微生物学报
Acta Microbiologica Sinica
2017, 57(10): 1493-1503
http://journals.im.ac.cn/actamicrocn
DOI: 10.13343/j.cnki.wsxb.20160442

Research Article 研究报告

原核生物全基因组中 16S rRNA 基因的识别

闫文凯[#]，许明敏[#]，张广乐，乔宁，徐炜娜，陈园园，张良云[*]

南京农业大学理学院，江苏 南京 210095

摘要：【目的】识别原核生物全基因组中的 16S rRNA 基因。【方法】本文依据基因序列的 GC 碱基含量、碱基 3-周期性和马尔可夫链 3 个方面的特性，构建了识别原核生物全基因组中 16S rRNA 基因的三层过滤模型。【结果】经检验，模型的特异性、敏感性和马修斯相关系数分别为 99.58%、91.60%和 91.49%。【结论】结果表明，本文所提出的方法可以高效、准确地识别出 16S rRNA 基因。

关键词：16S rRNA 基因，GC 碱基含量，碱基 3-周期性，马尔可夫链

细菌的系统分类学研究中最常用的分子标记是 rRNA，它是核糖体的基本组成成分。rRNA 的基因按 5′-16S-23S-5S-3′方式排列，被 2 个非编码间隔区序列所分开，16S rDNA、23S 和 5S rDNA 三部分组成一个操纵子，作为一个单位进行转录，转录后处理成为成熟的 16S、23S、5S rRNA[1]。由于其种类少、含量大，并且存在于所有的生物中，既具有保守性又具有高变性，被广泛用于微生物分子差异与遗传特征的研究。16S rRNA 基因约 1.5 kb 左右，大小合适，含有高度保守的片段，同时在不同的菌株间也含有变异的片段。既能体现相同菌种之间的相似性，又能体现不同菌种之间的差异性，因此是常用于细菌分类与鉴定的分子标记[2]。

基因预测是生物信息学领域的一个重要研究方向，是研究生物遗传、进化等工作的基础，其目的是对 DNA 序列中的功能性基因、调控元件等进行注释。许多统计概率方法及机器学习方法被用于此类研究工作中。如 Zhao 等[3]采用最大熵隐马尔可夫模型，基于 TATA、CAAT 框等启动子信号元件来识别启动子。Li 等[4]利用序列的组分特征及位点关联特征，并结合支持向量机对基因剪接位点识别等。目前，针对 16S rRNA 基因进行预测的算法及软件，如 RNAmmer[5]采用 HMMer-2 模型对原核生物的 16S rRNA 基因进行识别，该方法只适用于全基因组序列。Meta-RNA[6]是一个 Python 程序，可对宏基因组片段序列中 16S rRNA 序列进行筛选。rRNASelector[7]是以隐马尔可夫模

基金项目：国家自然科学基金(11571173)；江苏省自然科学基金(BK20141358)

[*]**通信作者**。Tel：+86-25-84396063；E-mail：zlyun@njau.edu.cn

[#]并列第一作者。

收稿日期：2016-10-14；**修回日期：**2016-11-24；**网络出版日期：**2017-02-20

型为基础,应用 Java 编写的 rRNA 基因预测软件。由于隐马尔可夫算法的初始参数是依据已知数据训练得出，因此 rRNA Selector 的预测结果受所选训练数据的影响较大。

为了快速、高效地对原核生物全基因组中的 16S rRNA 基因进行识别，本文提出一种集成算法。实验将真实的 16S rRNA 基因序列与对照序列(非 16S rRNA 基因序列)进行了 GC 碱基含量、序列碱基 3-周期性和序列的马尔可夫链模型 3 个方面对比分析。通过分析，并针对每个方面的数据设定出了一个阈值对待选序列进行判别。经过 3 个方面分析、3 种方法处理、3 层过滤手段，构建出一种准确、高效的筛选 16S rRNA 基因序列模型。

1 材料和方法

1.1 材料

本实验选取的 16S rRNA 基因序列与细菌全基因组序列来自 NCBI 数据库(http://www.ncbi.nlm.nih.gov/)。其中，原核生物 16S rRNA 基因序列共 41252 条,实验从中随机选取序列碱基缺失个数小于 50 的基因序列，共 9000 条。非 16S rRNA 基因序列从 50000 个菌株的全基因组序列中随机选取，长度为 1550 bp。训练集中，16S rRNA 基因序列选用 3000 条作为正集，非 16S rRNA 基因序列选用 3000 条作为对照负集。测试集中，16S rRNA 基因序列选用 6000 条作为正集，非 16S rRNA 基因序列选用 6000 条作为负集。

1.2 序列 GC 含量

GC 含量是指所研究的 DNA 序列中鸟嘌呤与胞嘧啶两种碱基所占比例。在 DNA 双链结构中，G 和 C 以 3 个氢键配对,与 DNA 链的稳定性密切相关。因此，GC 含量是 DNA 序列碱基组成的重要特征，蕴含基因结构、功能和进化信息。对于功能性基因或特定的 DNA、RNA 序列，其 GC 碱基的含量有相对固定比例，利用这一特性可对其进行预测和识别。

1.3 序列碱基 3-周期性

信号频谱分析通常利用离散傅里叶变换对信号进行离散处理，构建功率谱和信噪比，该方法在各领域取得了重要研究成果。Berryman 等[8]在编码序列识别问题中引入了信号频谱分析的方法，基于离散傅里叶变换，给出了功率谱的定义。大量的研究发现，对 DNA 序列进行数值化映射和傅里叶变换后得到的功率谱，编码序列的功率谱在 1/3 处具有较大的峰值,而非编码序列却没有类似的峰值，故把这种统计现象叫做碱基的 3-周期性[9]。3-周期性是生物长期演化后形成的本质属性，是 DNA 序列重要的生物特征。目前，已有较多关于基因 3-周期性方法的创新与拓展。为使得到的数值序列具有更多的生物信息，许多学者对数值化方法进行研究。如 Voss 映射[10−11]、Z-Curve 映射[12]、复数法映射[13]、实数法[14]等。人们定义了信噪比 R，用来衡量 3-周期性的强弱，且大量研究表明特殊碱基序列的 R 通常大于或小于某个阈值 R_0。很多研究者给出了阈值 R_0 的确定方法，并提出了基因识别的算法[15]。基因序列的数值化映射、功率谱和信噪比的详细定义如下。

对基因序列进行数值化映射(公式 1)，令 $I=\{A,T,G,C\}$，现对任意确定的 $b\in I$,

$$u_b[b]=\begin{cases}1, & S[n]=b\\ 0, & S[n]\neq b\end{cases},\quad n=0,1,2,...,N-1 \quad \text{公式(1)}$$

其中，$S[n]$是所取得长度为 N 的碱基序列，n=0,1,2,⋯,N–1。映射(1)称为 Voss 映射，生成相应的 4 个二进制序列 $\{u_b[n]\}$: $u_b[0]$, $u_b[1]$, ...,

$u_b[N-1]$, $b \in I$ 称为 DNA 序列的指示序列。如，给定一段 DNA 序列为 S = ATCTCACTGGT，则：$u_A = \{1,0,0,0,0,1,0,0,0,0,0\}$, $u_T = \{0,1,0,1,0,0,0,1,0,0,1\}$, $u_G = \{0,0,0,0,0,0,0,0,1,1,0\}$, $u_C = \{0,0,1,0,1,0,1,0,0,0,0\}$.

对指示序列进行离散 Fourier 变换(DFT)[16]，如公式(2)。得到 4 个长度均为 N 的复数序列 $\{u_b[k]\}$, $b \in I$。

$$U_b[k] = \sum_{n=0}^{N-1} u_b[n] e^{-j\frac{2\pi nk}{N}},\ k = 0,1,...,N-1 \quad \text{公式(2)}$$

复序列的二范数之和 $\{U_b[k]\}$，得到整个 DNA 序列 S 的功率谱序列 $\{p[k]\}$ [17](公式 3)。

$$p[k] = |U_A[k]|^2 + |U_C[k]|^2 + |U_G[k]|^2 + |U_T[k]|^2, \quad k = 0,1,...,N-1 \quad \text{公式(3)}$$

每给定一段 DNA 序列，通过公式(3)作出其对应的功率谱曲线，图 1 分别表示沙门氏菌 16S rRNA 基因序列和非 16S rRNA 基因序列的功率谱曲线。

在图 1 中，16S rRNA 基因序列的频谱图在三分之一处无峰值出现，而非 16S rRNA 基因序列的功率谱在三分之一处有较大的峰值(公式 4)。

$$R = \frac{p(\frac{N}{3})}{\overline{E}} \quad \text{公式(4)}$$

为 DNA 序列信噪比(Signal to noise ratio，SNR)[18]，其中 $\overline{E} = (\sum_{k=0}^{N-1} p[k]/N)$ 是给定 DNA 序列的频谱均值。

信噪比值的大小是给定 DNA 序列在 N/3 处的频谱峰值的大小的表征，即，16S rRNA 基因序列或非 16S rRNA 基因序列的 3-周期性的强弱，显然，16S rRNA 基因序列的信噪比值较小，而非 16S rRNA 基因序列的信噪比较大。对于一段 15 kb 的 DNA 序列，选取一个最优的 R_0 值，使尽可能多的非 16S rRNA 基因序列的信噪比大于 R_0，而 16S rRNA 基因序列的功率谱或信噪比小于 R_0。根据这个特征选定某个适当的阈值，通过信噪比和阈值的大小来判断待选序列是否为 16S rRNA 基因序列。若信噪比小于阈值，判断此序列为 16S rRNA 基因序列；若信噪比大于阈值，判断此序列为非 16S rRNA 基因序列。

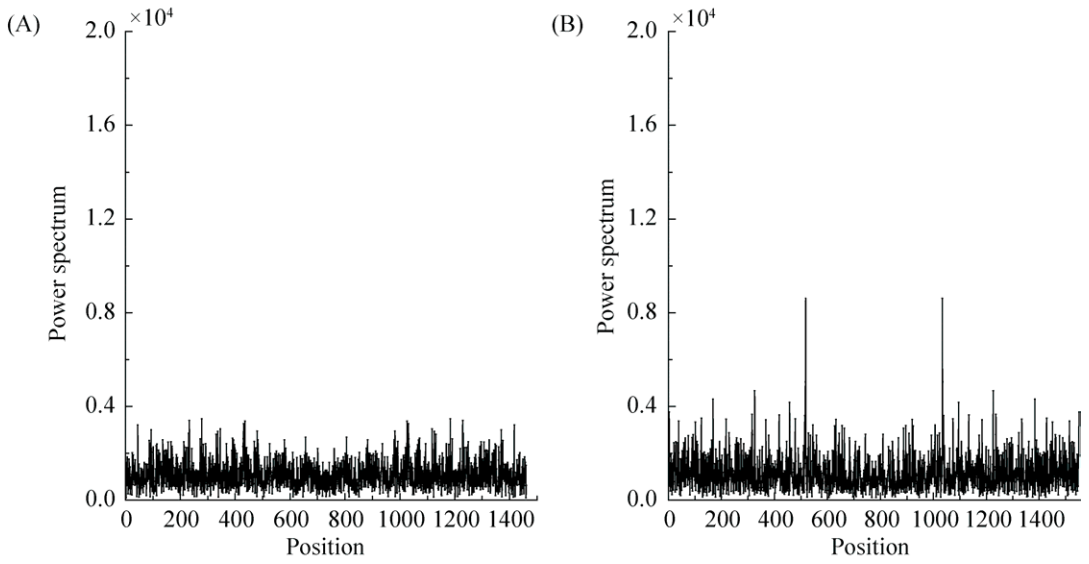

图 1. 16S rRNA 序列(A)与非 16S rRNA 序列(B)频谱图

Figure 1. Power spectrum curve of 16S rRNA sequence (A) and non-16S rRNA sequences (B).

1.4 马尔可夫模型

马尔可夫链是一种有效的概率模型。此模型已广泛应用于生物学领域，特别是在预测模型中有着深入的应用。如 DNA 序列的聚类研究中[19]，遗传与进化分析[20]，基于 RNA-seq 数据、甲基化数据以及拷贝数变异数据等构建遗传网络[21]等。马尔可夫过程[22–23]是一个无后效性的随机过程，即 t_m 时刻所处状态的概率仅和 t_{m-1} 时刻的状态有关，而与 t_{m-1} 时刻之前的状态无关。马尔可夫过程中的时间和状态可以是连续的，也可以是离散的。其中，时间离散、状态离散的马尔可夫过程为马尔可夫链。

马尔可夫模型应用于 16S rRNA 基因序列分析需要建立能够体现核酸序列的模型，其中的主要工作是构建转移概率矩阵 A。本文建立的概率模型由两条马尔可夫链组成，这两条马尔可夫链即为 16S rRNA 基因序列模型和非 16S rRNA 基因序列模型。通过计算待选序列在两个序列模型出现的概率来对其所属类别进行判定。出现的概率越大，说明序列内碱基状态转移模式最贴合相应的概率模型所生成的序列。即，待选序列在此概率模型出现的概率最大，则待选序列判定为此概率模型下的序列。转移概率按照公式(5)计算。

$$a_{st}=\frac{c_{st}}{\sum c_{st'}} \qquad \text{公式(5)}$$

其中 a_{st} 是状态 s 到状态 t 的转移概率，c_{st} 是对应概率模型的概率转移矩阵中元素 st 二元组的概率，$c_{st'}$ 是对应概率模型的概率转移矩阵中以 s 开头的二元组的概率。

对于本实验的马尔可夫模型，长度为 L 的待选序列 $X=\{x_1,x_2,...,x_L\}$，依据概率转移矩阵，对应的马尔可夫模型(简称模型 T)所产生的概率按照公式(6)计算。

$$P(X|T)=\prod_{i=1}^{L-1}a_{x_ix_{i+1}} \qquad \text{公式(6)}$$

其中，$P(X|T)$ 表示序列 X 由模型 T 产生的概率，i 是随机过程中的 i 时刻，$a_{x_ix_{i+1}}$ 是模型在 i 时刻产生状态 x_i，并且在 i+1 时刻产生状态 x_{i+1} 的概率，即，x_i 到 x_{i+1} 的转移概率。

设待选序列为 $X=\{x_1,x_2,...,x_L\}$，X 由 16S rRNA 基因序列模型所产生的概率(公式 7)。

$$P_{16S-rRNA}=P(X|T_{16S\ rRNA})=\sum_{i=1}^{L-1}\lg(a_{x_ix_{i+1}}) \qquad \text{公式(7)}$$

待选序列由非 16S rRNA 基因序列模型所产生的概率(公式 8)。

$$P_{non-16S-rRNA}=P(X|T_{non-16S\ rRNA})=\sum_{i=1}^{L-1}\lg(a_{x_ix_{i+1}}) \qquad \text{公式(8)}$$

$P_{16S-rRNA}$ 表示待选序列在 16S rRNA 基因序列碱基的转移概率矩阵条件下生成 16S rRNA 的概率，$P_{non-16S-rRNA}$ 表示待选序列在非 16S rRNA 基因序列碱基的转移概率矩阵条件下生成非 16S rRNA 的概率。若待选序列的生成概率设为 P，则 P 值的求解公式如公式(9)。

$$P=P_{16S-rRNA}-P_{non-16S-rRNA} \qquad \text{公式(9)}$$

1.5 方法流程图

任意 1 条待选序列经 3 层过滤，最终得到预测结果。流程图如图 2 所示。

1.6 评价指标

本文通过敏感度(Sensitivity，S_n)、特异度(Specificity，S_p)、马修斯相关系数(MCC)来衡量模型的优劣，见公式(10)、(11)、(12)。

$$S_n=\frac{TP}{TP+FN}\times 100\% \qquad \text{公式(10)}$$

$$S_p=\frac{TN}{TN+FP}\times 100\% \qquad \text{公式(11)}$$

$$MCC=\frac{TP\times TN-FN\times FP}{\sqrt{(TP+FN)\times(TN+FP)\times(TP+FP)\times(TN+FN)}} \qquad \text{公式(12)}$$

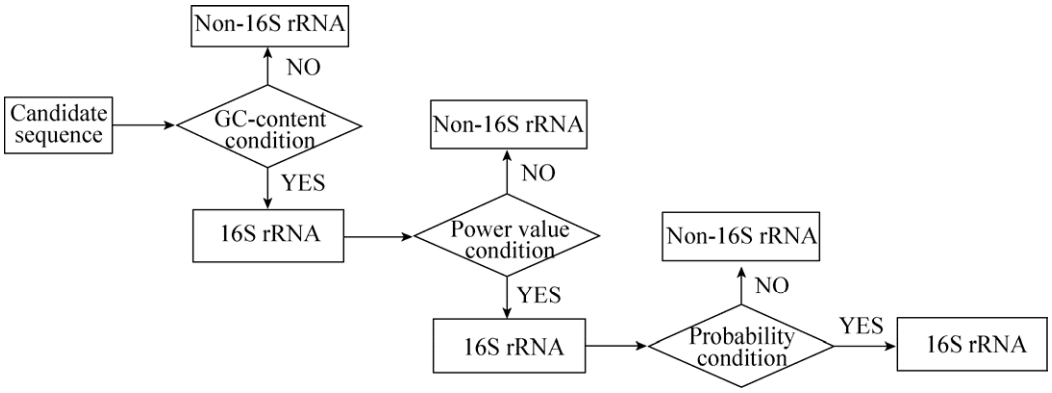

图 2. 实验流程图

Figure 2. Flow chart.

其中 *TP*、*TN*、*FP*、*FN* 分别表示真阳性、真阴性、假阳性和假阴性的数量。具体来说，*TP* 是 16S rRNA 基因序列并被识别为 16S rRNA 基因序列的数量；*TN* 是非 16S rRNA 基因序列并被识别为非 16S rRNA 基因序列的数量；*FP* 是非 16S rRNA 基因序列但是被识别为 16S rRNA 基因序列的数量；*FN* 是 16S rRNA 基因序列但是被识别为非 16S rRNA 基因序列的数量。

2 结果和分析

本文运算环境为 Matlab 2015b 版本。获取实验原始数据、运算程序等资源，请访问：http://www.yucetianxia.com/ywk/。

2.1 序列 GC 含量的分析

对正负集样本序列的 GC 碱基含量进行统计。结果显示，3000 条正样本的序列 GC 碱基含量百分比取值范围小，样本点聚集分布(图 3-A)；3000

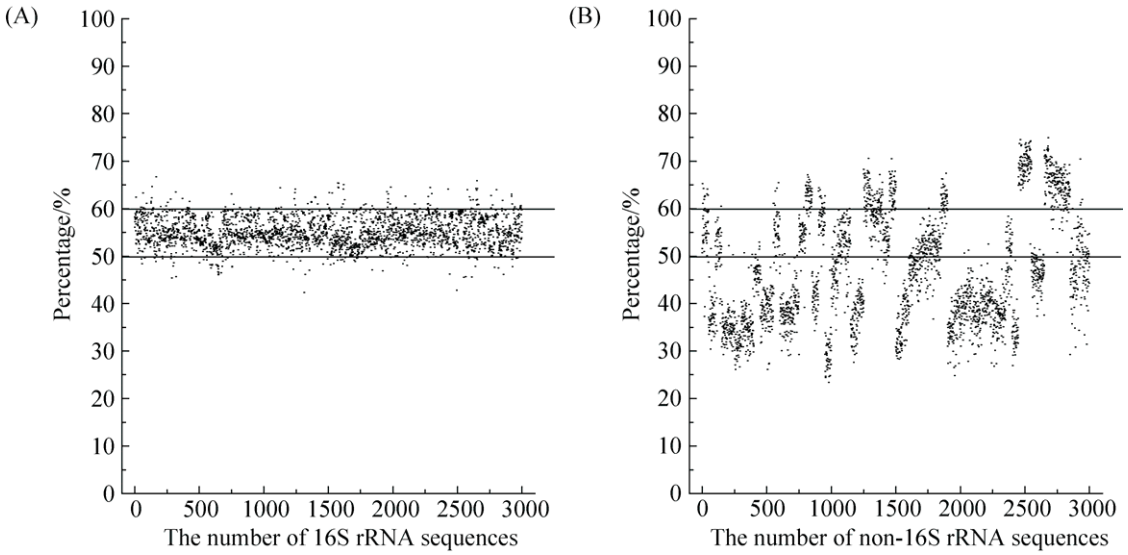

图 3. 16S rRNA 序列(A)和非 16S rRNA 序列(B) GC 含量分布

Figure 3. The GC content distribution of 16S rRNA sequences (A) and non-16S rRNA sequences (B). Horizontal straight line is GC_content=50% and GC_content=60%.

条负样本的序列 GC 碱基含量百分比取值范围大，样本点离散分布(图 3-B)。正样本集中序列的 GC 含量在 50%–60%之间的样本数高达 92.63% (图 4)，此区间之外的样本数为 7.37%。而负样本集的统计结果与正样本集有较大差异，GC 含量在 50%–60%之间的样本数仅占 21.87%，区间之外的样本数为 78.13% (图 4)。

以碱基序列的 GC 碱基含量作为第 1 个筛选标准，根据以上统计结果，若待选序列的 GC 碱基含量在 50%–60%区间内,则判定其为 16S rRNA 基因序列，在此区间之外，则判定为非 16S rRNA 基因序列。经此筛选后，正样本数量为 2779，负样本数量为 656。

2.2 序列碱基 3-周期性分析

对正负集样本的序列信噪比进行统计。结果显示，2779 条正样本与 656 条负样本的信噪比存在较大差异。正样本的信噪比值较小，呈规律性分布，一般在 0–5 区间内取值(图 5-A)；负样本的信噪比值较大，分布离散，且大部分样本的信噪比大于 5 (图 5-B)。正样本集中序列的信噪比在 0–5 区间的样本数高达 99.60%，此区间之外的样本数仅占 0.40% (图 6)。而负样本集的序列的信噪比在 0–5 区间的样本数仅占 14.63%，区间之外的样本数为 85.37% (图 6)。

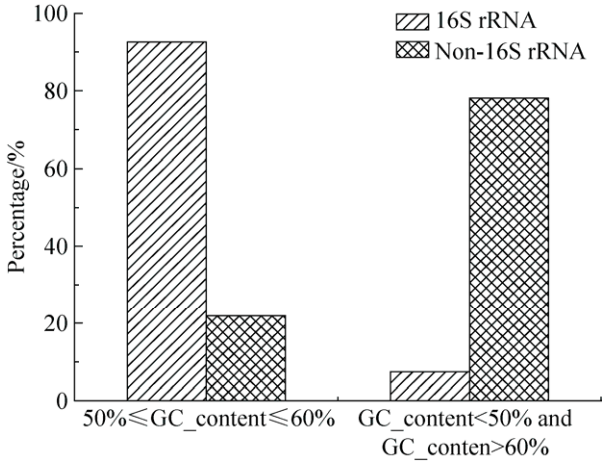

图 4. 以阈值为基础，16S rRNA 序列与非 16S rRNA 序列样本分布图

Figure 4. The distribution of 16S rRNA sequences and non-16S rRNA sequences based on the GC content threshold. The GC content threshold is 50%–60%.

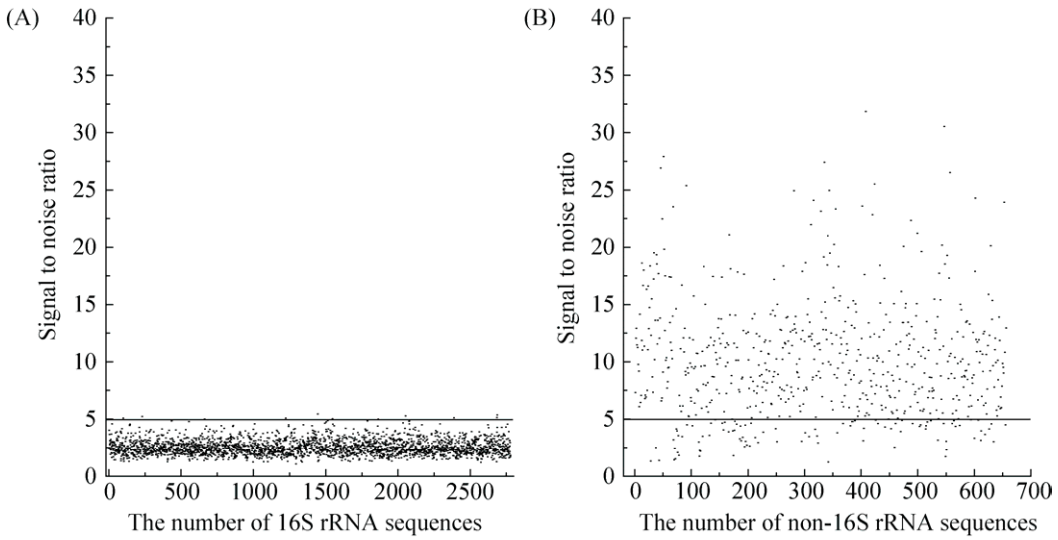

图 5. 16S rRNA 序列(A)和非 16S rRNA 序列(B)信噪比散点图

Figure 5. The SNR distribution of 16S rRNA sequences (A) and non-16S rRNA sequences (B). Horizontal straight line is R_0=5.

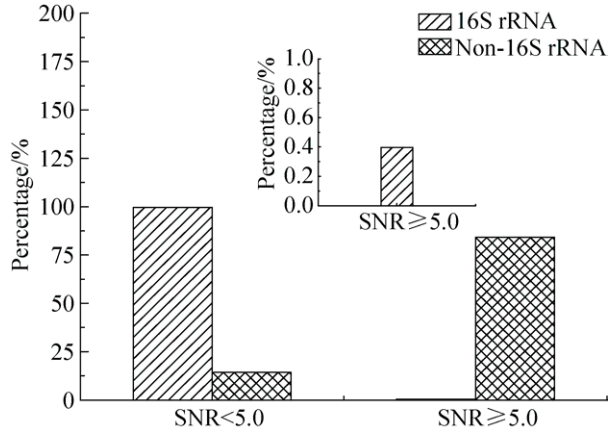

图 6. 以 SNR 阈值为基础的 16S rRNA 序列与非 16S rRNA 序列样本分布图

Figure 6. The distribution of 16S rRNA sequences and non-16S rRNA sequences based on the SNR threshold. The SNR threshold is 5.0; Sub-map is the ratio of 16S rRNA sequences that the SNR greater than 5.0.

以碱基序列的信噪比作为第 2 个筛选标准，依据上述统计结果，若待选序列的信噪比小于 5，则判定其为 16S rRNA 基因序列；若待选序列的信噪比大于等于 5，则判定为非 16S rRNA 基因序列。经此进一步筛选后，正样本数量为 2768，负样本数量为 96。

2.3 基于马尔可夫模型对序列进行分析

以初始的正负样本为训练数据，得 16S rRNA 基因序列碱基转移概率矩阵和非 16S rRNA 基因序列碱基转移概率矩阵如表 1、表 2 所示。

表 1. 16S rRNA 基因序列碱基的转移概率矩阵

Table 1. The transition probability matrix of 16S rRNA gene sequence

Transition probability	A	C	G	T
A	0.280272	0.244106	0.296359	0.179264
C	0.236400	0.237846	0.309018	0.216737
G	0.234274	0.232156	0.321973	0.211597
T	0.238288	0.208642	0.356803	0.196267

表 2. 非 16S rRNA 基因序列碱基的转移概率矩阵

Table 2. The transition probability matrix of non-16S rRNA gene sequence

Transition probability	A	C	G	T
A	0.325077	0.1894242	0.205089	0.280393
C	0.279231	0.2339100	0.254084	0.232775
G	0.249651	0.3011530	0.233852	0.215343
T	0.211117	0.2183440	0.245791	0.324749

以 16S rRNA 基因序列和非 16S rRNA 基因序列碱基转移概率矩阵为基础，计算得出 2768 条正样本与 96 条负样本的 *P* 值。统计结果显示，*P* 值大于 20 的正样本数较多(图 7-A)，而负样本中 *P* 值小于 20 的样本点为多数(图 7-B)。正样本集中序列的 *P* 值大于 20 的样本数所占比例为 96.48%，*P* 值小于 20 的样本数仅占 3.52% (图 8)。而负样本集中序列的 *P* 值小于 20 的样本数占 85.42%，*P* 值大于 20 的样本所占比例为 14.58% (图 8)。

以碱基序列的 *P* 值作为第 3 个筛选标准，依据上述统计结果，若待选序列的 *P* 值大于等于 20，则判定其为 16S rRNA 基因序列；若待选序列的 *P* 值小于 20，则判定为非 16S rRNA 基因序列。经此筛选后，正样本数量为 2708，负样本数量为 17。

2.4 基于测试数据对模型进行分析

本文通过对训练数据的碱基组分、碱基 3-周期性以及序列马尔可夫性 3 方面的分析，得出 GC 碱基含量、信噪比和序列生成概率 3 个阈值，以此为条件构建出筛选模型。实验应用测试数据，通过 3 种评价指标对模型性能进行评价。

测试数据为 6000 条 16S rRNA 基因序列，6000 条非 16S rRNA 基因序列。实验中测试数据分为 3 组，每组由 2000 条 16S rRNA 基因序列和 2000 条非 16S rRNA 基因序列组成。经过上述三级筛选过程，结合公式(10)、(11)及(12)可以计算出该模型的敏感性、特异性和马修斯系数，如表 3 所示。

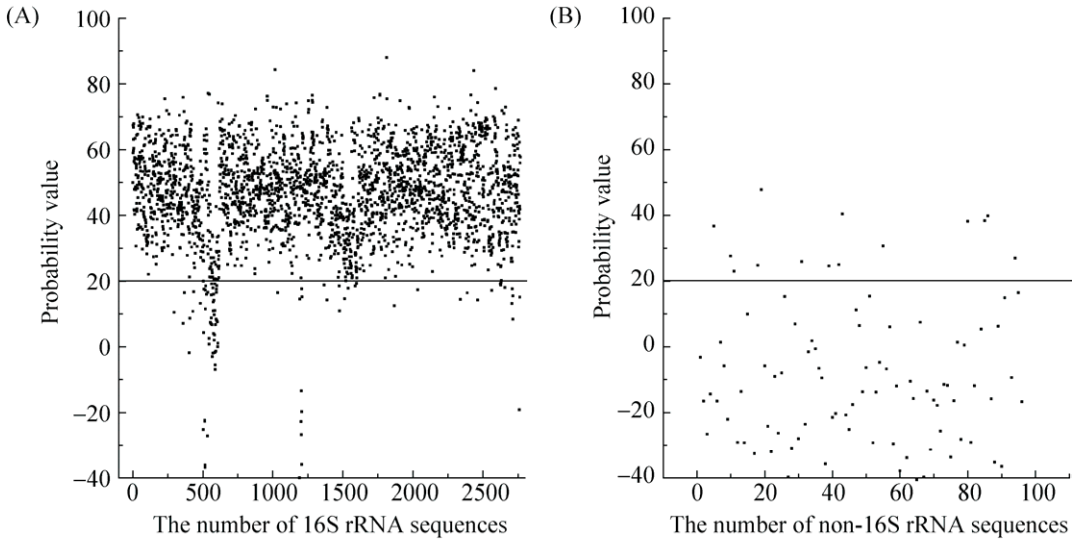

图 7. 16S rRNA 序列(A)和非 16S rRNA 序列(B) *P* 值分布的散点图

Figure 7. The *P* value distribution of 16S rRNA sequences (A) and non-16S rRNA sequences (B). Horizontal straight line is *P*=20.

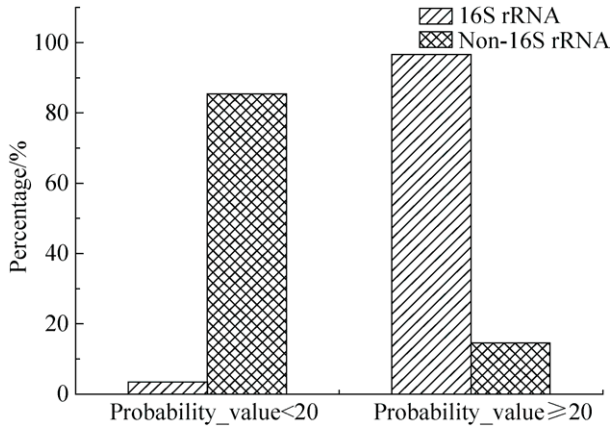

图 8. 以 *P* 阈值为基础，16S rRNA 序列与非 16S rRNA 序列样本分布图

Figure 8. The distribution of 16S rRNA sequences and non-16S rRNA sequences based on the *P* value threshold. The *P* value threshold is 20.

以表 3 中 3 组平行实验的敏感性、特异性以及马修斯相关系数的平均值作为评价模型的指标，可得本实验所构建模型的敏感性、特异性以及马修斯相关系数分别为 91.60%、99.58%和 91.49%。通过数据表明，本文所提出的模型在筛选 16S rRNA 基因序列中是可行的，并且十分有效。

2.5 基于全基因组数据对模型进行分析

为进一步说明模型的有效性与实用性，本实验将在原核生物全基因组中用滑窗法检测模型对 16S rRNA 基因序列的识别能力，以起到对基因组进行注释的作用。本实验所用的滑窗法是指以 DNA 序列的第一个碱基为起始点，每隔一定数量的碱基取出一段连续的、固定长度的序列，带入本文所构建的模型进行预测，判断其是否为特征序列。此实验从 NCBI 数据库随机选取 3 株细菌，对其基因组序列进行实验，预测结果如表 4 所示。

表 4 中包括细菌名称，模型预测出的 16S rRNA 基因序列的起始和终止位置与真实的 16S rRNA 基因序列的起始和终止位置，覆盖率以及注释比例。覆盖率是指模型预测出的 16S rRNA 基因序列与真实的 16S rRNA 基因序列的重叠区域在真实的 16S rRNA 基因序列中所占比例。例如，模

表 3. 模型的敏感性、特异性及马修斯相关系数

Table 3. The sensitivity, specificity and MCC of the model

Test set	True positive	True negative	False positive	False negative	Sensitivity/%	Specificity/%	MCC/%
Data set 1	1876	1990	10	124	93.80	99.50	93.45
Data set 2	1777	1994	6	223	88.85	99.70	89.06
Data set 3	1843	1991	9	157	92.15	99.55	91.95
Average value	–	–	–	–	91.60	99.58	91.49

表 4. 模型的覆盖率和注释比例

Table 4. The coverage percentage and annotation ratio of the model

Bacteria name	16S rRNA		Predicted 16S rRNA		Coverage rate/%	Annotation rate
	Start site	Termination site	Start site	Termination site		
Leuconostoc mesenteroides ID：116617174	22663	24222	22601	24151	95.45	2/4
	148398	149957	148401	149951	99.42	
Aeromonas hydrophila ID：117617447	85174	86723	84901	86451	82.44	6/10
	163372	164921	163051	164601	79.34	
	214907	216456	214601	220801	100	
	349946	351495	349701	351251	84.25	
	803843	805392	803601	805151	84.44	
	933020	934569	932801	934351	85.93	
Bacillus subtilis ID：740748848	20041	21595	19601	24251	100	7/8
	330891	332446	330551	335201	100	
	3563690	3565244	3563301	3567951	100	
	3584089	3586543	3585021	3588301	62.02	
	3644289	3645843	3644001	3648651	100	
	3650177	3651731	3649451	3654101	100	
	3715138	3716692	3714801	3719451	100	

型预测肠膜明串珠菌的第一个 16S rRNA 基因序列的起始与终止碱基位置分别为 22601 和 24151，而该 16S rRNA 基因序列真实的起始与终止碱基位置分别为 22663 和 24222。重叠区域长度为 1489，真实的 16S rRNA 基因序列长度 1560，则覆盖率为 95.45%。

注释比表示模型在某一细菌全基因组中识别出的 16S rRNA 基因的个数与该细菌全基因组中 16S rRNA 基因的总个数的比值。例如，模型预测出枯草芽孢杆菌全基因组中 16S rRNA 基因的个数为 7，该细菌全基因组中 16S rRNA 基因的总个数为 8，则注释比为 7/8。

上述 3 组实验的平均覆盖率都在 80%以上，因此，该模型可较为准确的定位 16S rRNA 序列所在位置。注释比结果中，以枯草芽孢杆菌的预测结果最为突出，全基因组中 8 个 16S rRNA 基因，通过本文所构建的模型可识别出 7 个。通过数据进一步表明，本文所提出的模型在筛选 16S rRNA 基因序列中是可行的，并且十分有效。

3 讨论

本文首次提出应用序列分析相关算法构建模型对 16S rRNA 基因进行识别。通过对序列 GC 碱

基含量，序列碱基 3-周期性以及马尔可夫链 3 种方法的有效结合，实现了对 16S rRNA 基因的识别。首先，对序列 GC 碱基含量进行统计，并设定初步筛选的阈值区间为 50%–60%。其次，对 GC 碱基含量在 50%–60%之间序列进行碱基 3-周期性分析。由于 16S rRNA 基因序列属于非编码序列，因此，这类序列不具有 3-周期性。依据统计训练数据的信噪比值，本文设定筛选阈值为 5。最后，通过构建两种马尔可夫模型，对满足 GC 碱基含量在 50%–60%之间，并且序列信噪比值小于 5 的序列进行 *P* 值求解。若待选序列 *P* 值大于 20，则此序列被判定为 16S rRNA 基因序列；反之，则被判定为非 16S rRNA 基因序列。经过上述步骤对待选序列进行最终识别。

本文所构建的模型可对全基因组中的 16S rRNA 基因进行快速注释。与 RNAmmer 和 Meta-RNA 不同，本模型不仅可以识别原核生物全基因组中的 16S rRNA 基因，同样可以对片段型序列进行识别。相较于 rRNASelector 应用一种统计学算法对 16S rRNA 基因进行识别，本文集成了三种序列统计方法来构建基因识别模型，因此，预测结果更具可靠性。但此方法同样存在不足之处，主要有以下三个方面：(1) 基因组中 16S rRNA 基因的数量较少且模型存在弃真行为，目标序列未能全部找出；(2) 5S rRNA 和 23S rRNA 基因与 16S rRNA 基因序列性质相似，试验中未采取有效的剔除方法，筛选结果存在此类噪声；(3) 试验中选取的阈值组合有待进一步修正。如果上述问题能得到有效解决，模型识别的准确率将会进一步提高。

参 考 文 献

[1] Yu C, Guo HY, Wei JL, Qian AD. Application of 16S to 23S rRNA intergenic spacer region in identification of bacteria. *China Animal Husbandry & Veterinary Medicine*, 2012, 39(2): 57–60. (in Chinese)
于超, 郭海勇, 魏嘉良, 钱爱东. 16S–23S rRNA 基因序列在细菌鉴定中的应用. 中国畜牧兽医, 2012, 39(2): 57–60.

[2] Liu C, Li JB, Rui JP, An JX, Li XZ. The applications of the 16S rRNA gene in microbial ecology: current situation and problems. *Acta Ecologica Sinica*, 2015, 35(9): 2769–2788. (in Chinese)
刘驰, 李家宝, 芮俊鹏, 安家兴, 李香真. 16S rRNA 基因在微生物生态学中的应用. 生态学报, 2015, 35(9): 2769–2788.

[3] Zhao XY, Zhang J, Chen YY, Li Q, Yang T, Pian C, Zhang LY. Promoter recognition based on the maximum entropy hidden Markov model. *Computers in Biology and Medicine*, 2014, 51: 73–81.

[4] Li JL, Wang LF, Wang HY, Bai LY, Yuan ZM. High-accuracy splice site prediction based on sequence component and position features. *Genetics and Molecular Research*, 2012, 11(3): 3432–3451.

[5] Lagesen K, Hallin P, Rødland EA, Stærfeldt HH, Rognes T, Ussery DW. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 2007, 35(9): 3100–3108.

[6] Huang Y, Gilna P, Li WZ. Identification of ribosomal RNA genes in metagenomic fragments. *Bioinformatics*, 2009, 25(10): 1338–1340.

[7] Lee JH, Yi H, Chun J. rRNASelector: a computer program for selecting ribosomal RNA encoding sequences from metagenomic and metatranscriptomic shotgun libraries. *The Journal of Microbiology*, 2011, 49(4): 689–691.

[8] Berryman MJ, Allison A. Review of signal processing in genetics. *Fluctuation and Noise Letters*, 2005, 5(4): R13–R35.

[9] Yin CC, Yau SST. Prediction of protein coding regions by the 3-base periodicity analysis of a DNA sequence. *Journal of Theoretical Biology*, 2007, 247(4): 687–694.

[10] Voss RF. Evolution of long-range fractal correlations and 1/*f* noise in DNA base sequences. *Physical Review Letters*, 1992, 68(25): 3805–3808.

[11] Sharma SD, Shakya K, Sharma SN. Evaluation of DNA mapping schemes for exon detection//Proceedings of 2011 International Conference on Computer, Communication and Electrical Technology. Tamilnadu: IEEE, 2011: 71–74.

[12] Zhang R, Zhang CT. Z curves, an intutive tool for visualizing and analyzing the DNA sequences. *Journal of Biomolecular Structure and Dynamics*, 1994, 11(4): 767–782.

[13] Anastassiou D. Frequency-domain analysis of biomolecular

sequences. *Bioinformatics*, 2000, 16(12): 1073–1081.
[14] Chakravarthy N, Spanias A, Iasemidis LD, Tsakalis K. Autoregressive modeling and feature analysis of DNA sequences. *EURASIP Journal on Advances in Signal Processing*, 2004, 2004(1): 952689.
[15] Kwan HK, Kwan BYM, Kwan JYY. Novel methodologies for spectral classification of exon and intron sequences. *EURASIP Journal on Advances in Signal Processing*, 2012, 2012(1): 50.
[16] Yan M, Lin ZS, Zhang CT. A new Fourier transform approach for protein coding measure based on the format of the Z curve. *Bioinformatics*, 1998, 14(8): 685–690.
[17] Coward E. Equivalence of two Fourier methods for biological sequences. *Journal of Mathematical Biology*, 1997, 36(1): 64–70.
[18] Silverman BD, Linsker R. A measure of DNA periodicity. *Journal of Theoretical Biology*, 1986, 118(3): 295–300.
[19] Jääskinen V, Parkkinen V, Cheng L, Corander J. Bayesian clustering of DNA sequences using Markov chains and a stochastic partition model. *Statistical Applications in Genetics and Molecular Biology*, 2014, 13(1): 105–121.
[20] Zhao L, Lascoux M, Waxman D. An informational transition in conditioned Markov chains: applied to genetics and evolution. *Journal of Theoretical Biology*, 2016, 402: 158–170.
[21] Wan YW, Allen GI, Baker Y, Yang E, Ravikumar P, Anderson M, Liu ZD. XMRF: an R package to fit Markov networks to high-throughput genetics data. *BMC Systems Biology*, 2016, 10(S3): 69.
[22] Komorowski T, Peszat S, Szarek T. On ergodicity of some markov processes. *The Annals of Probability*, 2010, 38(4): 1401–1443.
[23] Arns M, Buchholz P, Panchenko A. On the numerical analysis of inhomogeneous continuous-time Markov chains. *Informs Journal on Computing*, 2010, 22(3): 416–432.

Recognition of 16S rRNA genes in prokaryotic genomes

Wenkai Yan[#], Mingmin Xu[#], Guangle Zhang, Ning Qiao, Weina Xu, Yuanyuan Chen, Liangyun Zhang[*]

College of Sciences, Nanjing Agricultural University, Nanjing 210095, Jiangsu Province, China

Abstract: [Objective] We identified 16S rRNA genes in genomes of prokaryotes. **[Methods]** We constructed a 3-layer filtering model based on the three features of GC bases content of the gene sequences, 3-base periodicity and Markov chain to recognize the 16S rRNA genes from prokaryotic genomes. **[Results]** The specificity, sensitivity and Matthews correlation coefficients of the model were 99.58%, 91.60% and 91.49%, respectively. **[Conclusion]** The results showed that the 16S rRNA genes can be identified efficiently and accurately by using our model.

Keywords: 16S rRNA gene, GC base content, 3-base periodicity, Markov chain

(本文责编：张晓丽)

Supported by the National Natural Science Foundation of China (11571173) and by the Natural Science Foundation of Jiangsu Province (BK20141358)
[*]Corresponding author. Tel: +86-25-84396063; E-mail: zlyun@njau.edu.cn
[#]Those authors contributed equally to this work.
Received: 14 October 2016; Revised: 24 November 2016; Published online: 20 February 2017