm using a stack[J]. IBM J Res Develop, 1969, 13: 675—685.
[9] KROGH A, BROWN M, MIAN I S, et al. Hidden Markov models in computational biology: applications to protein modeling[J]. Journal of Molecular Biology, 1994, 235: 1501—1531.
[10] Washington University in St Louis. The pfam database of protein families and HMMs [EB/OL]. http://pfam.wustl.edu/2004—04—12.
[11] ALEXEY G, MURZI N. Structural classification of proteins [EB/OL]. http://scop.mrc-lmb.cam.ac.uk/scop/2004—04—12.

Multiple Alignment Analysis Based on Hidden Markov Models

LUO Ze-ju[1], *ZHU Si-ming*[1], *HE Miao*[2]
(1. Department of Mathematics, Sun Yat-sen University, Guangzhou 510275, China;
2. School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Hidden Markov models (HMMs) are statistical models of sequential data that have been used successfully in many machine learning applications, especially in protein family recognition in the last few years. This is due to the vast increase of data in biology that makes two fields (computational science and biology) associate with each other. To study multiple sequence alignment and profile HMMs, one essential algorithm of HMMs and how to build an HMMs are discussed. By E values and training scores an experiment for recognition and classification of protein family is given.
Key words: hidden Markov models (HMMs); multiple sequence alignment; protein classification

---

(上接第 8 页)

Information Hiding Based on DNA Sequences

ZHENG Guo-qing[1], *LIU Jiu-fen*[2], *HUANG Da-ren*[3], *XU An-long*[1]
(1. Department of Biochemics, Sun Yat-sen University, Guangzhou 510275, China;
2. Department of Information Research, PLA Information Engineering University, Zhengzhou 450002, China;
3. Department of Scientific Computing and Computer Applications, Sun Yat-sen University, Guangzhou 510275, China)

Abstract: Information hiding in DNA sequence is discussed, and an information hiding algorithm is proposed. In this algorithm secret messages are transformed into stochastic DNA sequences by beforehand managing, coding and modulating, and then the locations of secret messages embedded in the high complexity regions are decided by a stochastic sequence. Finally, the covered characters at embedded locations are translated from the location of secret message along with hidden information. This information hiding with the algorithm is demonstrated by robust, secure and of biologically significant. The efficiency of the algorithm is demonstrated by experiment results also.
Key words: DNA sequences; information hiding; cover