PIBB 生物化学与生物物理进展
Progress in Biochemistry and Biophysics
2009, 36(8): 1035~1040
www.pibb.ac.cn

核小体定位与 RNA 剪接 *

陈 伟 罗辽复 ** 张利绒 邢永强
(内蒙古大学物理科学与技术学院，呼和浩特 010021)

摘要 根据核小体定位序列和缺失序列的碱基分布特征，应用多样性增量二次判别方法(IDQD)构建模型对这两类序列进行了区分，受试者操作特性曲线下的面积达到了 0.958. 应用这一模型研究了核小体在人类基因组剪接位点(GT/AG)邻近序列中的分布方式，发现外显子所对应的 DNA序列通常倾向参与核小体的形成，并且由它所转录的 RNA 统计上具有较强的刚性，而剪接位点及其邻近的内含子对应的 DNA 序列则避免参与核小体的形成，所转录的 RNA 统计上具有较强的柔性. 进一步还发现，DNA 序列的核小体定位 / 缺失和 RNA 的刚性 / 柔性具有统计相关性，为从机制上解释为何前体 RNA 剪接事件与 DNA 序列中的核小体定位信息有关提供了依据.

关键词 多样性增量二次判别方法，核小体定位，剪接位点，RNA 柔性
学科分类号 Q61 **DOI:** 10.3724/SP.J.1206.2008.00816

核小体是染色质结构的基本重复单元，由 DNA 和组蛋白以特殊的方式相连接而组成. 每个核小体包括 200 bp 左右的 DNA 超螺旋，一个组蛋白八聚体(由 4 种核心组蛋白 H2A、H2B、H3 和 H4 各 2 个单体构成)以及一个单体组蛋白 H1，其中核心区的 DNA 序列长度约为 150 bp[1]. 由于真核生物基因组中的大部分 DNA 序列都在组蛋白八聚体上缠绕参与核小体的形成，因此基因组中核小体定位的研究成为现代分子生物学的前沿和焦点.

基因组中的核小体定位是指在特定的时空条件下 DNA 序列急剧弯曲并以超螺旋的方式在组蛋白八聚体上缠绕. 核小体的精确定位决定了 DNA 分子的可接近性，阻断了许多蛋白质因子与 DNA 分子的接触机会，因此对基因的表达和调控起着重要的作用[2~4]. 核小体在基因组中的定位在一定程度上要受到序列信息的影响，不同碱基组成的 DNA 序列之间核小体形成能力的强弱差别可以达到 1 000 倍[5]. 一些 DNA 序列，尤其是基因编码区的序列，能够长期稳定地与核小体结合，并在漫长的进化过程中逐渐优化 DNA 序列组分以利于与核小体的结合[6].

虽然已经有研究人员对基因组中的核小体分布方式及其对基因表达的影响进行了分析，但是大部分工作都是在启动子区展开的. 本文以 Dennis 等[7]提供的人类基因组的生物芯片数据为基础，利用多样性增量二次判别算法[8,9]构建了分类器，成功地对核小体定位序列进行了识别. 应用该分类器对人类基因组中可变剪接和组成性剪接位点邻近序列进行分析，发现基因组中外显子所对应的 DNA 序列表现出极强的核小体形成能力，而剪接位点及其邻近的内含子所对应的 DNA 序列的核小体形成能力则比较弱，对组蛋白八聚体表现出了极强的排斥行为，避免在组蛋白八聚体上缠绕，处于核小体缺失状态. 同时，我们还定义并研究了核小体定位 / 缺失序列所转录的 RNA 序列的柔性，发现 DNA 序列的核小体定位 / 缺失和 RNA 刚性 / 柔性具有统计相关性，初步解释了为何 RNA 剪接事件与 DNA 序列中的核小体定位信息有关. 这将有助于 RNA 剪接机制的更深入探索和基因组中剪接位点的识别.

1 数 据

1.1 核小体定位和缺失序列

1 000 条核小体定位序列(nucleosome positioning sequence)和 1 000 条核小体缺失序列(nucleosome

* 国家自然科学基金资助项目(90403010, 10447003).
** 通讯联系人.
Tel: 0471-4992676, E-mail: lolfcm@mail.imu.edu.cn
收稿日期：2008-11-27，接受日期：2009-03-09

inhibiting sequence)取自 Dennis 等提供的芯片数据[7, 10]，所用探针长度为 50 mer，覆盖了 42 个基因的转录起始位点上游 20 kb 到下游 5 kb 的区域[7]. 为了避免对剪接位点邻近区域分析时出现外显子偏好现象，把这 2 000 条长度为 50 bp 的序列同人类 cDNA 文库(http://atidb.cshl.org/maize/Homo_sapiens.NCBI36.50.cdna.all.fa)进行 BLAST 比对分析，剔除了与 cDNA 存在相似性(相似性阈值 80%)的序列后，剩余的 762 条核小体定位序列和 916 条核小体缺失序列分别作为核小体定位序列预测时的训练正集和训练负集.

1.2 剪接位点邻近序列

从 ASD 数据库[11](http://www.ebi.ac.uk/asd/altsplice)人类基因组 release 3 的数据资料中获取所有组成性外显子剪接位点对(AG/GT)的定位信息. 考虑到核小体核心区的 DNA 序列长度约为 150 bp，因此挑选出了长度分布在 140～160 bp (909 对)、160～180 bp (899 对)、190～210 bp (538 对) 和 210～230 bp (284 对)范围内且其紧邻的内含子长度大于 1 000 bp 的外显子所对应的剪接位点对(AG/GT). 然后分别以受体端(AG)和供体端(GT)剪接位点为中心截取 1 000 bp 长的片段(AG/GT 上下游各取 500 bp)作为剪接位点邻近序列.

2 方法和序列特征

使用多样性增量二次判别方法(IDQD)[8, 9]对核小体定位序列和核小体缺失序列进行分类预测.

统计核小体定位(缺失)序列中 k-mer(k=1,2,⋯,6)的出现频数[10～12]，定义 4^k (k=1, 2, ⋯, 6)维的向量 $X_{1\text{-}6}$，构建核小体定位(缺失)序列的 6 个多样性源. 相应地由训练正集(G_1)和训练负集(G_2)的全部序列分别建立 6 个多样性源，作为正集标准源和负集标准源. 对于任意一个序列 Y，通过计算它和正负集标准源的多样性增量[13]，得到一个 12 维的特征向量 R (x_1, x_2, ⋯, x_{12})作为序列 Y 的分类参数.

当 $Y \in G_i$ (i=1, 2)时，则可以得到训练集中序列的分类参数，对训练集内全部序列计算平均值，便可得到 12 维的平均向量 μ_i(i=1, 2)和协方差矩阵 $\sum_i$(i=1, 2). 这样对于一个待识别的序列 Y，通过二次判别函数 ξ 就能够给出该序列的分类判别.

$$\xi=\log_2\frac{N_1}{N_2}-\frac{\delta_1-\delta_2}{2}-\frac{1}{2}\log_2\frac{|\sum_1|}{|\sum_2|}\quad \delta_i=(R-\mu_i)^T\sum_i^{-1}(R-\mu_i) \tag{1}$$

式中 N_1 和 N_2 分别为训练正负集中的样本数目，$|\sum_i|$ (i=1, 2)是协方差矩阵行列式的值. 在通常情况下，正负集样本可在 ξ 空间的 ξ_0=0 附近分得很开，如果 $\xi > \xi_0$，则序列 Y 被识别为真(核小体定位序列)，否则识别为假(核小体缺失序列). 由于正负集样本数量的有限性，两者可能对正态分布有偏离，正负集样本的分界点不能保证为 0，最佳分界值 ξ_0 要由经验选取[13].

3 结　果

3.1 核小体定位序列的识别

IDQD 算法在核小体定位序列和缺失序列识别方面的性能可以通过定义敏感性(S_n)，特异性(S_p)和总体预测成功率(Acc)来评价.

$$S_n = \frac{TP}{TP + FN} \tag{2}$$

$$S_p = \frac{TN}{TN + FP} \tag{3}$$

$$Acc = \frac{TP + TN}{TP + FN + TN + FP} \tag{4}$$

把训练集中所有序列计算得到的 ξ 值排序后，我们确定出最佳分界点为 ξ_0=−0.65. 当 ξ_0=−0.65 时，IDQD 模型对核小体定位序列和缺失序列识别的敏感性和特异性分别为 95.28%和 96.07%，总体预测成功率达到了 95.71%.

由于在 IDQD 方法中 ξ_0 是给出分类判别的唯一参数，算法的敏感性和特异性要随着 ξ_0 的变化而改变，通过计算受试者操作特性(receiver operating characteristic, ROC)曲线下的面积(area under ROC，auROC)可以实现对算法性能的客观评价[14]. 图 1 给出了由 IDQD 算法构建的分类器对核小体定位和缺失序列进行识别时的 ROC 曲线，10 交叉检验的 auROC 平均值达到了 0.958，说明我们的模型具有较高的可靠性，能够在基因组尺度上展开核小体定位序列的预测.

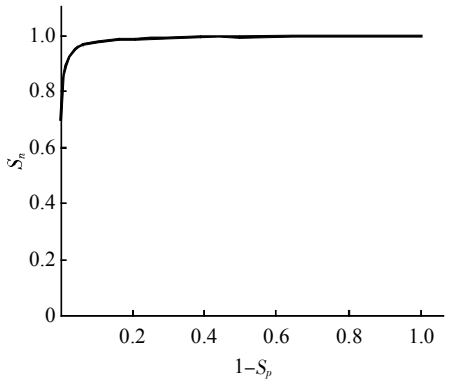

Fig. 1 The ROC curve for the prediction of nucleosome positioning and inhibiting sequences

3.2 剪接位点邻近的核小体分布

为考察核小体在剪接点邻近序列中的分布，引入核小体形成能力得分 S 来衡量 DNA 序列核小体形成能力的强弱．训练集中的每一个序列经过二次判别函数处理后都会得到一个无量纲的参数 ξ，把它们按照大小顺序排列，找出最大值 ξ_{max} 和最小值 ξ_{min}．对于一个待判别的序列，求得参数 ξ 后，核小体形成能力得分 S 定义为，

$$S = \frac{\xi - \xi_{min}}{\xi_{max} - \xi_{min}} \quad (5)$$

令 $\xi = \xi_0 = -0.65$ 时的得分为阈值 S_0，如果序列的得分 S 大于 S_0，则说明该序列是核小体定位序列，否则为核小体缺失序列．

由于芯片数据中探针的长度为 50 mer，因此，以窗口宽度为 50 bp，步长为 10 bp 分别计算了 4 类不同长度范围的组成性外显子剪接位点的供体端(GT)和受体端(AG)邻近序列的核小体形成能力得分 S，并绘制了分值 S 按照位置变化的曲线，如图 2 所示．

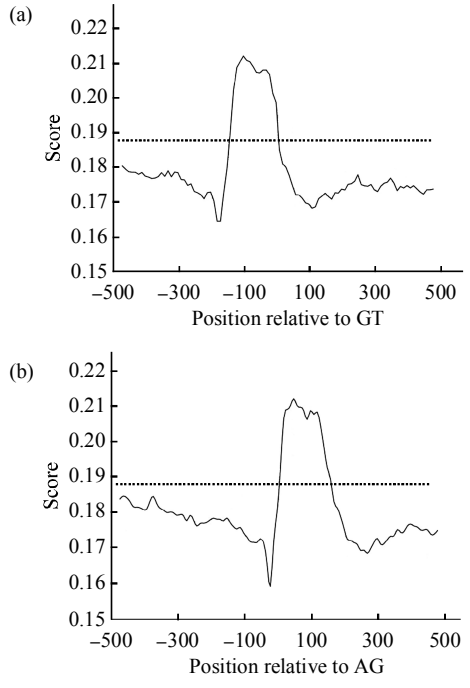

Fig. 2 The nucleosome formation potential around splice sites (GT/AG) of constitutive splicing

The data were smoothed with a 50 bp sliding window in 10 bp increments from −500 to 500 bp relative to the splice sites of constitutive exons that ranged from 140 to 160 bp and given for the donor (a) and acceptor sites (b), respectively. The cutoff value S_0 = 0.188 is pointed out by the horizontal dash line in the graph. The x-axis gives the position of the sliding window labeled by its center relative to donor site GT (denoted as 0) or to acceptor site AG (0) and the y-axis represents the nucleosome formation score. —: S; ·····: S_0.

结果表明，外显子对应的 DNA 序列的核小体形成能力得分 S 总是高于阈值 S_0，说明这些 DNA 序列更偏好于核小体的形成，而剪接位点(GT/AG)及其邻近的内含子所对应的 DNA 序列的得分 S 则明显低于阈值 S_0，极力地避免参与核小体的形成，并且靠近位点 GT/AG 的内含子区对核小体表现出了极强的排斥行为．通常剪接位点 GT/AG 邻近的内含子序列中包含了剪接过程发生所必需的共有序列(供体位点、分支位点、多嘧啶区和受体位点)[15]，这些区域对核小体的排斥性有利于剪接复合体识别并结合共有序列．核小体形成能力得分的最低值出现在受体端(AG)上游约 25 bp 的分支位点邻近，说明核小体的定位可能影响剪接过程中套索结构的形成以及 U2 snRNP 同分支位点序列的结合．由于篇幅所限，文中只列出了长度 140～160 bp 的外显子所对应剪接位点对(AG/GT)邻近序列的统计结果(图 2)．对于另外 3 种长度的组成性外显子(160～180 bp、190～210 bp 和 210～230 bp)，得到的结果完全类似．

3.3 核小体定位与 RNA 柔性

内含子序列的切除需要前体 RNA 序列形成同剪接复合体相匹配的高级结构并在适当位置断裂，该过程的实现通常与 RNA 序列的柔性和结构稳定性有关．

迄今为止，还没有描述 RNA 序列动力学柔性的合适参数．因此，仿照 DNA 序列柔性的描述方法[16, 17]，引进 RNA 序列柔性的概念，用序列折叠自由能的均方偏差来定义 RNA 的柔性．在给定温度下，一条 RNA 序列通常存在多种可能的折叠方式，不同方式的折叠自由能之间存在差异，折叠自由能的差异越大，RNA 序列的柔性就越大．

由于 RNA 序列折叠自由能的大小和序列的长度相关[18]，我们对四类长度范围的外显子逐一进行了分析．在剪接位点对 AG/GT 之间分别取长度 150、170、200 和 220 bp 的片段作为核小体定位序列所转录的 RNA，记为 *Exon*．在 AG 上游和 GT 下游分别截取相应长度的片段(150、170、200 和 220 bp)作为核小体缺失序列所转录的 RNA，记为 I_{up} 和 I_{down}．

Vienna 软件包中的 RNA structure 程序[19]是预测 RNA 二级结构的常用程序之一，使用该程序可以计算三类序列(I_{up}，*Exon* 和 I_{down})中任意一条 RNA 序列可能出现的二级结构及其折叠自由能．将 RNA 二级结构数目设置为 20，温度参数使用默认

值 37℃，计算出每一条 RNA 序列折叠自由能的平均值 $\overline{E}$ 和标准偏差 Sd_E

$$Sd_E = \sqrt{\frac{\sum_{j=1}^{n}(E_j-\overline{E})^2}{(n-1)}} \tag{6}$$

$$\overline{E} = \frac{\sum_{j=1}^{n} E_j}{n} \tag{7}$$

其中 E_j 为第 j(j=1, 2, …, n)种结构所对应的折叠自由能，n=20 为二级结构的数目.

利用折叠自由能在单位均值上的离散程度定义柔性系数 *CF* (coefficient of flexibility)，

$$CF = 100 \times \frac{Sd_{Energy}}{|\overline{E}|} \tag{8}$$

柔性系数 *CF* 值越大，意味着 RNA 序列的结构变化就越大，因此序列的柔性就越大.

表 1 给出了不同长度范围的外显子所对应的三类序列间(I_{up}、*Exon* 和 I_{down})柔性系数比值的均值(< *CF* (I_{up}) / *CF* (*Exon*) > 和 < *CF* (I_{down}) /*CF* (*Exon*)>)及其标准偏差. 从表 1 中的结果可以看出，外显子上游内含子序列的柔性比外显子序列的柔性大约高 25%，而外显子下游内含子序列的柔性比外显子大约高 19% (其中 210～230 bp 组的序列数较少，所以数据偏差较大). 由于所考虑的 I_{up}，*Exon* 和 I_{down} 三类序列的核小体形成能力是有明显差异的，故所得结果说明 DNA 序列的核小体形成能力和其所转录的 RNA 序列的柔性相关. 核小体形成能力强(弱)的 DNA 序列转录的 RNA 序列的平均柔性较弱(强).

Table 1　The comparison of *CF* for RNA transcribed from nucleosome positioning and inhibiting sequences

Exon size/bp	<*CF* (I_{up})/ *CF* (*Exon*)>		<*CF* (I_{down})/ *CF* (*Exon*)>	
	Mean	Standard deviation	Mean	Standard deviation
140～160	1.26	0.54	1.19	0.52
160～180	1.24	0.55	1.19	0.53
190～210	1.26	0.64	1.19	0.44
210～230	1.25	0.58	1.12	0.58

为了进一步说明核小体在剪接位点附近的空间排布方式对 RNA 序列柔性的影响，我们分析了 DNA 序列的核小体形成能力和其所转录的 RNA 序列柔性系数 *CF* 间的关系. 由于 RNA 折叠自由能是在长度为 *L* (*L*≈150, 170, 200, 220 bp)的序列中计算得到的，而核小体形成能力得分 *S* 是针对长度为 50 bp 的序列定义的，所以在分析时需用窗口宽度为 50 bp、步长为 10 bp 的方法求出长度为 *L* bp 的 DNA 序列核小体形成能力的平均得分 $\overline{S}$，并以此来表示该序列的核小体形成能力. 图 3 给出了 DNA 序列的核小体形成能力的平均得分和 RNA 柔性系数间的关系. 结果表明，二者存在显著的负相关性(r=–0.498，P < 0.0001)，DNA 序列的核小体形成能力越弱，由它转录的 RNA 序列的柔性系数越大. 因此，剪接位点附近的 DNA 序列对核小体的偏好决定了由它所转录的 RNA 序列的柔性特征，核小体形成能力弱的 DNA 序列所转录的 RNA 具有较强的折叠柔性，而核小体形成能力强的 DNA 序列所转录的 RNA 则具有较强的折叠刚性.

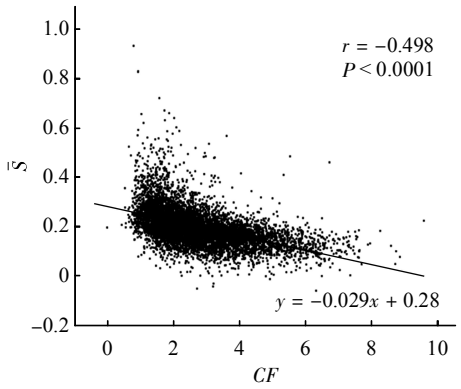

Fig. 3　Correlation between coefficient of flexibility and nucleosome formation potential

The relation between the coefficient of flexibility and the nucleosome formation potential is plotted for four groups of exons and introns considered in the text. The coefficient of flexibility *CF*(x-axis) is negatively correlated with the average nucleosome formation score $\overline{S}$(y-axis) calculated with a 50 bp sliding window in 10 bp increments. The regression equation is $y = -0.029x + 0.28$.

4　讨　　论

依据核小体定位序列和缺失序列中的碱基分布特征，以 k-mer (k=1, 2, …, 6)作为 IDQD 方法的输入参数，成功地对这两类序列进行了识别，并构建了核小体定位序列的预测模型. 使用该模型对人类基因组中组成性剪接位点邻近序列的核小体分布方式进行了研究. 结果表明，编码区 DNA 序列的核小体形成能力得分较高，倾向于参与核小体的形成，而剪接位点和其邻近的内含子所对应的 DNA 序列的核小体形成能力得分总是低于阈值 S_0，极力地避免参与核小体的形成.

通过定义柔性系数 *CF*，实现了序列间柔性的

比较，发现外显子上下游内含子序列的柔性比外显子序列的柔性高．核小体形成能力弱的 DNA 序列(对应于内含子)所转录的 RNA 序列的柔性大于核小体形成能力强的 DNA 序列(对应于外显子)所转录的 RNA 序列的柔性，并且 DNA 序列的核小体形成能力的平均得分 $\overline{S}$ 和 RNA 序列的柔性系数 CF 之间呈现负相关性．

上述研究是以组成性剪接的数据为基础进行的，为了更深入地探究核小体在剪接位点邻近序列中的分布，我们还进一步分析了可变剪接的情况．利用 **3.2** 中的方法，对 ASD 数据库中[11]外显子长度分布在 140～160 bp(716 对)、160～180 bp(789 对)、190～210 bp (485 对)、210～230 bp (305 对)范围内的盒式外显子进行分析，发现核小体在剪接位点邻近序列中的分布方式与组成性剪接的情况相同:核小体形成能力得分的峰值出现在外显子区域，而波谷则分布在外显子两侧剪接位点邻近的内含子区域，虽然核小体形成能力得分在远离剪接位点的内含子区域中逐渐呈上升的趋势，但是内含子区域中的得分仍然低于阈值．

对于这四类不同长度范围的盒式外显子，DNA 序列核小体形成能力的平均得分 $\overline{S}$ 和 RNA 序列的柔性系数 CF 之间同样也表现出了负相关性($P < 0.0001$)，相关系数分别为–0.528、–0.485、–0.478 和–0.466．考虑到盒式外显子是可变剪接的主要类型，这个结果说明，无论是组成性剪接还是可变剪接，核小体在剪接位点邻近 DNA 序列中的定位都和该区域所转录的 RNA 序列的柔性相关，即核小体形成能力强的 DNA 序列所转录的 RNA 具有较强的折叠刚性，而核小体形成能力弱的 DNA 序列所转录的 RNA 序列则具有较强的柔性，结构多变，容易发生折叠和弯曲．

染色质结构的高度保守，DNA 序列在核小体上的缠绕和断裂基因的存在是真核生物基因组的特有现象．我们有理由相信，核小体在剪接位点邻近序列中的分布方式及其和 RNA 序列柔性的关联是真核生物基因组在进化过程中形成的，这种关系的存在可能为剪接复合体在内含子序列上的准确定位、剪接位点的识别和内含子序列切除提供了有效的机制：a．剪接过程需要剪接复合体和前体 RNA 序列结合．剪接位点邻近内含子序列的较大柔性使得这些区域更容易招募各种蛋白质分子形成剪接复合体，通过 RNA-RNA、RNA- 蛋白质、蛋白质 -蛋白质等多重相互作用有效地切除内含子．b．外显子对应的 DNA 序列参与核小体形成则保证了剪接位点识别的准确性．这些 DNA 序列所转录的 RNA 具有较强的折叠刚性和内部应力，因此剪接过程中 RNA 序列发生折叠和弯曲时，外显子整体不容易跟随发生结构变化，但易于在它的两个端点发生断裂，从而为准确剪接提供了进一步的保障．

本文仅对人类基因组中的剪接位点邻近序列进行了分析，下一步工作将探究这种机制在其他物种中的推广．

参 考 文 献

1 Thåström A, Bingham L M, Widom J. Nucleosomal locations of dominant DNA sequence motifs for histone-DNA interactions and nucleosome positioning. J Mole Biol, 2004, **338**(4): 695～709

2 Narlikar G J, Fan H Y, Kingston R E. Cooperation between complexes that regulate chromatin structure and transcription. Cell, 2002, **108**(4): 475～487

3 Boeger H, Griesenbeck J, Strattan J S, *et al*. Nucleosomes unfold completely at a transcriptionally active promoter. Mol Cell, 2003, **11** (6): 1587～1598

4 Reinke H, Hörz W. Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. Mol Cell, 2003, **11**(6): 1599～1607

5 Segal E, Fondufe-Mittendorf Y, Chen L, *et al*. A genomic code for nucleosome positioning. Nature, 2006, **442**(7104): 772～778

6 Chen K, Meng Q, Ma L, *et al*. A novel DNA sequence periodicity decodes nucleosome positioning. Nucl Acid Res, 2008, **36** (19): 6228～6236

7 Dennis J H, Fan H, Reynolds S M, *et al*. Independent and complementary methods for large-scale structural analysis of mammalian chromatin. Genome Res, 2007, **17**(6): 928～939

8 Laxton R R. The measure of diversity. J Theor Biol, 1978, **70** (1): 51～67

9 Zhang L R, Luo L F. Splice site prediction with quadratic discriminant analysis using diversity measure. Nucl Acid Res, 2003, **31**(21): 6214～6220

10 Gupta S, Dennis J, Thurman R E, *et al*. Predicting human nucleosome occupancy from primary sequence. PLoS Comput Biol, 2008, **4** (8): e1000134

11 Thanaraj T A, Stamm S, Clark F, *et al*. ASD: the alternative splicing database. Nucl Acid Res, 2004, **32**(Database issue): D64～D69

12 Mai X, Chou S, Struhl K. Preferential accessibility of the yeast *his3* promoter is determined by a general property of the DNA sequence, not by specific elements. Mol Cell Biol, 2000, **20**(18): 6668～6676

13 吕 军, 罗辽复. 人类 pol Ⅱ 启动子的识别. 生物化学与生物物理进展, 2005, **32**(12): 1185～1191
Lu J, Luo L F. Prog Biochem Biophys, 2005, **32**(12): 1185～1191

14 Lu J, Luo L F. Prediction for human transcription start site using diversity measure with quadratic discriminant. Bioinformation, 2008, **2**(7): 316～321

15 Benjamin L. GENES Ⅷ. New Jersey: Pearson Prentice Hall, 2004. 705～736

16 Tsai L, Luo L F. A statistical mechanical model for predicting B-DNA curvature and flexibility. J Theor Biol, 2000, **207**(2): 177～194

17 Luo L F. Theoretic-physical Approach to Molecular Biology. Shanghai: Shanghai Science and Technology Publisher, 2004. 408～428

18 Carlini D B. Context-dependent codon bias and messenger RNA longevity in the yeast transcriptome. Mol Biol Evol, 2005, **22**(6): 1403～1411

19 Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucl Acid Res, 1981, **9**(1): 133～148

Nucleosome Positioning and RNA Splicing*

CHEN Wei, LUO Liao-Fu**, ZHANG Li-Rong, XING Yong-Qiang

(*School of Physical Science and Technology, Inner Mongolia University, Hohhot* 010021, China)

Abstract Based on the characteristic of nucleotide distribution in nucleosome positioning and inhibiting sequences, the method of Increment of Diversity with Quadratic Discriminant (IDQD) was applied to the classification of these two types of sequences. The mean area under ROC curve archives 0.958. By using this model, the nucleosome formation potential was analyzed in the regions around the splice sites (GT/AG). The results show that coding regions have a high potential to form the nucleosome and the primary RNA transcripts are rigid, while DNA sequences corresponding to the splice sites and their adjacent intron regions tend to be nucleosome free and the primary transcripts from these regions are relative flexible. Moreover, the negative correlation between nucleosome positioning/inhibiting of DNA sequences and RNA flexibility/rigidity is demonstrated around the splice sites, providing a mechanism for understanding the correlation between the nucleosome positioning of DNA and the splicing of transcribed RNA sequences.

Key words increment of diversity with quadratic discriminant analysis, nucleosome positioning, splice sites, RNA flexibility

DOI: 10.3724/SP.J.1206.2008.00816

*This work was supported by grants from The National Natural Science Foundation of China (90403010,10447003).

**Corresponding author.

Tel: 86-471-4992676, E-mail: lolfcm@mail.imu.edu.cn

Received: November 27, 2008 Accepted: March 9, 2009