隐马尔科夫过程在生物信息学中的应用*

周海廷

(西南科技大学 生命科学与工程学院,中国四川 绵阳 621000)

摘 要:隐马尔科夫过程(hidden Markov model,简称HMM)是20世纪70年代提出来的一种统计方法,以前主要用于语音识别[1].1989年Churchill[2]将其引入计算生物学.目前,HMM是生物信息学中应用比较广泛的一种统计方法[3~7],主要用于:线性序列分析、模型分析、基因发现等方面.对HMM进行了简明扼要的描述,并对其在上述几个方面的应用作一概略介绍.

关键词:隐马尔科夫过程;序列搜索;模型估计;基因识别
中图分类号:Q811.4 文献标识码:A 文章编号:1007-7847(2002)03-0204-07

An Introduction to the Hidden Markov Models for Bioinformatics

ZHOU Hai-ting

(*Southwest University of Science & Technology*, *Mianyang* 621000, *Sichuan*, *China*)

Abstract: The Hidden Markov Model (HMM) is a statistical model, which is very well suited for many tasks in molecular biology, although they have been mostly developed for speech recognition since the early 1970's. The most popular use of the HMM in molecular biology is as a "probabilistic profile" of a protein family, which is called a profile HMM. From a family of proteins (or DNA) a profile HMM can be made for searching a database for other members of the family. The HMM can be applied to other types of problems. It is particularly well suited for problems with a simple "grammatical structure", such as gene finding.

Key words: hidden Markov models; sequence search; model estimation; gene-finding

(*Life Science Research*, 2002, 6(3): 204~210)

1 隐马尔科夫过程方法描述

1.1 马尔科夫过程(Markov model)

在分子遗传学中,Markov model(简称MM)主要用于描述某一核苷酸序列从其特定的祖代遗传而来的概率,换言之,从现有的核苷酸序列来推测最有可能出现的祖代核苷酸序列.我们可以将MM想象成用棋盘玩的游戏,其游戏规则为:1)棋盘上的每个正方形格子用字母标记(例如,用*ACTG*四个字母表示DNA的四种核苷酸,用20个字母来标记组成蛋白质的氨基酸).格子中的字母以不同的比例出现(例如,某些格子中绝大部分时间得到的是字母*A*和*C*,但有时也可能是字母*G*出现,重要的是不能有两个相同的格子同时出现).每个字母都有一个分值(取值范围在0与1之间),游戏结束时将这些分值相乘得到总分;2)你在同一格子中停留的时间愈长,得到的模型就愈好.因此,每次你从一个格子移动到另一个格子时要受罚;3)你可以随时删除或插入格子,而且基本上不受罚,但是这些格子不会给出字母,因而你

* 收稿日期:2001-12-18;修回日期:2002-05-20
作者简介:周海廷(1957-),男,四川罗江县人,西南科技大学副教授,硕士,主要从事生物信息学研究,Tel: +86-0816-6332006,E-mail: zhouhaiting@263.net.

也不会得到任何分值.

MM 游戏的目的是, 对于给出的某一核苷酸序列, 尽可能找出其祖代核苷酸序列, 得到最高分, 同时使受罚降至最低. 下面是一个只有两个格子的最简单的游戏的一个例子(见图 1) . 假如核苷酸序列为: *AAATATTATACTATCGGCGAGCGCG*. 根据上述游戏规则, 这轮游戏的最佳玩法应该是, 在第 15 个字母之前待在格子 1 中, 从 15 个字母移动到格子 2 中. 如果在碰到第一个字母 *C*(第 11 个字母) 就移动到格子 2 中的话, 紧接着遇到字母 *A* 又要返回格子 1, 虽然字母 *T* 和 *A* 在格子 1 中, 字母 *C* 在格子 2 中得到了高分, 但移动两次的罚分远比得到的分值高, 这种玩法是不可取的.

格子 1: 主要用于字母 *T* 和 *A*, 偶而也用于字母 *C* 和 *G*. In Square 1, you get *A*' s and *T*" s nearly all of the time, but you can get the Rare *C* or *G*.	格子 2: 主要用于字母 *C* 和 *G*, 偶而也用于字母 *T* 和 *A*. In Square 2, you get *C*' s and *G*' s nearly all of the time, but you can get the Rare *A* or *T*.

图 1　MM 游戏示例

Fig. 1　An examle of the Markov model game

1. 2　隐马尔科夫过程

如图 2 所示的 DNA motif , 如果用常规的方法表达, 应当写成:

[*AT*] [*CG*] [*AC*] [*ACGT*]* *A* [*TG*] [*GC*],

其意思是第一个位点出现的核苷酸为 *A* 或 *T*, 第二个位点的核苷酸为 *C* 或 *G*, 余类推. 而 [*ACGT*]* 则表示 *ACGT* 四个核苷酸中的任何一个, 都可以多次出现. 这种表达方式既正规又简单, 但是用来区分核苷酸序列或蛋白质结构就可能出现问题. 例如: 核苷酸序列 *TGCT- - AGG* 与图 2 所示的 DNA motif 相差甚远, 而 *ACAC- - ATC* 则与图 2 所示的 DNA motif 非常接近. 用上述的正规表达式却无法将这两段差异很大的核苷酸序列区分开, 人们不得不寻求新的表达方法.

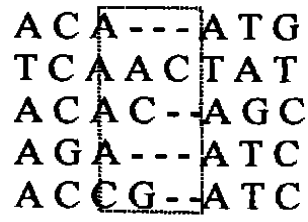

图 2　DNA Motif 示意图

Fig. 2　Sketch map of DNA motif

引入概率的概念, 对每个位点可能出现的核苷酸进行打分. 例如图 2 中, 第一位点出现 A 的概率为 4/ 5= 0. 8, 出现 *T* 的概率为 1/ 5= 0. 2. 第二位点出现 *C* 的概率为 4/5= 0. 8, 出现 *G* 的概率为 1/ 5= 0. 2, 余类推. 第四位点到第六位点(图 2 中的虚框部分) , 5 个序列中有 3 个插入了长度不等的核苷酸, 定义插入的概率为 3/5= 0. 6, 无插入的概率为 2/ 5= 0. 4. 结果用图 3 表示.

图 3 实际上就是一个 HMM. 与正规表达式一样, 图中每个方框代表一种状态, 方框中的黑色柱子表示该位点核苷酸出现的概率, 概率的计算如上所述. 插入状态中核苷酸的概率是用插入区域(图 2 中的虚框部分) 中所出现的核苷酸进行计算的, 即: *A* 的概率为 1/ 5= 0. 2, *C* 的概率为 2/ 5= 0. 4, *G* 的概率为 1/ 5= 0. 2, *T* 的概率为 1/ 5= 0. 2. 图 2 中的第 2 序列有 3 个插入, 第 3 和 5 序列分别有一个插入, 因而插入状态中有 5 次转移, 其中在插入状态内部转移的次数为 2, 概率为 2/5= 0. 4, 从插入状态转移到正常状态的次数为 3, 概率为 3/5 = 0. 6.

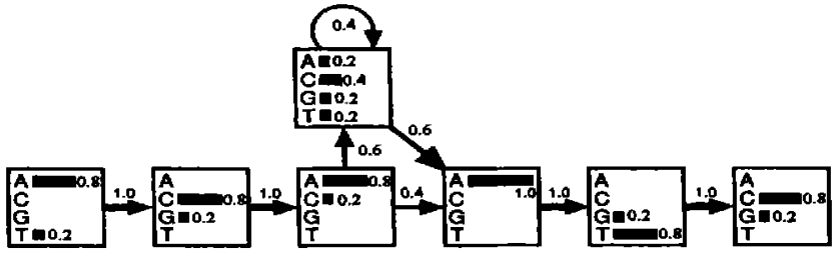

图 3　由图 2 DNA Motif 构成的 Hidden Markov model

Fig. 3　A hidden Markov model derived from the alignment discussed in the text

现在用 HMM 来区别核苷酸序列 *TGCT – – AGG* 和 *ACAC – – ATC* 就容易多了. 序列 *ACAC – – ATC* 出现的概率为:

$$P(ACACATC) = 0.8 \times 1 \times 0.8 \times 1 \times 0.8 \times 0.6 \times 0.4 \times 0.6 \times 1 \times 1 \times 0.8 \times 1 \times 0.8 = 4.7 \times 10^{-2}$$

序列 *TGCT – – AGG* 出现的概率则为:

$$P(TGCTAGG) = 0.0023 \times 10^{-2}.$$

核苷酸序列 *TGCT – – AGG* 和 *ACAC – – ATC* 出现的概率相差了 2 000 倍. 图 2 中各核苷酸序列出现的概率见表 1.

表 **1** 图 **2** 中各核苷酸序列出现的概率及对数值

Table 1 Probabilities and Log-odds scores for the 5 sequences in the alignment

	核苷酸序列 Sequence	概率 × 100 Probability × 100	对数值 Log-odds
与原始序列非常相似 Consensus	*ACAC – – ATC*	4.7	6.7
原始序列 Original Sequences	*ACA – – – ATG*	3.3	4.9
	TCAACTATC	0.0075	3.0
	ACAC – – AGC	1.2	5.3
	AGA – – – ATC	3.3	4.9
	ACCG – – ATC	0.59	4.6
与原始序列相差甚远 Exceptional	*TGCT – – AGG*	0.0023	– 0.97

核苷酸序列出现的概率大小在很大程度上取决于核苷酸序列的长度. 因而, 用概率值给核苷酸序列打分, 并不十分恰当. 用公式(1) 将表 1 中的概率值转换为对数值, 原始序列之间的对数值与概率值相比, 显然对数值之间的差异要小得多. 与原始序列非常相似的核苷酸序列的对数分值与原始序列的对数分值也十分接近, 而与原始序列相差甚远的核苷酸序列的对数分值与原始序列的对数分值差异很大. 可见对数分值对核苷酸序列的区分度, 要比概率值的区分度好得多.

设核苷酸序列 S 对数值为 $W(S)$, 核苷酸序列 S 的概率为 $P(S)$, 则

$$W(S) = \log(P(S)/(0.25)^{L}) = \log P(S) - L\log 0.25 \quad (1)$$

式中, 0.25^{L} 是无效模式.

1.3 Profile 隐马尔科夫过程(Profile HMM)

Profile[8~10] 是一个位置特定的评分矩阵, 它包含了一个序列对比结果中每个位置的所有残基信息. 这一点与共有序列不同, 共有序列中只包含每个位置的保守残基的信息. Profile 做好后可用于搜索数据库、数据库划分或在一个集合中搜索与原始对比结果中的序列相似的序列. 它也可以用于把一条单独的序列与一个对比结果进行对比.

与标准的 Profile 相比, Profile HMM[3, 4, 6, 11] 有正规的概率作基础, 对于序列的间隙和插入状态的记分也有较为可靠的理论依据. 而标准的 Profile 纯粹是一种启发式的方法. HMM 用统计方法估计序列某一位点核苷酸残基出现的真正概率, 而标准的 Profile 却是用自身的观察频率给核苷酸残基指派分数. 这就意味着用 Profile HMM 方法从 10 至 20 个核苷酸序列构成的队列中提取的信息, 相当于用标准的 Profile 从 40 至 50 个核苷酸序列构成的队列中提取的信息.

图 4 是 Profile HMM 的结构转移示意图, 图中最下层的方形表示主状态(匹配状态), 每个状态实际上就是一个 DNA motif 中的一列. 中间层的菱形为插入状态, 常用于构建队列中变化较大的区域中的模型, 其功能就象图 3 中上层方框的作用. 最上层的圆称为删除状态, 这是与 HMM 不同的地方, 称为哑状态或无效状态. 它们不与队列中的任何核苷酸残匹配, 而是在建模的过程中如果遇到队列中的“ – ”区域可以跳过去.

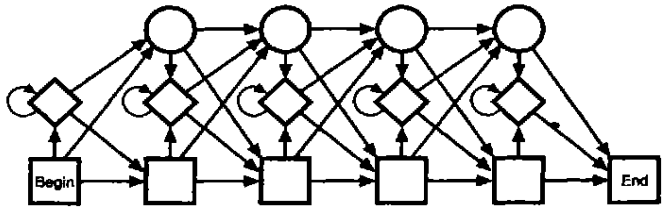

图 **4** **Profile** 隐马尔科夫过程结构轮廓图

Fig. 4 The transition structure of a profile hidden Markov model

2　隐马尔科夫过程的应用

2.1　核苷酸序列搜索

前面论述了特定 DNA motif 中核苷酸序列概率的计算方法. 然而, 对于不是特定 DNA motif 中的核苷酸序列, 我们并不知道计算概率的路径, 因而下一个问题就是如何对这一类核苷酸序列进行打分. 如果某一个核苷酸序列与已知的核苷酸队列匹配得比较好的话, 就可以应用已知核苷酸队列概率的计算方法对新的核苷酸序列记分[6, 12].

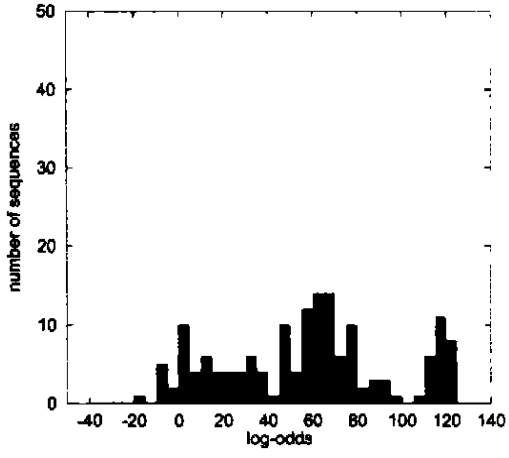

图 5　用 **Swissprot** 软件对 **SH3** 域进行 **Profile HMM** 搜索得到的对数分的分布图

Fig. 5　The distribution of Log-odds scores from a search of swissprot with a profile HMM of the SH3 domain

例如, 我们已知某个蛋白质序列是由氨基酸 A_1、A_2、A_3、……构成的, 在 HMM 中用 M_1、M_2、M_3、……表示匹配状态, 用 I_1、I_2、I_3 表示插入状态. 然后在队列中, A_1 与 M_1, A_2 和 A_3 与 I_1 匹配, A_4 与 M_2 匹配, A_5 与 M_6 匹配(经过了三个删除状态) 等等. 通过指定的路径, 可以计算出序列的概率值或对数分. 这样就可以找出最好的核苷酸队列, 即概率值最大的队列. 用这种方式得到的对数分, 可用于在数据库中搜寻相似的家系. 图 5 是数据库搜寻的一个典型例子, 方柱型图中的黑色部分是经过注释的 SH3 域, 浅色部分是没有注释过的队列. 与大多数情况一样, 图中真正的正值与虚假的正值并没有明确的分界线, 在对数值 0 左右波动的序列需要作进一步的研究.

核苷酸序列记分的另一种方法是求构成模型序列的全部可能列阵的概率总和. 其概率可以用一种称为向前算法(forward algorithm) 求得, 不过在序列比较中很少用这种方法给序列打分.

2.2　模型估计

Profile HMM 的另一个作用是可以用它来估计模型[13,14], 从没有形成队列的序列(*unaligned sequences*) 估计所有的概率参数, 进而构成多重队列(multiple alignment of the sequences) . 与其它多重队列方法一样, Profile HMM 建模也有一个迭代的过程. 从某一队列开始构建的模型或多或少都带有一些随机概率的成份, 随着全部序列加入模型形成队列后, 就可以用队列去改进模型中的概率, 新的概率可能会使队列产生轻微的差异. 重复上述过程, 不断修正概率值, 直到队列稳定为止, 最终形成多重队列.

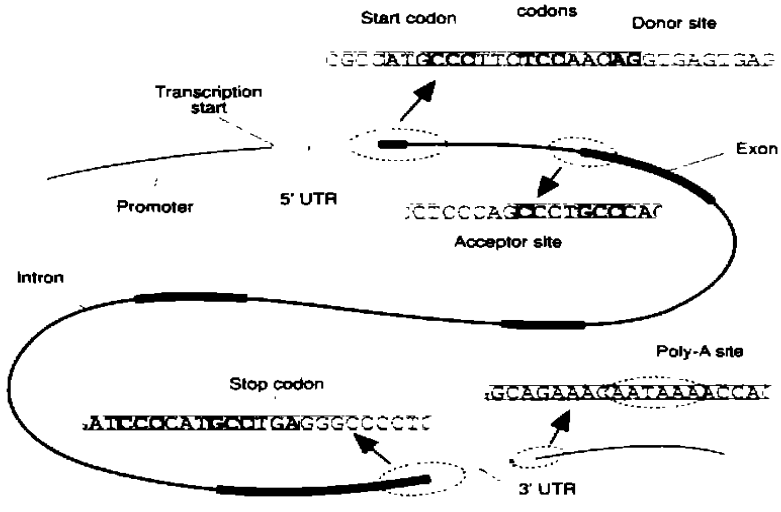

图 6　带有重要信号的基因结构图

Fig. 6　The structure of a gene with some of the important signals shown

2.3 基因识别

在序列分析的许多问题中都涉及到语法问题, 或者说是真核细胞基因的结构问题. 如果我们把外显子(exons) 和内含子(introns) 比成语言中的单词, 句子的形式则为: 外显子(内含子(外显子(内含子…内含子(外显子. 一个句子决不会用一个内含子来结尾, 当一个基因完成时, 起码决不可能出现两个外显子中间不带内含子的情况. 显然, 这种语法非常简单, 因为对于基因结构还有其它几个约束条件, 比如外显子必须和有效编码区同时存在. 如图 6 所示.

将正规的语法理论用于解决生物学问题已经不是新的发明, David Searls[15] 已将它用于基因寻找. 从形式上说, HMM 只是将谓正规的语法[3] 简单化了, 从而使基因寻找[16~20] 及其它问题显得相对容易一些.

2.3.1 信号传感器

图 7 是人类 DNA 接收器位点周围的某些核苷酸构成的一个列阵. 这个列阵有 19 列, 刚好用 HMM 的 19 种状态进行表示(不存在插入和删除状态) . 由于该列阵不存在间隙, 所以 HMM 与加权矩阵是等价的.

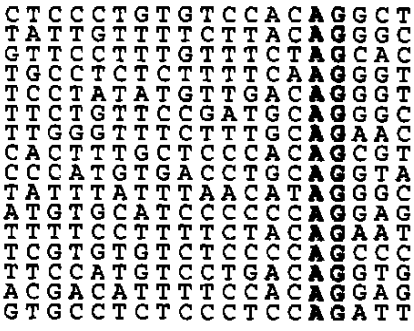

图 7 人类接收器位点的例子(核苷酸 5′ 端与外显子结合)

除极少的情况外, 内含子由 *AG* 结束(图中阴影部分) .

Fig. 7 Examples of human acceptor sites(The splice site 5′ to the exon)

Except in rare cases, the intron ends with *AG*

在 DNA 中普遍都存在二核苷酸(dinucleotide) 偏爱问题, 而我们建立的模型是基于核苷酸独立存在的基础之上的, 因而可能无法捕捉到二核苷酸. 这个问题可以通过将原来每种状态下的 4 个概率参数调整为 16 个参数而得到解决. 在图 7 中, 以第二列为例, 我们可以将第一列是 *A* 所对应的第二列的 4 种核苷酸进行计数, 进而计算出第一位点为 *A*, 第二位点 4 种核苷酸出现大小的条件概率. 同理, 可以分别计算出位点一为 *C*、*G*、*T*, 第二位点 4 种核苷酸出现大小的条件概率. 进而得到两种 HMM 状态的全部 16 个概率值. 所有其它状态的概率值如法炮制. 例如, 可以用公式 2 计算序列 *ACTGTC* …的概率:

$$P(ACTGTC\ldots) = p_1(A)\times p_2(C\mid A)\times p_3(T\mid C)\times p_4(G\mid T)\times p_5(T\mid G)\times p_6(C\mid T)\times \ldots \quad (2)$$

式中, p_1 是 4 个核苷酸在状态 1 的概率, $p_2(x\mid y)$ 是第前一个核苷酸为 y 的核苷酸 x 在状态 2 的条件概率, 等等.

带有条件概率的状态称为一阶状态, 因为它可以捕获相邻核苷酸的一阶相关. 很容易将这种情况扩展到高阶状态.

2.3.2 编码区

密码子(codon) 结构是编码区中最重要的特征. 图 8 表示用 3 种状态构建的三联密码模型. 从图中看出, 编码区模型可以用非绞接的基因的简单模型来表示, 该基因由一个开始密码子(*ATG*) , 和若干个密码子及一个结束密码子组成. 因为一个密码子是由 3 个基本长度构成的, 因此密码子模型的最后状态至少应在第二位上, 这样才可能获得正确的密码子统计数. 在一套已知的编码区中, 通过对每个密码子进行计数, 可以估计出 64 种概率. 例如, 通过对 *CAA*、*CAC*、*CAG* 和 *CAT* 密码子进行计数, 可以得到以下概率:

$$p(A\mid CA) = c(CAA)/(c(CAA)+c(CAC)+c(CAG)+c(CAT))$$

$$p(C\mid CA) = c(CAC)/(c(CAA)+c(CAC)+c(CAG)+c(CAT))$$

$$p(G\mid CA) = c(CAG)/(c(CAA)+c(CAC)+c(CAG)+c(CAT))$$

$$p(T\mid CA) = c(CAT)/(c(CAA)+c(CAC)+c(CAG)+c(CAT))$$

式中 $c(xyz)$ 是密码子 xyz 的计数.

编码区的特征之一是不存在结束子. 由于 $p(A\mid TA)$、$p(G\mid TA)$ 及 $p(A\mid TG)$ 与 3 个结束子 *TAA*、*TAG* 和 *TGA* 相对应, 其值自动为 0.

在密码子统计学建模过程中, 通常用普通状态(0 阶) 作为密码子模型的第一状态, 第 1 阶作为密码子模型的第二状态. 不过, 相邻密码子间实际上也可能是独立, 因此, 有可能需要更高阶的状态[21] .

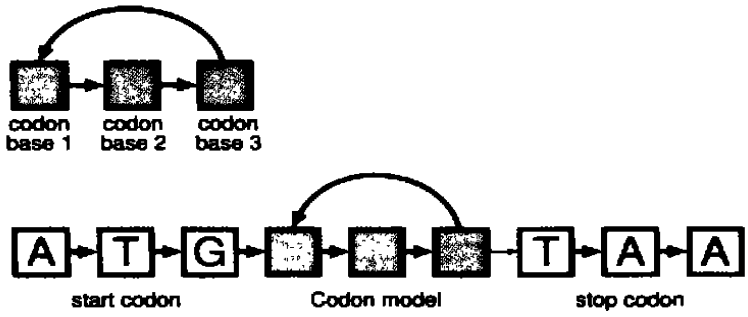

图 8　编码区模型

顶部:编码区模型,状态 1、2、3 分别与第 1、2、3 密码子位置相匹配.底部:非绞接的基因的一个简单模型,前面 3 个状态与开始密码子匹配,后面 3 个与结束密码子匹配.

Fig. 8　A model of coding regions

Top: A model of coding regions, where state one, two and three match tne first, second and third codon positions respectively. A coding region of any length can match this model, first three states matching a start codon, the next three of the form shown to the left, and the last three states matching a stop codon(Only one of the three possible stop codons are shown).

2.3.3　模型的结合

为了发现基因,我们需要用某种确保符合基因文法的方式将模型联合起来.首先,我们来考虑非绞接基因的模型结合.如果忽略了基因的交迭和相当接近的空间状态,它的模型可以认为像图 9 这种形式.它是由基因间区域、围绕启动子区域的模型、编码区模型、围绕结束子区域的模型等组成的.围绕启动子区域的模型非常象前面描述过的接受子模型.启动子区域的模型由 8 个上行比特为一组的碱基打头,接着是 *ATG* 启动子,紧随其后是第一个密码子,整个模型就是一个大的 HMM.

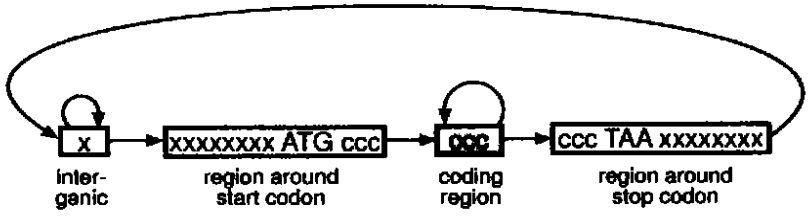

图 9　非绞接基因的 HMM

图中 x 表示非编码 DNA,c 表示编码 DNA

Fig. 9　A hidden Markov Model for unspliced genes

"x" means a atate for non-coding DNA, and "c" means a state for coding DNA. Only one of the three possible stop codons are shown in the model of the region around the stop codon.

有了这样的模型,我们怎样用它预测匿名序列上的基因呢?这是一个十分简单的问题,用 Viterbi 算法发现最有可能通过模型的路径,当通过路径走到 *ATG* 状态时,启动子就被发现了.同理,当路径通过密码子区域时,密码子也就被发现了,等等.

该模型预测基因不可能百分之百地准确,但起码可以预测出符合文法、容易察觉的基因.一个基因总是以启动子开始,以结束子结尾,基因的长度能被三整除,基因的阅读框架中决不可能含有结束子,这些是一个非绞接基因最起码的条件.

构建一个完全符合基因拼接规律的模型有一定的难度,因为拼接可以在 3 种不同的阅读框架中进行(见图 10),而一个外显子的阅读框必须与下一个外显子的阅读框相匹配.

模型中顶上的线条是表现出现在两个密码子之间的内含子.在内含子开始匹配外显子末尾的密码子之前,该模型有一个三态(用字母 *ccc* 表示).首先内含子模型的二态与 *GT* 匹配,其次六态立即与 *GT* 之后的6个碱基配对.这是为供体

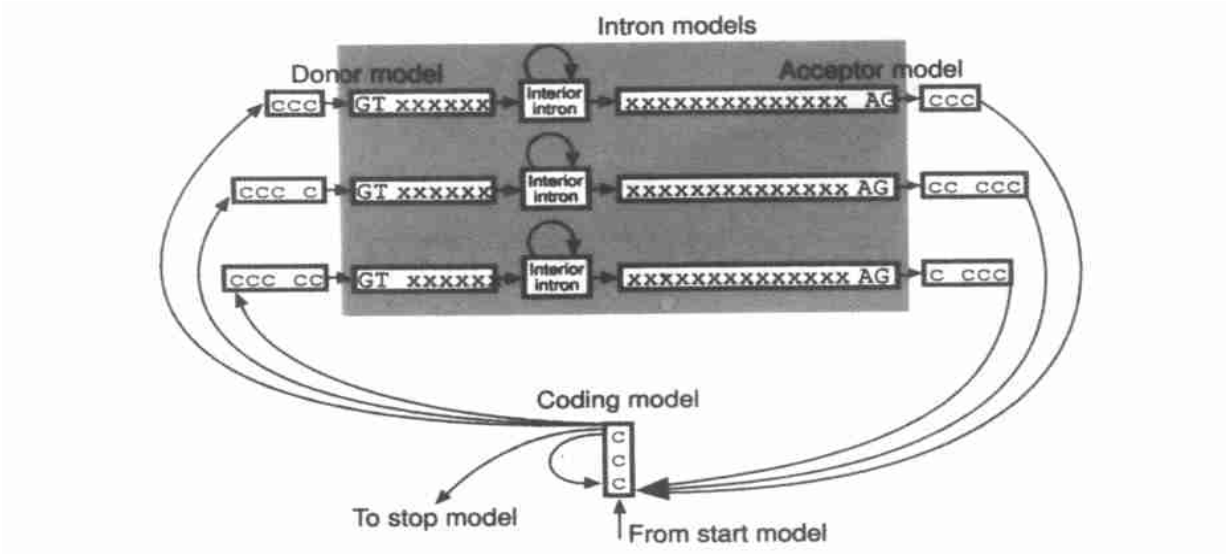

图 **10** 3 种不同框架基因拼接示意图

为了保证框架不出错, 还必须在内含子模型前后加上“间隔”.

Fig. 10 The structure of gene splicing in three different frames

To get the frame correct “spacer states” are added before and after the intron models.

位点构建的模型, 其概率可以用前述计算受体位点概率的方法进行计算. 紧接着是一个单态为内含子的内部情况进行建模. 随之是受体模型, 其中的三态与下一个外显子的第一个密码子相匹配.

模型中第二条线是为了表示出现在密码子第一个碱基之后的内含子. 与第一条线的差异是, 就一个编码碱基在内含子前有一个多态, 在内含子之后有两个多态. 这就保证了阅读框架与两个相邻的外显子相匹配. 相应的, 模型中第 3 条线, 在两个内含子之间有两个额外的编码状态, 这样就可以与在密码子中出现在第 2 个碱基之后的内含子相匹配.

参考文献(**References**):

[1] LAWRENCE R A. Tutorial on hidden Markov models and selected applications in speech recognition[J]. Proceedings of the IEEE, 1989, 77(2): 257-286.

[2] CHURCHILL G A. Stochastic models for heterogeneous DNA sequences[J]. Bulletin of Mathematical Biology, 1989, 51: 79-94.

[3] DURBIN S R. Biological sequence analysis: Probalistic models of proteins and nucleic acids[M]. Cambridge: Cambridge University Press, 1998. 121-138.

[4] EDDY S R. Hidden markov models[J]. Current Opinion in Structurel Biology, 1996, 6: 361-365.

[5] EDDY S R. Profile hidden markov models[J]. Bioinformatics, 1998, 14: 755-763.

[6] KROGH M. Hidden Markov models in computational biology: Applications to protein modeling[J]. J Mol Biol, 1994, 235: 1501-1531.

[7] STEVEN S. In Computational Methods in Molecular Biology[M]. Holand: Elsevier Science, 1998. 45-63.

[8] GRIBSKOV M. Profile analysis: Detection of distantly related proteins[J]. PNAS, 1987, 84: 4355-4358.

[9] LUTHY R. Assessment of protein models with three-dimensional profiles[J]. Nature, 1992, 356: 83-85.

[10] TAYLOR W R. Identification of protein sequence homology by consensus template alignment[J]. Journal of Molecular Biology, 1986, 188: 233-258.

[11] BALDI P. Hidden Markov models of biological primary sequence information[J]. PNAS, 1994, 91: 1059-1063.

[12] HUGHEY R, KROGH A. Hidden Markov models for sequence analysis: Extension and analysis of the basic method[J]. CABIOS, 1996, 12: 95-107.

[13] EDDY S R. Multiple alignment using hidden Markov models. Proc. of Third Int. Conf. on Intelligent Systems for Molecular Biology[C]. Cambridge England: AAAI/MIT Press, 1995. 3: 114-120.

[14] DONG S, SEARLS D B. Gene structure prediction by linguistic methods[J]. Genomics, 1994, 23: 540-551.

[15] KULP D. A Generalized hidden Markov model for the recognition of human genes in DNA, Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology[C]. Menlo Park CA: AAAI Press, 1996. 134-142.

[16] KROGH A. Two methods for improving performance of a HMM and their application for gene finding. Proceedings of the Fifth International Conference on Intelligent Systems for Molecular Biology[C]. Menlo Park CA: AAAI Press, 1997. 179-186.

[17] FICKETT J W. Finding genes by computer: the state of the art? [J]. Trends Genet, 1996, 12: 316-320.

[18] KROGH A. A hidden Markov model that finds genes in *E. coli* DNA[J]. Nucleic Acids Research, 1994, 22: 4768-4778.

[19] HENDERSON J, SALZBERG S, FASMAN K H. Finding genes in DNA with a hidden Markov model[J]. Journal of Computational Biology, 1997, 4: 127-141.

[20] BORODOVSKY M, MCININCH J. Gene mark: Parallel gene recognition for both DNA strands[J]. Computers and Chemistry, 1993, 17: 123-133.