文章编号:1001 - 2486(2006)01 - 0051 - 06

# 基因短序列模式分析及其在 5′剪接位点识别中的应用*

晏　春,杜耀华,王正志

(国防科技大学 机电工程与自动化学院,湖南 长沙　410073)

摘　要:短序列模式分析是基因序列分析的一个重要组成部分,在进行生物信号识别的时候,一般都会利用到短序列模式的信息。通常短序列模式的数目很多,如果每个都应用到生物信号识别中,会产生大量的参数,而且无法体现信号的主要特征。为了找出在识别信号位点中起关键作用的短序列模式,以信息增益作为评价依据,按照逐步选择的策略,将模式进行排队。根据排队结果,选取信息增益突出的短序列模式作为识别生物信号的关键依据,这样可以用较少的模式得到较好的结果。结合选取的短序列模式,用最大熵模型作为信号序列真实分布的估计,从而对给定序列进行识别。最后将这个方法用于 5′剪接位点的识别,得到了满意的结果。

关键词:5′剪接位点识别;模式分析;最大熵模型

中图分类号:Q52　　文献标识码:A

# Analysis of Short Sequence Motifs with Applications to 5′Splice Sites Identification

YAN Chun, DU Yao-hua, WANG Zheng-zhi

(College of Mechatronics Engineering and Automation, National Univ. of Defense Technology, Changsha 410073, China)

**Abstract**: Analysis of short sequence motifs is an important component of gene sequence analysis. Information of motifs is usually used for identifying biological signals. However, the number of short sequence motifs is very large. If all of them are used for signal identification, there will be too many parameters, thus covering the main characteristics of the signal. To find out the key short sequence motifs for signal identification, in this paper, a stepwise strategy was adopted to rank motifs by their information gain. As a result, the motifs were selected orderly for signal identification. In so doing, good results were achieved with fewer motifs. Consisted with the selected motifs, maximum entropy model was used as the approximation of the true distribution of the signal sequences, thus realizing the identification of a given sequence. Finally, the model was used to identify 5′splice sites, and approving experiment results were achieved.

**Key words**: 5′splice sites identification; motif analysis; maximum entropy model

随着人们开始利用计算机对生物序列进行功能分析,生物信息学研究又向前迈进了一大步。如何把信号序列和背景序列区分开来,是对生物序列进行功能分析的关键。通常人们认为具有相同功能的基因序列应该具有一定的相似性,因此有很多统计学习方法应用到这一领域,常见的方法有 Markov 模型(MM)[1]、隐 Markov 模型(HMM)[2]和支持向量机(SVM)[3]等等。这些方法通过统计得到信号序列的特征概率分布,从而对信号进行识别。为了进一步分析模式信息在信号序列识别中的作用,Yeo 和 Burge 提出用最大熵模型(MEM)对短序列模式进行分析[4]。本文根据已知序列统计出短序列模式分布的边缘约束,结合这些约束,采用最大熵模型对信号序列的分布进行最小无偏估计。分别基于信号序列和背景序列,可以得到两个不同的分布,由此可以计算出给定序列分别属于信号序列和背景序列的概率,根据这两个概率的比值,对给定序列的类别进行判断。采用最大熵模型的优越性就在于,不必对序列分布提出更多的前提假设,仅仅利用从序列中得到的经验分布作为约束。这个方法在自然语言识别

* 收稿日期:2005 - 09 - 12
基金项目:国家自然科学基金资助项目(60471003)
作者简介:晏春(1979—),女,博士生。

和数据库模式提取[6]的应用中得到了较好的结果。

短序列模式多种多样,数目很大,而我们对生物信号进行分析的目的是找出能够反映其核心特征规律的模式。因此,我们需要定量描述短序列模式在信号识别中的作用。一般认为,短序列模式提供的信息越多,序列的熵就会减少得越多,我们把这样的模式称为强模式。因为模式之间并非是独立的,不同的模式组合方式会产生不同的信号序列描述。为了更合理地描述信号序列,与文献[4]中的贪婪搜索算法不同,本文采用逐步搜索算法,把短序列模式按照信息增益的大小,由强到弱进行排队。

最后,我们以5'剪接信号的识别为例,采用多种短序列模式对剪接位点类型进行识别。识别的结果与采用的短序列模式有很大的关系;"好的"模式集合体现了信号序列的主要特征。根据DNA序列数据计算得到的结果表明,本文介绍的方法能够较好地区分剪接信号和干扰信号,这也说明我们选取的模式能够较好地体现信号序列的特征。

# 1 方法

## 1.1 最大熵原理

定义 $x$ 是长度为 $\lambda$ 的序列变量,$x=\{x_1,x_2,\cdots,x_\lambda\}$,其中每一位是一个碱基。$s=\{s_1,s_2,\cdots,s_\lambda\}$ 是某个确定的序列,把序列变量 $x$ 取某个确定的序列 $s$ 的概率 $P(x=s)$ 简记为 $p(s)$。所谓的序列相等就是指两个序列长度相等,并且序列每个位上的碱基相同。我们把需要满足的某些概率分布叫做"约束"。最大熵原理最早由Jaynes提出[7],当对象空间的分布满足一系列的约束时,对空间真实分布的最佳近似是最大Shannon熵 $H$,其表达式为

$$H(p)=-\sum_{s}^{all}p(s)\log_2(p(s)) \tag{1}$$

Shannon熵是对自由变量 $X$ 平均不确定性的一种度量。在计算最大熵的时候,约束集的选择十分重要,必须保证估计出的分布是可信的。但是,这些约束可能会是不一致的,不能同时被满足,例如 $\{P(AA)=3/4,P(GG)=1/2\}$。在本文中,所取的约束是通过统计得到的序列经验分布中的一些边缘分布,因此都是一致的,不会出现冲突现象。

利用最大熵原理进行计算,我们最终可以得到两个分布:作为信号序列的 $p^+(x)$ 和作为背景序列的 $p^-(x)$。对于一个给定序列 $s$,判断其是否为信号序列的依据是它作为信号序列与作为背景序列的概率比值 $L$:

$$L(x=s)=\frac{p^+(x=s)}{p^-(x=s)} \tag{2}$$

当 $L(x=s)\geqslant C$ 时,认为 $s$ 是信号序列,否则认为其是背景序列。其中 $C$ 是门限值。

## 1.2 边缘约束

我们定义的约束有两种,一种是"完全"约束,一种是"特定"约束。因为序列中模式的功能分析不仅与模式的组成相关,同时与模式所处的位置也有很大的关系,因此我们所建立的约束都包含位置信息。

### 1.2.1 "完全"约束

把"完全"约束集合记为 $C_x$,它是对于序列中所有的短序列模式(包括位置信息)而言。以 $\lambda=3$ 为例:

$$C_x=\{p(x_1),p(x_2),p(x_3),p(x_1,x_2),p(x_2,x_3),p(x_1,x_3)\} \tag{3}$$

另外定义 $C_s^m\subseteq C_x$,其中 $m$ 表示短序列模式的长度,$s$ 表示短序列中碱基之间相隔的碱基数,称之为跨度。可以看出,当仅仅计算 $C_0^1$ 的时候,对应的最大熵模型等同于权矩阵模型(WMM)[8],当仅仅计算 $C_0^2$ 的时候,等同于1阶Markov模型。因此通过取不同短序列模式,可以得到不同的模型。

### 1.2.2 "特定"约束

"特定"约束就是"完全"约束中的某些特定的短序列模式的观察概率。例如 $p(x_1,x_3)$ 的"特定"约束就是 $\{ANA,ANC,\cdots,TNG,TNT\}$ 的观察概率 $\{p(ANA),p(ANC),\cdots,p(TNG),p(TNT)\}$,其中 $N$ 表示

为任意碱基。在进行序列信号识别的时候,通常只考虑一些关键的短序列模式。

### 1.3　迭代计算最大熵分布(MED)

包含约束集的最大熵分布的计算可以由一个迭代过程来实现[9-10]。迭代起始于一个均匀分布 $p(x)^0 = 4^{-\lambda}$,即所有序列出现的概率相等。然后建立一组完全约束和对应的特定约束。特定约束集合中的元素记为 $Q_i(i = 1,2,\cdots,n)$。第 $j$ 步的迭代方程为

$$p^j = \begin{cases} p^{j-1}\dfrac{Q_i}{\widehat{Q}_i^{j-1}} & if \quad Q_i \in x \\ p^{j-1}\dfrac{1-Q_i}{1-\widehat{Q}_i^{j-1}} & else \end{cases} \tag{4}$$

其中 $Q_i \in x$ 表示 $Q_i$ 对应的短序列模式在 $x$ 中出现,$p^{j-1}$和 $p^j$ 分别是迭代时第 $j-1$ 和第 $j$ 步的概率。$\widehat{Q}_i^{j-1}$ 是根据第 $j-1$ 步的分布概率得到的第 $i$ 个特定约束的取值。

$$\hat{Q}_i^{j-1} = \sum p^{j-1}(x) \quad \forall x, if \quad Q_i \in x \tag{5}$$

重复迭代过程,直到得到的最大熵分布没有明显的改变,结束迭代。这样就得到了序列真实分布的近似估计。

### 1.4　短序列模式排队

在计算最大熵分布的时候,随着迭代的进行,$H$ 将从 $2\lambda$ 下降到 MED 下的 Shannon 熵。这是因为加入的短序列模式约束越多,带来的信息越多,序列的不确定性将下降,因而熵会减小。

不同的短序列模式提供的信息量不同,因而对信息增益的作用也不同。我们根据其提供的信息增益,将各个短序列模式进行排队。排队的策略有很多,文献[4]中采取的是贪婪搜索算法,它其实是一种向前选择的算法,变量一旦被选中,将永远被留在模型中。然而,由于变量之间并不是独立的,而是存在相互传递的关系,随着其他变量的引入,一些先前被选中进入模型的变量的解释作用可能会变得不再显著。

为了克服上述问题,我们采取了逐步搜索策略,它是一个边进边退的方法。对于被剔除出模型的变量,只要后来它又可以提供显著的解释信息,就可以再次进入模型;而对于已选中的变量,一旦它提供的解释信息变得微弱,就可以被剔除。具体实现过程如下:

初始化:

(1) 定义两个集合:$G_r$ 和 $G_b$。$G_r$ 最开始为空,$G_b$ 包括所有的特定约束。

(2) 定义两个门限,$C_i$ 和 $C_o$。$C_i$ 是选入特征时的门限值,$C_o$ 是剔除特征时的门限值。

(3) 初始为均匀分布,熵值为 $2\lambda$。

(4) 逐个取 $G_b$ 中的约束 $Q_i$,计算它们对应的熵减 $\Delta H_i$,找出使得 $\Delta H_i$ 值最大的约束,比如说是 $Q_k$,将 $Q_k$ 加入到 $G_r$ 中,并且从 $G_b$ 中去除 $Q_k$,$r = 1$。

迭代开始:

(5) 初始为均匀分布,熵值为 $2\lambda$。

(6) 计算由 $G_r$ 中所有约束得到的 MED。

(7) 逐个取 $G_b$ 中的所有约束,在第(6)步确定的分布概率条件下,分别计算它们对应的熵减 $\Delta H_i$,取大于 $C_i$ 中 $\Delta H_i$ 值最大的约束,将其加入到 $G_r$ 中,并且从 $G_b$ 中去除。$r = r + 1$。

(8) 计算 $G_r$ 中任意 $r-1$ 个约束得到的 MED,再计算加入剩下一个约束后得到的 MED,如果 $\Delta H_i$ 小于 $C_o$,将这个约束从 $G_r$ 中去掉,回放到 $G_b$。$r = r - 1$。

重复(5)-(8)步,直到 $G_b$ 中的约束不再转移到 $G_r$ 中。为了避免变量的进出循环,决策的两个门限值应该满足 $C_i > C_o$。

## 2 剪接位点识别数据

文中采用的剪接数据是文献[1]中提供的,它是从 EXON – INTRON 数据库[11]中提取的有实验验证的全注释人类基因序列。其中包括 1115 条基因,有 5733 条真实的 5 '剪接位点数据和 478 983 条虚假的 5 '剪接位点数据。所有的剪接位点都符合 GT 规则,即 5 '剪接位点的 1 和 2 位置上的碱基分别是 G 和 T,生物的剪接方式绝大部分都满足这个规则。定义序列中满足 GT 规则,但是没有被标注为剪接位点的地方是虚假位点。将数据分为两部分,其中一部分作为训练数据,剩下的一部分作为测试数据,具体分割如表 1。

表 1 训练集和测试集中的序列数目

Tab.1 Number of sequences in train and test sets

| | True site | False site |
|---|---|---|
| Train | 3841 | 320918 |
| Test | 1892 | 158065 |
| Total | 5733 | 478983 |

## 3 实验结果和分析

图 1 给出了用 Sequences logos[12]对 5 '剪接位点附近的碱基分布做出的量化和可视化分析结果。可以看到 5 '剪接位点在 – 3 到 + 6 的位置具有明显的碱基偏好性,因此在对 5 '剪接位点进行分析的时候,取 – 3 到 6 的窗口作为研究对象。

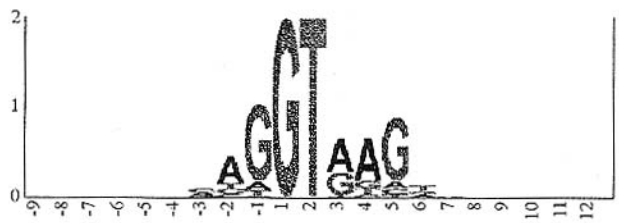


图 1 5 '剪接位点的位置偏好 logo 图

Fig.1 Logos of position bias of 5 'splice sites

根据(2)式可以看出,通过调整 $C$ 值,可以得到不同的识别结果。为了分析 $C$ 值对识别结果的影响,图 2 给出了采用不同短序列模式得到的 ROC(receiver operating curve)分析[13]结果,为了和其他方法相比较,图 2 也给出了 MDD 方法[11]得到的结果。其中 me(·,·)中的 me 表示最大熵,第一个参数表示短序列模式的长度,第二个参数表示短序列模式中碱基之间的跨度,其中第二个参数允许有多个取值。图 2 中的曲线是取不同 $C$ 值时,敏感性($Sn$)和特异性($Sp$)的对应关系,可以看出特异性和敏感性两个指标不能同时达到理想值。当特异性指标提高的时候,敏感性指标就会下降。在 ROC 图中,越靠近左上的曲线,对应的模型精确度越高。为了将特异性指标和敏感性指标综合起来考虑,表 2 列出了各个模型得到的相关系数($CC$)值。

定义 $TP$ 为真阳性的数目,$TN$ 为真阴性的数目,$FP$ 为假阳性的数目,$FN$ 为假阴性的数目。可见 $TP$ 和 $TN$ 都是正确的判断,而 $FP$ 和 $FN$ 都是错误的判断。定义 $Sn$ 是敏感度,$Sp$ 是特异性,$CC$ 是相关系数。有

$$Sn = \frac{TP}{TP + FN} \tag{6}$$

$$Sp = \frac{TP}{TP + FP} \tag{7}$$

$$CC = \frac{(TP * TN) - (FN * FP)}{\sqrt{(TP + FN) * (TN + FP) * (TP + FP) * (TN + FN)}} \tag{8}$$

**表 2　基于不同最大熵模型的剪接位点识别结果**

Tab.2　Identification results based on different MEM

| 5′splice site | |
|---|---|
| Models | *CC* |
| me(1,0) | 0.809459 |
| me(2,1) | 0.823753 |
| me(4,0) | 0.845357 |
| me(2,(1 – 2)) | 0.853458 |
| me(2,(1 – 3)) | 0.857254 |
| me(3,0) | 0.862597 |
| me(2,(1 – 4)) | 0.862791 |
| me(2,0) | 0.863527 |
| me(2,(1 – 5)) | 0.864650 |

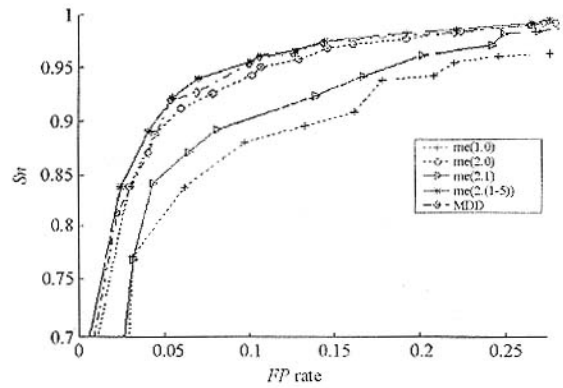


**图 2　5′剪接位点的 ROC 分析**

Fig.2　ROC analysis of 5′splice sites

**由表 2 可以看到,对于 5′剪接位点,me(2,(1 – 5))得到的结果最好,me(2,0)与 me(2,(1 – 5))的性能比较接近,得到的结果要好于 me(1,0)和 me(3,0)。用前面介绍的方法分别对 5′剪接位点附近的短序列模式进行排队,得到的部分结果如表 3 所示。为了观察排队后短序列模式对 5′剪接位点识别的影响,在 me(2,0)模型下,按照短序列模式的排队顺序,依次将它们用于位点识别,得到的结果如图 3 和图 4 所示。图 3 是随着模式的加入,信息增益的上升曲线,图 4 是随着模式的加入,*CC* 值的上升曲线。由图 3 和图 4 可以看到,当采用少量相同数目的短序列模式时,按照排队后顺序选取短序列模式的性能要远远好于按照随机顺序选取短序列模式的性能(前者提供的信息量和识别相关系数都上升得很快)。这说明,强模式已经体现了剪接位点的绝大部分特征,弱模式提供的信息几乎可以忽略不计。**

**表 3　关于 5′剪接位点的 me(2,0)和 me(2,(1 – 5))模式排列结果**

Tab.3　Rank result of me(2,0) and me(2,(1 – 5)) for 5′splice sites

| me(2,0) | | me(2,(1 – 5)) | |
|---|---|---|---|
| rank | $\Delta H_i$ | rank | $\Delta H_i$ |
| 1 | · · · g t · A G · | 1 | · · G g t · · G · |
| 2 | · A G g t · · · · | 2 | · · G g t · A · · |
| 3 | · · G g t A · · · | 3 | · A · g t · A · · |
| 4 | · · · g t · · G T | 4 | · A · g t A · · · |
| 5 | · · G g t G · · · | 5 | · · · g t · A T · |
| 6 | · · · g t A A · · | 6 | · · · g t A · · T |
| 7 | · · · g t G A · · | 7 | A · · g t A · · · |
| 8 | · · · g t T A · · | 8 | · · · g t C · G · |
| 9 | · · · g t C A · · | 9 | · · · g t T · G · |
| 10 | C A · g t · · · · | 10 | C · · g t · · G · |
| 11 | · · · g t · G T · | 11 | · · T g t · · A · |
| 12 | · · · g t G T · · | 12 | · · C g t · · A · |
| 13 | · · · g t · G A · | 13 | · · C g t · · C · |
| 14 | · · · g t C C · · | 14 | · · C g t · · T · |
| 15 | · · · g t · G C · | 15 | · · T g t · · T · |
| 16 | · · · g t · T C · | 16 | · · A g t · · A · |
| 17 | · · · g t G C · · | 17 | · · A g t · · T · |
| 18 | · · · g t · T A · | 18 | · · T g t · · C · |
| 19 | · T C g t · · · · | 19 | T · G g t · · · · |
| 20 | · · · g t · T C · | 20 | · · A g t · · C · |

**注:小写的字母表示 5′剪接位点的 gt,大写的字母表示短序列模式。**

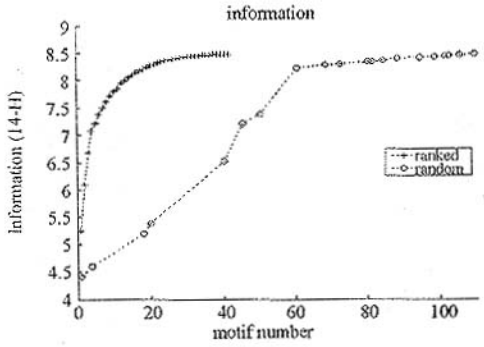


**图3 模式选取和信息增益的对应关系**
Fig.3 The relationship between motifs selection and information gain

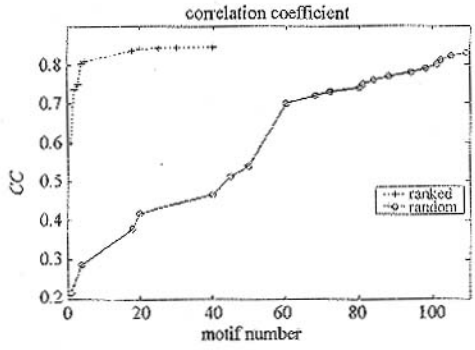


**图4 模式选取和 *CC* 值的对应关系**
Fig.4 The relationship between motifs selection and *CC* value

## 4 结论

包含短序列模式约束的最大熵分布是对序列实际分布的一种最小无偏估计,根据在不同类型训练数据下估计出的各种不同分布,我们可以对测试数据进行识别。同时通过采用各种不同的短序列模式,可以建立不同复杂程度的分类模型来适应实际情况的需要。为了分析各个短序列模式在信号序列识别中的重要性,本文依据信息增益,采用逐步搜索算法对这些短序列模式进行排队。通过与随机排列的短序列模式进行比较可以发现,采用排队后的模式可以很快地提高序列集的信息,选取较少的强模式就可以得到很好的识别结果。通过实验验证可以看到,我们的方法可以很好地识别 5′剪接位点,其敏感性和特异性都可以达到 90% 以上。同时这个方法也可以解决其他生物信号序列的模式识别问题。

## 参考文献:

[1] Pertea M, Lin X Y, Salzberg S L. GeneSplicer: a New Computational Method for Splice Prediction[J]. Nucleic Acides Research, 2001, 29(5): 1185 – 1190.

[2] Durbin R, Eddy S, Krough A, et al. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids[M]. Cambridge University Press, 1988.

[3] Zien A, Rötsch G, Mika A, et al. Engineering Support Vector Machine Kernels that Recognize Translation Initiation Sites[J]. Bioinformatics, 2000, 16(9): 799 – 807.

[4] Yeo G, Burge C. Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals[A]. RECOMB '03, 2003, Berlin, Germany.

[5] Berger A, Pietra S, Pietra V. A Maximum Entropy Approach to Natural Language Processing[J]. Computational Linguistics, 1996, 22(1): 39 – 71.

[6] Krogh A, Mitchsion G. Maximum Entropy Weighting of Aligned Sequences of Proteins or DNA[A]. In Proceeding of Intelligent system for molecular biology, 1995, 215 – 221.

[7] Jaynes E. Information Theory and Statistic Mechanics[J]. Physics Review, 1957, 106: 620 – 630.

[8] Burge C. Modeling Dependencies in Pre-mRNA Splicing Signals[A]. Salzberg S L, Searls D B, Kasif S (Eds.). Computational Methods in Molecular Biology, Elsevier Science, 1998, 129 – 164.

[9] Brown D. A Note on Approximations to Discrete Probability Distributions[J]. Information and Control, 1959, 2: 386 – 392.

[10] Lewis P. Approximating Probability Distributions to Reduce Storage Requirements [J]. Information and Control, 1959, 2: 214 – 225.

[11] Saxonov S, Daizadeh I, Fedorov A, et al. An Exhaustive Database of Protein-coding Intron-containing Genes[J]. Nucleic Acids Research, 2000, 24: 3439 – 3452.

[12] Thomas D, Schneider R, Stephens M. Sequence Logos: a New Way to Display Consensus Sequences[J]. Nucleic Acids Res., 1990, 18: 6097 – 6100.

[13] Swets J. Measuring the Accuracy of Diagnostic Systems[J]. Science, 1988, 240(4857): 1285 – 1293.