文章编号:1004-9037(2014)01-0141-05

一种基于改进 Segal 模型的核小体定位方法

黄 星 王加俊

(苏州大学电子信息学院,苏州,215006)

摘要:核小体预测是目前遗传学研究的重要内容,但现有的预测算法大部分仅依据核小体的统计特性,定位准确性很受局限。另一方面,经研究发现,DNA 连接序列作为两个核小体的连接纽带,存在一定的统计特性。基于此事实,本文对 Segal 模型做了改进,通过核小体和连接序列的二核苷酸位置频率建立了核小体和连接序列两组得分函数,并以其差值作为核小体的定位依据。利用该算法模型对酵母染色体中核小体进行定位预测,发现定位准确性得到明显提高。

关键词:核小体定位; 二核苷酸位置频率; 连接 DNA 序列; 核心 DNA 序列

中图分类号:Q34 **文献标志码**:A

Nucleosome Positioning Profile Based on Improved Segal′s Model

Huang Xing, *Wang Jiajun*

(School of Electronics and Information Engineering, Soochow University, Suzhou, 215006)

Abstract: Genome-wide nucleosome prediction has been an important research area in genetics so far. However, most existing nucleosome prediction algorithms are based on the statistical features of nucleosome-bound DNA sequences, usually resulting in low accuracies. Basides our statistical studies in linker DNA sequences, each of which connects with two nucleosome-bound DNA sequences, show that linker DNA sequences have some specific statistical properties. Concerning the fact, improvement of the Segal model is presented, where two score functions are constructed, based on the dinucleotide position frequencies of the nucleosome-bound and linker DNA sequences respectively. Nucleosome positions are predicted according to the difference between the above two score functions. Experimental results on the yeast′s chromatin demonstrate that the improved algorithm can significantly increase the accuracy in positioning nucleosomes.

Key words: nucleosome positioning; dinucleotide position-frequency; linker DNA sequence; nucleosome-bound DNA sequence

引 言

核小体作为真核生物的基本单位,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[1-3]。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[3-5]。

当下核小体定位研究重点在核小体核心 DNA 序列的特性研究,而且大部分的理论预测算法都是基于核心 DNA 序列统计特性的基础上建立起来的[6-10],预测结果很不理想。于是在研究过程中有部分研究人员开始重视连接 DNA 序列,认为它带

基金项目:国家自然科学基金(60871086)资助项目;江苏省自然科学基金(BK2008159)资助项目。

收稿日期:2012-04-19;**修订日期**:2012-08-21

有一定的统计特性，建立几种结合连接 DNA 序列特性理论预测算法[10-12]，预测核小体定位结果比仅仅考虑核小体核心 DNA 序列的准确性有所提高。基于此理论，本文研究发现，连接序列也存在一定的统计特性。在对染色体序列通过一阶马尔科夫分别建立核小体核心序列和连接序列的得分函数后，以其差值作为定位酵母染色体中核小体的位置的依据，并采用最大谱连续小波对得分函数的峰值进行锐化，寻找其峰值位置作为本算法预测的核小体中心位置。本文中所得结果与实验测得核小体图谱[13-14]相比，比原 Segal 算法的准确性有显著提高。

1 核小体定位方法

1.1 核小体定位的数据来源

核小体定位方法是基于酵母基因组的方法，酵母染色体序列取自于酵母基因数据库（http://www.yeastgenome.org）。核小体核心序列的二核苷酸位置频率来自于 Segal 实验室，而连接序列的二核苷酸位置频率是依据 Segal 算法[6]通过 296 条长度超过 100 bp 的连接序列与 199 条根据标准核小体位置[13-14]识别的超过 100 bp 的连接序列获得，其中 296 条连接序列由 Yuan Guo-Cheng 通过基因芯片获得，其中包含一条超过 100 bp 的核小体自由区，为使连接序列中心对齐且长度与核小体一致，充分反映连接序列在不同位置上的统计规律，Yuan 和 Liu[15]对其进行中心对称扩展，而 199 条标准连接序列依照引文进行对称扩展。Segal 算法所预测核小体中心位置同样来自于 Segal 实验室。验证理论方法定位核小体位置准确性的核小体图谱实验数据出自于文献[13,14]。

1.2 核小体定位模型的理论基础

本文对核小体连接 DNA 序列建立二核苷酸位置频率（依据 Segal 模型），如图 1 所示，仅以 TA 为例，表 1 列出了 16 种二核苷酸在连接序列和核小体序列中的平均差异。假设核小体连接 DNA 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 大小徘徊。由图 1 可知，TA 位置频率除极少点的值小于 1/16，大部分都远大于 1/16，而且表 1 可以看出所有

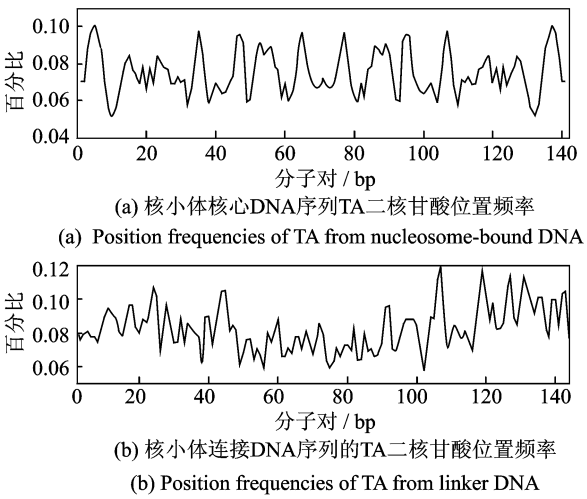

图 1 核小体核心及连接 DNA 序列的 TA 二核苷酸位置频率

Fig. 1 Position freuencies of TA from nucleosome-bound and linker DNA

表 1 16 种二核苷酸在连接序列与核小体中的平均占有率

Table 1 Averaged dinucleotide position-frequencies for nucleosome-bound and linker DNA

	AA	AT	AG	AC	TA	TT	TG	TC
连接序列	0.124 0	0.089 4	0.055 4	0.051 6	0.082 6	0.120 0	0.055 8	0.057 6
核小体	0.018 1	0.085 4	0.064 2	0.054 2	0.075 2	0.097 9	0.059 7	0.062 8
	GA	GT	GG	GC	CA	CT	CG	CA
连接序列	0.057 0	0.049 8	0.034 2	0.038 3	0.057 3	0.056 1	0.034 1	0.036 9
核小体	0.062 3	0.050 6	0.037 8	0.042 5	0.067 3	0.060 5	0.028 7	0.040 6

二核苷酸在连接序列与 DNA 核小体序列中并没有平均分配，这表明连接序列中存在一定统计特性。

1.3 二核苷酸位置频率的得分函数

对连接 DNA 序列（由于连接序列长度变化不定，与核小体核心 DNA 序列相对应，本文取 147 bp）建立二核苷酸位置频率，实验发现其中在相同位置不同二核苷酸出现的概率不同，不同位置相同二核苷酸出现的频率也不同，虽然没有核小体核心 DNA 序列某种统计特性[16]明显，但并不是随机产生。类似于文献[17]中建立两种概率得分模型，应用连接 DNA 序列与核小体核心序列二核苷酸位置频率建立理论算法模型。

对染色体中每一段长度为 147 bp 序列用概率模型建立得分函数。模型总共分两部分，第一部分由核小体核心 DNA 序列二核苷酸位置频率构建得分函数

$$\mathrm{score_n}(s_i) = \log \frac{p_n(s_i)}{p_b(s)} \tag{1}$$

式中：$p_n(s_i)$表示核小体与 DNA 的相互关系，$p_b(s)$表示 DNA 序列的随机基底。

$$p_n(s_i) = p(s_i) \prod_{k=2}^{147} p(s_{i+k} \mid s_{i+k-1}) \tag{2}$$

$$p_b(s) = \prod_{i=1}^{147} p_b(x_i) \tag{3}$$

式中：$p(s_{i+k} \mid s_{i+k-1})$由核小体核心 DNA 序列 $i+k-1$位置的二核苷酸位置频率除以当前位置单核苷酸概率获得。$p_b(x_i)=1/4$，其中 x_i 为 A/T/G/C。第二部分由连接 DNA 序列二核苷酸位置频率构建得分函数，依照式(1,2,3)。其中 $p_l(s_i)$表示连接序列，其类似于 $p_n(s_i)$，通过连接 DNA 序列的二核苷酸位置频率获得，$p_b(s)$表示 DNA 序列的随机基底，保持不变。模型得分函数定义为

$$\mathrm{score}(s_i) = \mathrm{score_n}(s_i) - \mathrm{score_l}(s_i) \tag{4}$$

模型以两组得分函数差值为核小体定位的审判依据，对其去噪、滤波。定位滤波后的信号中峰值位置为核小体中心位置。

1.4　最大谱连续小波

最大谱连续小波(Maximum spectrum continuous wavelet transform，MSCWT)[7,18]依据周边的环境对信号的峰值和谷值实施锐化。为更加准确方便地找到得分函数中峰值位置，MSCWT 被用于检测峰值，确定理论算法预测核小体的位置。

对于一时间信号 $f(t)$，其连续小波(Continuous wavelet transform，CWT)表达为

$$\mathrm{Wf}(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} f(t) \varphi^* \left(\frac{t-b}{a}\right) \mathrm{d}t \tag{5}$$

式中：a 为尺度因子，b 为平移因子。

通过检测连续小波的最大模量来确定 MSCWT，其定义如下

$$\mathrm{MSCWT}(t) = \max(\mid \mathrm{Wf}(t) \mid) \tag{6}$$

在 MSCWT(t)中峰值位置与原信号峰值位置保持不变。检测峰值时，计算连续小波基底采用 Mexican 小波，尺度因子选择是[30：32]。

2　结果与讨论

2.1　核小体定位的评判标准

大规模核小体定位采取常用的评价指标有：敏感性(Sensitivity，Sn)、分辨率(Resolution，Rs)和准确率(Accuracy，Ac)。但文中对其中各个指标含义做了重新定义。

真阳性(True positive，TP)为理论算法预测的核小体中心位置与实验测得的核小体中心位置差距不大于 35 bp 的理论算法预测的核小体数目，假阳性(False positive，FP)为理论算法预测的核小体中心位置与实验测得的核小体中心位置差距大于 35 bp的理论算法预测的核小体数目，假阴性(False negetive，FN)为理论算法预测的核小体边缘位置与实验测得的核小体边缘位置差距大于 35 bp 的实验测得的核小体数目，sum 为染色体序列长度，则有如下定义

$$\mathrm{Sn} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \times 100\% \tag{7}$$

$$\mathrm{Rs} = \frac{\mathrm{sum}}{\mathrm{TP}+\mathrm{FP}} \tag{8}$$

$$\mathrm{Ac} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \times 100\% \tag{9}$$

2.2　实验结果

本次实验以酵母 1 号染色体为模板。通过 MSCWT 确定核小体精确位置过程中要注意几种峰值位置是不可取的。首先，通过 MSCWT 定位出的峰值位置的原得分函数值小于阈值，因为得分小于阈值，此位置是核小体的可能性比较小；其次，如果定位的连续两个峰值位置靠得太近(通过预测发现连续两个核小体位置相差不超过100 bp)，去掉其中峰值较小的，因为理论算法预测的核小体是考虑非重叠的，核小体实验图谱也是非重叠的。

以酵母 1 号染色体中 1 000～3 000 bp 的 DNA 序列为图例说明本模型。如图 2 所示，其中实线为本文模型的得分函数；虚线为 MSCWT 对得分函数处理后的信号；菱形为本模型预测的核小体中心位置；矩形为核小体实验图谱中的核小体中心位置[13-14]。由结果可以明显地看出，模型算法测得的核小体多数周围都有一个实验测得的核小体，而且可知一些峰值在算法模型中不可取也是符合逻辑和实践的。当然由于所取的特性都为统计特性，而且

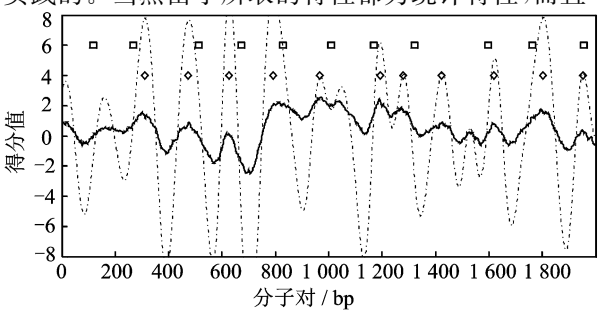

图 2　本文模型预测酵母 1 号染色体中 1 000～3 000 bp 的 DNA 序列的核小体中心位置

Fig. 2　Predicted nucleosome positions based on the paper's method in 1000～3000 bp of the first chromatin

算法模型只是尽可能地利用核小体核心 DNA 序列和连接 DNA 序列的统计特性，所以预测的结果还是存在一定的可扩展空间。

为说明算法的可行性，本模型在酵母 1 号染色体上与改进前 Segal 算法模型进行比较（见表 2），参照评价标准可知，在以少量减少预测的核小体数目与敏感性（减少的部分原因是因为去掉部分连续靠得太近核小体位置）为代价可以在很大程度上提高核小体预测的准确性。通过降低阈值大小来提高预测核小体中心位置的数目和敏感性，会使准确性有所减低，但是准确性依旧比 Segal 算法高。由此证明所建立的结合连接 DNA 序列统计特性的算法是可行的。

表 2 Segal 算法与本算法实验结果比较

Table 2 Results of segal's method and this paper's method

算法	染色体	准确率/%	敏感性/%	分辨率/bp
本文算法	1	45.66	79.82	241
Segal 算法	1	38.86	90.71	218

2.3 算法比较

为证明本文算法模型的可行性，将现行的部分算法列出进行比较，见表 3，二核苷酸位置频率算法参照文献[7]采用本实验所采用的数据编写（用其中预测结果最好的 AA/TT 二核苷酸位置频率作为基底），曲率模型算法的得分取自于网页软件（http://www.gri.seu.edu.cn/icons/），两种算法均采用 MSCWT 锐化其得分信号的峰值，其中的尺度因子在文献[7]的补充文献中均有说明。

表 3 其他两种算法[7]在酵母染色体中的预测结果

Table 3 Predicted results of the other two algorithms in yeast

算法	染色体	准确率	敏感性	分辨率/bp
位置频率法	1	26.93	70.06	250
曲率方法	1	39.82	66.47	206

由表 2,3 可知，在保证预测数目相差不多的前提下，无论准确性还是敏感性，本文算法模型在这几种算法中是最好的，也就是说在标准核小体 35 bp 范围内，本文算法预测准确的核小体最多，标准与预测核小体中心在 36～181 bp 范围内数目相差不大。同样可以微调阈值来改变预测结果，使所有评价标准都高于上面两种算法模型。通过概率模型能够很好地把核小体连接 DNA 序列和核心 DNA 序列的统计特性表达在得分函数上，结果比上述算法准确也在预想之中。综上所述，本文算法在预测核小体中心位置时，明显优于文中其他算法和原改进前的 Segal 算法模型。

3 结束语

核小体定位领域中往往注重核小体核心 DNA 序列特性研究，没有考虑核小体连接序列，核小体定位预测算法亦是如此。文中对核小体连接 DNA 序列进行统计研究，建立二核苷酸位置频率，发现其在每个位置并不是随机分配而是差异很大，表明核小体连接 DNA 序列存在一定的统计特性。为让连接 DNA 序列统计特性应用于核小体位置预测中，以 Segal 算法为基础，对染色体序列建立两组得分函数，以其差值判断这段序列是偏向于核小体核心 DNA 序列，还是连接 DNA 序列。通过最大连续小波锐化差值得分函数的峰值，此算法取得了明显效果，比改进前 Segal 算法和已有的一些算法更准确。本文算法预测结果同样说明，核小体连接 DNA 序列统计特性与预测核小体位置准确性息息相关，其研究前景应受到关注。

参考文献：

[1] Richmond T J, Davey C A. The structure of DNA in the nucleosome core[J]. Nature, 2003, 424:145-150.

[2] Luger K, Mader A W, Richmond R K, et al. Crystal structure of the nucleosome core particle at 2.8 Å resolution[J]. Nature, 1997, 389:251-260.

[3] Kornberg R D, Lorch Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryotic chromosome[J]. Cell,1999, 98:285-294.

[4] Lee W, Tillo D, Morse R H, et al. A high resolution atlas of nucleosome occupancy in yeast[J]. Nat Genet, 2007, 39:1235-1244.

[5] Vaillant C, Audit B, Arneodo A. Experiments confirm the influence of genome long-range correlations on nucleosome positioning[J]. Phys Rev Lett, 2007, 99:218-303.

[6] Segal E, Foudufe-Mittendorf Y, Chen L, et al. A genomic code for nucleosome positioning[J]. Nature, 2006, 443:772-778.

[7] Liu Hongde, Wu Jiansheng, Xie Jianming, et al. Characteristics of nucleosome core DNA and their applications in predicting nucleosome positions [J]. Biophyisical Journal, 2008, 94:4597-4604.

[8] Yuan G C, Liu J S. Genomic sequence is highly predictive of local nucleosome depletion[J]. Plos Computational Biology, 2008(4):164-174.

[9] Wu Q, Wang J, Yan H. Prediction of nucleosome

positions in the yeast genome based on matched mirror position filtering[J]. Bioinformation, 2009(3): 454-459.

[10] Field Y, Kaplan N, Fondufe-Mittendorf Y, et al. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals[J]. PLos Computational Biology, 2008, 4:e1000216.

[11] Peckham Heather E. Nucleosome positioning signals in genomic DNA[J]. Genome Res, 2007, 17:1170-1177.

[12] Ogawa R, Kitagawa N, Ashida H, et al. Computational prediction of nucleosome positioning by calculating the relative fragment frequency index of nucleosomal sequences[J]. FENS Letter, 2010, 584: 1498-1502.

[13] Jiang C, Pugh B F. A compiled and systematic reference map of nucleosome positions across the saccharomyces cerevisiae genome[J]. Genome Biol, 2009, 10:R109.

[14] Jansen A, Verstrepen K J. Nucleosome positioning in saccharomyces cerevisiae[J]. Microbiology and Molecular Biology Reviews, 2011,75(2):301-320.

[15] Yuan J S, Liu G C. Genomic sequence is highly predictive of local nucleosome depletion[J]. PLos Computational Biology, 2008, 4:164-174.

[16] Satchwell S C, Drew H R, Travers A A. Sequence periodicities in chicken nucleosome core DNA[J]. J Mol Biol, 1986, 191:659-675.

[17] 陆安南. 一种长基线组合二维角测向方法[J]. 数据采集与处理,2012,27(3):385-388.
Lu Annan. A method of direction finding by combination of long baselines[J]. Data Acquisition and Processing, 2012, 27(3):385-388.

[18] Lu X Q, Liu H D, Kang J W, et al. Wavelet frequency spectrum and its application in analyzing oscillating chemical system[J]. Anal Chim Acta, 2003, 484:201-210.

作者简介:黄星(1987-),男,硕士研究生,研究方向:生物信息处理,E-mail: huangxing523888@163.com;王加俊(1969-),男,教授,博士生导师,研究方向:图像处理、模式识别、生物信息学等。