doi:10.3969/j.issn.1672-5565.2013.14.

基于核小体位置预测的酵母进化印迹研究

肖建平,丰继华*,卢 英,单秋甫
(云南民族大学电气信息工程学院,云南 昆明 650500)

摘 要:在利用核小体定位实验数据训练支持向量机(SVM)对任意酵母DNA序列的核小体形成能力进行预测的过程中,发现染色质结构对基因组DNA分子进化过程有着显著影响。我们观察到核小体DNA比连接DNA的平均预测准确率低15%,这种普遍存在的局部预测准确率差异性代表了酵母核小体定位的进化印迹(Evolutionary footprint),它揭示了核小体组织在基因组的整个进化过程中所具有的保守性。

关键词:核小体;转录起始位点;预测准确率;进化印记
中图分类号:Q-3 **文献标识码**:B **文章编号**:1672-5565(2013)-02-150-04

Study of yeast evolutionary footprints based on the prediction of nucleosome positioning

XIAO Jian-ping, FENG Ji-hua*, LU Ying, SHAN Qiu-fu
(*School of Electrical and Information Technology, Yunan University of Nationalities, Kunming* 650500, *China*)

Abstract: When we used DNA sequences of Saccharomyces cerevisiae whose nucleosome positioning data have been experimentally determined to train a support vector machine to predict the nucleosome formation potential of any given sequence of DNA, we observed that chromatin structure has an impact on the evolution of genomic DNA molecules. We have found, on average, 15% lower predictive accuracy rates in nucleosomal DNA than in linker DNA. This widespread local rates heterogeneity represents an evolutionary footprint of nucleosome positions and reveals that nucleosome organization is a genomic feature conserved over evolutionary timescales.

Keywords: Nucleosome; Transcriptional Start Sit; Predictive Accuracy Rates; Evolutionary Footprints

真核细胞中的DNA与相关蛋白质的结合称作染色质。在所有的真核生物中,染色质的基本亚基都是相同的,这些亚基又称为核小体(nucleosome)[1-2]。细胞中大多数DNA被包装进核小体,每个核小体包装大约147bp的DNA,核小体之间的DNA称为连接DNA(linker DNA)[3]。但无论是在分裂间期细胞核中的常染色质或异染色质里,还是在有丝分裂期的染色体中,核小体都是其中不变的组分。

由于蛋白质-蛋白质、蛋白质-DNA之间的相互作用,以及一些复杂大分子复合物的形成,导致真核生物的转录调控是一个多级的复杂过程。核小体作为真核生物染色体中基因调控的重要一级,是表观遗传机制的重要组成部分。因此,研究核小体在基因组里的统计位置是理解核小体如何通过自身定位影响DNA进化的前提[4]。

本文利用支持向量机(SVM)模型,根据NCBI数据库下载的酵母基因组DNA序列,结合DNA物理特征和转录因子结合位点数据,以实验获得的酵母核小体定位数据作为参照,训练支持向量机对酵母基因组DNA的核小体包装能力进行了预测,通过对预测准确率进行分析,我们得出了一些有趣的关于染色质结构影响DNA进化的结论。

收稿日期:2012-12-05;**修回日期**:2012-12-28.
基金项目:国家自然科学基金资助项目(31160234);云南省应用基础研究计划项目(2011FB082)。
作者简介:肖建平,男,在读研究生,福建闽清人,主要从事信号与信息处理,E-mail:275939685@qq.com.
*通讯作者:丰继华,男,副教授,云南昆明人,主要从事生物信息处理,E-mail:fengjihua@21cn.com.

2　数据与方法

2.1　数据准备

数据包括三个部分,第一部分来源于 Lee 等的核小体定位实验数据[5],其中还包括转录因子结合位点数据、酵母 DNA 结构数据;第二部分数据来源于 NCBI 数据库中的酵母 16 条染色体 DNA 序列;第三部分是来源于 David 等人文献中 4792 个高置信度酵母基因实验数据[6],其中包括经验证的转录本起始位点(TSS)和转录终止位点(TTS)数据。由于以上数据的异质性,我们根据研究目的,对部分数据进行了重构。

2.2　数据处理

2.2.1　核小体定位实验数据的处理

由于 Lee 等人提供的核小体定位原始实验数据为每隔 4bp 采样的微阵列数据 (an Affymetrix tiling microarray with 4-bp resolution)[5]。因此,首先要通过插值方法得到一个覆盖全基因组每一位点的核小体占位率数据,以供后续对齐使用。另外,对于用隐马尔可夫模型辨识得到的核小体位置数据,把对应 DNA 序列上有核小体的地方置为 1,没有核小体的地方置为 0。这样就形成了一组对应于酵母 16 条染色体,以 0、1 代表核小体有无的二值数据(其中 0 代表连接 DNA,1 代表核小体 DNA)。

2.2.2　数据对齐实处理

根据 David 等人所提供的高置信度转录本实验数据中给出的 4 792 条基因的 TSS 坐标,在每个基因上选取 TSS 上、下游各 800bp 数据进行对齐(对于 C 型基因所对应的数据还要进行反向)后,再进行平均。这样就得到一个以每条基因 TSS 对齐的全基因组平均分布图谱[6]。

2.2.3　转录因子结合位点数据的处理[7]

为了后续研究,采用与核小体位置数据相似的方法,处理了 126 个转录因子结合位点数据,即以 0、1 代表转录因子结合位点在 DNA 上的有无,为了分析提取方便,将分布在 16 条染色体上的转录因子数据融合为了一组数据,其代表了酵母全基因组 126 个转录因子结合位点的分布。

2.2.4　酵母 DNA 结构(物理特征)数据的处理

我们使用了 16 个酵母 DNA 结构(物理特征)数据,主要包括了 GC content、Melting temperature、Enthalpy change、Free energy 等特征。通过插值方法得到了覆盖全基因组每 1bp 的 16 个 DNA 物理特征数据,并同时对原始数据进行了归一化处理。

2.3　支持向量机预测

2.3.1　全基因组核小体位置预测

在对基因组任意 DNA 的核小体形成能力进行预测过程中,我们以实验获得的核小体定位数据、酵母 DNA 结构数据及经过融合的转录因子结合位点数据作为支持向量机的训练数据集。

在全基因组范围内随机选取了 10 000 段长度为 4 000bp 的 DNA(预测窗口长设为 50bp,算法迭代次数为 1 000 次)进行了 DNA 的核小体形成能力预测。这些 DNA 片断随机分布于 16 条染色体长。我们观察到不同片段间的平均预测准确率变化不大。但是核小体包装区间与间隔区预测准确率存在显著差异,而这种差异与 DNA 所在的染色体无关[8]。经统计分析后,发现核小体 DNA 的预测准确率要普遍低于连接 DNA($p < 10^{-7}$)。为了更深入了解核小体预测准确率与基因间的关系,我们对每个基因 TSS 上、下游区域作了进一步研究。

2.3.2　在基因 TSS 周围的预测

选取 DNA 结构特征中的冲击强度(Clash_strength)、熵(Entropy)、Rise(碱基抬升度)、Tip(碱基倾斜度),以及经过融合的转录因子结合位点数据,然后以 4 792 个高置信度基因的 TSS 为中心,对其上、下游各 800bp 范围内 DNA 的核小体形成能力进行了预测。预测中,为了验证算法的稳健性,选取了不同的预测窗口(长度分别为 30bp、40bp、50bp、60bp),窗口滑动步长为均为 1bp,然后将所得到的预测准确率数据进行对齐平均,得到了以基因 TSS 对齐的预测准确率平均分布图谱(见图 1)。

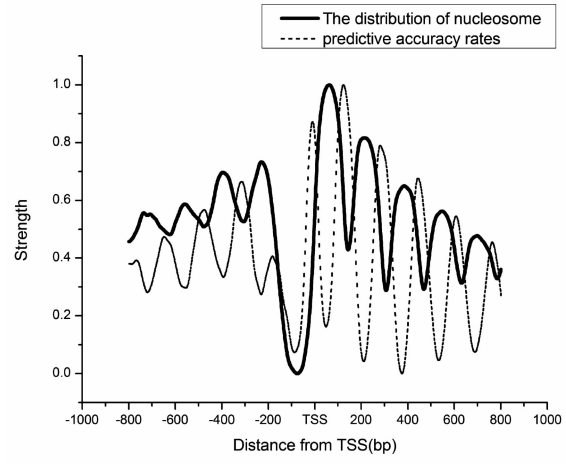

图 1　基因起始位点周围核小体占位率与预测准确率分布图谱

Fig. 1　The distribution of nucleosome and the predictive accuraty rates from TSS

图中黑色粗线代表实验上获得的核小体分布(The distribution of nucleosome);细虚线部分代表支持向量机所得到的预测准确率(predictive accuracy rates)。

为了验证实验结果的可重复性,我们选用了不同的酵母 DNA 结构数据做了交叉实验,结果均与图 1 中虚线部分相似。由此可以说明实验结果具有普遍性。

3 结果分析

3.1 预测准确率与核小体位置间的关联性

从图 1 中可以看到,核小体占位率与核小体预测准确率间存在着很强的关联性,即在总体上,预测准确率曲线与核小体位置曲线呈现出负相关。即:在核小体连接区域预测率较高,而在核小体包装区域预测率较低。这一归律与基因组中随机选取范围的实验结果一致。但在基因的 TSS 周围的核小体缺失区域(NFR),则表现出一个有趣的例外现象,即核小体占位率低的地方,预测准确率也相对较低。

据此,我们认为,由于核小体连接 DNA 上编码了大量转录因子结合位点,是转录机器组装的重要场地,其对生命的维系至关重要[9]。因此,这些 DNA 片段较为保守,在进化中不能轻易改变,否则将产生严重后果。这一点体现在了支持向量机的预测准确率上,正因为这些 DNA 较为保守,其编码中含进化出了大量不利于核小体形成的 DNA 物理特征,因此较易于被支持向量机捕获。与之相反,核小体包装区在进化过程中受到的约束较少,在进化上表现相对活跃,这种编码的变化导致了支持向量机难以找出判断核小体位置规律的信息。这个结论与 Washietl 等人的结论是一致的[10],他们在研究中发现,核小体连接 DNA 的替换率(Substitution rates)要低于核小体包装 DNA,得出了与本文相同的结论。

3.2 核小体缺乏区域具有较低预测准确率

在图 1 中,我们观察到核小体缺失区域的预测准确率明显偏低,这与其他区域的情况明显不同。这一区域是聚合酶Ⅱ与其他转录蛋白共同形成转录机器的场所,尽管其编码了启动子、转录起始位点和增强子等基因组信息,但由于其对核小体的驱离主要是招募蛋白导致的,其机理与基因调控有关,因此原因较为复杂,同样造成支持向量机在此处表现不佳,难以获得有效的预测信息。因此,在核小体缺失区域,我们所获得的预测准确率较低。

3.3 转录因子结合位点分布与预测准确率的关系

由图 2 中可知,在核小体缺失区域,预测准确率与核小体占位率都较低,但是转录因子结合率较高。这个区域是形成转录机器的场所。真核生物转录起始十分复杂,往往需要多种蛋白因子的协同。所以,在这个区域上核小体的缺失主要是由转录因子结合造成的。由于这个区域的特殊性,此处的 DNA 对核小体的包装能力成为次要因素,因此支持向量机难以获得预测信息。这进一步说明了我们上面得出的核小体缺乏区域具有较低预测准确率产生的原因。

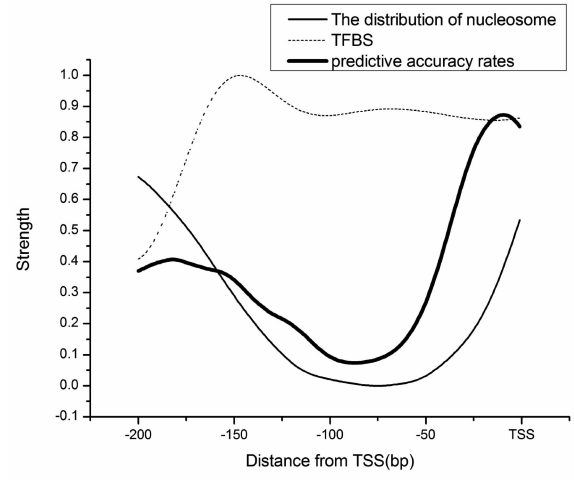

图 2 核小体缺失区域(NFR)上的预测准确率与核小体占位率分布谱

Fig. 2 The distribution of nucleosome and the predictive accuraty rates in NFR

4 结 语

我们用支持向量机对酵母菌基因组 DNA 进行核小体形成能力预测,结果表明编码在 DNA 中的核小体形成信息在包装区域与连接区域具有显著的差异性,这代表了 DNA 在局部进化上的不同。在此,我们从另一个角度证明了核小体在进化过程中的保守性,这一结果有着重要的生物学意义。

参考文献(References)

[1] Chen, S H, Sun, J, Dimitrov, L, Turner, A R, Adams, T S, Meyers, D A, Chang, B L, Zheng, S L, Gronberg, H, Xu, J, Hsu, F C, A Support Vector Machine Approach for Detecting Gene – gene Interaction[J]. Genet Epidemiol, 2008, 32(2): 152-167.

[2] Zhenhai Zhang and B. Franklin Pugh, High-resolution Genome-wide Mapping of the Primary Structure of Chromatin[J]. Cell, 2011. 144(2): 175-186.

[3] Sasaki, S, Mello, C C, Shimada, A, Nakatani, Y, Hashimoto, S, Ogawa, M, Matsushima, K, Gu, S G, Kasahara, M, Ahsan, B, Sasaki, A, Saito, T, Chromatin-associated Periodicity in Genetic Variation Downstream of Transcriptional Start Sites[J]. Science, 2009. 323(5912): 401-404.

[4] JANSEN, A. and K. J. Verstrepen, Nucleosome Positioning in Saccharomyces Cerevisiae[J]. Microbiol Mol Biol Rev, 2011, 75(2): 301-320.

[5] William Lee, Desiree Tillo, Nicolas Bray, Randall H Morse, Ronald W Davis, Timothy R Hughes Corey Nislow, A high-resolution Atlas of Nucleosome Occupancy in Yeast[J]. Nat Genet, 2007, 39(10): 1235-1244.

(下转至第 160 页)

$$K = \frac{aa_2k_1k_3k_5 + aa_1k_1k_3k_6 + ak_1k_3k_4}{k_1k_3k_4 + a_2k_1k_3k_5 + a_1k_1k_3k_6 + a_1a_2k_2k_4k_6 + a_1k_3k_4k_6} \quad (4)$$

(4)式表明[ADP]和[Pi]对 F 的影响是不一样的。F 随[Pi]增大而减小,但 F 随[ADP]的增大而增大。Zhe Lu 在对兔腰肌纤维的实验中指出,当 ADP 的浓度从 0.05mm ~ 0.5mm 变化到后,肌纤维产生的张力增加约 19%[8]。为了进一步证明模型的结果,我们计算一下(4)式中力随 ADP 变化的情况。

依据文献[8]、[9],上式中的速率常数及有关物质浓度取值如下:

$k_1 = 8s^{-1}, k_2 = 1M^{-1}s^{-1}, k_3 = 2s^{-1}, k_4 = 25s^{-1}$, $k_5 = 10^6M^{-1}s^{-1}, k_6 = 10^6M^{-1}s^{-1}, a_1 = 5 \times 10^{-3}M, a_2 = 0.05 \times 10^{-3}M, a_3 = 0.05 \times 10^{-3}M$。

我们将数据代入(4)式中力 F 的方程得到当 ADP 的浓度从 0.05mM 变化到 0.5mM 后,肌纤维产生的张力从 0.060 997 3Ka 增加到 0.066 048 7Ka,增加了 8.28%,与实验结果基本相符。

为了了解处于四个态中肌球蛋白分子浓度的比例,将上面的数值代入(3)式,得到

$$\begin{aligned} [A \cdot M \cdot D \cdot Pi] &= x_1 = \frac{0.25}{1.3304}a \\ [A \cdot M \cdot D] &= x_2 = \frac{1}{1.3304}a \\ [A \cdot M^2 \cdot D] &= x_3 = \frac{0.08}{1.3304}a \\ [A \cdot M^*] &= x_4 = \frac{0.0004}{1.3304}a \end{aligned} \quad (5)$$

从(5)式可以看出,在所有与肌动蛋白丝处于结合态的肌球蛋白分子中,强结合态的分子约占 6%。$[A \cdot M^*]/[A \cdot M^* \cdot D] = 1/200$,处于僵直态的肌球蛋白远少于能发生动力冲程态的肌球蛋白,这种分布有利于肌肉的收缩。这也说明了 Houdusse 和 Sweeney 模型的合理性。

参考文献(References)

[1] Mooseker MS, Cheney RE. Unconventional myosin[J]. Annu Rev Cell Dev Biol, 1995,11:633.

[2] Rayment I, Rypniewski WR, Schmidt-Base K, Smith R, Tomchik DR, Benning MM, Winkelmann DA, Wesenberg G, Holden HM. Three-dimensional structure of myosin subfragment-1: a molecular motor[J]. Science, 1993,261:50.

[3] Chen Y, Yan B, Chalovich JM, Brenner B. Theoretical kinetic studies of models for binding myosin subfragment-1 to regulated actin: Hill model versus geeves model[J]. Biophys J, 2001,80: 2338.

[4] Spudich JA. How molecular motor works[J]. Nature, 1994,372: 515.

[5] 郭维生,罗辽复. 肌球蛋白工作循环的一个新模型[J]. 生物化学与生物物理进展,2003,30:216.

[6] Houdusse A, Sweeney Hl. Myosin motors: missing structures and hidden springs[J]. Current Opinion in Structural Biology, 2001, 11:182.

[7] Suzuki Y, Yasunaga T, Ohkura R, Wakabayashi T, Sutoh K. Swing of the lever arm of a myosin motor at the isomerization and phosphate – release steps[J]. Nature, 1998,396:380.

[8] Zhe lu, Darl R Swartz, Joseph M Metzger, Richard L Moss, Jeffery W Walker. Regulation of force development studied by photolysis of caged ADP in rabbit skinned psoas fibers[J]. Biophys J, 2001,81:334.

[9] Eisenberg E, Hill TL, Chen YD. Cross-bridge model of muscle contraction[J]. Biophys J, 1980,12:195.

(上接第 152 页)

[6] David, LHuber, W, Granovskaia, M, Toedling, J, Palm, C J, Bofkin, L, Jones, T, Davis, R W, Steinmetz, L M, A high-resolution Map of Transcription in the Yeast Genome[J]. Proc Natl Acad Sci U S A, 2006,103(14): 5320-5325.

[7] VEERLA, S., M. Ringner and M. Hoglund, Genome-wide Transcription Factor Binding Site/Promoter Databases for the Analysis of Gene Sets and Co-occurrence of Transcription Factor Binding Motifs [J]. BMC Genomics, 2010,11: 145-150.

[8] Heather E. Peckham Robert E. Thurman, Yutao Fu, John A. Stamatoyannopoulos, William Stafford Noble, Kevin Struh and Zhiping Weng, Nucleosome Positioning Signals in Genomic DNA[J]. Genome Res, 2007,17(8): 1170-1177.

[9] Ronen Sadeh, and David Allis, Genome-wide "re"-modeling of Nucleosome Positions[J]. Cell, 2011,147(2): 263-266.

[10] Stefan Washietl, Rainer Machne and Nick Goldman Evolutionary Footprints of Nucleosome Positions in Yeast[J]. Trends Genet, 2008,24(12): 583-587.