二代测序技术454测序仪模拟测序算法

陈 伟[1,2] 程咏梅[1] 张绍武[1] 潘 泉[1]
(西北工业大学自动化学院 西安 710072)[1] (耶鲁大学医学院 纽黑文 06510)[2]

摘 要 随着环境基因组学及深度测序技术的发展,基于16S rRNA基因序列研究微生物种群结构取得了长足进展。然而,由于环境样本的复杂性,尤其缺少真实背景信息,定量研究环境微生物种群结构仍是当前的研究难点。测序算法仿真平台研究,不仅有助于定量、定性分析微生物种群组成及结构,而且有助于建立基准数据库来评价当前微生物数据分析算法。分别基于易错PCR误差模型和正态分布过程,模拟454测序仪乳液PCR过程及边合成边测序过程,提出454测序仪模拟测序算法(Tsim)。仿真结果表明:该模拟算法能较好地模拟454测序过程。

关键词 16S rRNA基因,模拟测序算法,PCR误差模型,正态分布过程,454测序仪
中图法分类号 TP311.52 **文献标识码** A

Simulation Algorithm for 454 Pyrosequencing Sequencers

CHEN Wei[1,2] CHENG Yong-mei[1] ZHANG Shao-wu[1] PAN Quan[1]
(School of Automation, Northwestern Polytechnical University, Xi'an 710072, China)[1]
(Medicine School of Yale University, New Haven 06510, USA)[2]

Abstract Recent advance of environment genome and deep sequencing technologies has expanded our understanding of composition and structure of microbial community based on 16S rRNA gene sequences. However, the complexity and difficulty of separation of the environmental samples and lack of ground-truth make it difficult to analyze the microbes quantificationally. Thus, simulation datasets will be useful in developing novel softwares because it not only helps us explore the microbial structure quantitatively, but also allow us to construct benchmark studies for evaluating existing methods for processing 16S rRNA sequences data. In the present work, based on error-prone PCR model and making use of the normal distribution model, a simulation algorithm for 454 sequencer (Tsim) was established to simulate the process of sequencing by synthesis. The simulation results show that the simulator can effectively simulate 454 sequencing process.

Keywords 16S rRNA Gene, Simulation algorithm, PCR model, Normal distribution process, 454 sequencer

1 引言

微生物是地球上种类最多、分布最广的生物类型,与人类生活密切相关。它不仅仅在维持生态平衡方面具有重要作用,且与疾病密切相关,目前在医学健康、食品卫生、军事装备养护、农业等方面均具有重要研究及应用价值[1-3]。然而,由于传统培养技术限制,仅有少量的微生物(<1%)可以通过分离培养技术予以描述[4]。近年来,随着宏基因组学的建立,尤其是二代测序技术的发展,突破了传统微生物种群研究的分离培养技术,从整体上研究微生物种群组成与结构成为可能。与传统测序技术相比,它二代测序技术的核心思想是边合成边测序,具有高通量、低成本等优点。自2005年Margulies等人[5]提出焦磷酸测序以来,越来越多的公司和学者开展了相关领域的研究,相继出现了454 FLX、Illumina及SOLID等测序仪。这些工作极大地促进了环境基因组学研究,拓展了人们对土壤[6]、肠道[7]、太空舱[8]、海洋[9]等环境下微生物分布特性的认识。然而,由于当前测序样本大都来自于未知混合环境DNA(Mix Environment Samples)样本,也就说缺少样本真实背景信息,研究人员在处理这些海量测序数据时不得不面对下面问题:如何获知混合物真实组成及结构,如何评价当前各种处理测序数据算法的有效性与精确性。因此,发展二代测序技术仿真算法显得必不可少,它不仅可以扩展我们对微生物种群的认识,还可建立标准数据库用以评价各种微生物应用及处理算法的有效性。

Richter等人[10]对二代测序技术进行研究,提出基因组学模拟算法MetaSim。Balzer等人[11]研究454测序仪数据分

到稿日期:2013-04-23 返修日期:2013-08-03 本文受国家自然基金重点项目(61135001),国家自然科学基金(61170134,60775012),航空基金(20100853010),西北工业大学博士论文创新基金(cx201017)资助。

陈 伟(1984-),男,博士,主要研究领域为信息融合、模式识别、生物信息学,E-mail:chenwei903@gmail.com;程咏梅(1960-),女,博士,博士生导师,主要研究领域为信息融合、证据推理、动态系统建模、组合导航和相对导航中的应用;张绍武(1964-),男,博士,博士生导师,主要研究领域为模式识别理论方法及应用、机器学习、复杂网络、计算生物学,E-mail:zhangsw@nwpu.edu(通信作者);潘 泉(1961-),男,博士,博士生导师,主要研究领域为估计辨识和信息融合理论及应用、无人机导航、避撞及对地探测技术、计算生物学。

布特性，提出一种基于经验分布的 454 测序模拟算法 Flowsim。Lysholm 等人[12]提出一种多线程 454 测序模拟算法。Marth 等人[13]从二代测序数据统计特性出发，提出一种算法模拟 454、Illumina、SoLiD 测序仪测序过程。宣黎明等提出了一种基于 GPU 技术的二代测序技术模拟软件，与 Meta-Sim 相比，它进一步提高了算法速度[14]。虽然上述算法一定程度上加深了我们对测序过程的认识，促进了微生物种群研究，但是这些算法主要针对全基因组测序，难以完成对特定可变区域测序的模拟过程，如 16S rRNA 基因的 V4/V5/V6 区域的模拟测序。与此同时，上述算法仅考虑部分误差，如测序误差，未考虑其它误差因素，如乳液 PCR (ePCR) 误差。研究表明：ePCR 误差也是影响二代测序精度的一个重要因素，由于删除/插入错误、碱基替换错误的存在，当前聚类算法在操作分类单元(OTUs)计算时普遍存在过估计问题[15,16]。而基于 16S/18S rRNA 基因可变区域基因序列(如 V4)是当前研究微生物种群组成、结构的基础。针对上述问题，本文充分考虑乳液 PCR 及测序误差，对 454 测序仪模拟测序算法进行研究，仿真生成 16S/18S rRNA 基因特定区域序列片段。

2 模拟原理

454 测序仪在二代测序技术中占有重要地位，当前许多环境样本都是基于 454 测序仪完成测序的。454 测序仪主要包括模板制备分离、ePCR、合成测序等过程。针对 454 测序仪测序的上述各个步骤，本文 454 测序模拟测序算法流程图如图 1 所示。

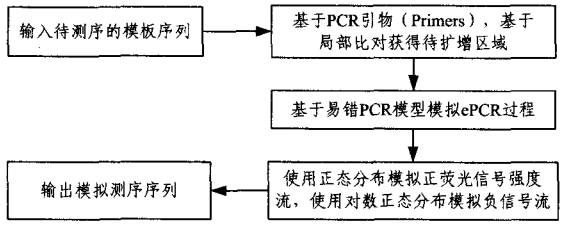

图 1 454 模拟测序算法流程图

2.1 PCR 误差模型

在 454 测序过程中，测序误差主要包括碱基删除/插入错误及替换错误。其中碱基替换错误主要是在 ePCR 阶段产生的。当前 454 测序仪模拟算法主要是针对全基因组序列。随机片段化全基因组序列，一般通过两个随机数(伪随机数)确定片段化的起始位置与克隆长度，即通过随机化过程确定待测序序列，因此其难以模拟针对特定区域的测序过程，同时该过程未考虑 PCR 误差。针对上述问题，本文提出了一种可以模拟测序 16S rRNA 特定区域的 454 测序仪模拟算法。首先通过局部搜索比较，获得全基因组序列中待扩增区域序列，然后基于易错 PCR 模型[17]模拟 ePCR 过程中的碱基替换错误。假定初始核苷酸替换矩阵为：

$$R=\begin{bmatrix} R_{AA} & R_{AC} & R_{AG} & R_{AT} \\ R_{CA} & R_{CC} & R_{CG} & R_{CT} \\ R_{GA} & R_{GC} & R_{GG} & R_{GT} \\ R_{TA} & R_{TC} & R_{TG} & R_{TT} \end{bmatrix}$$

其中，R_{ij} 表示碱基 i 被碱基 j 替换的概率，$R_{ij}\in\{A,C,G,T\}$，且 PCR 过程不受其它因素，诸如温度、碱基浓度差异等因素影响，则第 n 轮 PCR 的碱基突变概率矩阵可表示为 $M=[M_{ij}^n]_{4\times4}$，其中 M_{ij}^n 为：

$$M_{ij}^n=\begin{cases} R_{ij}, & n=1 \\ \sum\limits_{k=A,C,G,T} M_{ij}^{n-1}\cdot R_{kj}, & n\geqslant 2 \end{cases} \tag{1}$$

假定 PCR 初始模板序列数量为 S_0，扩增效率为 θ，则 N 轮 PCR 扩增之后总的序列数 S_N 及经过 l 次扩增延伸的序列数 $L_{N,l}$ 为：

$$S_N=S_0(1+\theta)^N \tag{2}$$

$$L_{N,l}=C_N^l S_0\theta^l \tag{3}$$

假定序列突变的位置是随机分布的，即 PCR 过程中碱基替换位点出现的位置不受任何环境因素约束限制，则 N 轮 PCR 扩增过程的平均累积突变概率矩 P_{ij}^N 为：

$$P_{ij}^N=\frac{\sum\limits_{n=1}^{N} M_{ij}^n L_{N,l}}{S_N} \tag{4}$$

其中，累积突变概率矩阵 P_{ij}^N 表示 N 轮 PCR 循环的总产物中碱基 i 被碱基 j 替换的概率，对每一个模板序列，基于 P_{ij}^N 产生的模拟序列等价从 N 轮 PCR 产物中随机选择一条序列。

假定每条满足引物条件的目标区域模板被选择并扩增的机率相同，即独立于浓度等因素（实际应用中，模板被选择扩增概率与浓度有关），根据上述定义，ePCR 过程可表述为：

1. 遍历 16S rRNA 基因混合文库，根据给定的引物(前向/后向引物，前向引物不允许错配，后向引物允许 1-2 个碱基位置错配)原则，基于 blast 局部搜索比对，获得满足特定引物条件的 16S rRNA 基因序列的目标扩增区域子序列，构成扩增模板序列集合$\{T_1,T_2,\cdots,T_m\}$；

2. 对第一步产生的扩增模板集合中的每一条扩增模板序列基于 PCR 误差模型模拟 ePCR 过程，产生扩增序列集合：

```
For i=1:m //m:总的扩增模板数
  基于式(2)获取 N 轮 PCR 之后总的产物数量 S_N
  For j=1:S_N
      基于模板序列 T_i 及累积突变概率(4)产生扩增序列。
      //即遍历模板序列每个位置，基于式(4)计算该位置碱基是否突变，然后产生模拟序列
  End For
End For
```

3. 直至所有的扩增模板序列模拟 ePCR 扩增完毕，否则，重复步骤 2。

2.2 454 边合成边测序模型

454 是当前一种重要的二代测序技术，其主要是通过检测荧光信号强度达到测定同聚物长度的目的。在测序过程中，4 种不同的 DNA 碱基$\{A,C,G,T\}$依次输入到 PTP 板上的磁珠，在 DNA 聚合酶等的协同作用下，将引物上每一个 dNTP 聚合作用与荧光信号释放偶联起来，通过检测释放的荧光信号强度达到实时测定 DNA 序列的目的。然而由于化学反应的特殊性，如化学反应速率问题等，以及一些技术原因，如 CCD 相机抖动及分辨率限制，通常检测的光信号存在抖动起伏，进而导致了测序误差。Margulies 等人研究发现焦磷酸测序过程中，正信号流服从正态分布 $N(\mu_1,\sigma_1)$[5]。本文中，为了更好地模拟边合成边测序过程，分别基于正态分布与对数正态分布模拟正信号流与负信号流。

454 测序中，连续相同的碱基长度即聚合体长度是通过检测光信号强度予以估计的。研究表明，同聚体越长，输出光

信号密度误差越大。也就是说，同聚体越长，测序序列出现插入/删除错误的概率越大。鉴于此，本算法拟基于高斯分布模拟合成测序中的荧光信号强度。即假定同聚体（Homopolymer）的长度为 l，则输出的荧光信号 s 的输出强度使用下列正态分布过程描述：

$$s \sim N(\mu,\sigma) \tag{5}$$

其中，均值 $\mu=l$，标准差 $\sigma=k\cdot\sqrt{r}$，k 为比例因子，可根据用户需要进行调整以模拟不同的正态分布。本文中，采用文献[5]中统计值 $k=0.15$ 作为默认设置。

在 454 测序过程中，进入 PPT 板输入碱基（$\{A,C,G,T\}$ 中的一种）与待测序序列上的碱基不能配对时，正常情况下输出的荧光信号强度为零，然而由于实验误差的存在，输出荧光信号通常不完全为零。非正常的碱基序列延伸释放的荧光信号会产生负信号流（Negative Flow）。对于负信号流，为了方便起见，将其视为长度为 1 的同聚体。本文中参考 Margulies 等人的研究，基于对数正态分布[5]模型模拟负信号流。假定负信号流为 X，则 X 服从分布：

$$In(X) \sim N(\mu,\sigma) \tag{6}$$

其中，默认参数设置为 $\mu=0.2$，$\sigma=0.15$。

根据上述定义，454 边合成边测序过程可描述为：

1. 扫描 ePCR 过程产生的模拟克隆序列池，对目标序列顺序进行边合成边测序模拟，确定当前位置的同聚体的长度 l；

2. 根据同聚体长度选择合适的信号流模拟模型 $s\sim N(l,\sigma)$ 或 $In(X)\sim N(\mu,\sigma)$ 模拟合成测序过程中的荧光信号释放强度；

3. 根据接收的即模拟产生的荧光信号强度估计同聚体长度及碱基类型；

4. 直至所有模拟克隆序列模拟合成测序完毕，否则转步骤 1。

3 结果分析

本文基于 Matlab 语言环境开发了特定目标区域的 454 模拟算法。由于当前缺少针对特定基因区域的 454 模拟测序算法，本文分别通过模拟效率及模拟测序误差率等指标来分析仿真算法的性能。首先采用算法默认参数，在 HP 笔记本电脑（主频：1.79GHz，CPU：2Duo T5670，RAM：2G）上分别模拟产生了不同数量的 16S rRNA 基因序列，所需要的时间统计如表 1 所列。从表中可以看出，454 仿真算法具有较好的性能，能在有限的时间内产生大量的 16S rRNA 序列，例如产生 1000000 序列所需要的时间<0.1h，平均每小时大约能模拟 1.14×10^7 个序列。从图 2 中可以看出，模拟算法具有线性复杂度。

表 1 模拟序列时间复杂度

序列数目	模板数量	模板长度	时间(s)	平均效率（read/h）
100000	10	80～160nt	31	1.16×10^7
200000			66	1.09×10^7
400000			128	1.13×10^7
600000			188	1.15×10^7
800000			250	1.15×10^7
1000000			317	1.14×10^7

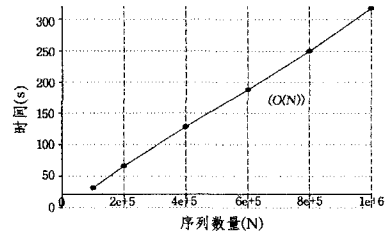

图 2 模拟算法时间复杂度

为了进一步说明模拟算法的有效性，基于默认参数，模拟扩增了 10 个物种的 V6 区域。首先从 RDP 数据库[18]中选取了 10 个已注释种（species）信息的全长 16S rRNA 序列（当前版本 RDP 10 包含 2765278 条 16S rRNA 基因序列，但只有不到 10%的序列注释到了种层次），然后基于 V6 区域引物提取了该 10 个物种的 V6 区域作为待扩增模板序列，其中 10 个初始物种的系统发生树如图 3 所示，模拟算法产生的测序集合序列误差率如表 2 所列。

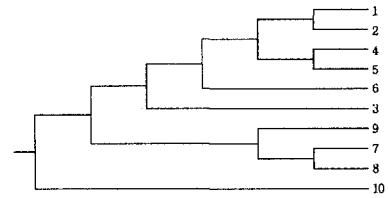

图 3 10 个待扩增序列的进化树

表 2 模拟序列误差率

待扩增物种数量	10（模板来自 16S rRNA 基因序列）
扩增区域	V6
模拟扩增产生的序列总数	100000
总测序误差率	1.02%
删除/插入误差率	1%
突变错误率	0.02%

注：对每一个测序序列，将其与相应模板进行序列比对，然后基于比对序列计算相应的误差率。

从表 2 中可以看出，模拟序列的总的误差率为 1.02%，其中删除/插入错误率为 1%，突变错误率为 0.02%。实际应用中，454 测序仪的误差率为 1%～2%，其中主要为删除/插入错误，该实验结果与实际应用中 454 测序仪测序误差率极为相似。可见该模拟算法具有良好的模拟性能，能较好地模拟 454 测序过程。同时，该算法参数易于调整，适当改变突变概率矩阵及高斯分布中的参数，即可改变突变误差率与删除/插入误差率。

表 3 与其他算法比较

算法	能否扩增特定基因区域	误差类型	算法效率（read/second）
MetaSim	否	删除/插入	30000/s
ART	否	删除/插入	1500/s
Tsim	是	删除/插入、ePCR 误差	2800/s

与现有的一些基于全基因组的 454 模拟算法相比，本文算法 Tsim 效率有所降低，如模拟 100000 条长度为～230nt 的序列，本文算法所需要的时间为 5 分钟，而传统的算法如 MetaSim 等所需时间<1 分钟，但 Tsim 算法效率要高于 ART 算法。其中 Tsim 所增加的复杂度主要是由模拟 ePCR 过程产生的。MetaSim、ART 等传统的 454 模拟算法采用随机片段化全基因组序列模拟 ePCR 过程，未考虑 ePCR 误差

（下转第 284 页）

[J]. 计算机仿真,2007,24(7):229-234
[6] Lowe D G. Distinctive image features from scale-invariant keypoints[J]. International Journal of Computer Vision, 2004, 60(2),91-110
[7] Mikolajczyk K, Schmid C. Scale & Affine Invariant Interest Point Detectors[J]. International Journal of Computer Vision, 2004,60(1):63-68
[8] 唐朝伟,肖健,邵艳清,等. 一种改进的 SIFT 描述子及其性能分析[J]. 武汉大学学报:信息科学版,2012,37(1):11-16
[9] 范志强,赵沁平. 一种基于数据聚类的鲁棒 SIFT 特征匹配方法[J]. 计算机研究与发展,2012,49(5):1123-1129
[10] 程邦胜,唐孝威. Harris 尺度不变性关键点检测子的研究[J]. 浙江大学学报:工学版,2009,43(5):855-859
[11] 黄帅,吴克伟,苏菱. 基于 Harris 尺度不变特征的图像匹配方法[J]. 合肥工业大学学报:自然科学版,2011,34(3):379-382
[12] 钟金琴,檀结庆,等. 基于二阶矩的 SIFT 特征匹配算法[J]. 计算机应用,2011,31(1):29-32
[13] 张海燕,李元媛,储晨昀. 基于图像分块的多尺度 Harris 角点检测方法[J]. 计算机应用,2011,31(2):356-357
[14] 王鹏,王平,沈振康,等. 一种基于 SIFT 的仿射不变特征提取方法[J]. 信号处理,2011,27(1):88-93
[15] 于丽莉,戴青. 一种改进的 SIFT 特征匹配算法[J]. 计算机工程,2011,37(2):210-212
[16] 程德志,李言俊,余瑞星. 基于改进 SIFT 算法的图像匹配方法[J]. 计算机仿真,2011,28(7):285-289
[17] 张海燕,李元媛,储晨昀. 基于图像分块的多尺度 Harris 角点检测方法[J]. 计算机应用,2011,31(2):356-357
[18] Manjunath B S, Shekhar C, Chellappa R. A new approach to image feature detection with application[J]. Pattern Recognition, 1996,29(4):627-640
[19] Yasein M, Agathoklis P. A feature-based image registration technique for images of different scale[C]// IEEE International Symposium on Circuits and Systems. May 2008:3558-3561
[20] Daubechies I. Ten Lectures on Wavelets[M]. Fourth Printing, Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania, 1992
[21] Mikolajczyk K. Detection of local features invariant to affine transformations[D]. Institut National Polytechnique de Grenoble, France, 2002
[22] Qian Wei, Fu Zhi-zhong, et al. Voting-strategy-based approach to image registration[J]. Opto-Electronic Engineering, 2008, 35(10):86-91

(上接第 263 页)

及特定测序目标区域不同,本文算法不仅能扩增特定的基因目标区域(如 16S/18S rRNA),而且能有效模拟 ePCR 误差。

结束语 本文在现有的一些研究基础之上,提出了一种新的 454 模拟算法,它基于易错 PCR 模型模拟乳液 PCR 过程,分别采用正态分布模型及对数正态分布模拟边合成边测序过程中的正信号流与负信号流过程。实验结果表明,该算法具有较好的性能,能在有限的时间内模拟大量的测序序列,所需要的时间与模拟序列数量成正比。与此同时,该算法具有易推广至其它二代测序平台的优点,通过引物局部序列比对获得待扩增区域,ePCR 扩增过程结合现有的一些算法,如 ART 即可完成对基因特定目标区域的跨平台模拟测序。由上述分析可见,该算法不仅能完成特定目标基因区域模拟测序,帮助我们定量、定性分析微生物种群,还有助于验证评价当前 454 测序数据分析软件,促进二代测序数据分析。

参 考 文 献

[1] Grice E A, Kong H H, Conlan S, et al. Topographical and temporal diversity of the human skin microbiome [J]. Science, 2009,324(5931):1190-1192
[2] 蒋德明,孙玉华,李丹,等. 基于 16S rRNA 基因序列分析受砷和硫酸盐污染的土壤细菌多样性 [J]. 微生物学通报,2011,38(10):1592-1601
[3] Oakley B B, Fiedler T L, Marrazzo J M, et al. Diversity of human vaginal bacterial communities and associations with clinically defined bacterial vaginosis [J]. Appl Environ Microbial, 2008, 74(15):4898-4909
[4] Kellenberger E. Exploring the unknown: The silent revolution of microbiology [J]. Embo Rep, 2001, 2(1):5-7
[5] Margulies M, Egholm M, William E, et al. Genome sequencing in microfabricated high-density picolitre reactors [J]. Nature, 2005,437(7057):376-380
[6] 刘玮琦,茆振川,杨宇红,等. 应用 16S rRNA 基因文库技术分析土壤细菌群落的多样性[J]. 微生物学报,2008,35(10):1344-1350
[7] Ley R E, Backhed F, Turnbaugh P, et al. Obesity alters gut microbial ecology [J]. PNAS, 2005, 102: 11070-11075
[8] 曹波,杨红,许强华,等. 基于 16S rRNA 技术的长江口微生物分子生物学鉴定与分析 [J]. 上海大学学报,2011(2):191-197
[9] Benoit M R, Li W, Stodieck LS, et al. Microbial antibiotic production aboard the International Space Station [J]. Appl Microbial Biotechnology, 2006, 70:403-411
[10] Richter D C, Ott F, Auch A F, et al. MetaSim—A Sequencing Simulator for Genomics and Metagenomics [J]. PLoS ONE, 2008, 3(10): e3373
[11] Balzer S, Malde K, Lanzén A, et al. Characteristics of 454 pyrosequencing data — enabling realistic simulation with flowsim [J]. Bioinformatics, 2010, 26(18): i420-i425
[12] Lysholm F, Andersson B, Persson B. An efficient simulator of 454 data using configurable statistical models [J]. BMC Research Notes, 2011, 4:449
[13] Huang Wei-chun, Li Le-ping, Myers J R, et al. ART: a next-generation sequencing read simulator [J]. Bioinformatics, 2012, 28(4):593-594
[14] 宣黎明,韦朝春,李亦学. 基于 GPU 运算的宏基因组第二代测序模拟软件 [J]. 华东理工大学学报,2012(4):472-476
[15] Huse S M, Welch D M, Morrison H G, et al. Ironing out the wrinkles in the rare biosphere through improved OTU clustering [J]. Environ Microbial, 2010, 12(7): 1889-1898
[16] Sharpton mail T J, Samantha, et al. PhylOTU: A High-Throughput Procedure Quantifies Microbial Community Diversity and Resolves Novel Taxa from Metagenomic Data [J]. PLoS Comput Biol, 2011, 7(1): 1-13
[17] Pritchard L, Corne D, Kell D, et al. A general model of error-prone PCR [J]. J Theor Biol, 2005, 234(4): 497-509
[18] Wang Q, Garrity G M, Tiedje J M, et al. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy [J]. Appl Environ Microbiology, 2007, 73(16): 5261-5267