Hans Journal of Biomedicine 生物医学, 2018, 8(4), 55-61
Published Online October 2018 in Hans. http://www.hanspub.org/journal/hjbm
https://doi.org/10.12677/hjbm.2018.84007

Hans汉斯

The Analysis of Human Nucleosome Location Sequences

Yun Jia[1], Hong Li[2], Jun Lv[1], Jingfeng Wang[1]

[1]Department of Physics, College of Science, Inner Mongolia University of Technology, Hohhot Inner Mongolia
[2]School of Physical Science and Technology, Inner Mongolia University, Hohhot Inner Mongolia
Email: yunbao2004haijun@163.com

Received: Sep. 20th, 2018; accepted: Oct. 4th, 2018; published: Oct. 11th, 2018

Abstract

High-throughput experiments *in vitro* have confirmed that DNA sequences are important factors influencing nucleosome localization, and differences between DNA sequences can affect the ability of nucleosomes to localize. In this paper, we analyzed the sequence features of the nucleosomal localization sequences, k-mer position preference, and so on. The results showed that the content of nucleotides G and C was significantly higher than that of A and T in the nucleosome mapping sequence. The GC content in the nucleosome localization sequence was significantly higher than that in the AT, and the lower frequency motifs may be the characteristic motif of the nucleosome localization sequence.

Keywords

Nucleosomal Localization Sequences, k-mer

人类核小体定位序列特征分析

贾　芸[1]，李　宏[2]，吕　军[1]，王景峰[1]

[1]内蒙古工业大学理学院物理学系，内蒙古　呼和浩特
[2]内蒙古大学物理科学与技术学院，内蒙古　呼和浩特
Email: yunbao2004haijun@163.com

收稿日期：2018年9月20日；录用日期：2018年10月4日；发布日期：2018年10月11日

摘　要

高通量体外实验证实DNA序列是影响核小体定位的重要因素，DNA序列之间的差异能够影响核小体定位

文章引用: 贾芸, 李宏, 吕军, 王景峰. 人类核小体定位序列特征分析[J]. 生物医学, 2018, 8(4): 55-61.
DOI: 10.12677/hjbm.2018.84007

的能力。本工作以人类基因组序列为研究对象，分析了人类核小体定位序列的组分特征，k-mer位置偏好等，统计结果显示核小体定位序列单核苷酸G和C的含量明显高于A和T，发现核小体定位序列中GC含量显著高于AT含量，并且出现频率较低的模体可能是核小体定位序列的特征模体。

关键词

核小体定位序列，k-mer

Copyright © 2018 by authors and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

核小体作为真核生物染色质高级结构的基本单位，是由 DNA 与组蛋白结合而成的典型生物大分子，由约 147 碱基对的 DNA 分子盘绕组蛋白八聚体上形成的核心 DNA 序列与长度约 10~50 碱基对的连接序列两部分组成[1] [2] [3]。组蛋白八聚体是由高度保守的 H2A，H2B，H3 和 H4 各二分体组成，在组蛋白 H1 的连接作用下，形成一个高级分子结构[2]。核小体的特殊结构限制了负责基本生命过程的蛋白质与围绕组蛋白上的 DNA 接触，所以在基因表达过程中它的形成以及在染色质上的精确定位在基因表达过程中起着无可替代的作用，直接或间接地影响转录等基本生物过程[3] [4] [5]。

本研究利用实验检测的高分辨率的人类 CD4 + T 细胞中休眠状态下核小体占据率数据，分析和比较了人类基因组核小体定位与缺乏序列的一些特征。

2. 材料与方法

2.1. 材料

人类 CD4 + T 细胞全基因组的核小体占据数据来自于 Schones 等[6]所做的工作。数据通过 MNasese 方法获得。该数据为人类 CD4 + T 细胞在休眠状态下和被 CD3 抗原激活后的全基因组核小体占据 bed 数据。网址为 http://dir.Nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.aspx。

人类全基因组序列数据来源于 UCSC 基因组数据库 hg18 版本 http://hgdownload.cse.ucsc.edu/。

本工作中数据集构建包括核小体占据序列数据集(简写为 nucleosome)与核小体缺乏序列数据集(简写为 null)两部分。核小体占据序列数据集即为基因组上 DNA 被核心组蛋白占据的部分。首先，扫描基因组上碱基被目标探针覆盖的次数，经过统计占据率分布后选取分数 13 为阈值进行初步筛选，即将基因组上被探针覆盖 13 次及 13 次以上的区域作为定位备选集。而后，进行长度筛选，取连续覆盖范围 201 bp 的序列。最后，利用 blast 软件进行序列比对，将完全重复或高度相似的序列进行归类，每类中只保留其中一条序列，最终筛选得到核小体占据序列 36,777 条。缺乏序列数据集为核小体缺乏 DNA 序列数据集，它的构建与定位序列数据集的构建的相似。首先在基因组上进行探针覆盖次数扫描，只选取 0 覆盖的区域作为缺乏备选集(为了保证可靠性，覆盖次数在 0 与 13 之间的被认为是模糊定位区域而放弃不用)。在进行序列长度筛选时,选用连续 0 覆盖范围在 101 bp 的序列来构建缺乏区数据，经过挑选后获得 332,772 条长度为 101 bp 的序列。为了避免占据、缺乏数据集间因序列长度造成的统计偏差，又将定位区每条序列拆成 1~101 bp，101~201 bp 两部分。

2.2. 方法

2.2.1. K-Mer 出现频率

如果把长度为 k 的核甘酸片段看作是一种“字”(k-mer)，那么 k-mer 的频数就是长度为 k 的窗口在核苷酸序列上顺次移动时出现的次数，即“字”在序列上出现的次数。当 k 较大时，k-mer 的频数分布构成了基因组的一个“等价表示”，即 k-mer 的频数分布可以唯一地确定基因组序列[7] [8]。组成 DNA 序列的核苷酸有 4 种：A (腺嘌呤)，G (鸟嘌呤)，T (胸腺嘧啶)和 C (胞嘧啶)，k-mer 共有 4^k 种。对于序列总长度为 LE 的 DNA 序列 k-mer 出现频率定义为 P_i。

$$P_i = \frac{N_i}{\sum_{i=1}^{4^k} N_i} \left(i = 1,2,\cdots,4^k\right) \tag{2.2.1}$$

其中 N_i 为第 i 个 k-mer 在序列中出现的频数，$\sum_{i=1}^{4^k} N_i = LE$ 。

2.2.2. GC 含量、CpG 相对丰度

G + C 含量(GC ($content$))是基因组结构中的一个重要的因素[9]，定义

$$GC\left(content\right) = \frac{N_G + N_C}{N_A + N_C + N_G + N_T} \tag{2.2.2}$$

Kariln 等人通过对基因中寡聚核苷酸的相对丰度的研究，发展出一套用数值比较物种间不同基因以及物种基因的不同部位的方法[10]。相对丰度的思想认为由于 DNA 序列中碱基相邻的频率并不是独立的，也就是说碱基的分布不是随机产生，相邻碱基的频率不等于单个碱基频率的乘积。CpG 相对丰度(ρ_{CG})描述 CpG 二核苷酸的实际频率与从其组成核苷酸的频率估算的理论频率的差别，其定义如下：

$$\rho_{CG} = \frac{P\left(CpG\right)}{P\left(C\right)P\left(G\right)} \tag{2.2.3}$$

2.2.3. K-Mer 频数分布率

为了反映各 k-mer 在序列中出现的频数的分布状况，我们首先分别计算了每条序列中各 k-mer 出现的频数，统计每个 k-mer 在不同频数区间出现的条数，则一定区间出现的各 k-mer 的条数率定义为 k-mer 频数分布率 FA_j：

$$FA_j = \frac{L_{ij}}{\sum L_i} \tag{2.2.4}$$

L_{ij} 为第 j 个 k-mer 出现在第 i 个频数区间的序列条数，分母为该数据集序列总条数。

2.2.4. 二核苷酸位置频率

对于核小体定位序列数据集建立二核苷酸位置频率 SA_j：

$$SA_j = \frac{L_{ij}}{\sum L_i} \left(i = 1,2,\cdots,4^2, j = 1,2,\cdots,LE-1\right) \tag{2.2.5}$$

L_{ij} 为第 i 个 2-mer 出现在第 j 个位点上的序列条数，分母为该数据集序列总条数。

3. 结果与讨论

3.1. 核小体定位序列偏好 C、G

根据(2.2.1)、(2.2.2)式我们对核小体定位序列数据集、核小体缺乏序列数据集的序列总数、碱基频率、2-mer 频率以及 GC 含量做了统计，结果见表 1 和表 2。

Table 1. The frequency of single-base, GC content and CpG relative abundance in nucleosome and null sequences
表 1. 核小体定位、缺乏序列单碱基频率、GC 含量、CpG 相对丰度数据统计

sequences	序列条数	P_A	P_C	P_G	P_T	$GC(content)$	ρ_{CpG}
nucleosome	73,554	0.2519	0.2466	0.2542	0.2473	0.5008	0.2266
null	332,772	0.2905	0.2039	0.2054	0.3002	0.4093	0.3008

Table 2. The frequency of dinucleotides in nucleosome and null sequences
表 2. 核小体定位、缺乏序列 2-*mer* 出现频率统计

频率	AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
nucleosome	0.063	0.053	0.080	0.054	0.083	0.070	0.014	0.078	0.064	0.061	0.074	0.053	0.040	0.061	0.084	0.061
null	0.097	0.049	0.067	0.078	0.071	0.052	0.013	0.068	0.056	0.045	0.054	0.051	0.067	0.058	0.072	0.104

统计结果显示核小体定位序列单核苷酸 *G*、*C* 的含量明显高于 *A*、*T*。*A* + *T* 含量越高，序列的刚性越强，越不利于 *DNA* 的弯曲。因此，定位序列的单核苷酸 *A* 和 *T* 的含量低有助于核小体 DNA 缠绕组蛋白八聚体。通过计算亦证实定位序列的整体 *G* + *C* 含量(*G* + *C* = 0.5008)显著高于缺乏序列区(*G* + *C* = 0.4093)。核小体定位序列中出现频率最高的八个二联体依次为 *TG*、*CA*、*AG*、*CT*、*GG*、*CC*、*AA*、*TT*；缺乏序列中出现频率最高的八个二联体依次为 *TT*、*AA*、*AT*、*TG*、*CA*、*CT*、*AG*、*TA*。定位序列中出现频数最高的八个 4-*mer* 依次为 *CAGG*、*CCTG*、*CTGG*、*CCAG*、*GCAG*、*TGTG*、*CACA*、*GCTG*；缺乏序列中出现频数最高的八个四联体依次为 *TTTT*、*AAAA*、*AAAT*、*ATTT*、*TTTA*、*TAAA*、*ATAT*、*AATT*。缺乏序列高频次出现的 4-mer 由碱基 *A* 和 *T* 组成。核小体定位序列中 *GC* 含量显著高于 *AT* 含量而高频率的 k-mer 并不完全是只包含 G、C 碱基。

二核苷频率 *AA*、*TT* 在两类序列的频率差别显著(图 1)。图 2 显示 *CG* 二核苷在两类数据集出现的频率都最小，定位区仅为 0.017，缺乏区占 0.0199，均低于随机水平 0.0625，*CG* 的低频现象也与基因组中 *CG* 缺乏的现象吻合。

3.2. *k-mer* 频数分布率特征

利用 2.2.3 方法得 GC 含量及各单碱基频数分布率，由图 3 显示核小体定位序列数据集中 *GC* 含量变化范围是在 5 到 85 之间，其中含量在 35 到 70 之间的序列占据了序列总数的 98%。核小体缺乏序列的 *GC* 含量多数集中在 0 到 65 之间，多数缺乏序列的 *GC* 含量低于定位序列。单碱基的频数分布在两类序列集中没有显著区别(图 4)。

3.3. 二核苷酸位置频率特征

我们利用长度是 201 bp 的核小体定位序列数据集计算每个二核苷酸出现在各位点的概率作图 5，发现在相同位置不同二核苷酸出现的频率不同，不同位置相同二核苷酸出现的频率也不同。表 3 列出了 16 种二核苷酸在核小体定位序列中的平均占据率。假设核小体定位序列不存在统计特性，那么二核苷酸位置频率的值在任意位置应该是随机产生的，在 1/16 大小浮动。而且表 3 可以看出所有二核苷酸在定位序列中并没有平均分配，这表明定位序列中二核苷酸位置频率存在一定统计特性。

4. 结论

分析核小体定位序列统计特征发现人类核小体定位序列与缺乏序列在 GC 含量和 k-mer 分布存在显著差异。核小体定位序列偏好单碱基 *G* + *C*，偏好二核苷 CC、CG、GG、GC；核小体缺乏序列偏好单碱

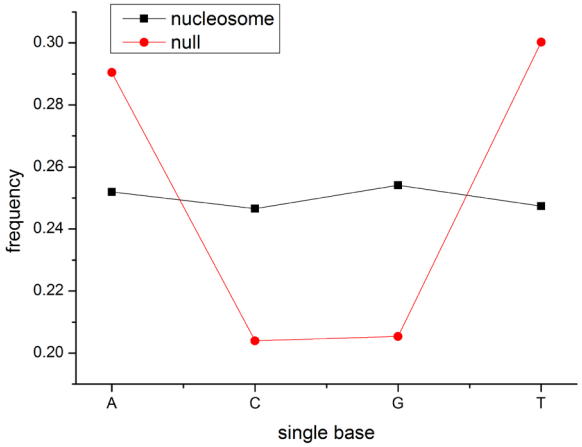

Figure 1. The frequency of single-base in nucleosome and null sequences
图 1. 核小体定位、缺乏序列单碱基出现频率

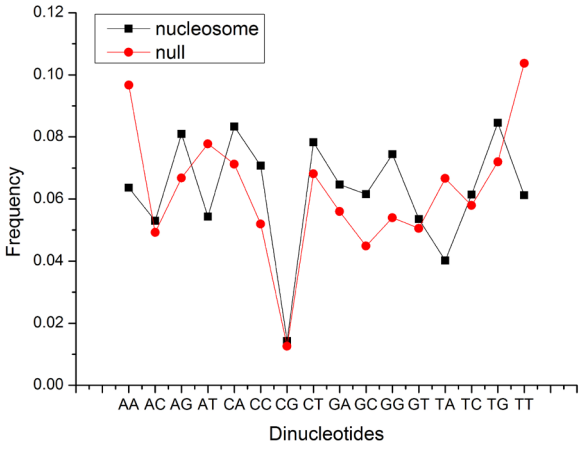

Figure 2. The frequency of dinucleotides in nucleosome and null sequences
图 2. 核小体定位、缺乏序列 2-*mer* 出现频率

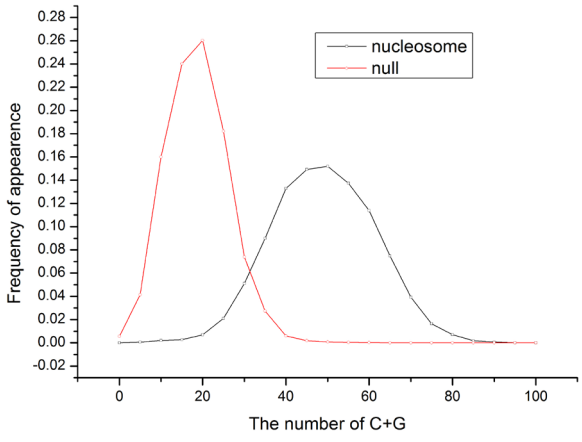

Figure 3. The frequency of distribution on GC content in nucleosome and null sequence
图 3. GC 含量在核小体定位缺乏序列分布率

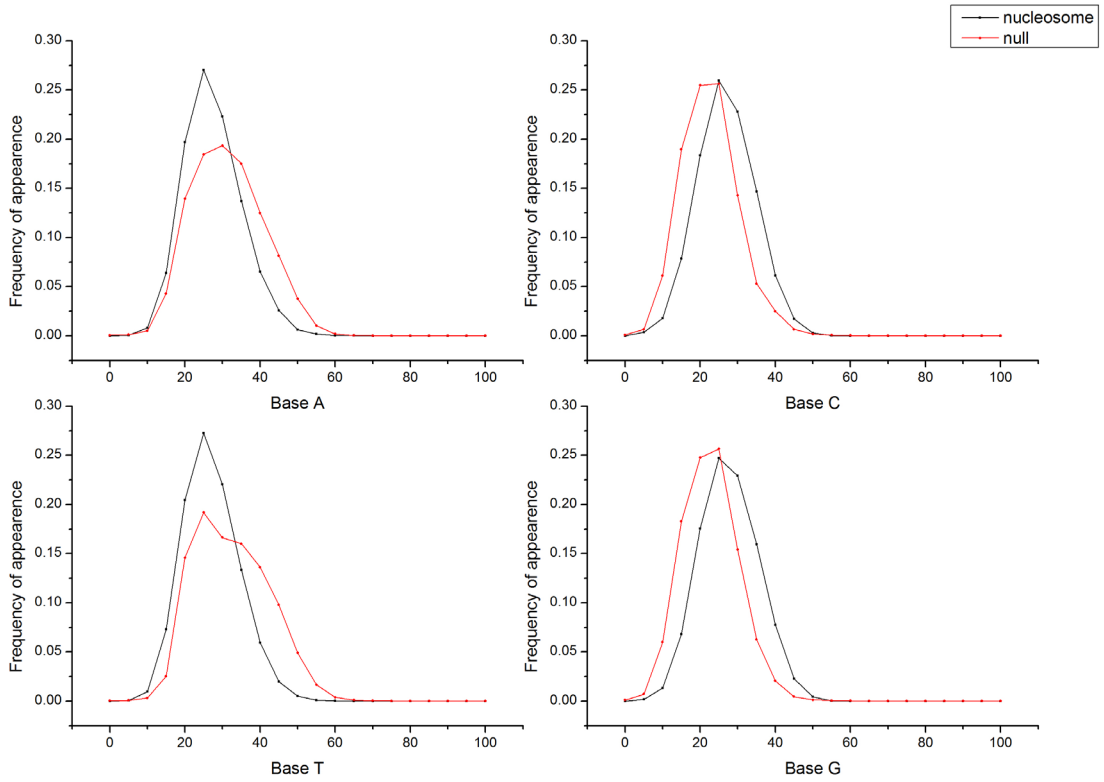

Figure 4. The frequency of distribution on single-base in nucleosome and null sequence
图 4. 单碱基在核小体定位、缺乏序列分布率

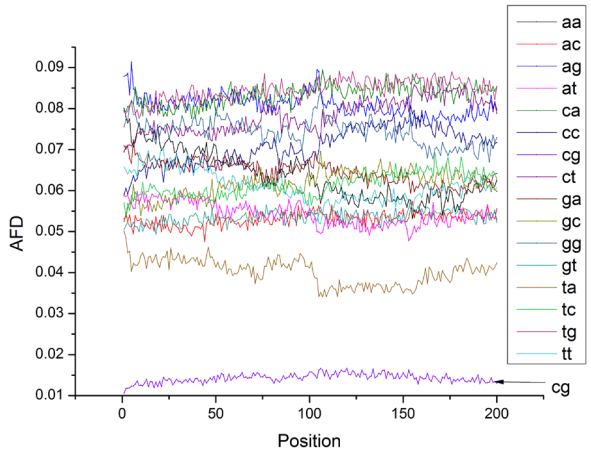

Figure 5. Nucleosome location sequence dinucleotide location frequency
图 5. 核小体定位序列二核苷酸位置频率

Table 3. Average occupancy rate of 16 dinucleotides in nucleosome localization
表 3. 16 种二核苷酸在核小体定位中的平均占有率

dinucleitides	aa	ac	ag	at	ca	cc	cg	ct
Average occupancy rate	0.0636	0.0530	0.0809	0.0543	0.0833	0.0707	0.0142	0.0782
dinucleitides	ga	gc	gg	gt	ta	tc	tg	tt
Average occupancy rate	0.0646	0.0615	0.0744	0.0535	0.0402	0.0614	0.0845	0.0612

基 $A+T$，偏好二核苷 TT、AA、AT、TA，表明与 C 和 G 相关的二核苷、甚至 k 核苷(k > 2)在核小体形成及其功能行使上以基本的功能单位参与作用；16 种二核苷在核小体定位序列和核小体缺乏序列上的平均占有率同样存在上述偏好性，这不仅支持了与 C 和 G 相关的二核苷、甚至 k 核苷(k > 2)作为基本功能单位的观点，同样表明此类二核苷及 k 核苷对基因转录及进化同样具有深刻的意义。本文只是从 k 核苷偏好性及其组成偏好性出发进行了分析，进一步研究将对此类探索具有重要意义。

基金项目

内蒙古工业大学重点研究项目(ZD201614)。

参考文献

[1] Richmod, T.J. and Davey, C.A. (2003) The Structure of DNA in the Nucleosome Core. *Nature*, **424**, 145-150. https://doi.org/10.1038/nature01595

[2] Luger, K., Mader, A.W., Richmond, R.K., *et al.* (1997) Crystal Structure of the Nucleosome Core Particle at 2.8 Resolution. *Nature*, **389**, 251-260. https://doi.org/10.1038/38444

[3] Kornberg, R.D. and Lorch, Y. (1999) Twenty-Five Years of the Nucleosome, Fundamenta Particale of the Eukaryotic Chromosome. *Cell*, **98**, 285-294. https://doi.org/10.1016/S0092-8674(00)81958-3

[4] Lee, W., Tillo, D., Morse, R.H., *et al.* (2007) A High Resolution Atlas of Nucleosome Occupancy in Yeast. *Nature Genetics*, **39**, 1235-1244. https://doi.org/10.1038/ng2117

[5] Vaillant, C., Audit, B. and Arneodo, A. (2007) Experiments Confirm the Influence of Genome Long-Range Correlation on Nucleosome Positioning. *Physical Review Letters*, **99**, 218-303. https://doi.org/10.1103/PhysRevLett.99.218103

[6] Schones, D.E., Cui, K., Cuddapah, S., *et al.* (2008) Dynamic Regulation of Nucleosome Positioning in the Human Genome. *Cell*, **132**, 887-898. https://doi.org/10.1016/j.cell.2008.02.022

[7] Xie, H. and Hao, B. (2002) Visualization of K-Tuple Distribution in Prokaryote Complete Genomes and Their Randomized Counterparts. *CSB*2002 *Proceedings* (*C*), Los Alamitos, California, 2002, 31-42.

[8] 罗辽复. DNA 序列信息内容的普适关系[J]. 合肥学院学报, 2005, 15(1): 1-7.

[9] Naimuddin, M., Kurazono, T. and Nishigak, L.I.C. (2002) Commonly Conserved Genetic Fragments Revealed by Genome Profiling Can Serve Tracers of Evolution. *Nucleic Acids Research*, **30**, e42. https://doi.org/10.1093/nar/30.10.e42

[10] Karlin, S. (2001) Detecting Anomalous Gene Clusters and Pathogenicity Islands in Diverse Bacterial Genomes. *Trends in Microbiology*, **9**, 335-343. https://doi.org/10.1016/S0966-842X(01)02079-0

Hans汉斯

知网检索的两种方式：

1. 打开知网页面 http://kns.cnki.net/kns/brief/result.aspx?dbPrefix=WWJD
   下拉列表框选择：[ISSN]，输入期刊 ISSN：2161-8976，即可查询
2. 打开知网首页 http://cnki.net/
   左侧“国际文献总库”进入，输入文章标题，即可查询

投稿请点击：http://www.hanspub.org/Submission.aspx
期刊邮箱：hjbm@hanspub.org