

基于 AFM 成像的 DNA 长度的计算机统计处理

陈新庆² 张 益¹ 胡 钧¹ 吴世英¹ 黄一波¹ 艾小白¹ 邱锡钧² 李民乾¹

1 (中国科学院上海原子核研究所 上海 201800)

2 (上海大学 上海 201800)

摘要 利用计算机图像处理技术,对自然沉积 DNA 和拉直 DNA 分子 AFM 图像的长度进行了自动统计测量,并与 DNA 分子的理论长度进行了比较。该方法在 DNA 分子的 AFM 图像的基础上,将 DNA 分子拟合成曲线,用插值法求得 DNA 分子长度,整个过程实现了自动化。结果表明,该方法测得自然沉积的某 DNA 分子长度为 $(180.4 \pm 16.4) \text{ nm}$,与理论长度 185.0 nm 相当符合。该方法与拉直法和直接法求 DNA 分子长度相比是一次改进,也为其它基于 AFM 图像的线状物体求长度提供了一种新的测量方法。对另一组拉直 DNA 测量的结果表明 拉直后 DNA 分子长度为 $(343.6 \pm 20.7) \text{ nm}$,与理论长度 307.0 nm 相比明显变长,这表明对自然沉积 DNA 进行长度测量的必要性,保证了 DNA 分子的长度不受拉直后有所伸长的影响。

关键词 图像处理, DNA, AFM, MATLAB

中图分类号 Q523, N32

目前测量 DNA 长度的标准的化学方法是琼脂糖或聚丙烯酰胺凝胶电泳法^[1],该方法可同时对大量 DNA 分子染色后进行宏观测量。但电泳法对大片段 DNA 的长度测量不灵敏,而且实验中用到的化学物质是有毒的。

发展单分子水平的生物大分子特性统计方法是当前科学研究的一个前沿,其中一个基础性的研究工作是对 DNA 分子的几何、物理状态作单分子水平上的统计测量^[2]。这一工作对双螺旋结构 DNA 分子弹性模型^[3]的建立以及空间构象^[4]的研究具有重要意义。

原子力显微镜 (Atomic Force Microscope, AFM) 是 80 年代中期新兴起来的原子级分辨率的显微镜^[5],为单分子水平上测量 DNA 长度提供了一个全新的方向。原子力显微镜有许多优点,可以研究各种导电与非导电样品、能在液体条件下工作、可以提供高度信息、可进行无损成像等,使得它一产生就被应用于生物学领域^[6]。目前利用 AFM 图像测量 DNA 分子长度有两种方法^[7]:一种是将 DNA 分子拉直后测量,另一种是沿自然状态下 DNA 分子的走向测量。这两种方法都是在 AFM 离线处理系统中沿着 DNA 分子走向测量两点间距后相加而得。测量方法是用手工作将直线段代替曲线,结果比较粗糙,自动化程度较低,而单分子统计需要大量的生物分子样品,手工操作速度太慢,误差太大,显然已经不能满足统计的需要。并且,在拉直操纵的影响下, DNA 分子有所拉伸,结果与 DNA 的自然长度有明显差异。

通过计算机编程对实验数据进行自动处理近年来已十分普遍。如何从复杂图像中提取有

中国科学院生物科学与生物技术特别支持项目 (STZ - 00 - 07)、国家自然科学基金委九. 五重大交叉资助项目 (19890385)、国家杰出青年基金资助项目 (19725415)、中国科学院九. 五重点基础资助项目 (KJ952 - J1 - 469)、上海优秀学科带头人资助计划 (00XD14029)

第一作者 陈新庆,男,1971 年出生,1996 年毕业于山东师范大学物理系,硕士研究生

收稿日期 2000 - 06 - 19, 修回日期 2000 - 12 - 11

用的数据,近年来成为对图像数据进行处理的热点之一^[8,9]。因为对于一般的实验数据,无论是数据矩阵,亦或是数据向量,都可以直接进行运算,而图像实验数据需要进行复杂的预处理才能从中提取有用的信息。将图像数据中有用信息向量化是图像数据预处理的重要一步^[9],如本文中介绍的细化、寻找起始点和关键点,都是为了将图像实验数据中有用的信息(DNA 分子的中轴)向量化,都可以归为数据的预处理过程。而其中的数据处理(求长度)过程倒非常简单。

Matlab(Matrix laboratory)语言是近年来十分流行的一门主要用于科学计算的语言。特别是它具有图像处理工具箱(image processing toolbox),能对图像进行整体的形貌处理^[10]。但将图像数据矩阵中的有用信息向量化的技术还不成熟,直接对 DNA 分子图像数据向量化的计算还不多见。本工作在这方面进行了探讨,应用 Matlab 语言编程,实现了 DNA 分子图像数据的向量化,对自然状态下 DNA 分子的长度进行了自动统计。

目前,在物体测量方面可以通过二维空间中一簇点拟合一条曲线。例如对一个曲线型物体的细化骨架拟合一条中轴,但这种拟合受到某种限制,就是只能对单调向量进行拟合,也就是对任意的图像 $F(x)$ 必须是单调的,这可能会妨碍拟合曲线恰当的穿过中轴所在点^[8]。作者认为,首先将采样点向量化,可以避免这种缺陷。

基于 AFM 图像的计算机统计测量,只要 AFM 图像的分辨率足够,就可以对任意 DNA 分子进行测量。更重要的是,该方法是基于单分子水平的统计测量,可以在单分子水平上对 DNA 提供更为详尽的描述,这是该方法不同于电泳法的主要特点,为在单分子水平上研究 DNA 分子提供了一种有力的工具。

1 材料与方法

1.1 DNA 分子的 AFM 图像获取

新剥离的云母表面用 1% 的 APTES 试剂处理 2min,水冲洗后空气吹干。自然沉积样品 DNA 长 544bp,通过 PCR 技术扩增后经纯化得到 拉直样品 DNA 长 903bp,也是通过 PCR 技术扩增后经纯化得到。两种样品 DNA 均用 TE 缓冲液(40mmol/L Tris - HCl,1mmol/L EDTA,pH8.0)稀释为 3ng/ μ L 备用。

取 2 μ L 上述 544bp 长的 DNA,滴于 Para 膜上,将约 1cm² 大小的 APTES 处理过的云母片覆盖于上面,静置 2min,然后用水冲洗云母表面,空气吹干。

长为 903bp 的 DNA 由改进的分子梳方法拉直于云母表面^[11]。

实验仪器为美国 Digital 公司生产的 Nanoscope III AFM 系统,样品由轻敲模式 AFM 成像,所有操作均在大气条件下进行,实验结果为高度模式图像。仪器测量系统由标准样品标定,X 和 Y 方向误差均不超过 3%。

1.2 DNA 长度的计算机统计处理

将 AFM 图像以 Tiff 的格式输出到 photoshop 图像处理软件中。调节图像的灰度阈值,选取恰当的阈值将图像二值化(二值化是指在 Photoshop 图像处理软件中,根据背景与图像的灰度差异,选取恰当的灰度阈值,将图像与背景分开的一种图像处理手段)。利用编好的 MATLAB 软件对图像中的 DNA 进行识别,自动测量出每条 DNA 分子的长度。

2 实验结果

2.1 AFM 成像

用原子力显微镜获得的 DNA 分子的 AFM 图像见图 1、2。图 1 为自然沉积样品 DNA 的 AFM 图像 图 2 为拉直后样品 DNA 的 AFM 图像。

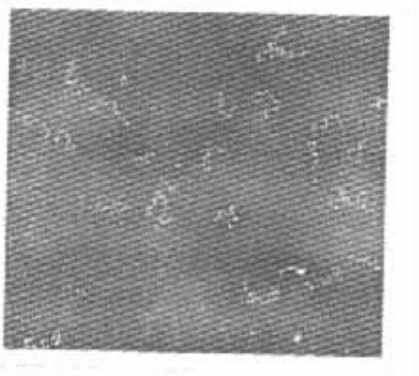


图 1 自然沉积 DNA 的 AFM 图像
Fig.1 Image naturally deposited DNA

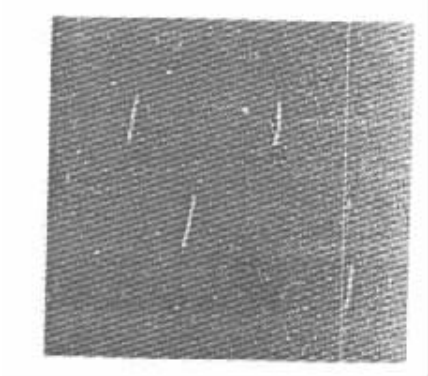


图 2 拉直 DNA 的 AFM 图像
Fig.2 Image of straight DNA

2.2 图像处理

样品 DNA 长度的统计分布如图 3—6 所示。图 3 为自然沉积 DNA 长度频数分布图 图 4 为拉直 DNA 长度频数分布图 图 5 是自然沉积 DNA 的长度分布曲线 图 6 是拉直 DNA 长度分布曲线。此次统计我们共取自然沉积 DNA 样品 1005 条,拉直 DNA 样品 381 条。有关的统计数据列于表 1 中。

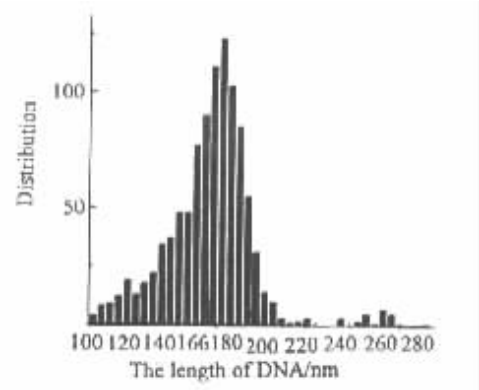


图 3 自然沉积 DNA 频数分布
Fig.3 Length distribution of naturally deposited DNA

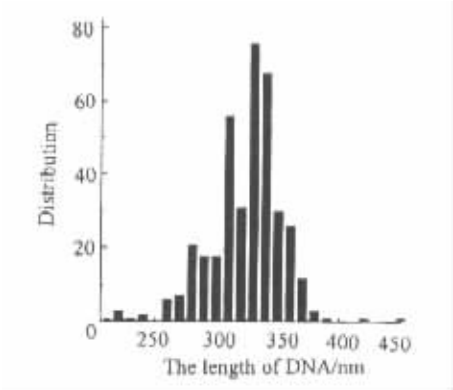


图 4 拉直 DNA 频数分布
Fig.4 Length distribution of straight DNA

从表 1 可知,自然沉积 DNA 样品长度主要集中在 (174.5 ± 16.4) nm 左右, DNA 分子长度频数分布曲线的半宽度为 16.4nm。过长和过短的 DNA 都比较少。DNA 的长度分布相对比较集中,但与理论值 185.0nm 之间尚不能完全吻合。校正后的 DNA 长度为 (180.4 ± 16.4) nm,与理论值 185.0nm 符合得相当好,误差仅为 2.5%。

拉直 DNA 样品长度为 (335.8 ± 20.7) nm,校正值为 343.6nm,与理论长度 307.0nm 相比明

显偏长,说明拉直操纵对 DNA 长度有明显影响。

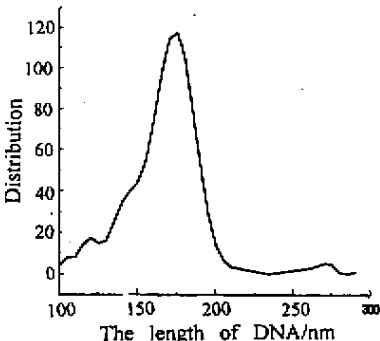


图 5 自然沉积 DNA 长度分布曲线
Fig.5 The length distribution curve of deposited DNA

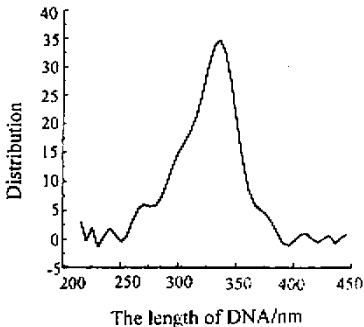


图 6 拉直 DNA 长度分布曲线
Fig.6 The length distribution curve of straight DNA

表 1 自然沉积 DNA 和拉直 DNA 统计数据

Table 1 The data of deposited DNA and straight DNA (nm)

统计量 Statistical parameters	自然沉积(544bp) Deposited DNA	拉直 DNA(903bp) Straight DNA
峰 值 Peak	174.5	335.8
平均值 Mean	168.5	324.3
校正值 Modification	180.4	343.6
理论值 Theoretical length	185.0	307.0
半宽度 Half - width	16.4	20.7
标准差 Standard deviation	26.3	29.0
误 差 Error/%	2.5	11.9

从图 5、6 中可以看到,无论是自然沉积 DNA 样品,还是拉直 DNA 样品,DNA 长度频数分布的最大值都大于平均长度,呈负偏态分布。

3 讨论

自然沉积 DNA 分子的长度集中在(174.5 ± 16.4) nm 之间,比理论值 185.0nm 偏小的原因可能有：

(1)在将 DNA 分子向量化的过程中,由于关键点的选取以 3 个像素为步长,所以起点选取和终点确定都可能造成误差,而且这种误差只可能导致 DNA 分子长度的计算值小于其实际长度。但对终点确定而言,这种误差是可以修正的。确定终点时有三种可能：①终点的确定与 DNA 分子的末端恰好重合,这时就关键点的选取而言正好取到了 DNA 分子的实际长度,不造成误差；②终点的确定与 DNA 分子末端相差一个像素；③相差两个像素。后两种可能因终点与 DNA 分子末端的差距小于步长(3 个像素),这种差距无法在关键点的选取中体现,从而被舍弃,这是不合理的。根据概率论的基本原理,这三种情况出现的概率是相同的。因此,用该方法测量 DNA 分子长度时,在终点确定上,计算值平均比实际长度小一个像素(5.9nm)。考虑到 DNA 终点确定的影响,矫正的 DNA 长度为(180.4 ± 16.4) nm,而理论值为 185.0nm,误差缩小到 2.5%。这说明程序中端点长度的矫正对结果中的误差有很大程度的修正作用。

(2)AFM 成像时,样品 DNA 本身的小的卷曲、折叠,由于受仪器分辨率的限制,AFM 成像后无法显示这种形变,导致计算机测量长度偏小。这也是该方法在测量 DNA 长度方面受仪器

分辨率的限制而无法避免的缺点。

拉直后 DNA 分子长度明显变长,说明拉直操纵对 DNA 长度有明显影响。这从另一角度说明对自然沉积 DNA 进行长度测量的必要性,保证了 DNA 分子的长度不受拉直后 DNA 有所伸长的影响。

从图 5、6 中可以看到,无论是自然沉积 DNA,还是拉直后 DNA,DNA 长度频数分布的最大值都略大于平均长度,DNA 长度频数分布呈负偏态。从统计学角度讲,样品 DNA 的折叠对长度的影响应当是随机的,只能影响到平均值和标准差,而不应改变样品的分布状态。DNA 的长度频数分布不对称的原因还有待于进一步探讨。

4 结论

(1)基于 AFM 图像的自动统计方法实现了 DNA 长度测量的自动化。统计结果表明,用该方法测得的 DNA 分子长度与理论长度相当符合。同时该方法也为其它基于 AFM 图像的线状物体求长度提供了一种新的方向。

(2)统计结果表明,无论是自然沉积 DNA 样品,还是拉直 DNA 样品,DNA 长度频数分布的最大值都略大于平均长度,呈负偏态分布。

参 考 文 献

- 1 Sambrook J, Fritsch E F, Maniatis T. Molecular cloning, Cold Spring Harbor Laboratory Press,1989
- 2 Julia D, Daniel C. Science, 1999, **283**(5408) :1667
- 3 ZHOU Haijun, ZHANG Yang, OU - YANG Zhongcan. Phys Rev Lett, 1999, **82** 4560—4563
- 4 Axelle T, Giovanni D. Preliminary Proceedings of STM 99. 1999. 12—13
- 5 Binnig G, Quate C F, Gerber Ch. Phys Rev Lett, 1986, **56** 930—933
- 6 Ruger D, Daniel R, Paul H. Phys Today, 1990, **43** 23—30
- 7 ZHANG Yi, XU Lei, LI Minqian, *et al.* Micro Nanometer Sci Technol,1996, **2**(1) :12—17
- 8 Kenneth R C. Digital image processing. Prentice Hall Press, 1996
- 9 Pedro A R R, Jesus D L S. IEEE International Conference on Image Processing. 3rd. 1996. 319—322
- 10 CHEN Guiming, ZHANG Mingzhao, QI Hongyu. Digital signal and digital image processing of MATLAB. Science Press, 2000
- 11 ZHANG YANG, XIONG Q H, LOU S T. *et al.* Probe Microscopy, 2000, **2** 31—36

Statistical length of DNA based on AFM image measured by a computer

CHEN Xinqing² ZHANG Yi¹ HU Jun¹ WU Shiyong¹ HUANG Yibo¹

AI Xiaobai¹ QIU Xijun² LI Minqian¹

1 (Shanghai Institute of Nuclear Research, the Chinese Academy of Sciences, Shanghai 201800)

2 (Shanghai University, Shanghai 201800)

Abstract

Taking advantage of image processing technology, the contour length of DNA molecule was measured automatically by a computer. Based on the AFM image of DNA, the topography of DNA was simulated into a curve. Then the DNA length was measured automatically by inserting mode. It was shown that the experimental length of a naturally deposited DNA ($180.4 \pm 16.4\text{nm}$) was well consistent with the theoretical length (185.0nm). Comparing to other methods, the present approach had advantages of precision and automatism. The stretched DNA was also measured. It was shown that the experimental length ($343.6 \pm 20.7\text{nm}$) was much longer than the theoretical length (307.0nm). This result indicated that the stretching process had a distinct effect on the DNA length. However, the method provided here avoided the DNA – stretching effect.

Key words Image processing, DNA, AFM, MATLAB

CLC Q523, N32