doi:10.3969/j.issn.1672-5565.2016.04.01

基因组序列 8-mer 频次使用规律及与物种进化的关系

朱孝先,杨 镇,段成妍,吕文萍,李 宏*
(内蒙古大学物理科学与技术学院,呼和浩特 010021)

摘 要:基因组序列 k-mer 的非随机使用规律及包含的生物学意义一直是人们关注的问题,目前还没有根本性进展。本文以七个物种的全部基因序列为样本,得到各物种基因组序列的 8-mer 频谱分布。发现狗和牛的频谱有三个峰,而斑马鱼、青鳉鱼、秀丽线虫和酿酒酵母的频谱只有一个峰,鸡的频谱分布形状介于两者之间。将 8-mer 集合按照 XY 二核苷含量分类,结果显示只有 CG 二核苷分类下 0CG、1CG 和 2CG 三类子集的频谱形成各自独立的单峰分布。对照随机序列,发现 0CG 模体是随机进化的,1CG 和 2CG 模体是定向进化的,它们的使用频次远小于随机频次,且这种独立进化分离规律具有物种普适性。三个 CG 子集频谱之间的距离是产生单峰或多峰现象的根本原因。将七个物种基因组序列标准化到 10^9bp,比较发现 1CG 和 2CG 子集频谱与物种进化显著相关,0CG 子集频谱与物种进化无显著关系。可以认为三种 CG 模体各自执行着不同的生物学功能。基因组序列 8-mer 的独立分离规律为揭示基因组结构、基因组进化以及模体的生物功能提供了一种新的思维方式。

关键词:基因组序列;8-mer 频谱;CG 二核苷分类;独立分离规律;基因组进化
中图分类号:Q343.1 **文献标志码**:A **文章编号**:1672-5565(2016)04-195-08

Rules of 8-mer usage in genome sequences and its relation to genome evolution

ZHU Xiaoxian, YANG Zhen, DUAN Chengyan, LÜ Wenping, LI Hong*
(*School of Physical Science and Technology*, *Inner Mongolia University*, *Huhhot* 010021, *China*)

Abstract: The rules of k-mer non-random usage in genome sequences and its biological significance are important problems and its mechanism is still not clear. Based on seven genome sequences, the distributions of 8-mer frequency spectra were gotten. Results show that 8-mer spectra of dog and cow are trimodal and of zebra fish, medaka, nematode and yeast are unimodal. For chicken genome, the 8-mer spectrum is a medium between the two models. When the 8-mer set were classified into three subsets according to XY dinucleotide content, results show that only if in CG dinucleotide classification, the 0CG, 1CG and 2CG subsets form independent and unimodal distributions respectively. Compared with random sequences, it is found that 0CG motifs are the result of the random evolution, 1CG/2CG motifs are the result of the directed evolution and their frequencies are far low from the random frequencies. The rules of independent separation for the three CG subsets have species universality. Results indicate that the prime reasons about unimdals or multimodals of 8-mer spectra in different species are the distance differences of the three CG spectra. When seven genome sequences are normalized into 10^9 bp, results show that the spectra of 1CG and 2CG motifs are correlated significantly with genome evolution and of 0CG motifs has not obvious relation to genome evolution. We think that the three CG motifs have different biological functions. The rules of independent separation for the three CG subsets will provide a novel idea to research genome structures and evolutions and provide a method to reveal the functional elements in genome sequences.

Keywords: Genome sequence; 8-mer spectrum; CG dinucleotide classification; Independent separation rule; Genome evolution

---

收稿日期:2016-06-23;**修回日期**:2016-08-24.
基金项目:国家自然科学基金项目(No.31260219);国家级大学生创新训练计划项目(No.201512149)。
作者简介:朱孝先,男,本科生,研究方向:理论生物物理;E-mail:zhxixa@qq.com.
*通信作者:李 宏,男,教授,博士生导师,研究方向:理论生物物理;E-mail: ndlihong@imu.edu.cn.

基因组 DNA 序列 k-mer 非随机使用的已有许多研究报道。研究主要关注 k-mer 字频非随机使用的生物学意义[1-2]、k-mer 分布的概率模型[3-4]、稀有 k-mer 和富含 k-mer 的片段及其在基因组序列上的分布[5-8]。一些工作分析了特殊 k-mer 片段与 CpG 岛序列的关系[9-11],从 k-mer 使用入手寻找功能位点的调控片段,预测 RNA 功能片段,给出基因组组装方法[12-15]。最近,一些研究者特别关注了基因组序列 k-mer 的频谱分布,期望揭示基因组序列的构成特性和进化。Chor[16]研究了近百个物种基因组序列 k-mer 频谱的分布特征,发现不同生物基因组序列或不同类型 DNA 序列的 k-mer(k>7)频谱分布呈现单峰或多峰现象。低等生物基因组祖列一般是单峰分布,高等生物如四爪哺乳动物基因组序列呈现三峰分布,所有物种编码序列是单峰分布。研究组 Bao[17-18]等分析了人类 1 号染色体各类序列的 8-mer 频谱,结果显示基因间序列和内含子序列呈三峰分布,编码序列是单峰分布,5' UTR 和 3' UTR 序列分布介于两者之间。研究发现 8-mer 集合按照 CG 二核苷含量分成三类后,三个 CG 模体子集的频谱各自形成独立的单峰分布,而其它 15 种 XY 二核苷分类则没有此现象,由此揭示了产生单峰和多峰频谱的根本原因并推测了三个 CG 模体子集的生物学功能。为了验证基因组序列中三个 CG 模体子集独立进化规律的普适性,以七个物种基因组序列为研究对象,分析每个物种中三个 CG 模体子集频谱的分布形式,由此验证独立进化规律的普适性。同时,通过对物种间的对比,分析这种进化分离现象与物种进化之间的关系。

1 数据与方法

1.1 七种生物全基因组序列的提取与总长度

七个物种狗、牛、鸡、斑马鱼、青鳉鱼、秀丽线虫、酿酒酵母的全基因组序列以及注释信息均来自 UCSC Genome Browser Home (http://genome.ucsc.edu)。所有的物种没有考虑性染色体,因为性染色体与物种的进化关系还不明确。其相关信息见表 1。

表 1 七种生物基因组序列信息

Table 1 Sequences information of seven genomes

物种名称		基因组总长度/(bp)	染色体数/条	发布版本
Dog (狗)	*H. hyaena*	2 318 226 206	38	May. 2005 (canFam2)
Cow (牛)	*B. taurus*	2 545 896 661	29	Oct. 2007 (bosTau4)
Chicken (鸡)	*G. gallus*	957 020 481	28	May.2006 (galGal3)
Zebra fish (斑马鱼)	*D. rerio*	1 322 655 876	25	Dec. 2008 (danRer6)
Medaka (青鳉鱼)	*O.latipes*	712 924 927	24	Oct. 2005 (oryLat2)
Nematode(秀丽线虫)	*C. elegants*	82 553 665	5	May 2005 (apiMel3)
Yeast (酿酒酵母)	*S. cerevisiae*	12 071 326	16	Feb. 2013 (ce11)

1.2 8-mer 按照二核苷含量分类

8-mer 模体集合共有 $4^8 = 65\,536$个。理论上,包含 0 个二核苷 XY(X,Y = A、T、C、G)的 8-mer 有 40 545个,记为 0XY;包含 1 个二核苷 XY 的 8-mer 有21 468个,记为 1XY;包含两个或两个以上 XY 二核苷的 8-mer 有3 523个,记为 2XY。若两个核苷酸相同,则三类集合的数目分别是44 631、14 930 和 5 974个。按照上述约束,可将全体 8-mer 集合分为 0XY、1XY 和 2XY 三个模体子集。这种分类称为 XY 二核苷分类,这种分类一共有 16 种。

1.3 8-mer 相对模体数

对于给定的 DNA 序列,以 8 bp 作为窗口,1 bp 作为步长,统计得到每个 8-mer 在该序列中出现的频次。8-mer 相对模体数 (Frequency of Appearance, *FA*) 定义如下:若频次区组 i 中出现的 8-mer 个数为 N_i,则该区组上 8-mer 的相对模体数为:

$$FA = \frac{N_i}{4^8} \tag{1}$$

以 8-mer 使用频次作为横坐标,相对模体数 *FA* 作为纵坐标,得到 8-mer 相对模体数随频次的分布。

1.4 序列长度标准化

对于不同长度的 DNA 序列,为了方便比较它们的 8-mer 使用频次之间的关系,对序列进行标准化。对于本文研究的基因组序列,将所有基因组序列的长度标准化到 10^9 bp。若某序列长度为 N bp,则对该序列 8-mer 出现的频次乘以一个权重系数 λ。

$$\lambda = \frac{10^9}{N} \qquad (2)$$

2　结果与分析

2.1　各物种基因组序列的 8-mer 频谱

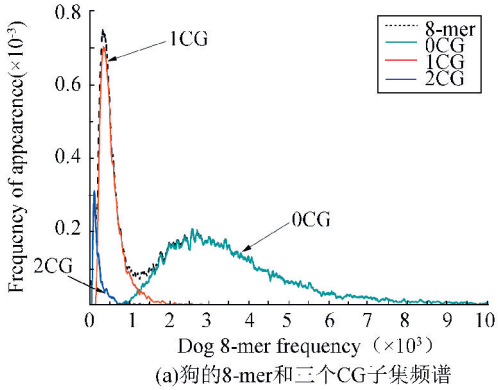
统计得到七个物种全部常染色体 DNA 序列的 8-mer 频次,按照公式(1)得到各物种 8-mer 相对模体数随频次的分布,结果见图 1 中用“8-mer”标注的外围曲线。图 1 展示的频谱分布经过了光滑处理。分析频谱发现:狗和牛的 8-mer 频谱有三个峰,而斑马鱼、青鳉鱼、秀丽线虫和酿酒酵母的 8-mer 频谱只有一个峰,鸡的频谱分布形状介于两者之间。对于不同的物种,8-mer 频谱展现的分布形式有所不同。

2.1　XY 二核苷分类下各模体子集的频谱

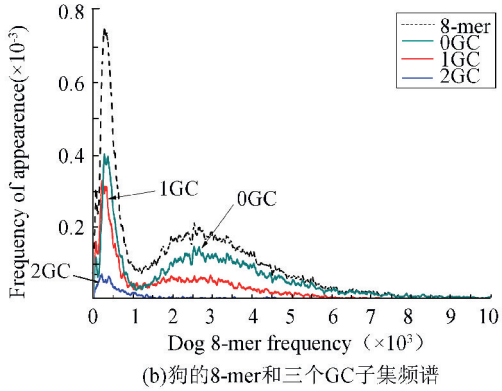
为了进一步研究不同物种 8-mer 频谱分布的差异和 8-mer 非随机使用的规律,把各基因组的 8-mer 做 16 种 XY 分类,部分结果见图 1。图 1 中只给出了 CG 和 GC 分类下各个 8-mer 子集的频谱。分析 16 种 XY 分类下 0XY、1XY 和 2XY 子集的 8-mer 频谱,发现所有物种在 CG 分类下三个子集的 8-mer 各自形成完全独立的单峰分布,而其他 15 种 XY 分类下的三个子集均不能形成独立的单峰分布。

分析图 1,可以发现所有物种中,基因组在 CG 分类下 0CG、1CG 和 2CG 三个模体子集的分布具有相同的规律。首先,1CG 模体出现的频次小于 0CG,2CG 模体出现的频次小于 1CG。其次,2CG 和 1CG 模体出现的频次具有保守性,即它们的频次分布在一个很窄的范围之内。而 0CG 模体的频次分布范围很广。其次,三个 CG 子集模体分布的最概然相对模体数具有相同的规律,即 1CG 的最概然相对模体数明显高于 0CG 和 2CG。使用其他 15 种二核苷来分类得到的图谱中,频数分布并不是完全独立的单峰,每个子集的频数分布与全部 8-mer 分布形状相似。第三,三个 CG 子集之间的距离随着进化水平的提高而增加,1CG 和 2CG 的峰越来越高,峰的宽度越来越窄,显示了这两个子集模体的使用保守性越来越强。第四,CG 模体的独立进化现象具有物种普适性,这种进化规律在低等真核生物酵母基因组中就已经显现出来。

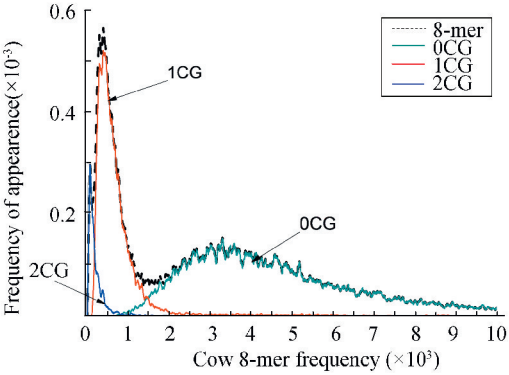
对于狗和牛而言,三个 CG 子集分布的中心与总体 8-mer 多峰分布中心相对应,由于三个 CG 子集分布中心距离很远,其叠加效果呈现三峰形式。对于单峰分布的物种,由于三个 CG 子集分布中心距离很近,其叠加效果呈现单峰形式。对于鸡这个物种,因为三个 CG 子集分布中心相对单峰分布的物种要远一些,叠加之后形成的总体分布会出现介于单峰和三峰之间的分布形状。之所以不同物种的 8-mer 图谱会有单峰和多峰的现象,是因为每个物种的三个 CG 子集分布之间的距离不同造成的。研究结果圆满解释了不同物种 8-mer 频数单峰和多峰分布现象产生的原因。



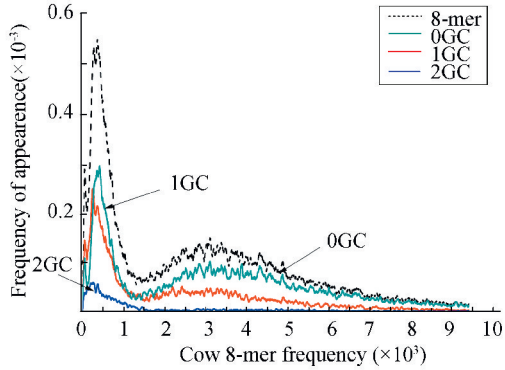
(a)狗的8-mer和三个CG子集频谱



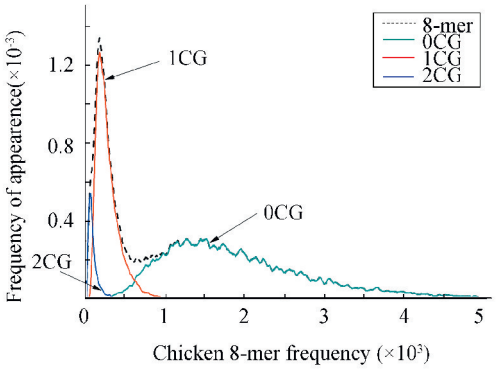
(b)狗的8-mer和三个GC子集频谱



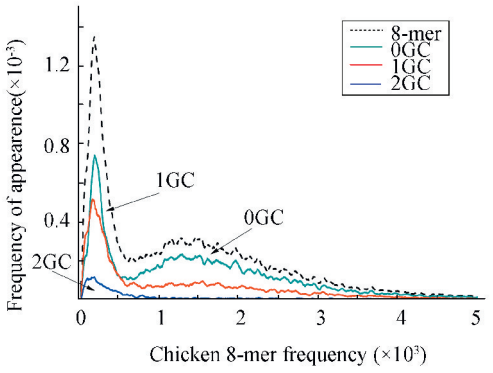
(c) 牛的8-mer和三个CG子集频谱



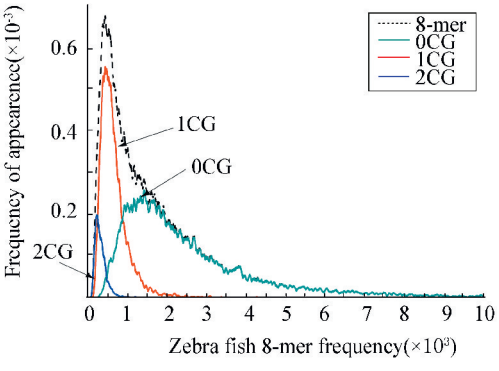
(d) 牛的8-mer和三个GC子集频谱



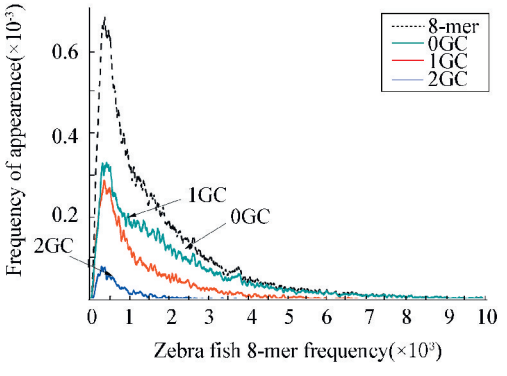
(e) 鸡的8-mer和三个CG子集频谱



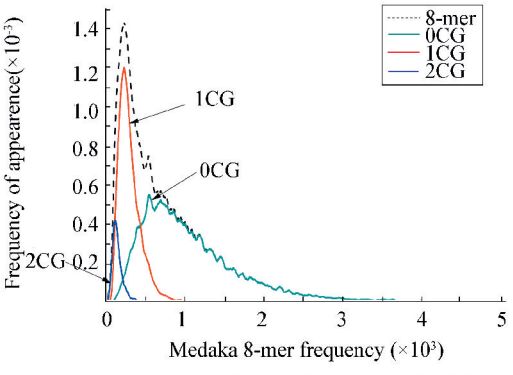
(f) 鸡的8-mer和三个GC子集频谱



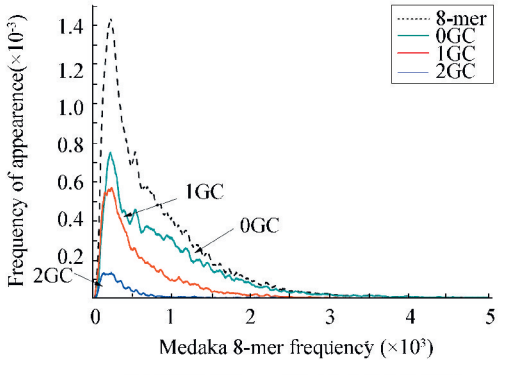
(g) 斑马鱼的8-mer和三个CG子集频谱



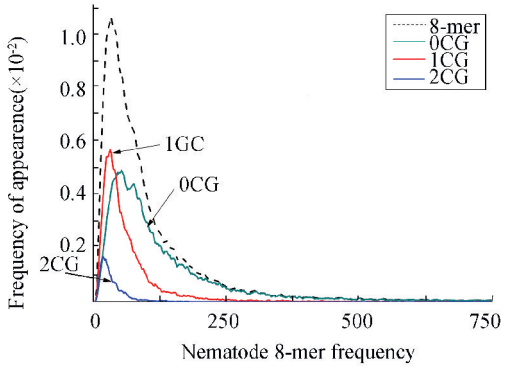
(h) 斑马鱼的8-mer和三个GC子集频谱



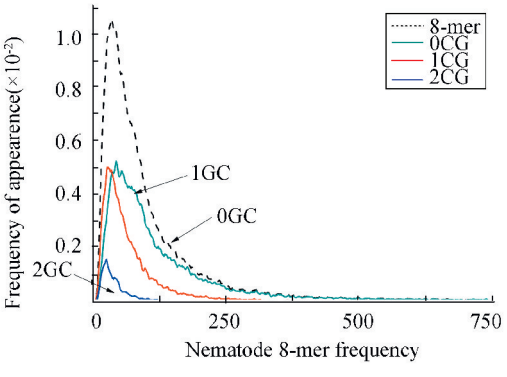
(i) 青鳉鱼的8-mer和三个CG子集频谱



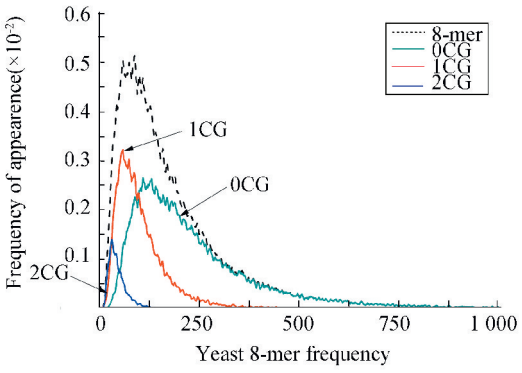
(j) 青鳉鱼的8-mer和三个GC子集频谱



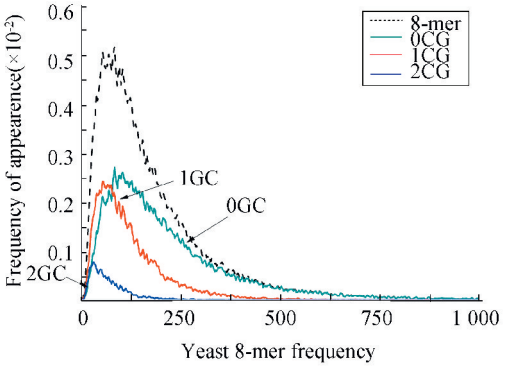
(k) 线虫的8-mer和三个CG子集频谱



(l) 线虫的8-mer和三个GC子集频谱



(m) 酵母的8-mer和三个CG子集频谱



(n) 酵母的8-mer和三个GC子集频谱

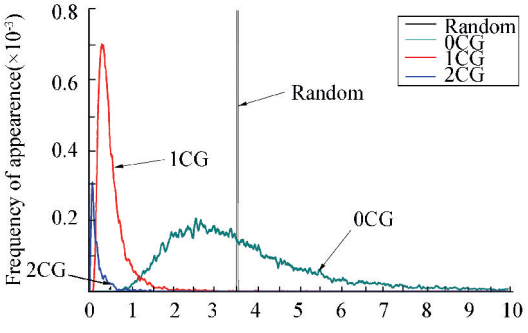
图 1　七个物种基因组序列 8-mer 相对模体数(*FA*)随频数的分布图谱

Fig.1　Distributions of frequency of appearances (*FA*) with 8-mer frequencies in seven genome sequences

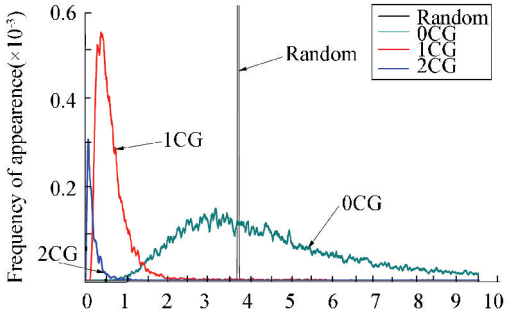
2.3　三个 CG 子集频谱分布与随机序列比较

为了揭示三个 CG 子集的使用与进化的关系,针对每一个基因组序列,使用长度相同,碱基组分完全随机的随机序列作为对照分析。选取完全随机序列的原因是我们只关心随机序列 8-mer 频次分布的最概然频次位置。若将碱基含量设定为与实际基因组序列组分相同的随机序列,则 k-mer 分布会出现 $k+1$ 个尖锐且离散的峰,比较其包络曲线的最概然频次很不直观,但效果相同。

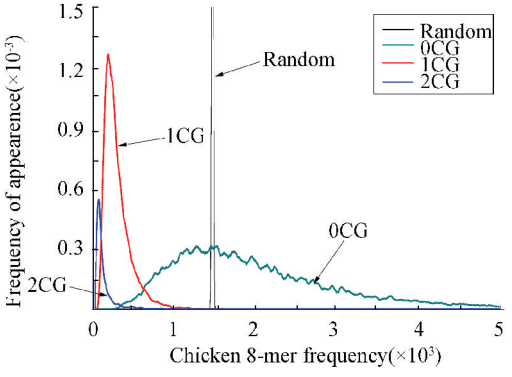
图 2 给出各物种三个 CG 子集频数分布与随机序列的分布。可以看出,各物种中 0CG 子集的频数分布中心最靠近随机分布中心,1CG 次之,2CG 子集频数分布中心距离随机分布中心最远。这说明 0CG 模体的使用频次基本上是围绕随机中心进化的,或者说 0CG 模体的频次分布主要是在随机压力下进化的。而 1CG 和 2CG 模体的频次使用远离随机中心,表明这两类模体是定向进化的,而且随着物种进化,其频次使用相对越来越小,且越来越保守。另外,在三类 CG 模体中,2CG 模体的 G+C 含量最高,0CG 模体的 G+C 含量最低,1CG 模体的 G+C 含量介于两者之间。就是说 G+C 含量越高的模体出现的频率越低。



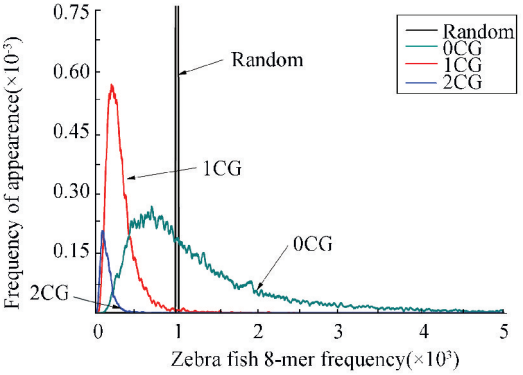
(a)狗的三个CG子集与随机序列的频谱



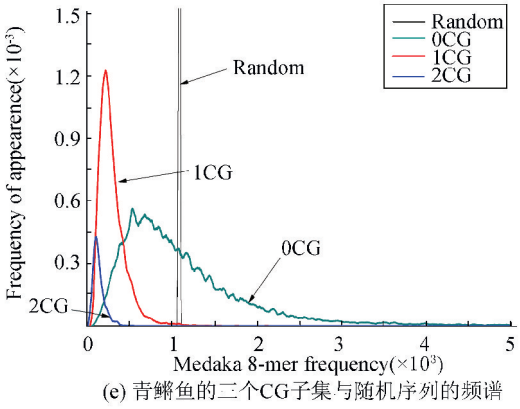
(b)牛的三个CG子集与随机序列的频谱



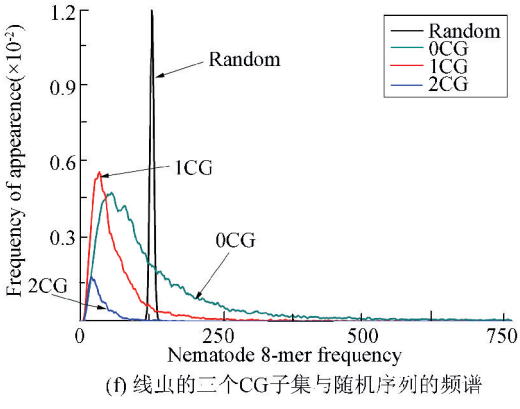
(c) 鸡的三个CG子集与随机序列的频谱



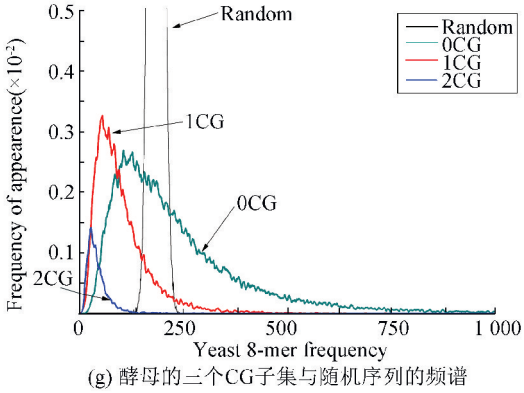
(d) 斑马鱼的三个CG子集与随机序列的频谱



(e) 青鳉鱼的三个CG子集与随机序列的频谱



(f) 线虫的三个CG子集与随机序列的频谱



(g) 酵母的三个CG子集与随机序列的频谱

图 2 七个物种的三个 CG 子集 8-mer 频谱与随机序列的频谱比较

Fig. 2 Comparison between three CG subsets' spectrums of seven genomes and 8-mer spectrum of random sequences

为了分析各物种之间三个 CG 模体使用的进化规律,将各物种基因组序列长度标准化到 10^9 bp 后做对照分析。长度标准化后各物种三个 CG 模体子集的最概然频次见表 2。以随机序列 8-mer 最概然频次为参考点,得到三个 CG 模体子集最概然频次与随机序列最概然频次的比值,记为 ρ,它们与物种进化的关系见图 3。

表 2 基因组序列标准化后三个 CG 模体子集最概然频数

Table 2 Most probable frequencies of the three CG subsets spectrums in the normalized genome sequences

Name	Random	0CG	1CG	2CG
Dog	15 259	10 295	1 551	281
Cow	15 259	15 496	1 393	309
Chicken	15 259	18 021	1 824	429
Zebra fish	15 259	9 058	3 350	1 452
Medaka	15 259	8 768	2 786	1 376
Nematode	15 259	6 856	4 252	2 495
Yeast	15 259	11 101	5 302	2 320

可以看出在序列标准化后,随着物种由低等向高等的进化,1CG 和 2CG 模体的最概然频次与随机序列的最概然频次的比值逐渐减小,1CG 和 2CG 之间频次比的差值基本保持不变。表明 1CG 和 2CG 模体使用与物种进化密切相关,1CG 和 2CG 之间其频次使用保持同步进化,说明这两类模体使用在生物进化过程中受到的进化压力是相同的。0CG 模体的最概然频次与物种进化没有明显的联系,说明 0CG 模体的使用主要是在随机压力下进化的。比如酵母和狗,两类物种其 0CG 模体的最概然频次与随机序列的比值基本相同。

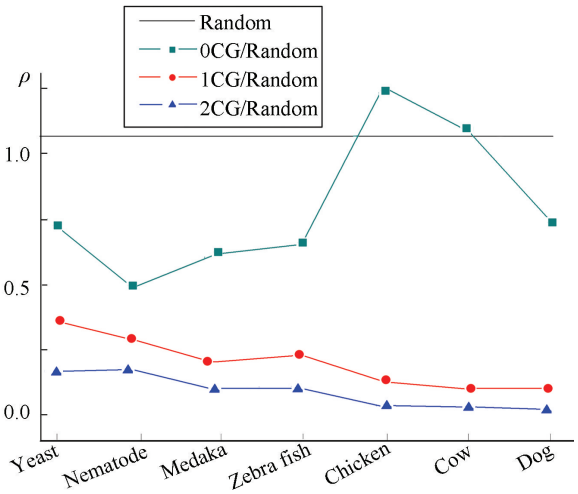


图 3　三个 CG 模体频谱的最概然频次与基因组进化之间的关系

Fig.3　Relations between the most probable frequencies of the three CG subsets spectrums and genome evolutions

注:纵坐标 ρ 为三个 CG 模体子集最概然频次与随机序列最概然频次的比值。

显然三类 CG 模体应该执行着不同的生物学功能。关于三类 CG 模体的功能,本研究组提出了一个猜想:2CG 模体是构成 CpG 岛序列的核心模体,也是 DNA 序列上各个功能位点的识别信号模体,1CG 模体与核小体结合模体紧密相关,0CG 模体与表观遗传多样性相关。关于 2CG 和 1CG 模体的功能,分别在人类和酵母基因组中得到初步验证。对于 0CG 模体的功能,Time Quante 推测富含 AT 片段的 DNA 序列(也就是 0CG 模体)是表观基因组的重要组成部分[19],该结论从另一个角度支持了我们的猜想。

3　讨论与展望

分析了七个物种基因组序列 8-mer 相对模体数随频次的分布规律,发现所有物种基因组序列在 16 种 XY 二核苷分类中只有在 CG 二核苷分类下三个 CG 模体集合的频次分布各自形成完全独立的单峰分布。从统计学上来讲,三个独立的单峰分布代表它们来自三个独立的总体,也就是说基因组序列是由这三类独立进化单元构成的。0CG 模体的分布中心在随机中心附近,1CG 与 2CG 模体出现的频次远低于 0CG 模体出现的频次,且它们的分布中心则远离随机中心。这表明 0CG 模体主要是在随机压力下进化的,而 1CG 与 2CG 模体是定向进化的。CG 模体的独立进化现象具有物种普适性,这种进化规律在低等真核生物酵母基因组中就已经显现出来。随着物种进化,1CG 和 2CG 模体使用保守性越来越强。0CG 模体使用与物种进化无明显关系,1CG 和 2CG 模体使用与物种进化显著相关。发现,三个 CG 模体分布中心的距离是产生 8-mer 图谱(或 k-mer 图谱)单峰或多峰现象的根本原因,因此这一现象也就得到了自然的解释。研究结论预示了这三类 CG 模体具有不同的生物学功能。

由于其它 15 中 XY 分类中各个子集均未显示出各自独立的进化现象,具有物种普适性的 CG 分类独立进化现象勾画出了基因组序列结构和进化的规律。在生命的早期,在随机进化压力下,形成以碱基 A 或 T 组成的原始 DNA 序列背景,出于功能的需要,以 CG 二核苷作为定向进化的中心,逐步形成含 CG 二核苷的功能片段,以满足越来越复杂的生物的需求。为何生物只选择 CG 二核苷作为进化的核心而不是其它二核苷呢?我们不得而知,但一个普遍现象是必须承认的,就是只有 CG 二核苷存在甲基化现象,而不是其它二核苷。两者之间应该存在必然的联系。

许多工作根据 DNA 序列的 k-mer 频次偏好来揭示物种之间的进化,他们往往关注序列中高频次出现的模体。根据以前的研究结论,高频次出现的模体属于 0CG 模体子集,这些模体与物种进化无明显的关系,而低频次出现的 1CG 和 2CG 模体与进化显著相关。如果用 1CG 和 2CG 模体来构造系统发生关系,得到的结果应该会更加可靠。另外,无论实验还是理论分析核小体中心序列上与组蛋白相互作用模体时,也存在这种倾向。其实,核小体结合模体往往是出现频次较低的模体。如何从大的背景噪音中将这些低频出现的功能模体提炼出来才是正确的思考方向。

参考文献(References)

[1] CSÜRÖS M, NOÉ L, KUCHEROV G. Reconsidering the significance of genomic word frequencies[J]. Trends in Genetics, 2007, 23(11): 543-546.DOI: 10.1016/j.tig.2007.07.008.

[2] D'HAESELEER P. What are DNA sequence motifs[J]. Nature Biotechnology, 2006, 24(4): 423-425. DOI: 10.1038/nbt0406-423.

[3] TULLER T, CHOR B, NELSON N. Forbidden penta-peptides[J]. Protein Science, 2007, 16(10): 2251-2259. DOI: 10.1110/ps.073067607.

[4] HAO Bailin, LEE H C, ZHANG Shuyu. Fractals related to long DNA sequences and complete genomes[J]. Chaos, Solitons & Fractals, 2000, 11(6): 825-836. DOI: 10.1016/S0960-0779(98)00182-9.

[5] SUBIRANA J A, MESSEGUER X. The most frequent short sequences in non-coding DNA [J]. Nucleic Acids Re-

search, 2010, 38(4): 1172-1181. DOI: 10.1093/nar/gkp1094.

[6] HAMPIKIAN G, ANDERSEN T. Absent sequences: Nullomers and primes[J]. Pacific Sympusium on Biocomputing, 2007(12): 355-366. DOI:10.1142/9789812772435_0034.

[7] HARIHARAN R, SIMON R, PILLAI M R, et al. Comparative analysis of DNA word abundances in four yeast genomes using a novel statistical background mode [J]. PLoS One, 2013, 8(3): e58038. DOI: 10.1093/bioinformatics/btn166.

[8] YU Hongjie. Segmented k-mer and its application on similarity analysis of mitochondrial genome sequences[J]. Gene, 2013,518(2): 419-424. DOI: 10.1016/j.gene.2012.12.079.

[9] CHAE H, PARK J, LEE S W, et al. Comparative analysis using k-mer and k-flank patterns provides evidence for CpG island sequence evolution in mammalian genomes [J]. Nucleic Acids Research, 2013, 41(9): 4783-4791. DOI: 10.1093/nar/gkt144.

[10] YANG Y, NEPHEW K, KIM S. A novel k-mer mixture logistic regression for methylation susceptibility modeling of CpG dinucleotides in human gene promoters [J]. BMC Bioinformatics, 2012, 13 (Suppl 3): S15. DOI: 10.1186/1471-2105-13-S3-S15.

[11] CHIKHI R, MEDVEDEV P. Informed and automated k-mer size selection for genome assembly [J]. Bioinformatics, 2013, 30(1): 31-37.DOI: 10.1093/bioinformatics/btt310.

[12] BINAA M, WYSSA P, SHERYL A, et al. Discovering sequences with potential regulatory characteristics[J]. Genomics, 2009, 93(4): 314-322.DOI: 10.1016/j.ygeno.2008.11.008.

[13] BINAA M, WYSSA P, RENB W, et al. Exploring the characteristics of sequence elements in proximal promoters of human genes[J]. Genomics, 2004, 84(6): 929-940. DOI:10.1016/j.ygeno.2004.08.013.

[14] XIE, Xiaohui, LU Jun, KULBOKAS E J, et al. Systematic discovery of regulatory motifs in human promoters and 3'UTRs by comparison of several mammals[J]. Nature, 2005, 434(7031): 338-345. DOI: 10.1038/nature03441.

[15] ZHANG Yi, WANG Xianhui, KANG Le. A k-mer scheme to predict piRNAs and characterize locust piRNAs[J]. Bioinformatics, 2011, 27(6): 771-776. DOI:10.1093/bioinformatics/btr016.

[16] CHOR B, HORN D, GOLDMAN N, et al. Genomic DNA k-mer spectra: Models and modalities[J]. Genome Biology, 2009, 10(10): R108. DOI:10.1007/978-3-642-12683-3_37.

[17] BAO T, LI H, ZHAO X Q, et al. Predicting nucleosome binding motif set and analyzing their distributions around functional sites of human genes[J]. Chromosome Research, 2012, 20(6): 685-698.DOI: 10.1007/s10577-012-9305-0.

[18] 周德良, 李宏, 杨小希. 人类 1 号染色体 DNA 序列 8-mer 的相对模体数分布及 8-mer 使用的进化分离[J]. 生物物理学报, 2015, 31(1): 53-64.DOI:10.3724/SP.J.1260.2015.40050.

ZHOU Deliang, LI Hong, YANG Xiaoxi. Frequency distributions of 8-mer and the evolution diversity of 8-mer usage in human DNA sequences[J]. Acta Biophysica Sinica, 2015, 31(1): 53-64. DOI: 10.3724/SP.J.1260.2015.40050.

[19] QUANTE T, BIRD A. Do short, frequent DNA sequence motifs mould the epigenome? [J]. Nature Reviews Molecular Cell Biology, 2016, 17(4): 257-262.DOI: 10.1038/nrm.2015.31.