生物化学与生物物理进展 *Progress in Biochemistry and Biophysics*, 2006, 33(7): 660~664
www.pibb.ac.cn

PIBB

一种基于单分子纳米操纵的有序化测序策略 *

吕 鸣[1)] 石宝晨[3)] 李雪玲[1)] 吕军鸿[2)**] 陈润生[3)**] 胡 钧[1,2)**]

([1)]上海交通大学生命科学技术学院 Bio-X 生命科学研究中心，上海 200030；
[2)]中国科学院上海应用物理研究所，上海 201800；
[3)]中国科学院生物物理研究所，北京 100101)

摘要 尽管包括人类在内的许多生物物种的基因组测序工作已经完成，但由于现有测序技术的限制，大部分复杂基因组还存在很多大大小小的缺口. 缺口的填补以及对其他重复序列区域的测序迫切需要全新的思路和技术. 基于在 DNA 单分子定位切割和拾取方面的实验进展，提出了一种基于原子力显微镜纳米操纵技术的单分子有序化测序策略. 计算机模拟的结果表明，这一方法和策略是可行的，有助于解决目前测序工作中所遇到的一些棘手问题.

关键词 DNA 单分子，原子力显微镜，纳米操纵，有序化测序
学科分类号 Q523

核酸是一切生物遗传信息的载体. DNA 测序技术的发展[1~3]，使得人们能够获得其中包含的信息. 自此，人类对生命的研究步入“基于序列的时代”. 由于如今测序技术的局限，一个测序反应一般只能对 800 bp 左右长度的片段进行测序，人们必须根据片段序列之间的重叠，把一个一个的片段拼接起来，从而得到最终的一致序列. 基于这种基本思想，人们陆续提出了一些可用于复杂基因组测序的策略[4]，如目前广泛采用的分级鸟枪法测序(clone-by-clone shotgun sequencing)策略和全基因组鸟枪法测序(whole-genome shotgun sequencing，WGSS)策略[5]. 国际人类基因组测序联盟(IHGSC)[6]和赛雷拉公司(Celera Genomics)[7]分别采用这两种策略完成了人类基因组的测序[8].

分级鸟枪法采用的是分而治之的思想，首先要构建物理图谱，图谱构建好后，整个基因组的测序就转化为对图谱的每个克隆的测序. 各个克隆的测序工作是相互独立的，适合于多个测序组织之间的合作. 但是，图谱的构建是一件费时费力的工作，全基因组鸟枪法测序的提出正是为了跳过图谱构建这一步，直接对整个基因组进行鸟枪法测序[9]. 当然，这样一来，序列的拼接势必比分级鸟枪法困难许多.

1 问题的提出

无论是分级鸟枪法还是全基因组鸟枪法，重复序列的存在都是鸟枪法测序所面临的最大问题. 尽管人类基因组计划已经宣告完成[10]，但报道完成的人类基因组序列并不完美，其中还包含着大大小小的缺口(gap)341 个 (308 个位于常染色质部分)，并且还有占全基因组约 6%的异染色质部分未进行测序. 而这些缺口总体上又分为两大类[11]：一类是已有其克隆库但是很难拼接的缺口(finishing gaps)；另一类是还无法成功获得其克隆库的缺口(clone gaps). 对这些缺口周边的序列特征研究发现其绝大部分都与重复序列有紧密联系[11,12]. 这些缺口代表了那些基于现有技术无法进行可靠的建图谱、克隆和测序的区域，并且大部分都处在着丝粒和端粒部分. 真核细胞染色体着丝粒区域被认为是基因组测序中最具挑战的部分，目前只有包括我国科学家在内的少数研究组[13~16]完成了几个比较简单的着丝粒

*国家自然科学基金重点资助项目(10335070).
** 通讯联系人. Tel/Fax: 021-59552394
E-mail: jhlu@sinap.ac.cn,crs@sun5.ibp.ac.cn, junhu22@hotmail.com
收稿日期：2006-01-23，接受日期：2006-02-28

区域序列的测定，但对于大部分着丝粒人们仍然无能为力.

重复序列所带来的问题是显而易见的，除了可能是导致克隆库无法成功建立的原因外，最直接影响到的就是最终序列架构 (Scaffold) 的建立和确认. 很多分子标记技术，如 STS、EST、RFLP、RAPD、AFLP 等等，被用来解决这一问题. 采取独特的重复序列处理算法，通过架构水平上的接近来进行组装也能有效克服“鸟枪法”全基因组测序组装过程中的困难[17~19]. 另外，其他如插入片段长度限制，比较基因组[19]等方法都被用来帮助序列架构的建立. 但这些方法还不能从根本上解决所有的问题，这就迫切需要发展新的技术来解决这些问题. 现有测序方法与新技术相结合，目前被认为是一种较好的选择. 譬如，利用光学作图(optical mapping)[20~22]技术，能够在单分子水平上确定较长片段的限制性酶切图谱，无需事先获悉序列的信息、PCR 扩增、建库等繁琐的操作，大大简化了整个序列基本架构 (Scaffold) 的建立，在一定程度上减弱了测序中特别是重复序列所带来的问题，但由于光镜的分辨率只有微米或亚微米水平，仅仅有助于解决 kb 以上的片段长度.

2 新的技术和策略

原子力显微镜 (atomic force microscopy，AFM) 除了拥有显微镜所共有的“眼”的功能外，还具有其独特的“手”的功能. 近年来，利用 AFM 在纳米尺度上对 DNA 分子进行操纵成为可能，不仅实现了对 DNA 分子的准确定位切割[23]，而且能够对特定位置的 DNA 分子片段切割后分离出来[24]. 此外，对单个 DNA 分子进行扩增和分析的相关技术也日渐成熟[24~26]. 在上述工作进展的基础上，本文提出一种在单分子水平上对 DNA 进行切割、分离、扩增和测序的策略，我们称之为有序化单分子纳米测序 (ordered single molecule sequencing based on nanomanipulation，OsmSN) 策略 (图 1). 具体思路是利用纳米定位切割技术，把拉直的长 DNA 片段按照顺序依次切割成小片段 (500 bp 左右)，然后通过纳米操纵分离出来单分子小片段，进行单分子 PCR 扩增，再进行常规的测序. 通过纳米操纵获得的这些小片段在长 DNA 链的位置和次序是已知的，所以不需要大规模拼接过程，从而从根本上克服了随机测序的弱点. 由于针尖切割时可能会对 DNA 片段造成一定的缺口损失，因此对于直径 20 nm (相当于 60 bp) 的针尖而言，只要经过两次切割，相互错开 250 bp，切割精度在 100 bp 以内即可. 拉直 DNA 的技术目前可以对 1 Mbp 以上的 DNA 片段进行拉直操纵，未来原则上可以发展到对整个染色体进行拉直操纵. 因此，也可以用于解决无法克隆片段的问题.

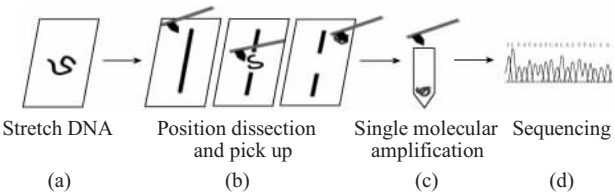

Fig. 1 Ordered single molecule sequencing based on nanomanipulation

(a) Stretching DNA molecule on substrate with molecular combing technology. (b) Image, cut and pick up DNA with AFM tip. (c) Amplification of the picked single molecular DNA.(d) Sequencing.

完全的 OsmSN 方法将不需要进行拼接. 不过，由于目前整个切割拾取过程还未能自动化，效率还不高，要成功完成这样的拾取，目前在技术上还存在困难. 但是，在目前的技术水平下，通过 OsmSN 策略可以获得某些特定区域 DNA 片段的序列信息，由于这些片段所在的位置是明确的，而且分辨率又可达到几百甚至几十碱基的水平，因此若以此为物理或遗传标记作为序列拼接的架构，与鸟枪法相结合 (图 2) 所能获得的部分序列信息，将有助于解决利用当前策略很难或者无法拼接的难题，这对于填补与重复序列相关的缺口将是特别有价值的.

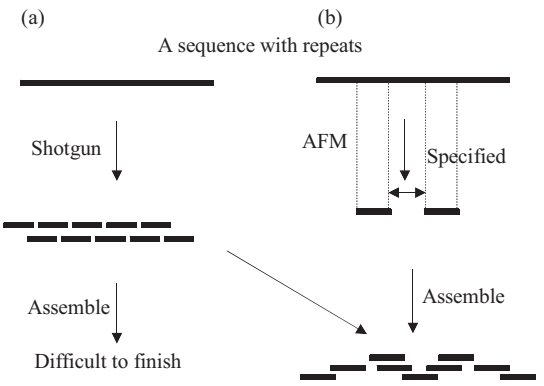

Fig. 2 Assemble sequences with OsmSN and shotgun

(a) Traditional shotgun sequencing, assembly of clones with repeats will be difficult. (b) OsmSN strategy, combined with shotgun method, getting some sequence information with known position guides the assembling process.

3 模拟试验及结果

为了验证这一策略的可行性，我们选取了一个简单的例子，进行计算机模拟. 其基本思想是：人为地构造一条含有重复片段的序列，由于其序列的重复性，单靠鸟枪法而不采取其他措施是无法正确确定其序列的. 采用 OsmSN 方法，事先获得几个指定位置的序列，并且这几个位置的前后顺序和它们之间的间隔都是已知的，然后以这几个序列作为分子标记，检验是否有助于最终序列的拼接.

整个模拟程序分为三部分：获得随机片段序列，获得指定某几个位置的标记序列，以标记序列指导随机片段拼接.

获取随机片段序列. 该部分模拟了对某一克隆进行鸟枪法随机测序的实际过程. 由于目前测序通常采用的是双桶法(double-barrelled)，对一个亚克隆进行前向与反向的双向测序，测得有一定间隔的一前一后两条序列. 基于此，实际模拟时采取如下步骤实现：按一定覆盖倍数在原序列上随机取得一些长度满足均值 3 000、标准差 500 的正态分布的亚克隆片段. 在每个亚克隆片段两端取得长度满足均值 600、标准差 100 的正态分布的序列. 所有取得的序列作为该克隆鸟枪法随机测序测得的片段库. 基于模拟的目的是要验证解决重复序列的问题，模拟所用的序列是连接起来的两条 lambda 噬菌体基因组序列(Accession Number: NC_001416)，总长 97 004 (48 502 × 2) bp.

获得指定的标记序列. 该部分模拟了用 AFM 在 DNA 分子上切割拾取指定位置的片段，然后扩增并最终测得标记序列. 切割的长度与现在一次测序反应所能测得的最大长度一致.

以标记序列指导随机片段拼接. 该部分综合利用了鸟枪法测得的随机片段序列和 AFM 切割拾取获得的标记序列，完成序列的拼接(图 3). 由于序列比对是寻找序列重叠的最基本的运算部分，原始的比对算法计算复杂度比较高，拼接时间势必很长. 目前的拼接程序在这一方面都采用以舍弃比对精确度来达到优化比对速度的方法. 这里也一样，采用的是类似于 Blast[27]的限分动态规划延伸(Gapped extension by score-limited dynamic programming).

作为比较，同时我们还采用 phrap (http://www.phrap.org/) 拼接程序对获得的随机片段序列进行了拼接.

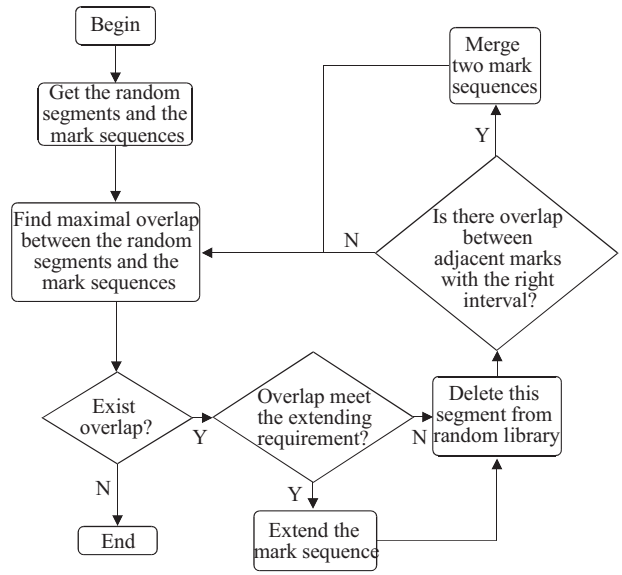

Fig. 3 The assembling procedure of our strategy

为验证拼接结果是否正确，我们分别对拼接成的序列、原始的 lambda 基因组序列、人工构造的两条相连的 lambda 序列和 phrap 拼接结果做了限制性酶切图谱.

从图 4 我们可以看出，单独使用鸟枪法对重复序列进行测序，采用 phrap 程序拼接出来的片段长

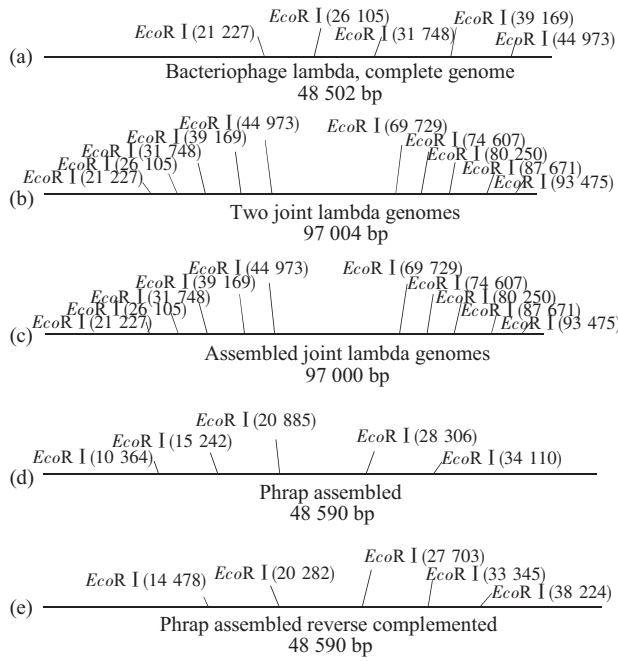

Fig. 4 The assembling results

(a) *Eco*R Ⅰ restriction map of lambda phage genome. (b) The map of two joint lambda phage genomes. (c) The map of the assembly sequence with our strategy. (d) The map of the assembly sequence with phrap. (e) The map of the sequence complemented with (d).

度仅为 48 590 bp (图 4d，4e)，只有原序列长度 97 004 bp 的一半稍多，而且拼接出来的片段也不正确. 通过 OsmSN 方法获得部分区域的序列信息，然后与鸟枪法获得的序列信息相结合，拼接获得的序列其长度 (97 000 bp) 与原序列长度基本一致；同时，其酶切图谱也与序列酶切图谱 (图 4b) 完全一致. 这些结果说明，通过 OsmSN 策略的确可以很好地解决富含重复序列的序列测序与拼接问题. 问题解决的关键在于，测序反应测得的 DNA 片段序列由随机变得有序化了. 相比之下，鸟枪法测序测得的片段位置的未知性正是其难以解决重复序列的主要原因.

4 讨　论

Schwartz 等提出的光学作图[20~22]方法大大简化了基因组物理图谱的构建，给在鸟枪法测序中所不能解决的一些问题提供了解决方案. 虽然与过去相比，物理图谱在分辨率上得到了很大的提高，但仍局限于光镜所能分辨的 kb 片段长度范围内，这对于低于此的重复区域是无能为力的. 事实上，常规方法难以解决的序列往往由长度处于约 1～10 kb 的重复元件组成，长于一次测序反应所能达到的最大长度，而又超出现有作图方法的分辨能力. 最近，Qu 等[28]报道了一种具有纳米分辨能力的荧光显微镜，这使得构建超高分辨 DNA 图谱成为可能，但还必须首先解决荧光探针的非特异性识别的问题. 而 AFM 所能观察和操纵的 DNA 长度却无此限制，尤其是近年来随着纳米管针尖的使用，其分辨率甚至可以达到几个纳米，这对于解决短重复序列是十分重要的. 纵观 OsmSN 策略，在以下几个方面存在潜在的优势：a. 常规的鸟枪法测序，需要对 DNA 片段随机打断，然后克隆、建库等多个步骤，将随机片段拿到测序仪上进行测序之前最少需要几天的时间. 而 OsmSN 方法利用 AFM 的精确定位能力，从而实现了测序的有序化. 同时，DNA 制样非常简单，需要样品的量也很少，可以直接在 AFM 下进行切割、分离和扩增，理想状况下整个测序过程不会超过一天时间. b. 人类基因组序列上的大部分缺口以及着丝粒等异染色质部分都属于当前 DNA 测序上的“禁区”. 由于某些未知原因，这些区域的测序往往十分困难，从而使得人们对它们的认识还十分有限，而这些区域却与细胞分裂等许多重要的生命活动紧密相关，这就迫切需要新的技术来研究它. 由于 AFM 可以在任意位置进行切割，因此能够仅对局部序列而不是整个基因组进行分析. 对 DNA 甚至是染色体的定位切割拾取，使得我们可以直接面对这些“禁区”，揭示其序列特征成为可能. c. 很多生物学现象都是单个分子水平上的因素造成的，单个分子的突变导致截然不同的结果[29]，如病毒的变异、细胞的癌变. 传统方法获得的序列都是一大批分子平均的性质. 以常规的 RNA 病毒的序列分析为例，由于建库所用的模板来自于不同的病毒，最后得到的病毒基因组序列是来自不同病毒的片段序列拼接而成的，因此很难得到病毒的“标准序列”. 而 OsmSN 策略是一种建立在单分子水平上进行测序的思想，因此其发展和延伸能够保证每个克隆库的插入片段都来自于同一病毒.

当然，目前 DNA 单分子操纵与定位切割拾取技术还处于发展之中. 就像自动化测序仪的发明大大推广了常规测序的应用一样，大批量快速廉价的拾取，也有待于自动化技术的引入. 另外，以分离的单个 DNA 分子为对象的后续研究也需进一步探索.

参 考 文 献

1 Sanger F, Nicklen S, Coulson A R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA, 1977, **74** (12): 5463～5467

2 Smith L M, Sanders J Z, Kaiser R J, *et al*. Fluorescence detection in automated DNA sequence analysis. Nature, 1986, **321** (6071): 674～679

3 Hunkapiller T, Kaiser R J, Koop B F, *et al*. Large-scale and automated DNA sequence determination. Science, 1991, **254** (5028): 59～67

4 Venter J C, Smith H O, Hood L. A new strategy for genome sequencing. Nature, 1996, **381** (6581): 364～366

5 Green E D. Strategies for the systematic sequencing of complex genomes. Nat Rev Genet, 2001, **2** (8): 573～583

6 Lander E S, Linton L M, Birren B, *et al*. Initial squencing and analysis of the human genome. Nature, 2001, **409** (6822): 860～921

7 Venter J C, Adams M D, Myers E W, *et al*. The sequence of the human genome. Science, 2001, **291** (5507): 1304～1351

8 Istrail S, Sutton G G, Florea L, *et al*. Whole genome shotgun assembly and comparison of human genome assemblies. Proc Natl Acad Sci USA, 2004, **101** (7): 1916 ～1921

9 Weber J L, Myers E W. Human whole genome shotgun sequencing. Genome Res, 1997, **7** (5): 401～409

10 IHGSC. Finishing the euchromatic sequence of the human genome. Nature, 2004, **431** (7011): 931～945

11 Eichler E E, Clark R A, She X. An assessment of the sequence gaps: unfinished business in a finished human genome. Nat Rev Genet, 2004, **5** (5): 345～354

12 She X, Jiang Z, Clark R A, *et al*. Shotgun sequence assembly and

recent segmental duplications within the human genome. Nature, 2004, **431** (7011): 927～930
13 Feng Q, Zhang Y, Hao P, *et al*. Sequence and analysis of rice chromosome 4. Nature, 2002, **420** (6913): 316～320
14 Nagaki K, Cheng Z, Ouyang S, *et al*. Sequencing of a rice centromere uncovers active genes. Nat Genet, 2004, **36** (2): 138～145
15 Wu J, Yamagata H, Hayashi-Tsugane M, *et al*. Composition and structure of the centromeric region of rice chromosome 8. Plant Cell, 2004, **16** (4): 967～976
16 Zhang Y, Huang Y, Zhang L, *et al*. Structural features of the rice chromosome 4 centromere. Nucl Acids Res, 2004, **32** (6): 2023～2030
17 Wang J, Wong G K, Ni P, *et al*. RePS: A sequence assembler that masks exact repeats identified from the shotgun data. Genome Res, 2002, **12** (5): 824～831
18 Yu J, Hu S, Wang J, *et al*. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science, 2002, **296** (5565): 79～92
19 Yu J, Wang J, Lin W, *et al*. The genomes of Oryza sativa: A history of duplications. PLoS Biol, 2005, **3** (2): e38
20 Cai W, Jing J, Irvin B, *et al*. High-resolution restriction maps of bacterial artificial chromosomes constructed by optical mapping. Proc Natl Acad Sci USA, 1998, **95** (7): 3390～3395
21 Jing J, Reed J, Huang J, *et al*. Automated high resolution optical mapping using arrayed, fluid-fixed DNA molecules. Proc Natl Acad Sci USA, 1998, **95** (14): 8046～8051
22 Lin J, Qi R, Aston C, et al. Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. Science, 1999, **285** (5433): 1558～1562
23 Hu J, Zhang Y, Gao H, *et al*. Artificial DNA patterns by mechanical nanomanipulation. Nano Letters, 2002, **2** (1): 55～57
24 Lü J H, Li H K, An H J, *et al*. Positioning isolation and biochemical analysis of single DNA molecules based on nanomanipulation and single-molecule PCR. J Am Chem Soc, 2004, **126** (36): 11136～11137
25 Rungpragayphan S, Kawarasaki Y, Imaeda T, *et al*. High-throughput, cloning-independent protein library construction by combining single-molecule DNA amplification with *in vitro* expression. J Mol Biol, 2002, 318 (2): 395～405
26 Li H K, Huang J H, Lü J H, *et al*. Nanoparticle PCR: Nanogold-Assisted PCR with Enhanced Specificity. Angew Chem Int Ed, 2005, **44** (32): 5100～5103
27 Altschul S F, Madden T L, Schaffer A A, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acids Res, 1997, **25** (17): 3389～3402
28 Qu X, Wu D, Mets L, *et al*. Nanometer-localized multiple single-molecule fluorescence microscopy. Proc Natl Acad Sci USA, 2004, **101** (31): 11298～11303
29 Hasty J, Collins J J. Translating the noise. Nat Genet, 2002, **31** (1): 13～14

A Strategy for Ordered Sequencing Based on Single Molecule Nanomanipulation*

LÜ Ming[1)], SHI Bao-Chen[3)], LI Xue-Ling[1)], LÜ Jun-Hong[2)**], CHEN Run-Sheng[3)**], HU Jun[1,2)**]

([1)] *Bio-X Life Science Research Center, Shanghai Jiaotong University, Shanghai* 200030, China;
[2)] *Shanghai Institute of Applied Physics, The Chinese Academy of Sciences, Shanghai* 201800, China;
[3)] *Institute of Biophysics, The Chinese Academy of Sciences, Beijing* 100101, China)

Abstract With the limitations of conventional sequencing technologies, there still remains many gaps in complex genomes including the completed human genome. The closure of gaps as well as the solution of other problems caused by repeat sequences call for the development of novel techniques and strategy. Thus an ordered sequencing strategy is proposed based on the experimental advancements in positioning dissection and isolation of single DNA molecules with AFM nanomanipulation. The preliminary results of computer simulation verify its feasibility and indicate the potentialities in overcoming some intractable problems in current sequencing projects.

Key words single DNA molecules, atomic force microscopy, nanomanipulation, ordered sequencing strategy

*This work was supported by a grant from The National Natural Science Foundation of China (10335070).

**Corresponding author . Tel/Fax: 86-21-59552394, E-mail: jhlu@sinap.ac.cn,crs@sun5.ibp.ac.cn, junhu22@hotmail.com

Received: January 23, 2006 Accepted: February 28, 2006