

一种基于卷积码模型的遗传序列分析方法

刘 晓¹, 田逢春¹, 李素芳², 王世元¹

(1 重庆大学 通信工程学院, 重庆 400044; 2 中国计量学院 生命科学学院, 浙江 杭州 310018)

[摘 要] **【目的】**研究通信纠错编码理论在生物学领域研究中的应用。**【方法】**基于纠错编码理论,将卷积码模型分析方法用于遗传序列的分析,结合密码子简并性以及碱基关联短程为主的特性,对卷积码模型的生成矩阵、编码长度和约束长度等参数进行讨论。对 2 种原核生物(*Escherichia coli* str. K-12 substr. MG1655 和 *Staphylococcus epidermidis* ATCC 12228)与 2 种真核生物(*Arabidopsis thaliana* chromosome 4 和 *Oryza sativa* (japonica cultivar-group) chromosome 10) DNA 序列翻译起始及终止的特征进行分析,并通过对不同卷积码模型比较,确定合适的分析模型。**【结果】**在适当的分析模型下,分析结果在开放阅读框(ORF)起始端和终止端附近,显示出与起始密码子和终止密码子位置紧密关联的明显码距变化,而不适当的分析模型则使分析结果与生物意义的对应关系下降;分析方法对 2 种原核生物和 2 种真核生物具有较好的统一性。**【结论】**通信纠错编码模型与蕴含纠错机制的生物序列编码特性可以较好地结合,这对将纠错编码理论与生物学研究相结合具有一定的指导意义。

[关键词] 纠错编码;卷积码;短程关联优势;密码子

[中图分类号] Q811.4

[文献标识码] A

[文章编号] 1671-9387(2010)04-0207-08

A method for analyzing genetic sequence based on a convolutional code model

LIU Xiao¹, TIAN Feng-chun¹, LI Su-fang², WANG Shi-yuan¹

(1 College of Communication Engineering, Chongqing University, Chongqing 400044, China;

2 College of Life Sciences, China Jiliang University, Hangzhou, Zhejiang 310018, China)

Abstract: **【Objective】** The objective of this research was to study the application of error-correction coding theory of communication engineering in biological field. **【Method】** The principle of convolutional code was employed to analyze genetic sequence based on error-correction coding theory. The parameters, such as generator matrix, code length and constraint length, of the convolutional code models were designed and compared based on the degeneracy of codons and short-range dominance of bases correlation. The coding features of the translation initiation and termination of the DNA sequences of two prokaryotes (*Escherichia coli* str. K-12 substr. MG1655 and *Staphylococcus epidermidis* ATCC 12228) and two eukaryotes (*Arabidopsis thaliana* chromosome 4 and *Oryza sativa* (japonica cultivar-group) chromosome 10) were studied to search appropriate analysis model. **【Result】** Obvious changes of average code distance were observed near the initiation site and the termination site of the open reading frame (ORF) in appropriate models, and the parallelism between analysis results and biology sense descended in inappropriate models. The proposed method presents universality for the two prokaryotes and the two eukaryotes. **【Conclusion】** The analysis suggests that the error-correction coding model of communication engineering can be used to ana-

* [收稿日期] 2009-09-01

[基金项目] 教育部高等学校博士学科点专项科研基金项目(20050611022);重庆市自然科学基金(重点)项目(2009BA2021)

[作者简介] 刘 晓(1975—),男,重庆人,讲师,在读博士,主要从事生物信息学研究。E-mail:liuxiao@cqu.edu.cn

[通信作者] 田逢春(1963—),男,重庆人,教授,博士,博士生导师,主要从事图像处理、通信信息技术、生物信息处理及生物医学电子研究。E-mail:FengchunTian@cqu.edu.cn

lyze genetic sequences which contain error-correction mechanism, and it provides instructions to use error-correction coding theory for further study in biological field.

Key words: error-correction coding; convolutional code; short-range dominance; codon

DNA 的一级结构由 4 种核苷酸的连接与排列顺序所确定,该结构表示了 DNA 分子的化学构成,其对 DNA 高级结构的形成有很大影响,其中 mRNA 上每 3 个核苷酸组合所形成的密码子在遗传信息的传递中具有重要作用。研究表明,遗传密码子以及同义密码子的使用对 DNA 高级结构的形成有很大影响,并且与生物进化有联系^[1-3],这体现了 DNA 序列一级结构所蕴含的丰富信息。在遗传信息传递过程中,基因编码错误的发生会引起遗传性状的变化,如同义突变、无义突变、错义突变和移码突变等。其中同义突变不会导致生物性状的变化,其他 3 种突变均会引起遗传性状的变化,这类似于通信编码中因使用纠错编码而得到的优良性能——即使有误码也不影响接收端的正确译码。

从上世纪 70 年代开始,研究人员将信息处理的原理、方法应用到 DNA 序列分析中,建立了若干相应的分析模型^[4-6],并提出了数字映射^[7]、图形表达^[8]及几何表达^[9-11]等碱基信号的表达方式。Bat-tail^[12]提出,生物学研究应该充分考虑纠错编码理论的应用。May 等^[13]将纠错编码理论引入 DNA 序列分析之中,并成功地分析了原核生物的翻译起始特性,为分析和理解遗传序列的编码特性及其翻译调控机制提供了一种新思路。May 等^[13]、Ponna-la 等^[14]分别使用分组码与卷积码模型完成了对 *E. coli* K-12 DNA 序列的分析,其中 May 等^[13]以 *E. coli* K-12 16SrRNA 的最后 13 个碱基作为模板,生成奇偶校验位以构建码字集,将待分析的 DNA 序列作为输入信号对其进行解码;而 Ponnala 等^[14]在此基础上对卷积码模型的生成矩阵算法进行了改进。他们工作的共同点是分析对象的某个子序列(16SrRNA 的最后 13 个碱基)作为处理信号码集的基础,此方法在处理不同分析对象时显得比较复杂,且卷积码分析效果并不理想。本研究拟寻找一种与分析对象序列子集无关的处理方法,使得其对多个对象分析时的处理更加简洁。而在纠错编码中,卷积码是一种具有较好性能的信道编码方式,理论和实际均已证明卷积码的性能至少不比分组码差,应该可以寻找较好的卷积编码模型来分析 DNA 序列的编码特性。

卷积码的特点是编码后的码元不仅与当前输入

信息有关,而且还与此前的若干信息有关,即码元相互间存在影响,这可以与密码子的上下文关系对遗传序列翻译、表达产生的影响及其具有的特点相联系^[2-3]。相对于长程相关性,研究人员对碱基之间的关联性提出了短程关联占优的普适特性^[15]。同时,由于密码子具有简并性——一种氨基酸可由几个不同密码子编码的现象,这使生物物种的变异更趋于稳定,因为由变异引起密码子中一个碱基的改变,其结果可能也只是变成了同一个氨基酸的另一个密码子(同义突变),进而合成出与原来没有区别的蛋白质。简并性具有防错作用,从而维持了物种的稳定性,这也可以看作遗传信息中蕴含纠错机制的一种体现,但目前有关这方面的研究工作还较少。

本研究采用卷积码模型,对 DNA 序列编码特性进行分析,考虑并使用密码子的简并性来设计模型参数,并将密码子作为基本的信息单元(而不是单个碱基)进行分析,又由于密码子前后碱基的使用(上下文关系)特点与蛋白质二级结构有关,故在分析模型设计时,以短程关联占优为基础,考虑了前后密码子之间的相互关联性,旨在为通信纠错编码理论在生物学领域研究中的应用提供参考。

1 相关编码概念简述

1.1 纠错编码

在数字通信系统中,信号处理通常包括信源编码和信道编码两部分,其后将经过调制的信号通过信道进行传送。其中信源编码将连续的消息变换为数字信号;信道编码则使数字信号与传输媒质相匹配,以提高传输的可靠性或有效性。

在生物系统的信息传递过程中,遗传信息会受到各种干扰从而影响到遗传信息的正常传递与表达。生物系统的高度稳定性已经表明其具有严格的错误纠正能力。因此,生物遗传信息在生物体中的传递过程,可以与信息在数字通信系统中传递的过程建立对应关系。由于密码子决定了 20 种氨基酸的选取,进而确定蛋白质的表达,因此可将生物遗传信息确定密码子与编码理论中的信源编码相结合。此外,大量核苷酸有序排列而成的 DNA 序列蕴含了丰富的遗传信息,包括确定翻译的起始、错误的修复以及大量的调控信息,因此可结合信道编码的概

念来分析 DNA 序列的调控、差错控制特性以及信息单元间的相互影响。

信道编码又称为纠错编码,其基本原理是在发送端被传输的信息序列上附加一些码元(称为监督码元),这些冗余的码元与信息码元之间以某种确定的规则相互关联(约束)。接收端按照预定的规则检验信息码元与监督码元之间的关系,如果在传输过程中发生差错,则信息码元与监督码元之间的关系将受到破坏。根据对应关系,可以使接收端发现错误乃至纠正错误。

通信信道编码定理为:每个信道具有确定的信道容量 C ,对任何小于 C 的码率 R ,存在码长为 n 、约束长度为 L 的卷积码 (n, k, L) ,用最大似然译码,则随着码长的增加译码错误概率 p 可以任意小,即:

$$p \leq A e^{-(L+1) \cdot n \cdot E(R)} \quad (1)$$

式中: A 为大于 0 的系数; $E(R)$ 为正实函数,也称为误差指数^[16]。

1.2 卷积码及基本分析模型的设计

根据信息码元和监督码元之间不同的约束方式,可将纠错码分为分组码和卷积码。卷积码 (n, k, L) 指编码器将 k 个 i 输入信息码元编为 n 个输出信息码元,其编码输出的 n 个码元不仅与当前 k 个输入信息码元有关,还与此前的 L 组信息码元(每组 k 个信息码元)有关,即编码过程中前后信息相互关联、相互影响,其中 $L+1$ 被称为约束长度。

卷积码模型中的参数可根据以下性质确定:

(1)密码子在遗传信息表达中具有重要作用,因此将密码子作为基本信息单元(而不是单个碱基),并将其长度 3 作为基本信息单元长度进行分析。

(2)由式(1)可知,码长 n 的增加可以减小译码错误概率。从信息传送的角度进行分析,若有附加的冗余信息对其进行监督和校正,则可能使生物遗传信息在传递过程中具有较高的查错纠错能力和较低的基因突变率,以保证物种可以保持较高的遗传稳定性。由于密码子使用的上下文关系与蛋白质二级结构有关,进而影响生物功能,因此本研究将其对应于卷积码模型中前后信息的相互约束关系,即在所设计的卷积码分析模型中,假设密码子之间存在相互影响、相互约束的关联性,使得其在数据量一定的条件下,可以获得更高的纠错能力。

在通信编码的实际应用中,卷积码的码长 n 和当前输入信息码元个数 k 均比较小,可结合编码效率 (k/n) ,选择码长 $n=6$ 进行分析,也即码长为 2 个密码子长度。对于约束长度 $L+1$,结合碱基关联的

短程为主特性,取 L 为 1,以着重考虑相邻密码子之间的相互联系与影响。

(3)对于卷积码而言,可用生成矩阵进行描述。在设计生成矩阵的过程中,首先考虑如下基本情况:输入信息的码元长度为 3 个碱基,编码输出信息的码元长度为 6 个碱基,同时考虑当前信息的前一个信息单元对应的 3 个碱基对当前输出的影响,则对应于 $(6, 3, 1)$ 卷积码。由当前输入信息 m_1 、前一时

$$C = m_1 g + m_0 g \quad (2)$$

刻输入信息 m_0 ,得到的卷积编码输出为:

式中:生成矩阵 $g = \begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}_{(6 \times 3)}$ 。

成矩阵 g 的元素 $g_{ij}^k (i=0, 1, 2; j=0, 1, 2, 3, 4, 5; k=0, 1, 2)$ 表示输入信息 m (长度为 3)的前溯第 i 时刻第 j 位输入对编码输出的第 k 位所产生的影响。结合密码子的简并性,分析中假设密码子的第 1, 2 位可由生物系统编码特性所确定,因此 j 为 0, 1, 3, 4 时使用了高权重的影响因子 1。而由于密码子第 3 位的相对灵活性,则假定其受到系统编码特性之外(例如生物进化特性等)的未确定因素的影响,因此 j 可取值 2, 5,即使用了低权重的影响因子 0。

2 试验方法

2.1 分析对象

选择了 4 种模式生物作为分析对象,其中原核与真核生物各 2 种(表 1)。从 NCBI 网站(<http://www.ncbi.nlm.nih.gov>)下载各自全序列数据,使用 GeneMark 对所选定的序列进行分析,寻找所有的 ORF,保存以 ATG 为起始密码子的碱基片段的起始、终止位点以便分析。

2.2 基本算法步骤

参考文献[13,17]中的计算方法,并结合卷积码模型特点进行计算。本研究的基本算法步骤如下:

1)碱基读取。对于每个 ORF 序列片段起始端,以起始密码子的第 1 个碱基作为 0 位点,前后各取 M 个碱基,以所获得的长 $2M$ 的碱基片段作为待分析的序列片段。在对 ORF 序列片段终止端进行分析时,以终止密码子第 3 个碱基作为 0 位点,前后各取 M 个碱基作为待分析的序列片段。

2)信号表达。为了将碱基序列表达为适合通信编码理论与信息处理进行分析的序列,需要对碱基序列进行合适的信号表达。本研究采用了较为简便的数字映射方式,分别用 0, 1, 2, 3 表示 A、G、C、T^[7],由

此可将碱基序列转化为数字序列。每个分析对象的 ORF 数量并不相同,若某分析对象的 ORF 数量为 r , 则其具有 r 个数字化的待分析编码序列。

表 1 供试的 4 种模式生物

Table 1 Four model organisms for analysis

分析对象 Analyzed object	分类 Classification	NCBI 参考序列登录号 NCBI RefSeq access number
<i>Escherichia coli</i> str. K-12 substr. MG1655	原核 Prokaryota	NC_000913
<i>Staphylococcus epidermidis</i> ATCC 12228	原核 Prokaryota	NC_004461
<i>Arabidopsis thaliana</i> chromosome 4	真核 Eukaryota	NC_003075
<i>Oryza sativa</i> (japonica cultivar-group) chromosome 10	真核 Eukaryota	NC_008403

3) 卷积运算。在 (6,3,1) 卷积码模型下,每个 ORF 对应的数字序列按如下步骤进行卷积运算: ①取该序列的第 1~6 个数据,将其中第 4,5,6 个数据作为当前输入 m_1 ,第 1,2,3 个数据作为前时刻的输入 m_0 ,根据式 (2) 计算卷积编码,并对结果进行模 4 运算得到当前时刻的编码输出。②依次取该序列的第 2~7 个数据,第 3~8 个数据,依此类推,并分别计算各个时刻的卷积编码 C 。③重复以上步骤,完成对各个 ORF 对应数字序列的卷积运算。当所利用的卷积码模型的参数改变时,对计算式和生成矩阵进行相应调整。

4) 码距计算。码距是 2 个长度相同的码字对应位置上取值不相同的个数,也称汉明距离。在分析时,分别采用了 2 种长度进行码距比较。(1)1 个密码子长度。在生成码长 $n \geq 3$ 时,将当前输出的卷积编码 C 的前 3 位与前一时刻输入信息 m_0 进行比较,得到此次计算的码距。(2)编码输出码长 n 。将卷积编码输出与对应长度的之前及以前时刻的输入信息进行比较,计算码距。在以上两种比较方式中,对每次卷积运算的卷积结果都进行码距计算,进而可得到每个 ORF 片段的码距序列。

5) 计算平均码距。对每个分析对象的所有 (r 个) ORF 片段的码距序列,按照分析位点计算平均码距,并对结果进行作图分析。

3 结果与分析

在不同的模型下,分析的序列片段长度均为 100 个碱基。

3.1 基于 (6,3,1) 卷积码模型的 4 种模式生物的遗传序列分析

在 (6,3,1) 卷积码模型下,对 4 种模式生物 ORF 起始密码子和终止密码子附近区域平均码距的分析结果如图 1,2 所示,其中用于码距计算的码长均为 3 个碱基(1 个密码子长度)。

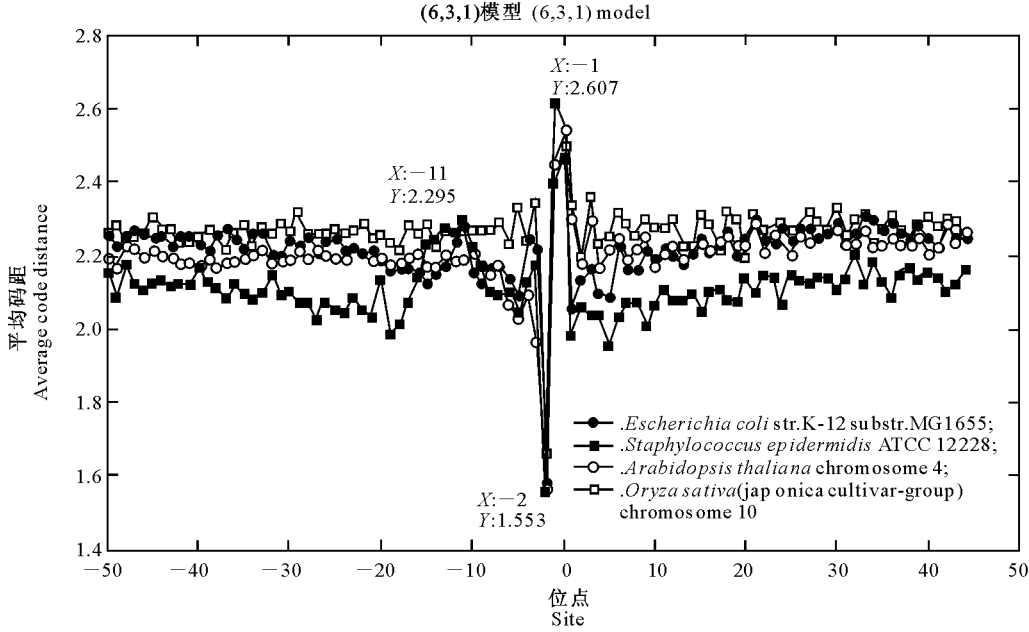


图 1 基于 (6,3,1) 卷积码模型的 4 种模式生物在 ORF 起始密码子附近区域的平均码距
Fig. 1 Average code distances of the 4 model organisms near the start codons of the ORFs based on (6,3,1) convolutional model

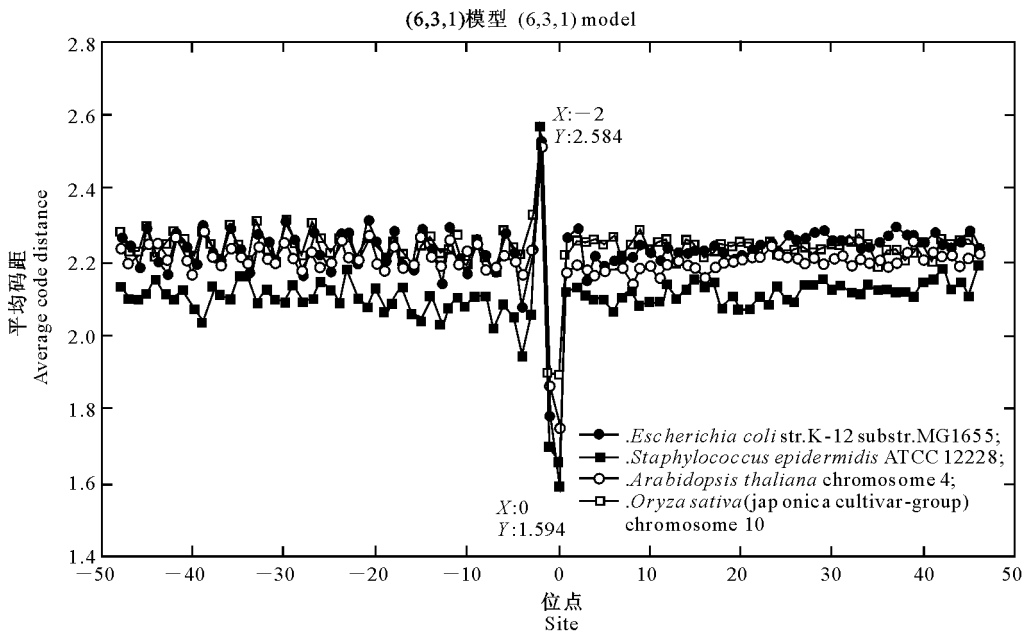


图 2 基于(6,3,1)卷积码模型的 4 种模式生物在 ORF 终止密码子附近区域的平均码距
Fig. 2 Average code distances of the 4 model organisms near the stop codons of the ORFs based on (6,3,1) convolutional model

由图 1 和图 2 可见,(1)在 ORF 起始端,4 种模式生物的平均码距在-2 和-1 位点均出现显著的变化,且具有较好的一致性。2 种原核生物平均码距在-11 位点附近均有较明显的变化,这与原核生物存在 SD(Shine-Dalgarno)序列,且 SD 区域与起始密码子的距离在 5~13 个碱基间变化的特性有关^[18],与通信系统中采用的前导码相似。2 种真核生物平均码距在-11 位点附近没有明显变化,这与真核生物不存在 SD 序列有关。(2)在 ORF 终止端,4 种模式生物的平均码距在-2 和 0 位点均出现显著变化,也体现出此模型的一致性。

3.2 基于几种不同卷积码模型的对比分析

根据上述试验结果,对基本分析模型的参数进行调整,在获得不同分析模型的基础上,再以 ORF 起始密码子为参考,对 4 种模式生物进行对比分析。本研究所采用的卷积码模型有:

- (1)在其他分析方法中,常将单个碱基作为基本信息单元,因此取 $k=1$,考虑编码效率(k/n),取码长为 2,以(2,1,1)卷积码模型进行分析。
- (2)按照 1.2 节所述思路,在输入信息不变的情况下增加码长 n ,利用(9,3,1)卷积码模型进行对比分析。
- (3)考虑过渡状态的对比,输入信息长度 $k=2$,并兼顾编码效率(k/n),取(3,2,1)卷积码模型进行

分析。

3.2.1 基于(2,1,1)卷积码模型的分析结果 在该模型下,选取生成矩阵 $g=[1 \ 0]$,以码长 2(码距计算的码长为 2 碱基)进行码距计算,其结果如图 3 所示。由图 3 可知,在 ORF 起始端附近的 0,1,2 位点,平均码距均有显著变化;2 种原核生物在-9 位点附近区域码距也有明显的变化。

3.2.2 基于(3,2,1)卷积码模型的分析结果 在该模型下,生成矩阵 $g=\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$ 。以码长 3(码距计算的码长为 3 碱基)进行码距计算,其结果如图 4 所示。由图 4 可知,在 ORF 起始端附近的-1,1 位点,平均码距均有显著变化;2 种原核生物,在-11,-12 位点附近区域码距也有明显变化。

3.2.3 基于(6,3,1)与(9,3,1)卷积码模型的分析结果 在(6,3,1)卷积码模型下,生成矩阵 $g =$

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$
;在(9,3,1)卷积码模式下,生成矩阵 $g =$

$$\begin{bmatrix} 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 & 0 & 1 & 1 & 0 \end{bmatrix}$$
。各自以码长 6,9(即码距计算的码长分别为 6,9 碱基)进

行码距计算,所得结果如图 5,6 所示。

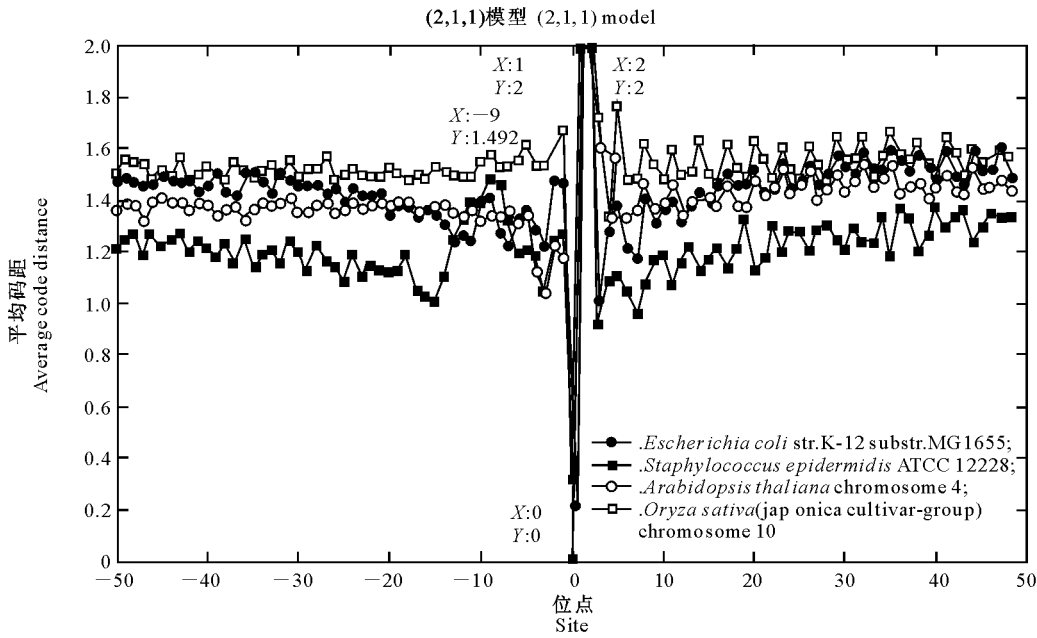


图 3 基于(2,1,1)卷积码模型的 4 种模式生物在 ORF 起始密码子附近区域的平均码距
Fig. 3 Average code distances of the 4 model organisms near the start codons of the ORFs based on (2,1,1) convolutional model

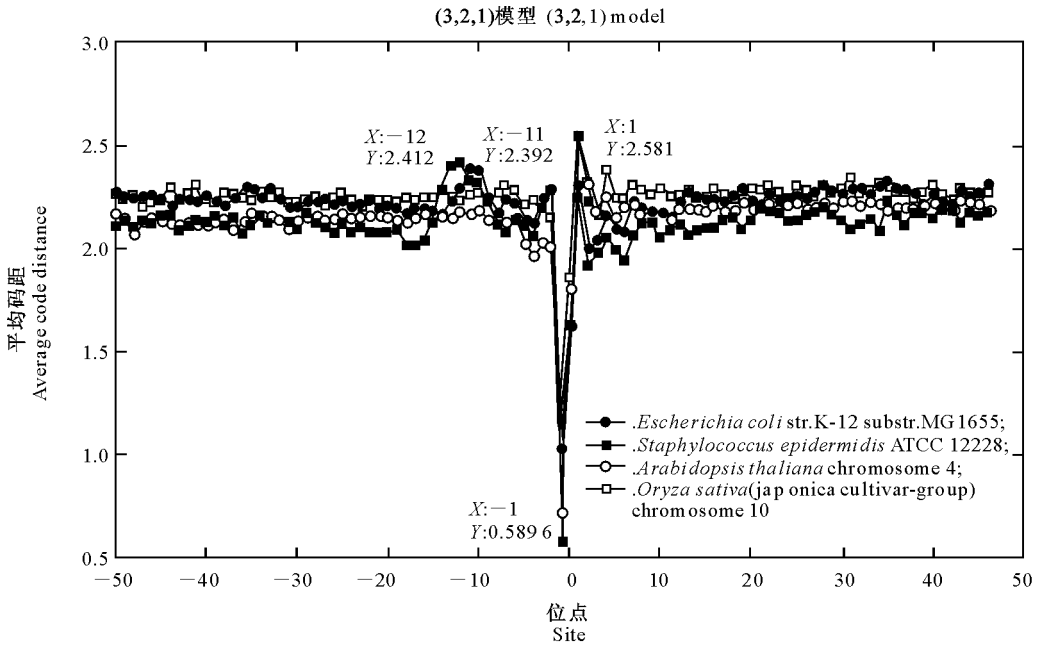


图 4 基于(3,2,1)卷积码模型的 4 种模式生物在 ORF 起始密码子附近区域的平均码距
Fig. 4 Average code distances of the 4 model organisms near the start codons of the ORFs based on (3,2,1) convolutional model

由图 5 和图 6 可知,在 ORF 起始端,4 种模式生物的平均码距虽然都在某些相同位点出现显著变化,但是曲线变化的峰谷个数随着比较码长的增加而增大,并不利于序列翻译起始生物特性的观察,说明比较码长增加会导致模型有效性下降。

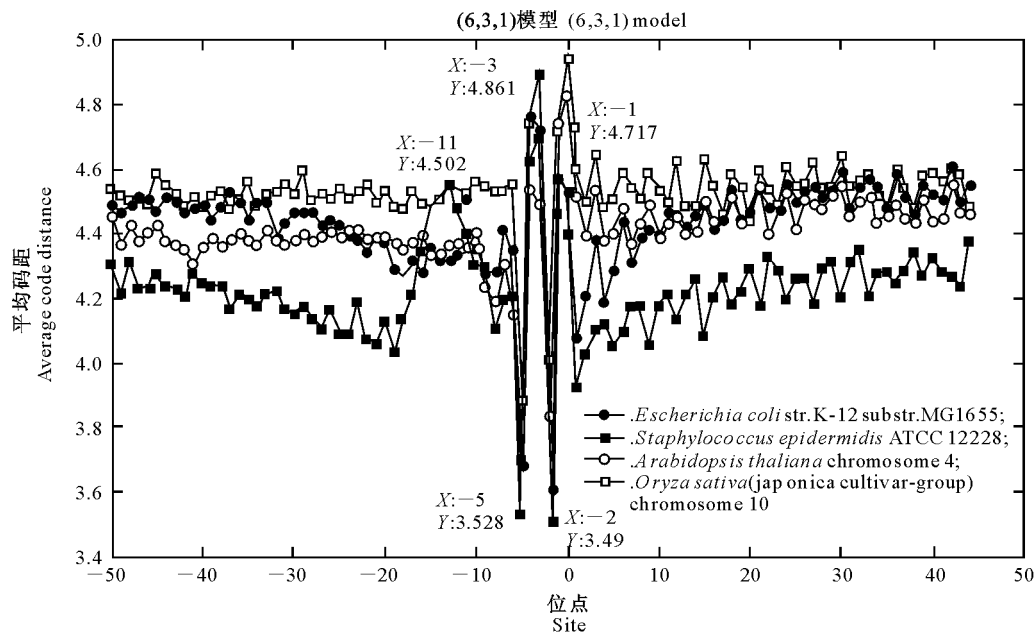


图 5 基于(6,3,1)卷积码模型的 4 种模式生物在 ORF 起始密码子附近区域的平均码距
Fig. 5 Average code distances of the 4 model organisms near the start codons of the ORFs based on (6,3,1) convolutional model

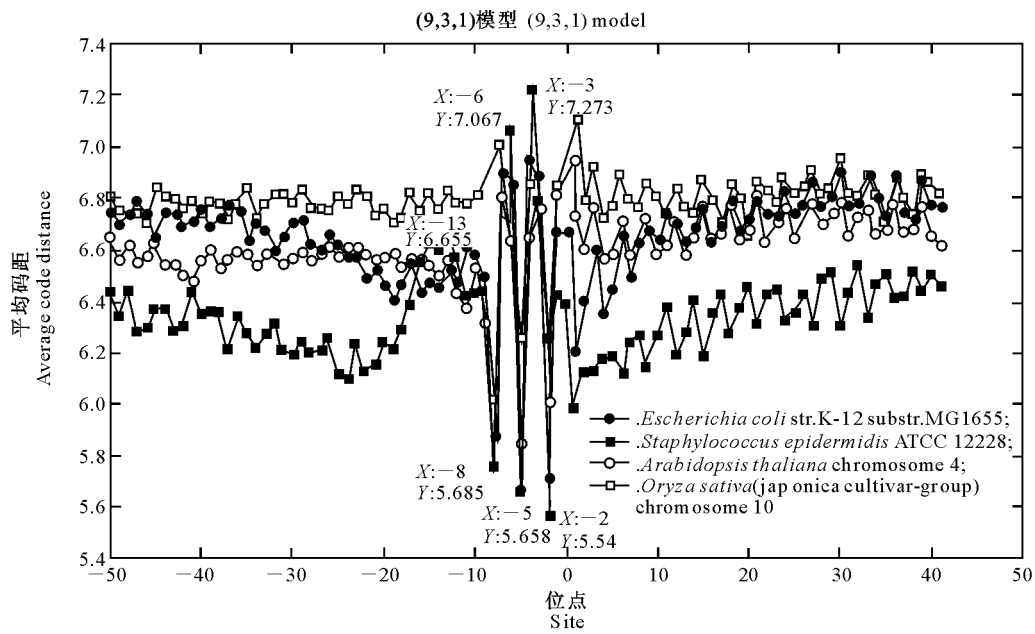


图 6 基于(9,3,1)卷积码模型的 4 种模式生物在 ORF 起始密码子附近区域的平均码距
Fig. 6 Average code distances of the 4 model organisms near the start codons of the ORFs based on (9,3,1) convolutional model

4 结论与讨论

本研究结合通信纠错编码概念,利用密码子的简并性和碱基短程关联为主的特性,设计了卷积码模型及其参数,对 2 种原核生物和 2 种真核生物遗传序列的分析结果显示,卷积码模型的分析结果与翻译起始、终止的特性关联较好;比较码长较小时卷

积码模型的分析结果较好,而比较码长增加时卷积码模型的分析效果变差,这可以与碱基短程关联为主性建立起对应联系。参数调整后的对比分析结果表明,适合于真核生物、原核生物遗传序列分析的模型有(6,3,1)、(3,2,1)和(2,1,1)卷积码模型,并且建立的卷积码模型对原核和真核生物的分析具有统一性。本研究结果表明,生物系统与通信编码系统

2 种体系在纠错机制上具有一定的关联性,这为进一步将信息论、通信编码理论与生物序列分析相结合提供了支撑,且本研究所设计的模型具有分析有效性。

本研究得到的结果是统计性的分析结果,需要在完成 ORF 分析的基础上开展研究,且仍然是对遗传序列特性的一种“描述”,如何利用相应的理论和方法对序列特性进行“预测”,应是下一步研究中关注的问题。在随后的研究中,应该考虑如何将更多的遗传序列特性以合理的方式加入到分析模型中,并对分析模型和模型参数进行调整,使模型具有更完备的生物意义。此外,本研究只对序列编码区附近区域进行了分析,而目前非编码序列的研究已经在生物信息领域引起了广泛的关注^[19],这是后续研究中的另一个重要内容。由于 DNA 序列具有多级结构,遗传信息在其中的表现形式也具有一维或多维的特点^[20],作为本研究的逆向研究,如何利用遗传序列的特性对通信编码的研究和改进提供有意义的借鉴,也是值得关注的研究方向^[21]。总之,将多种成熟的研究理论相结合,对仍然蕴含无穷未知信息的生物系统进行分析并向其学习,还需要开展更多的研究工作。

参考文献

- [1] 朱玉贤,李毅.现代分子生物学[M].2版.北京:高等教育出版社,2002:4-9.
Zhu Y X, Li Y. Modern molecular biology [M]. 2nd Edition. Beijing: Higher Education Press, 2002: 4-9. (in Chinese)
- [2] Wang F P, Li H. Codon-pair usage and genome evolution [J]. Gene, 2009, 433(1/2): 8-15.
- [3] Komar A A, Lesnik T, Reiss C. Synonymous codon substitutions affect ribosome traffic and protein folding during *in vitro* translation [J]. FEBS Letters, 1999, 462(3): 387-391.
- [4] Roman-roldan R, Bernaola-galvan P, Oliver J L. Application of information theory to DNA sequence analysis: A review [J]. Pattern Recognition, 1996, 29(7): 1187-1194.
- [5] Wang X H, Istepanian R S H, Song Y H, et al. Review of application of coding theory in genetic sequence analysis [C]//5th International Workshop on Enterprise Networking and Computing in Healthcare Industry. Santa Monica, Canada: IEEE, 2003: 5-9.
- [6] Dawy Z, Gonzalez, Faruck M G, et al. Modeling and analysis of gene expression mechanisms a communication theory approach [C]//2005 IEEE International Conference on Communications, Seoul, Korea: IEEE, 2005: 815-819.
- [7] Rosen G L. Examining coding structure and redundancy in DNA [J]. IEEE Engineering In Medicine And Biology Magazine, 2006, 25(1): 62-68.
- [8] Ashlock D, Golden J. Chaos automata: iterated function systems with memory [J]. Physica D, 2003, 181: 274-285.
- [9] Yao Y H, Nan X Y, Wang T M. Analysis of similarity-dissimilarity of DNA sequences based on a 3-D graphical representation [J]. Chemical Physics Letters, 2005, 411.
- [10] Liao B, Tan M S, Ding K Q. A 4D representation of DNA sequences and its application [J]. Chemical Physics Letters, 2005, 402: 380-383.
- [11] Wang S Y, Tian F C, Liu X, et al. A novel representation approach to DNA sequence and its application [J]. IEEE Signal Processing Letters, 2009, 16(4): 275-278.
- [12] Battail G. An engineer's view on genetic information and biological evolution [J]. Biosystems, 2004, 76: 279-290.
- [13] May E E, Mladen A V, Donald L B, et al. Coding theory based models for protein translation initiation in prokaryotic organisms, Biosystems [J]. Biosystems, 2004, 76(1/3): 249-260.
- [14] Ponnala L, Bitzer D L, Vouk M A. On finding convolutional code generators for translation initiation of *Escherichia coli* K-12 [C]//Proceedings of the 25th Annual International Conference of the IEEE. Cancun, Mexico: IEEE, 2003: 3854-3857.
- [15] Liaofu L, Weijiang L, Lijun J, et al. Statistical correlation of nucleotides in a DNA sequence [J]. Physical Review E, 1998, 58(1): 861-871.
- [16] 王新梅,肖国镇.纠错码:原理与方法[M].修订版.西安:西安电子科技大学出版社,2003:13-16.
Wang X M, Xiao G Z. Error correcting code: principle and method [M]. Revised edition. Xi'an: Xidian University Press, 2003: 13-16. (in Chinese)
- [17] 刘 晓,田逢春,李素芳.基于分组码模型的 DNA 序列编码特性分析 [J].生物信息学,2009,7(22):28-31.
Liu X, Tian F C, Li S F. Analysis of the encoding characteristic of the DNA sequences based on the block code model [J]. China Journal of Bioinformatics, 2009, 7(22): 28-31. (in Chinese)
- [18] Kozak M. Initiation of translation in prokaryotes and eukaryotes [J]. Gene, 1999, 234(2): 187-208.
- [19] 陈润生.与生物信息学相关的两个前沿方向:非编码基因和复杂生物网络 [J].生物物理学报,2007,23(4):290-295.
Chen R S. Two subjects in the field of genome research-non-coding sequence and biological network [J]. Acta Biophysica Sinica, 2007, 23(4): 290-295. (in Chinese)
- [20] Shultzaberger R K, Roberts L R, Lyakhov I G, et al. Correlation between binding rate constants and individual information of *E. coli* Fis binding sites [J]. Nucleic Acids Research, 2007, 35(16): 5275-5283.
- [21] Battail G. Should genetics get an information-theoretic education? [J]. IEEE Engineering in Medicine and Biology Magazine, 2006, 25(1): 34-45.