动物科学

DOI:10.11721/cqnuj20200305

秀丽隐杆线虫基因组编码区的碱基分布特征参数研究*

李　勃[1]，方　斌[2]，陶靖心[1]，刘明伟[3]
(1. 重庆师范大学 生命科学学院，重庆 401331；2. 武汉生物制品研究所有限责任公司，武汉 430207；
3. 重庆医科大学 检验医学院，重庆 401331)

摘要:【目的】考察秀丽隐杆线虫(*Caenorhabditis elegans*)基因组编码区(Coding sequences，CDSs)的碱基使用特点，并提炼出能够区分 CDSs 与非编码区的碱基成分偏移特征参数。【方法】在研究碱基成分偏移特性基础上，定义一个新的参数 d，探索 d 值的分布规律；采用从秀丽隐杆线虫基因组 6 类不同的 DNA 序列中随机抽样的方式，分析并验证该指标作为基因组 CDSs 特征参数的可行性。【结果】参数 d 经过线性变换后近似服从对数正态分布；抽样分析显示分别有 81.6%的 CDSs、70.7%的外显子、21.8%的内含子、4.7%的随机序列、17.8%的 5′非翻译区和 31.4%的 3′非翻译区落在 d 值变换后的特征取值区间内，即被预测为基因组 CDSs。【结论】碱基成分偏移指数 d 可以作为表征基因组编码区的特征参数，它的特定取值区间能很好地区分 CDSs(或开放阅读框及它的子片段)和其他非编码区。

关键词:秀丽隐杆线虫；编码区；碱基成分偏移；数据挖掘；特征参数
中图分类号:Q31　　**文献标志码**:A　　**文章编号**:1672-6693(2020)03-0014-07

秀丽隐杆线虫(*Caenorhabditis elegans*)是一种自由生活的小型土壤线虫，分类学上属于小杆亚纲(Rhabditia)小杆目(Rhabdtidia)小杆总科(Rhabditoidea)[1]。该物种生长周期短(3 d)，是具有神经细胞的多细胞生物，因而常被作为模式生物来探索个体与神经发育的遗传调控机制[2]。1993 年，科学家们通过对秀丽隐杆线虫的研究发现 lin-4 蛋白对胚胎发育事件的正常时间控制至关重要，进一步研究发现 lin-4 的 mRNA 存在着互补性的小 RNA[3]。Sydney Brenner 曾用秀丽隐杆线虫作为模型研究细胞凋亡的分子机制，并获得 2002 年度的诺贝尔生理与医学奖[4]。目前，秀丽隐杆线虫已经成为生命科学领域最流行的模式生物之一，它的作用越来越重要[5]。

作为一种经典的模式生物，秀丽隐杆线虫基因组结构的详细注释和解析对于线虫生物学研究来说极为重要，因此一直以来是学者关注的重点之一。秀丽隐杆线虫的全基因组测序始于 1992 年[6]，随着大规模并行测序[7]、基因结构和编码区(Coding sequences，CDSs)的计算预测[8]方法的兴起，秀丽隐杆线虫基因组结构越来越清晰。1998 年报道秀丽隐杆线虫基因组有 19 099 个 CDSs[9]，而到 2012 年报道的 CDSs 数目就增长到了 25 634 个(WS230 版)[10]。随着新的转录组数据的不断发布，先前那些不确定的编码基因逐步被显示出来，因此秀丽隐杆线虫基因组中 CDSs 的数目也在不断被调整[11]。显然，随着预测算法的不断改进，更多的 CDSs 会被发掘出来。

目前用于预测包括秀丽隐杆线虫在内的真核生物 CDSs 的常用算法有：从头预测算法[12]、判别分析算法[13]、神经网络算法[14]、隐马尔科夫链算法[15]、基于同源性的算法[16]、基于一致性的算法[17]等。这些算法主要涉及到开放阅读框(Open reading frames，ORFs)的定位以及内含子(Intron)或外显子(Exon)结构的描述，相关实现和程序已多有报道[18]。然而，大多预测程序较为复杂繁琐，同时对网络的依赖程度过高，因而不易推广。为进一步提高真核生物基因组 CDSs 预测的准确性，发掘更多潜在的 CDSs，本研究基于机器学习的策略[19]和碱基成分偏移特性[20]，结合基因组 CDSs 特征数学参数(d)[21]，运用 C 语言程序设计，以秀丽隐杆线虫染色体上已确定的 CDSs 作为训练数据进行统计计算，获得该特征参数的取值区间。随后以此参数的取值区间为基准，对秀丽隐杆线虫基因组中已确证的 CDSs、外显子、内含子、5′非翻译区(5′-Untranslation regions，5′-UTRs)、3′非翻译区(3′-Untranslation regions，3′-UTRs)和随机序列(Random sequences，RSs)等片段进行预测并验证。试验结果

* 收稿日期:2019-11-14　修回日期:2020-04-25　网络出版时间:2020-06-08 15:09
资助项目:国家自然科学基金面上项目(No.31871274)；重庆市自然科学基金面上项目(No.cstc2019jcyj-msxmX0527)；重庆市教育委员会科学研究项目(No.KJQN201800523)；重庆师范大学科研启动基金(No.17XLB017)
第一作者简介:李勃，男，讲师，博士，研究方向生物信息学，E-mail：libcell@cqnu.edu.cn；通信作者:刘明伟，男，副教授，博士，E-mail:liumw@cqmu.edu.cn
网络出版地址:https://kns.cnki.net/kcms/detail/50.1165.N.20200608.1405.004.html

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

表明,用 d 作为秀丽隐杆线虫基因组 CDSs 的特征参数是可行的,它的特定的取值区间可以很好地区分 CDSs(或 ORFs 及它的子片段)和其他非编码序列。本研究采用的以 C 编程实现秀丽隐杆线虫基因组 CDSs(或 ORFs 及它的子片段)预测的方法,可进一步推广至其他真核生物,为真核基因结构预测和基因功能注释等功能基因组学研究提供一种新颖简便的本地化方法。

1 材料与方法

1.1 数据和问题的定义

1.1.1 秀丽隐杆线虫全基因组序列数据的获取　本研究所使用的秀丽隐杆线虫全基因组数据来自 GenBank,文件保存为 *.gbk 格式,该数据集涵盖了 5 条常染色体和 1 条性染色体的 DNA 序列。

1.1.2 秀丽隐杆线虫基因组 CDSs 的数学特征参数 d　首先,为了便于统计在秀丽隐杆线虫基因组中单个基因编码区(由外显子拼接而来的编码蛋白产物的 DNA 序列,即 CDSs)上 A,T,C,G 等碱基的分布,采用生物统计中的"滑动窗口法"[22]:

1) 在给定的 DNA 片段上,选择 1 个窗口。1 个窗口即 1 段 DNA 序列,长度通常为 3 的倍数。

2) 分析该窗口对应的部分序列的碱基成分偏移量(D)。首先,对选定的窗口分别累加得到第 1 相位(即第 1,4,7,10,13,16,…位)上 4 种碱基 A,C,G,T 的总个数 C_{1A},C_{1C},C_{1G}和 C_{1T},第 2 相位(即第 2,5,8,11,14,17,…位)上 4 种碱基的总个数 C_{2A},C_{2C},C_{2G}和 C_{2T},以及第 3 相位(即第 3,6,9,12,15,18,…位)上 4 种碱基的总个数 C_{3A},C_{3C},C_{3G}和 C_{3T}。若窗口的长度为 $3n$,则

$$C_{1A}+C_{1C}+C_{1G}+C_{1T}=C_{2A}+C_{2C}+C_{2G}+C_{2T}=C_{3A}+C_{3C}+C_{3G}+C_{3T}=n。$$

接着将个数转换为频率 f_{ij}:

$$f_{ij}=\frac{100\times C_{ij}}{n}。\tag{1}$$

式中 i 代表第 1,2 或 3 相位,j 代表 4 种不同碱基(A,T,C 或 G),而 f_{ij}代表某个相位上特定碱基的频率。参考(1)式,可计算某个碱基在密码子 3 个不同位置的平均分布频率 M_j,即:

$$M_j=\frac{f_{1j}+f_{2j}+f_{3j}}{3}。\tag{2}$$

式中 j 的含义与(1)式相同。参考 Staden 提出的方法[23],根据(1),(2)式可得 D 值,即:

$$D=\sum_j\sum_i\sqrt{(M_j-f_{ij})^2}。$$

若将 D_0 作为秀丽隐杆线虫基因组选定的 CDSs 参数 D 的总体均值,则可定义该片段的数学特征参数 d 如下:

$$d=D-D_0。\tag{3}$$

3) 将窗口沿序列移动 1 个步长(通常也是 3 的倍数),得到 1 个新的 D 值。

4) 窗口从 5′端向 3′端逐步移动,得到一系列对应于给定的 DNA 片段上不同局部的 D 值,并将这些 D 值依此连成线,即 D 值曲线。

5) 分别以序列 5′端的第 1,2,3 位碱基为窗口移动的起始位点,得到对应于 3 个不同相位 D 值曲线。

在 D 值曲线上,"峰"对应于 DNA 序列中碱基成分偏移显著的区域,即可能的 CDSs(或 ORFs 及它的子片段)。

1.2 用于特征参数 d 估算的 C 语言程序设计

为计算基因组 CDSs 的特征参数 d,设计了多组 C 程序,主要实现以下数据分析过程:1) 提取 *.gbk 文件中已证实的每条 CDS 的起始位点和终止位点,并记录每个基因完整 CDS 中外显子的组数;2) 计算和记录上述的每段 CDS 的 D 值,并保存为 TXT 文档;3) 计算和记录相应 CDS 的中碱基 A,T,C,G 在密码子第 1,2,3 相位出现的个数及它们占相应位置碱基的比例,并显示每段 CDS 中每组外显子的起始和终止位点、组数及该 CDS 的 D 值;4) 计算由上述步骤得到的众多 D 值相应的 D_0 和 d 值,并计算 d 的均值和标准差;5) 计算每条待预测 DNA 片段的 d 值,并进行参数 d 的验证。即输入某个 DNA 片段的起点和终点位置,即可给出关于此片段是否为 CDS(或 ORF 及它的子片段)的预测结果。

1.3 基因组序列编码区的统计分析

首先对 *.gbk 格式的序列进行预处理,接着批处理计算选定的已被实验确证的 CDSs 的 D_0 值,以及每条

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

CDS的D值。然后,计算d值及其均值和标准差,进行后续分析。计算过程依据如下:

$$\overline{d}=\sum_i f(d_i)\times d_i,\ \sigma=\sqrt{\sum_i f(d_i)\times(d_i-\overline{d})^2}。\tag{4}$$

通过上述处理,即可获得CDSs的数学特征及相关参数。

1.4 参数d的检验与CDSs预测效果的评价

为了检验上述方法是否有效,随机选择大量的CDSs、外显子、内含子、5′-UTRs、3′-UTRs和RSs进行验证,以确定参数d能否作为衡量编码区与其他片段差异的指标而进一步用于编码区预测。假定:若通过参数d预测后,发现大多数CDSs的d值取值区间与其他序列的d值取值区间有明显差异(换言之,可以通过参数d的取值区间所来区分CDSs与其他片段),则认定参数d对于区分CDSs和非编码序列的差异是有效的,可用于秀丽隐杆线虫以及其他真核生物基因组CDSs(或ORFs及它的子片段)的预测。

2 结果

2.1 秀丽隐杆线虫基因组中密码子的使用和组成

秀丽隐杆线虫基因组大小为100.27 Mb,其中碱基A,T,C和G的比例分别为17.73%,17.73%,32.27%和32.27%。本实验统计所有已被确证的CDSs,结果提示5条常染色体的CDSs数目分别为1 664,2 250,1 553,1 800和2 600条,而性染色体上有1 619条CDSs。各条染色体上碱基组成比例见表1。

表1 秀丽隐杆线虫染色体上给定编码区中三联体密码子1,2和3位的碱基分布情况

Tab.1 The base distribution of site 1, 2 and 3 of codons for a given coding sequence on *C. elegans* chromosome

碱基在ORF中的位置	A的比例/%	T的比例/%	C的比例/%	G的比例/%
Site 1 in Ch 1	29.24	19.12	20.74	30.90
Site 2 in Ch 1	33.71	27.18	23.31	15.80
Site 3 in Ch 1	29.26	31.29	18.67	20.78
Site 1 in Ch 2	29.76	20.08	20.33	29.83
Site 2 in Ch 2	32.84	28.19	23.20	15.77
Site 3 in Ch 2	28.54	30.63	20.00	20.83
Site 1 in Ch 3	29.09	19.24	20.94	30.73
Site 2 in Ch 3	33.40	27.20	23.40	16.01
Site 3 in Ch 3	29.25	31.04	19.06	20.65
Site 1 in Ch 4	29.33	20.16	20.30	30.21
Site 2 in Ch 4	32.59	27.90	23.57	15.94
Site 3 in Ch 4	29.83	32.01	18.65	19.51
Site 1 in Ch 5	30.37	21.63	19.21	28.79
Site 2 in Ch 5	32.08	29.76	22.80	15.36
Site 3 in Ch 5	29.26	31.91	19.52	19.32
Site 1 in Ch X	30.09	19.77	20.75	29.39
Site 2 in Ch X	33.22	27.55	23.34	15.89
Site 3 in Ch X	28.78	30.02	21.34	19.86

注:表1中符号Ch代表染色体(Chromosome);6条染色体上所有确定的CDSs均被分析

Andrews曲线[24-25]是一种适用于多维数据结构可视化的方法。该方法的思想是根据三角变换方法将p维空间的点映射到二维平面上的曲线上,1条曲线代表1个样本点(即p维空间中的一个观测值),研究者可据此直观第找出数据的聚集情况[26]。将Andrews曲线用于密码子使用偏好性的可视化与探索性分析,发现已确定的CDSs上密码子不同位置的碱基使用偏好性在6条染色体间表现出高度相似(图1),这表明秀丽隐杆线虫所有染色体上的CDSs符合一个共同的碱基分布和使用规则,且碱基的使用并非均匀分布。因此,可选任意一条染色体上的CDSs去估算参数D和d的值。

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

2.2 秀丽隐杆线虫基因组的参数 D 和 d 值的分布规律

由于参数 D 表征了 CDSs 的碱基成分偏移程度，以 1 号染色体为基准计算它的 CDSs 的 D 值，并检查它的概率密度分布。利用 R 包 fitdistrplus 中的 descdist 函数[27]进行可视化统计检验的结果如图 2 所示，蓝色数据观察点最接近对数正态分布(Logarithmic normal distribution)；结合对数化 D 值的直方图(图 3)，可以基本确定 CDSs 区的 D 值近似服从对数正态分布，且据(4)式估算得到它的均值和标准差分别为 0.539 和 0.169。

结合(3)式，由于 D 近似服从对数正态分布，故 $(d+D_0)$ 也近似服从对数正态分布，将它对数化之后均值为 −0.663、标准差为 0.305。根据统计学规律可知，对 68.3% 的 CDSs 来说，它们的 $\ln(d+D_0)$ 值应落在目标区间 [−0.968，−0.358]。

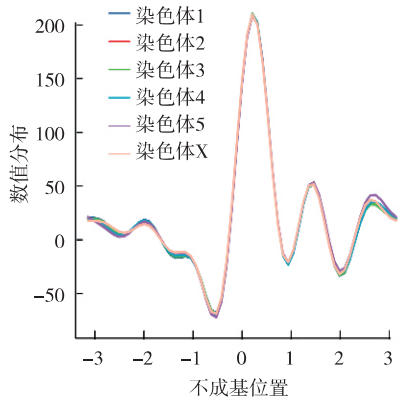


图 1　所有 CDSs 上密码子不同位碱基构成比例的 Andrews 曲线图

Fig.1　Different base composition ratios of codons on all CDSs

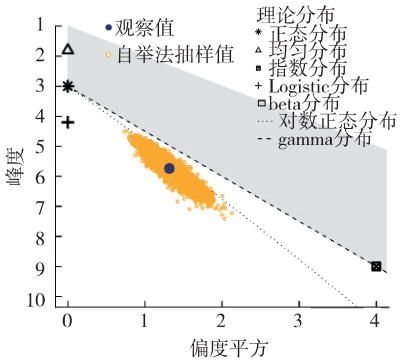


图 2　根据染色体上 CDSs 区统计结果进行参数 D 数据分布的估计

Fig.2　Estimation of distribution of parameter D based on statistical results in CDSs on chromosome

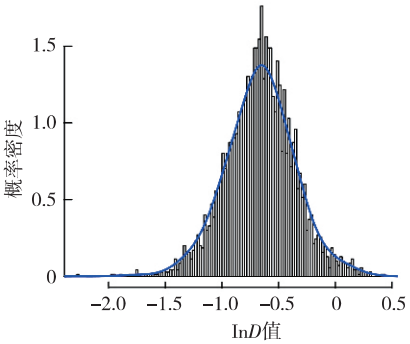


图 3　由随机选择的 CDSs 计算而来的 D 值分布直方图与概率密度图

Fig.3　Histogram and probability density diagram of D value calculated by randomly selected CDSs

2.3 基于 d 值的 CDSs 的预测验证

为检验 d 值否可以用来区分 CDSs 序列，对 d 值用于预测的效率进行研究。随机取若干条 CDS 序列、外显子、内含子、5′-UTR、3′-UTR 和 RS，将它们均假定为 CDSs 来计算它们相应的 d 值，并依此来推测它们为潜在 CDSs(或 ORFs 及它的子片段)的可能性：若某一片段的 $\ln(d+D_0)$ 值落在上述目标区间内，则该片段被预测为 CDS(或 ORFs 及它的子片段)，否则即被认定为非编码序列。结果显示有 81.6% 的 CDSs 被判定为 CDSs，有 70.7% 的外显子被认定为 CDSs(或 ORFs 及它的子片段)，另有 21.8% 的内含子、17.8% 的 5′-UTRs、31.4% 的 3′-UTRs 以及 4.7% 的 RSs 被误判为 CDSs(或 ORFs 及它的子片段)(表 2)。可以看出，CDSs 出现误判的概率最低。以上结果表明：参数 d 可以作为一个有效区分 CDSs(或 ORFs 及它的子片段，如外显子)和非编码区的特征参数，用于 CDSs(或 ORFs 及它的子片段)的初步预测。

为深入探讨参数 d 预测 CDSs(或 ORFs 及它的子片段)的效果，再次随机选择以上 6 类序列各 60 条，计算各自的 $\ln(d+D_0)$ 值并查看它们是否落入 CDSs 的目标预测区间内(图 3)。

图 4 清晰地显示，CDSs 和外显子由于 d 值大部分满足转换后落入 CDSs 的目标取值区间内，故很容易被预测为 CDSs(或 ORFs 及它的子片段)，显示了该方法能很准确地预测出 CDSs(或 ORFs 及它的子片段)。相反地，大多数内含子和随机序列不易被误判为 CDSs。而对于 UTRs 来说，因它们的 d 值转换后偶尔会落入 CDSs 的特征取值区间而间或被错误预测为 CDSs，特别是 3′-UTRs。有趣的是，对内含子和

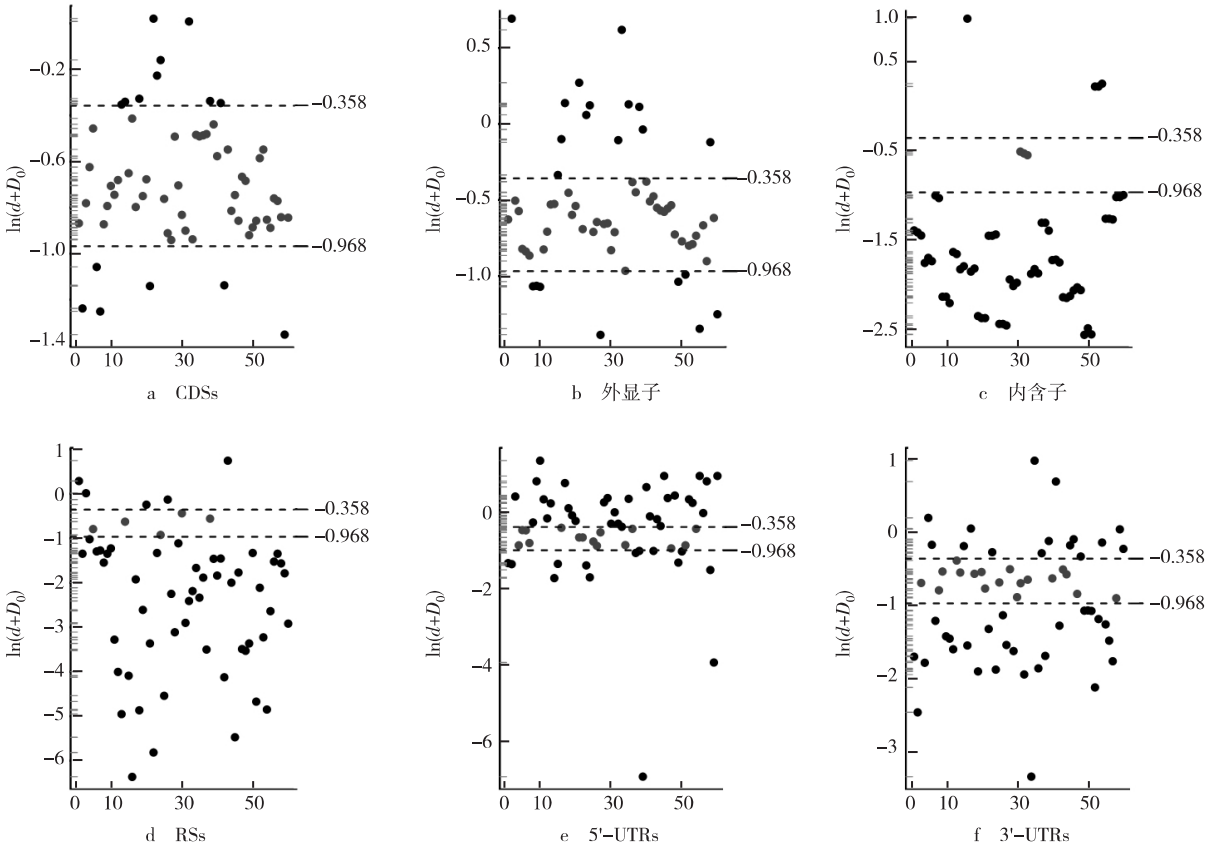
表 2　随机选取的序列数目、落入 d 取值区间的数目及被预测为 CDSs 的概率

Tab.2　The number of selected sequences and the sequences falling into the d value range, and the probability of being predicted as CDSs

序列类型	总数目	区间内的条数	所占比例
CDSs	1 600	1 306	81.6%
外显子	1 600	1 131	70.7%
内含子	1 600	349	21.8%
5′-UTRs	600	188	17.8%
3′-UTRs	600	107	31.4%
RSs	600	28	4.7%

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

RSs 而言,它们的 D 值往往接近 0 而 d 值在 0.5 左右,这说明这两类序列不具有强烈的碱基成分偏移特性。



注:实心点代表 60 个随机抽取的序列,虚线代表编码区特征参数 d 变换之后 $\ln(d+D_0)$ 的特征取值区间

图 4 60 条随机选取的不同序列 d 值的分布情况

Fig.4 Distribution of d values of 60 randomly selected different sequences

3 讨论

本研究结果显示,秀丽隐杆线虫的 6 条染色体上 CDSs 密码子不同位置的碱基组成极为相似(表 1 和图 1),具有明显的碱基使用偏好性。理论上讲,4 种碱基分布均匀的序列不太可能用来编码蛋白质或者多肽,提示对三联体密码子每个相位的碱基分布频率进行评估,可以用来区分 CDSs 和非编码区,进而精确预测基因结构和 CDSs(或 ORFs 及它的子片段)。显然,碱基成分偏移的检查和计算可能提供了一种可信的途径。在本研究中,笔者开发了多组 C 语言程序,分析了大量的 DNA 序列,并计算各类序列的 D 和 d 值从而比较 CDSs 和其他序列的差异,然后进行数据挖掘和验证。结果证实该方法可靠,用参数 d 可以表征 DNA 片段的差异性,其特定取值区间能很好地区分 CDSs(或 ORFs 及它的子片段)和其他非编码区。

然而,此法也存在一定的局限性。比如,它要求被预测的片段长度应该是 3 的倍数,否则预测将不太准确。另外,对于 UTRs 而言有一部分会被误判为 CDSs(或 ORFs 及它的子片段),尤其是 3′-UTRs 更容易被误判,推测原因可能在于某些情况下蛋白质的翻译被提前终止。也就是说,在进化的进程中某些 3′-UTRs 可能是潜在的蛋白质编码区。相比而言,5′-UTRs 往往不易被误判为 CDSs(或 ORFs 及它的子片段),可能因为 mRNA 上翻译的起始是非常精确的,细胞有一套极为精确的翻译起始机制(比如需要核糖体与 mRNA 上翻译起始位置的精准识别和结合,即真核生物翻译起始的"扫描模式"[28])。此外也有一小部分外显子被鉴定为非编码区,原因可能在于外显子长度太短,提示用此法预测 CDSs(或 ORFs 及它的子片段)时被预测片段的长度不能太短且尽可能是 3 的倍数。同时还有一部分内含子被鉴定为 CDSs(或 ORFs 及它的子片段),推测可能存在极少数的内含子具有可编码蛋白质的 ORFs,该情况在先前的研究中已有少数报道[29]。

综上所述,本研究提出了一种利用 C 语言编程结合碱基成分偏移来预测真核基因组中编码区的方法:首先

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

利用现存的 CDSs 作为训练数据，进行处理获得碱基成分偏移量 D 和 CDSs 的数学特征参数 d；然后借助机器学习的策略依据 d 的取值来判定某一序列是否为蛋白编码序列 CDSs(或 ORFs 及它的子片段)。通过对秀丽隐杆线虫基因组的测试，发现用它进行 CDSs(或 ORFs 及它的子片段)的预测，不但有效而且更精确。与以往类似研究相比，用 C 编程的方法对 *.gbk 格式的秀丽隐杆线虫基因组序列文件进行数据挖掘是本研究的一大特色。作为一种行之有效的本地化研究工具，它可以极大地提高 CDSs 甚至基因结构预测的效率，进一步加深人们在分子水平上对基因和蛋白质结构和生命活动的理解。此外，为了弥补可移植性的不足，本研究将设计的 C 语言程序编译为可执行文件，以获得较强的可移植性，降低了对操作系统的依赖性。

参考文献：

[1] BRENNER S. The genetics of *Caenorhabditis elegans*[J]. Genetics, 1974, 77(1): 71-94.

[2] BARGMANN C I. Neurobiology of the *Caenorhabditis elegans* genome[J]. Science, 1998, 282(5396): 2028-2033.

[3] LEE R C, FEINBAUM R L, AMBROS V. The *C. elegans* heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14[J]. Cell, 1993, 75(5): 843-854.

[4] FRIEDBERG E C. Sydney brenner[J]. Nature Reviews Molecular Cell Biology, 2008, 9(1): 8-9.

[5] STERKEN M G, SNOEK L B, KAMMENGA J E, et al. The laboratory domestication of *Caenorhabditis elegans* [J]. Trends in Genetics, 2015, 31(5): 224-231.

[6] SULSTON J, DU Z, THOMAS K, et al. The *C. elegans* genome sequencing project: a beginning[J]. Nature, 1992, 356 (6364): 37-41.

[7] HILLIER L W, MARTH G T, QUINLAN A R, et al. Whole-genome sequencing and variant discovery in *C. elegans*[J]. Nature methods, 2008, 5(2): 183-188.

[8] POP M, SALZBERG S L. Bioinformatics challenges of new sequencing technology[J]. Trends in Genetics, 2008, 24(3): 142-149.

[9] CONSORTIUM C E S. Genome sequence of the nematode *C. elegans*: a platform for investigating biology[J]. Science, 1998, 282(5396): 2012-2018.

[10] YOOK K, HARRIS T W, BIERI T, et al. WormBase 2012: more genomes, more data, new website[J]. Nucleic Acids Research, 2012, 40(D1): D735-D741.

[11] ALLEN J E, PERTEA M, SALZBERG S L. Computational gene prediction using multiple sources of evidence[J]. Genome Research, 2004, 14(1): 142-148.

[12] TER-HOVHANNISYAN V, LOMSADZE A, CHERNOFF Y O, et al. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training[J]. Genome Research, 2008, 18(12): 1979-1990.

[13] WANG Z, CHEN Y Z, LI Y X. A brief review of computational gene prediction menthods[J]. Genomics, Proteomics & Bioinformatics, 2004, 2(4): 216-221.

[14] PRADHAN M. Prediction using ANN-based classifier in DNA microarray[J]. International Journal of Applied Research on Information Technology and Computing, 2018, 9 (1): 1-14.

[15] CHAN K-L, ROSLI R, TATARINOVA T V, et al. Seqping: gene prediction pipeline for plant genomes using self-training gene models and transcriptomic data[J]. BMC Bioinformatics, 2017, 18(1): 1-7.

[16] MATHÉ C, SAGOT M F, SCHIEX T, et al. Current methods of gene prediction, their strengths and weaknesses [J]. Nucleic Acids Research, 2002, 30(19): 4103-4117.

[17] PAVLOVIĆ V, GARG A, KASIF S. A Bayesian framework for combining gene predictions[J]. Bioinformatics, 2002, 18(1): 19-27.

[18] GOEL N, SINGH S, ASERI T C. A comparative analysis of soft computing techniques for gene prediction[J]. Analytical Biochemistry, 2013, 438(1): 14-21.

[19] MIN S, LEE B, YOON S. Deep learning in bioinformatics [J]. Briefings in Bioinformatics, 2017, 18(5): 851-869.

[20] SHACKELTON L A, PARRISH C R, HOLMES E C. Evolutionary basis of codon usage and nucleotide composition bias in vertebrate DNA viruses[J]. Journal of Molecular Evolution, 2006, 62(5): 551-563.

[21] 张颖，李宏，吕军，et al. Yeast 基因组编码区特征参数的研究[J]. 生物物理学报, 2001, 17(3): 535-541.
ZHANG Y Y, LI H, LÜ J, et al. Resescrch in the characteristic parameter of gene conding region of yeast genome [J]. Acta Biophysica Sinica, 2001, 17(3): 535-541.

[22] GORODKIN J, HOFACKER I L, TORARINSSON E, et al. *De novo* prediction of structured RNAs from genomic sequences[J]. Trends in Biotechnology, 2010, 28(1): 9-19.

[23] STADEN R. Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes[J]. Nucleic Acids Research, 1984, 12(1Part2): 551-567.

[24] ZAWIEJA B, GLINA B. Application of multivariate statistical methods in the assessment of mountain organic soil transformation in the central Sudetes[J]. Biometrical Letters, 2017, 54(1): 43-59.

[25] PETRACKOVA A, HORAK P, RADVANSKY M, et al.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

Cross-Disease Innate Gene Signature: Emerging Diversity and Abundance in RA Comparing to SLE and SSc[J]. Journal of Immunology Research, 2019, 3575803.

[26] GARCIA-OSORIO C, FYFE C. Visualization of high-dimensional data via orthogonal curves[J]. Journal of Universal Computer Science, 2005, 11(11): 1806-1819.

[27] DELIGNETTE-MULLER M L, DUTANG C. Fitdistrplus: An R package for fitting distributions[J]. Journal of Statistical Software, 2015, 64(4): 1-34.

[28] NELSON D L, COX M M. Lehninger principles of biochemistry[M]. New York: W. H. Freeman & Company, 2008.

[29] PERLMAN P S, BUTOW R A. Mobile introns and intron-encoded proteins[J]. Science, 1989, 246(4934): 1106-1109.

Animal Sciences

Study on Characteristic Parameters for Genome Coding Sequences in *Caenorhabditis elegans*

LI Bo[1], FANG Bin[2], TAO Jingxin[1], LIU Mingwei[3]

(1. College of Life Sciences, Chongqing Normal University, Chongqing 401331;
2. Wuhan Institute of Biological Products, Co.LTD., Wuhan 430207;
3. College of Laboratory Medicine, Chongqing Medical University, Chongqing 401331, China)

Abstract: [**Purposes**]To investigate the characteristics of base usage in the coding sequences of *Caenorhabditis elegans* genome, and to extract the characteristic parameters of base composition deviation, for discriminating coding and non-coding sequences.[**Methods**] According to the characteristics of base component distribution, a new parameter d was defined, and then its distribution was investigated. In addition, the feasibility of using this index as a characteristic parameter of genomic coding region was analyzed and verified by random sampling from six types of different DNA sequences of *C. elegans* genome.[**Findings**]The results showed that the parameter d approximately obeyed the lognormal distribution after linear transformation. Sampling analysis indicated that 81.6% of CDSs, 70.7% of exons, 21.8% of introns, 4.7% of RSs, 17.8% of 5′-UTRs and 31.4% of 3′-UTRs fell within the range of characteristic values after d value transformation. That is, it is predicted to be the genomic coding region CDSs.[**Conclusions**]The index d, standing for base component offset, can be used as a characteristic parameter to characterize the coding sequences of the genome, and its specific value interval can well distinguish CDSs (or ORFs and its subfragments) from other non-coding sequences.

Keywords: *C. elegans* genome; protein coding sequence; base composition bias; data mining; characteristic parameter

(责任编辑 陈 乔)

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net