大肠杆菌基因组结构柔性的研究

蔡 禄[1,2],孙之荣[1]
(1. 清华大学生命科学和技术系,北京 100084;2. 包头钢铁学院生物与化学工程系,内蒙古包头 014010)

摘 要: 基于四核苷酸参数提出一个统计力学模型用于分析大肠杆菌基因组柔性. 结果表明:大肠杆菌基因组复制终止区柔性明显小于其他区域;柔性和 G+C 含量之间具有极强的关联;编码区的平均柔性明显大于非编码区的平均柔性. 对大肠杆菌基因组中三类 σ 类型启动子序列的结构柔性特征也作了分析.

关键词: 非紧邻碱基相互作用; 统计力学模型; DNA 柔性; 全基因组分析
中图分类号: Q617 **文献标识码**: A **文章编号**: 0372-2112 (2001) 12A-1753-03

A Study on Structural Flexibility in Escherichia Coli Genome

CAI Lu[1,2], SUN Zhi-rong[1]
(1. *Department of Biological Science and Technology, Tsinghua University, Beijing* 100084, *China*
2. *Department of Biological and Chemical Engineering, Baotou University of Iron and Steel Technology, Baotou, Neimenggu* 014010, *China*)

Abstract: Based on tetranucleotide parameters, a statistical mechanical model was suggested to analyze the flexibility of the *Escherichia coli* genome. The terminus region of replication displayed a low level of flexibility. A strong correlation between G+C content and flexibility can be seen. Average flexibilities in coding regions were found to be significantly larger than those in non-coding regions. The flexible characteristics in three class sigma promoter sequences in the *Escherichia coli* genome were also analyzed.

Key words: non-neighbor base-pair interaction; statistical mechanical model; DNA flexibility; whole-genome analysis

1 引言

依赖序列的 DNA 柔性在诸如复制、转录、重组和修复等生命过程中有重要作用. 尽管已有一些关于 DNA 柔性研究的理论工作[1,2],但这些模型并不完善和系统化. 迄今为止,人们引入诸如持久长度、扭曲刚性和环化概率等量来描述 DNA 柔性[3,4]. 尽管这些量在一定程度上能够描述 DNA 的结构柔性,但是均未强调 DNA 柔性本质上是由其动力学结构决定的这一实质问题.

从序列到结构,最终到功能的阐述,一直是生物信息学的主流研究思路. 今天,已经可以获得一系列全基因组序列. 这样,就有可能在全基因组水平理解生物学过程. 本文考虑碱基对之间的非紧邻相互作用、螺旋参数的非对称热涨落和依赖序列的涨落,提出一个由 DNA 序列预测其动力学结构的统计力学模型. 利用本模型和 DNA 柔性的新定义,可以方便地预测 DNA 柔性. 本文主要研究大肠杆菌基因组的柔性,其它基因组的研究可作类似处理.

2 材料和方法

2.1 晶体结构和基因组序列

本文使用的 DNA 双螺旋结构参数包括角参数 Ω、ρ 和 τ 及位移参数 D_y. 这些参数均由核酸数据库(NDB)中的数据计算而得. 这里仅研究 *B*-DNA. 为使讨论具有普遍性,使用的序列均不含误配碱基对、非 Watson-Crick 碱基对、不寻常碱基或裂口. 大肠杆菌 *K*-12 基因组全序列取自文献[5].

2.2 四联体中内碱基对的参数集

与以前大部分工作只研究碱基对二联体不同,这里研究碱基对四联体来分析碱基对的非紧邻相互作用. 由于目前尚无足够数据,四联体两侧用约化的嘌呤、嘧啶语言表示. 考虑到双螺旋的互补性,约化后的四联体共 36 个. $\Delta\Omega = <\Omega> - \Omega_0$, $\Delta\rho = <\rho> - \rho_0$, $\Delta\tau = <\tau> - \tau_0$, $\Delta D_y = <D_y> - D_{y0}$. 这里 $\Omega_0 = 36°$, $\rho_0 = 0.4°$, $\tau_0 = 0°$, $D_{y0} = 0.2$Å 是 NDB 中碱基对梯阶的平均螺旋参数. $<\Omega>$, $<\rho>$, $<\tau>$ 和 $<D_y>$ 是 NDB 中 36 个四联体内碱基对的螺旋参数的统计平均值. 36 个四联体内碱基对的 $\Delta\Omega$, $\Delta\rho$, $\Delta\tau$ 和 ΔD_y 如表 1.

2.3 B-DNA 螺旋结构参数的经验预规则

若第 j 个序列的实验螺旋扭角 Ω 为 $T_j(1)$, $T_j(2)$ ……$T_j(i)$ ……,第 j 个序列的理论相对螺旋扭角用 $Tw_j(1)$, $Tw_j(2)$ ……, $Tw_j(i)$ …表示. 定义

$$P = \sum_{j=1}^{N}\sum_{i=1}^{L_j-1}\mathrm{sgn}\left(\frac{Tw_i(i+1) - Tw_i(i)}{T_j(i+1) - T_j(i)}\right) \Big/ \sum_{j=1}^{N}(L_j - 1) \quad (1)$$

这里, N 是 NDB 库中的序列数; L_j 是第 j 个序列的长度. P 描述实验曲线 $T_j(i) - i$ 和预测曲线 $Tw_j(i) - i$ 之间的一致

收稿日期:2001-06-27;修回日期:2001-08-30
基金项目:国家自然科学基金资助(No. 199470006);中国博士后科学基金资助;内蒙古自然科学基金资助(No. 20001302)

表 1 *B*-DNA 36 个四联体内碱基对局部螺旋参数

Tetramer	RAAR	RAAY	YAAR	YAAY	RAGR	RAGY	YAGR	YAGY	RGAR	RGAY	YGAR	YGAY
$\Delta\Omega$(°)	0.0	1.0	-2.0	-5.0	-2.0	-2.0	-10.0	2.0	3.0	3.0	3.0	2.0
$\Delta\rho$(°)	0.2	-0.7	3.9	4.1	-2.0	-1.5	2.0	4.6	2.3	1.1	1.4	1.7
$\Delta\tau$(°)	0.0	0.0	-1.1	-0.4	1.5	1.4	-1.2	3.0	2.2	0.0	-0.5	-1.0
ΔD_y(Å)	-0.2	-0.2	0.2	0.0	0.0	0.2	-0.3	0.8	0.2	-0.4	-0.4	-0.3

Tetramer	RGGR	RGGY	YGGR	YGGY	RCAR	RCAY	YCAR	YCAY	RCGR	RCGY	YCGR	RTAR
$\Delta\Omega$(°)	4.0	1.0	4.0	-4.0	-3.0	-3.0	11.0	-4.0	-3.0	2.0	-4.0	1.0
$\Delta\rho$(°)	3.5	2.7	-4.0	4.3	2.2	7.6	-7.2	2.1	1.0	7.5	4.5	5.8
$\Delta\tau$(°)	-2.0	-0.4	-1.4	-1.4	2.2	-1.0	0.1	3.0	0.5	-0.9	2.2	4.2
ΔD_y(Å)	0.5	0.5	0.5	0.5	1.3	1.3	2.2	1.0	0.1	-0.2	0.7	0.2

Tetramer	RTAY	YTAR	RATR	RATY	YATR	RACR	RACY	YACR	YACY	RGCR	RGCY	YGCR
$\Delta\Omega$(°)	6.0	-5.0	-3.0	-4.0	-5.0	-5.0	2.0	-2.0	2.0	0.0	0.0	2.0
$\Delta\rho$(°)	2.9	3.3	-5.1	-4.6	-2.0	-2.0	-2.0	-5.3	-6.0	-1.2	3.0	-6.1
$\Delta\tau$(°)	3.0	1.6	-1.7	-0.3	2.0	0.0	-4.0	-1.0	5.0	-1.5	-3.2	0.2
ΔD_y(Å)	0.2	0.3	-0.5	-0.5	-0.2	-0.5	-0.5	-0.3	-0.3	0.5	-0.2	0.0

程度. $\mathrm{Sgn}(X) = X/|X|$. 沿 5′→3′方向阅读 DNA 序列,每一四联体的中央值取自表 1. 第 j 个序列的第 i 个碱基对螺旋扭角的理论相对预测值 $Tw_j(i) = \Delta\Omega_j(i) - f\Delta\Omega_j(i+1) - f\Delta\Omega_j(i-1)$. 这里,调节因子 $f(0 \leqslant |f| \leqslant 1)$ 表示两侧碱基对对四联体中内碱基对结构的调节,其值通过对 NDB 库中全部序列使 P 极大化而得. 第 j 个序列的第 i 个碱基对螺旋扭角的理论预测值 $\Omega_j(i) = \Omega_0 + Tw_j(i)$. 序列的 ρ, τ 和 D_y 值的预测过程与 Ω 的预测完全类似. 调节因子 f 对 Ω, ρ, τ 和 D_y 分别取 0.125, 0.125, 0.075 和 -0.125.

2.4 构象能模型

在此,首先对文献[6]中提出的统计力学模型作一简要回顾. 假设各螺旋结构参数的涨落相互独立,并且用简单的弹性能模型描述局部碱基对梯阶几何形状的涨落. i 碱基对的螺旋扭角位于 Ω_i 和 $\Omega_i + d\Omega_i$ 之间的概率为 $P(\Omega_i)\ d\Omega_i$. $P(\rho_i)$, $P(\tau_i)$ 和 $P(Dy_i)$ 可按同样方式导出.

$$P(\Omega_i) = \frac{\exp\left(\frac{k_\Omega}{2RT}(\Omega_i - \Omega_{i0})^2\right)}{\int_{-\infty}^{\infty}\exp\left(-\frac{k_\Omega}{2RT}(\Omega_i - \Omega_{i0})^2\right)d\Omega_i} = \sqrt{\frac{k_\Omega}{2\pi RT}}\exp\left(-\frac{k_\Omega}{2RT}(\Omega_i - \Omega_{i0})^2\right) \quad (2)$$

$$P(\Omega_i, \rho_i, \tau_i, D_{yi}) = P(\Omega_i)\ P(\rho_i)\ P(\tau_i)\ P(D_{yi}) \quad (3)$$

这里, R 是普适气体常数, T 是温度. 本文所有计算都是对室温而言的. 使用的力常数列于表 2. k_Ω, k_ρ, k_τ 的单位为 kJ/mol·rad^2, $k_{D_{yi}}$ 为 kJ/mol·Å^2, $R = 0.00831$kJ/mol·K.

表 2 模型中 10 个二联体的力常数

Step	AA	AC	AG	AT	CA	CG	GA	GC	GG	TA
k_Ω	109.5	111.6	192.3	94.5	141.7	139.2	99.5	116.2	244.5	166.8
$k_{\rho 1}$	120	405	360	330	165	135	210	360	555	90
$k_{\rho 2}$	60	75	120	60	195	285	90	90	75	210
k_τ	180	180	348	132	372	339	228	345	180	240
$k_{D_{yi}}$	7.6	9.0	19.5	11.1	18.5	17.6	13.8	5.8	19.3	7.9

由于 *B*-DNA 的碱基对平面接近垂直于整体螺旋轴,可以用连接相邻碱基对平面局域坐标系原点的一系列虚键矢量 V_i ($i = 1, 2 \cdots\cdots L-1$, L 是序列的长度) 来近似研究其动力学结构. 设每一碱基对梯阶的 D_x 和 D_z 分别等于 0Å和 3.4Å 那么,在局域坐标系中 $V_j = (0, D_{yj}, 3.4$Å$)$. 整体坐标系下的对应矢量称为碱基对矢量为 $a_j = \mathbf{R}_{12}\,\mathbf{R}_{23}\cdots\mathbf{R}_{j-1,j}V_j$. $\mathbf{R}_{j-1,j}$ 是与第 $(j-1)$ 和第 j 个碱基对有关的变换矩阵,其形式于文献[6]中相同. 让 $<r_j>$ 是第 j 个碱基对梯阶位置矢量 r_j 的期望值. 假设碱基对梯阶的结构涨落是独立的,那么

$$\begin{aligned}<r_j> &= <a_1> + <a_2> + <a_3> + \cdots + <a_j> \\ &= <V_1> + <\mathbf{R}_{12}>V_2 + <\mathbf{R}_{12}><\mathbf{R}_{23}>V_3 + \cdots \\ &\quad + <\mathbf{R}_{12}><\mathbf{R}_{23}>\cdots<\mathbf{R}_{j-1,j}>V_j \end{aligned} \quad (4)$$

使用公式(3),可得平均变换矩阵 $<\mathbf{R}>$ 为 Ω, ρ 和 τ 的正弦和余弦平均值的函数(见[6]). 定义长度为 L 的给定 DNA 序列的结构柔性为

$$f_L = \sqrt{<\Delta r_L^2>} = \sqrt{<r_L^2> - <r_L>^2} \quad (5)$$

在此定义下,用前面提到的统计力学模型可以方便地计算任意 DNA 序列的结构柔性.

3 结果和讨论

3.1 大肠杆菌 K-12 基因组的结构柔性

首先,按非重叠 1000 碱基对(bp)窗口将实际和搅乱的大肠杆菌基因组分割成一系列片断,然后用提出的统计力学模型按位置计算序列柔性. 接下来,构建显示基因组任意区域柔性的曲线图. 搅乱的大肠杆菌基因组序列的柔性值大多位于 1100Å到 1180Å之间,而实际大肠杆菌基因组序列的柔性值分布在更广的 1080Å到 1200Å之间的范围. 换句话说,实际大肠杆菌基因组序列比搅乱的大肠杆菌基因组序列柔性更加极端.

图 1 中对实际大肠杆菌基因组用 100kbp 窗口平滑后的柔性轮廓图表明:在 1.2Mbp 到 1.7Mbp 之间有一柔性极小区,极小值在 1.6Mbp 附近. 图 1 还显示在其它位置出现了一系列极大和极小值. 实验结果表明复制终止区和起始区分别位于 1.5Mbp 和 4Mbp 附近,理论预测结果和实验数据精确地吻合. 在功能重要位点处观察到的 DNA 柔性特殊模式在结构水平为基因组研究提供了信息.

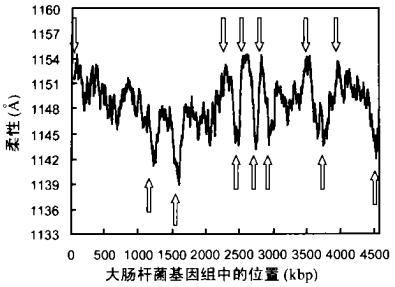

图 1　用 100kbp 滑动窗口计算大肠杆菌基因组的柔性轮廓图

为了理解基因组中柔性背后的机制,用非重叠 1000kbp 窗口计算每一位置下的 G+C 含量. G+C 含量和 DNA 柔性之间存在强烈的关联(相关系数 $R = 0.922$). 富含 G+C 区通常更加柔软. G+C 含量和 DNA 柔性之间强烈的关联可能与本模型下极端柔性二核苷酸 CA、AG、GC 和 CG 是富含 G 或 C 碱基的事实有关.

3.2　编码区、非编码区和搅乱的非编码区中的柔性

用 Genbank 中的注释信息从大肠杆菌基因组提取编码区和非编码区,然后分别结合成两个子序列研究基因组中编码区和非编码区的差别. 用非重叠 1000 碱基对窗口计算这两子序列的柔性,并构建显示柔性的曲线图. 图 2 显示:大肠杆菌基因组中编码区、非编码区和搅乱的非编码区中的平均柔性分布有很大不同. 编码区中 DNA 柔性明显大于非编码区中的相应值. 因此,我们推测非编码区中较低的 DNA 柔性可能是原核生物基因表达调控要求的普遍结构性质. 这一结果也可作为基因组中发现新基因的辅助指标.

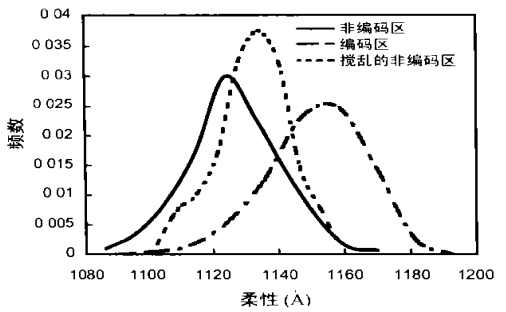

图 2　大肠杆菌基因组编码区、非编码区和搅乱的非编码区中柔性分布

3.3　启动子柔性

使用 Genbank 中的注释信息从大肠杆菌基因组提取三类 σ 类型启动子(σ-70,σ-54 和 σ-32)序列,按转录起始位点将全部启动子序列对齐,转录起始位点作为位置 0. 用 147bp 滑动窗口计算 -500bp 到 500bp 区域中每个位置柔性的平均值(每次窗口滑动 25bp). 图 3 计算结果显示:对 σ-70 和 σ-32 类型启动子序列,转录起始位点上游 20bp 到 200bp 有一柔性极小区,下游 24bp 到 124bp 有一柔性区. 注意到 σ-54 情形与前面提到的两类明显不同,这一结论与实验事实完全一致[7,8].

已有报道指出:许多蛋白质-DNA 复合物的形成与 DNA 柔性有关,DNA 柔性还会影响启动子活性. 内部或蛋白诱导的 DNA 柔性片断常常位于复制、转录起始重要的区域. 这些刚性或柔性区域行使功能的机制尚不清楚,由于缺乏实验数据目前无法给出生物学解释.

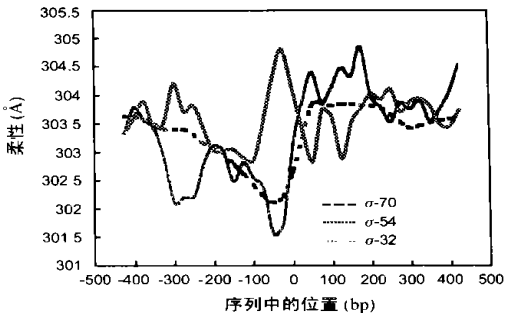

图 3　三类 σ 类型启动子序列转录起始位点上、下游各 500bp 区域中平均柔性

参考文献:

[1] Olson W K,et al. DNA sequence-dependent deformability deduced from protein-DNA crystal complexes [J]. Proc. Natl. Acad. Sci. USA 1998, 95:11163-11168.

[2] Baldi P,et al. Structural basis for triplet repeat disorders:a computational analysis [J]. Bioinformatics. 1999,15:918-929.

[3] Bhattacharyya D,et al. Sequence directed flexibility of DNA and the role of cross-strand hydrogen bonds [J]. J. Biomol. Struct. Dynam. 1999,17:289-300.

[4] Cognet J A H,et al. Static curvature and flexibility measurements of DNA with microscopy. A simple renormalization method:its assessment by experiment and simulation [J]. J. Mol. Biol., 1999,285:997-1009.

[5] Blattner F R,et al. The complete genome sequence of *Escherichia coli* K-12 [J]. Science 1997,277:1453-1474.

[6] Tsai L,Luo L F. A statistical mechanical model for predicting B-DNA curvature and flexibility [J]. J. Theor. Biol. 2000,207:177-194.

[7] Gross C A,Lonetto M,Losick R. Bacterial sigma factors [R]. In transcriptional regulation (Mcknight, S. and Yamamoto, Y. eds), Cold spring Harbor Laboratory Press,Cold Spring Harbor,NY. 1992:129-176.

[8] Merrick M J. In a class of its own-the RNA polymerase sigma factor σ^{54} (σ^{N}) [J]. Mol. Microbiol. 1993,10:903-909.

作者简介:

蔡　禄　男. 1964 年 8 月出生于内蒙古呼和浩特. 博士后,生物物理学会会员. 1987 年在内蒙古大学物理系获硕士学位,2000 年在内蒙古大学理工学院获博士学位. 现主要从事生物信息学方向的研究工作,发表论文 40 多篇.

孙之荣　男. 1947 年 4 月出生于江苏扬州. 1970 年清华大学热工测量及自动化专业毕业,现任清华大学生物科学与技术系教授、博士生导师,生物物理学会会员. 主要从事生物信息学方向的研究工作,发表论文 60 多篇.