基于仿射聚类的宏基因组序列物种聚类①

聂鹏宇[1,2], 潘玮华[1,2], 徐 云[1,2]

[1](安徽省高性能计算重点实验室, 安徽合肥 230027)
[2](中国科学技术大学 计算机科学与技术学院, 安徽 合肥 230027)

摘 要: 随着下一代测序技术的迅猛发展, 宏基因组学已经成为新的研究热点, 宏基因组学序列聚类问题使用无参考的方法, 对包含多个物种的宏基因组序列进行有效分离. 为此, 提出一种结合相似度信息和结构信息的宏基因组物种聚类算法, 并引入仿射聚类来进行序列物种聚类. 实验数据表明该方法聚类精度高、执行速度快. 我们也开发了基于该方法的宏基因组序列物种聚类软件.

关键词: 宏基因组学; DNA 序列; 物种聚类; 仿射聚类; 倒排索引

Metagenomic DNA Sequence Binning based on Affinity Propagation

NIE Peng-Yu[1,2], PAN Wei-Hua[1,2], XU Yun[1,2]

[1](Key Laboratory of High Performance Computing of Anhui Province, Anhui University, Hefei Anhui 230027, China)
[2](School of Computer Science and Technology, University of Science and Technology of China, Hefei Anhui 230027, China)

Abstract: Nowadays, with the rapid development of the next generation sequencing technologies, metagenomics have become a new hotspot,However research in metagenomics faces the issue of binning --- identification and taxonomic characterization of the NGS short reads. To solve this problem, this paper first analyzes the next generation sequencing technology characteristics, statistical characteristics of metagenomic sequence, then proposes a new clustering method for DNA sequence binning. Test results show that this method has a very good clustering accuracy. In the same time, we developed an software for metagenomic binning based on this algorithm MetaBinning.

Key words: metagenomics; DNA sequence; binning; affinity propagation; inverted index

近些年来, 下一代DNA测序技术的迅速发展使得对微生物进行进行更加深入的研究成为了现实, 也极大推动了宏基因组项目的发展[1].

自 1991 年 Pace 首次提出环境基因组学的概念并在同年构建了第一个通过克隆环境样品中 DNA 的噬菌体文库以来, 有关环境基因组学(也称微生物环境基因组学 Microbial Environmental Genomics, 宏基因组学、元基因组学 Metagenomics, 生态基因组学 Ecogenomics)的研究受到广泛关注. 宏基因组学(也叫随着环境基因组学,应用下一代测序技术来混合基因组样本, 对从一个环境样品获得或一系列相关的样本进行测序,并生产高通量的宏基因组 DNA 片段[2]. 对这些未培养微生物多样性及群落功能的研究将大大扩展科学家们对生命的了解, 包括了解生命如何耐受极端环境、新的生物能源、生命进化以及微生物与环境之间的相互作用. 此外, 自然界现存的物种数量以万亿计, 包括人类、大量的高等动物、植物和微生物. 这一研究超越了传统意义上单一物种的基因组学分析, 研究对象复杂、范围更广而方法要更新, 它要求在基因学、基因组学的基础上有新的学科来承担这一新的任务. 宏基因组学为最大限度地挖掘微生物资源带来了前所未有的机遇, 已成为国际生命科学技术研究和开发最重要的热点[3].

与传统的单一的基因组测序研究不同, 宏基因组序列研究同时研究一个环境中的所有物种的 DNA 片段,宏基因组测序采集的到数据集的到的 DNA 序列中包含不同物种的 DNA 片段. 对这些 DNA 序列进行数

① 基金项目:国家自然科学基金面上项目(60970085)
收稿时间:2013-04-22;收到修改稿时间:2013-05-13

据分析需要一个额外的分析步骤,称为宏基因组序列物种分析. 宏基因组物种聚类指使用无参考的方法将包含多个物种的DNA片段进行分离, 并将每个物种的DNA序列划分成一个类. 同时由于下一代测序技术产生的基因序列长度比较短, 仅为几十或几百个碱基, 这也给宏基因组序列物种聚类问题带来了新的挑战. 同时文献也指出多达 99%的 DNA 序列是目前未知的, 如何对他们的 DNA 序列进行有效聚类和分析也是目前亟待解决的问题.

为了研究宏基因组序列中物种聚类问题, 作者统计分析了宏基因组序列的规律, 将仿射聚类传播算法应用到宏基因组序列物种聚类问题中, 并通过试验结果和对比分析验证了算法的有效性. 本文将分成以下几个部分进行阐述, 第 2 节介绍了宏基因组序列中物种聚类问题模型并回顾了一些近期研究, 第 3 节对宏基因组序列进行了统计分析的到统计规律, 并给出基于仿射聚类算法的宏基因组序列物种聚类算法, 第 4 节使用了宏基因组测序序列对算法进行了测试, 并对试验结果进行简要分析. 第 5 节总结全文工作并给出未来的研究方向.

1 问题定义及研究回顾

在本小节中, 作者首先给出宏基因组序列物种聚类问题的定义, 并对这个问题的研究文章和方向进行了简短的回顾.

问题定义:

使用无参考的方法将包含多个物种的宏基因组序列进行聚类, 使得每个类仅包含一个物种的DNA序列, 并且同一物种的所有序列均出现在同一个类中. 例如对包含 k 个物种的序列 $\{S_{i1}S_{i2}...S_{ii}S_{i2}S_{i3}...S_{ir}S_{j1}S_{j2}...S_{js}.....S_{k1}S_{k2}...S_{kt}\}$, 其中任意序列 S_i 均是由 $ATCG$ 位点组成的序列. 这些序列包含 k 个物种, 其中 $S_{i1}S_{i2}S_{i3}...S_{ir}$ 属于物种 i, 而 $S_{j1}S_{j2}...S_{js}$ 属于物种 j. 宏基因组物种聚类问题就是设计使得每个物种的 DNA 序列聚集为一个类, 如上例中分离为类 $\{S_{i1}S_{i2}S_{i3}...S_{ir}\}$, 类 $\{S_{j1}S_{j2}...S_{js}\}$ …. 和类 $\{S_{k1}S_{k2}...S_{ks}\}$.

宏基因组序列物种聚类问题---针对序列的聚类是宏基因组研究中最迫切需要解决的问题, 而且这一直是研究热点, 也已经有很多学者针对宏基因组序列物种聚类问题设计了不少新的算法和工具. 针对宏基因组序列物种分析问题的研究根据基于进行聚类的依据可以分为基于相似度的和基于结构信息的, 根据研究方法可以分为有参考的和无参考的两类. 有参考的分类方法采用将序列和已知的物种的序列进行比对, 它常常可以获得更好的结果[4-6]. 但是因为很多的序列仍然是未知的(多达 99%), 有参考的分类方法无法处理这些未知的序列, 因此也已经有很多学者针对无参考的序列物种聚类问题提出了很多聚类方法[7]. 已有的工作中 Chatterji 首先使用 k-means 算法基于结构特性进行聚类, 并开发了宏基因组序列物种聚类工具 MetaCluster[8,9]. Shendure 基于结构特性, 对编码后连续 4 个碱基首先进行 PCA 分析, 然后再使用选取的特征值进行聚类, 并开发了宏基因组序列物种分类工具 CompostBin[10]. Olga 提出了基于相似度进行图聚类的宏基因组序列聚类算法, 并开发了基于图聚类的宏基因组序列物种聚类工具 TOSS[11]. 本文针对无参考的宏基因组序列物种聚类问题, 提出了结合相似度特性和结构特性的宏基因组物种聚类方法. 算法首先根据序列的公共子串信息进行初步聚类, 然后将仿射聚类算法应用到宏基因组序列序列物种聚类, 并开发了基于仿射聚类的宏基因组序列物种聚类工具 MetaBinning.

2 宏基因组序列分析及聚类算法设计

在本小节中, 作者首先针对宏基因组序列进行了统计分析. 首先, 针对宏基因组测序数据进行了分析, 然后给出了基于仿射聚类算法的宏基因组序列聚类方法.

2.1 宏基因组序列分析

(1) 宏基因组序列特性一:

相似度特性: 下一代测序技术针对同一个物种产生的不同测序序列间有一定的重复, 且测序得到的序列为长度短(约为 75～100)[12].

由于需要对测序后得到的序列进行拼接及测序技术存在错误率问题, 一般宏基因组序列测序都会对序列片段进行多次测序. 文献指出测序产生的同一物种的 DNA 序列间有一定的重复, 并且测序深度(对序列的平均测序重复度)越深, 序列间包含的重复子串数目越多.

(2) 宏基因组序列特性二:

两个物种间的DNA序列包含长度为 l 的公共子串的数目随着 l 的数值增大而迅速变小.

首先选取 50 个不同的物种距离进行了统计, 并给出了距离分别为种, 属, 类, 纲(Species, Genus, Family,

Order)时的微生物序列的所有 DNA 序列统计分析结果. 将连续的 m 位碱基进行编码, 统计物种间距离分别为种, 属, 类, 纲(Species, Genus, Family, Order), 两个物种的 DNA 序列中包含长度为 l 位碱基的相同的公共子串的数目占串的总长度的比例. 在文章随后介绍中, 选取了 l=20 并且, 提到的序列子串长度均为 l.

如图 1 所示, 可以发现随着 l 的数值增大, 不同物种间 DNA 序列包含的公共的长度为 l 的公共子串的概率逐渐降低, 而且物种间物种距离越远, 不同物种间包含长度为 l 的公共子串的比例越低. 同时文献中也指出下一代测序技术中, 不同物种测序序列间包含的公共子串的长度是有限的, 因此可以在实验中利用两个串包含的相同的 l 位子串的数目来决定两个串是否为同一个物种的序列.

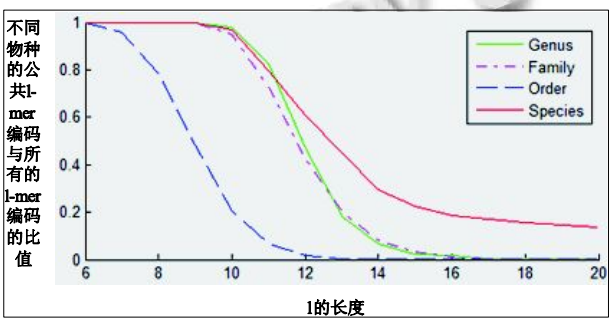

图 1 不同物种间的包含的连续个碱基的公共子串的占串长度比率

(3) 宏基因组序列特性三:

结构特性: 将宏基因组序列连续 4 个位点进行编码后的到的 2^4=256 维向量. 同一物种的序列编码后的到的 256 维向量类似, 而不同物种编码后的到的 256 维向量却很相差很大[12].

分别选取了长度为 5000 的不同物种的 DNA 片段和相同物种的不同位置的 DNA 片段进行统计, 并将它们连续的 4 个 DNA 位点编码. 将序列中碱基 A 编码为 00, 碱基 T 编码为 01, 碱基 C 编码为 10, 碱基 G 编码为 11, 这样对 DNA 序列连续的 4 位碱基编码后就可以得到 256 维的向量, 然后对这 256 维的向量进行归一化. 通过分析的到的编码, 可以发现不同物种的 DNA 序列的到的向量有着很大的不同, 而同一物种不同片段的 DNA 序列编码的到的向量却基本相似. 图 2 为两个不同物种序列进行编码后的到的分布, 图 3 为同一物种间不同片段的 DNA 序列进行编码后的到的序列分布. 基于此可以考虑使用 DNA 序列连续 4 个碱基位点编码后的到的 256 维向量对 DNA 序列进行聚类.

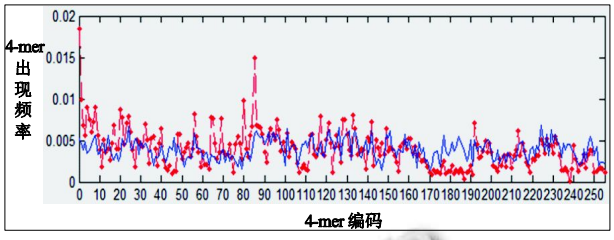

图 2 不同物种的 DNA 片段连续 4 个碱基位点编码分布

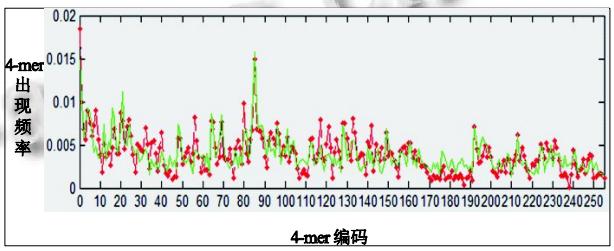

图 3 相同物种不同 DNA 片段的连续 4 个碱基位点的编码分布图

根据特性一及特性三, 在测序的 DNA 序列较短的时候可以首先利用特性一, 根据序列间的公共子串信息, 将那些相似度较高的串归为小类. 然后再利用特性三, 采用仿射聚类的算法对宏基因组序列进行聚类.

2.2 基于仿射聚类算法的宏基因组物种聚类研究

仿射传播聚类是由 Frey 等人在 Science 杂志上提出的一种新的聚类算法. 其优势是算法快速、有效. 它与 K 均值算法都是属于 K 中心聚类方法. 然而 K 均值算法需要用户指定聚类个数以及初始聚类中心, 且对初始聚类中心的选择敏感, 不同的初始聚类中心会导致不同的聚类结果[13]. 仿射传播聚类算法则克服了这些缺点, 其迭代过程不断搜索合适的聚类中心, 同时也使得聚类的适应度函数最大化, 能给出比较准确的聚类结果, 且运算速度快. 它根据 N 个数据点之间的相似度进行聚类, 这些相似度可以是对称的, 即两个数据点互相之间的相似度一样(如欧氏距离); 也可以是不对称的, 即两个数据点互相之间的相似度不等. 这些相似度组成 $N \times N$ 的相似度矩阵 S(其中 N 为有 N 个数据点).

仿射聚类算法以样本之间的相似性作为输入, 输出为簇类中心以及各样本与簇类中心的所属关. 算法引入了三个矩阵---相似度矩阵 $\mathbf{S}[N][N]$, 吸引度矩阵 $\mathbf{R}[N][N]$ 和归属度矩阵 $\mathbf{A}[N][N]$.

其中相似度矩阵 **S**[*N*][*N*] 中数据 *s*(*i*, *k*), 代表第 *i* 项数据和第 *j* 项数据的相似度. 吸引度矩阵 **R**[*N*][*N*] 中数据 *r*(*i*, *k*), 表示从点 *i* 发送到候选聚类中心 *k* 的数值消息, 反映 *k* 点是否适合作为 *i* 点的聚类中心. 归属度矩阵 **A**[*N*][*N*] 中数据 *a*(*i*, *k*)则从候选聚类中心 *k* 发送到 *i* 的数值消息, 反映 *i* 点是否选择 *k* 作为其聚类中心. 算法的执行过程中他们初始化如下:

$$s(i,j)=smiliarity(i,j) \tag{1}$$

$$r(i,k)=s(i,k)-\max_{k'\neq k}\{s(i,k')\} \tag{2}$$

$$a(i,k)=0 \tag{3}$$

迭代关系如下:

$$r(k,k)=s(k,k)-\max\{s(k,k')\} \tag{4}$$

$$a(i,k)=\min\left\{0,r(k,k)+\sum_{i'\notin\{i,k\}}\max\{0,r(i',k)\}\right\} \tag{5}$$

其中 *S* 矩阵的对角线上的数值 *s*(*k*,*k*)作为 *k* 点能否成为聚类中心的评判标准,这意味着该值越大, 这个点成为聚类中心的可能性也就越大, 这个值又称作参考度 *p*(preference). 聚类的数量受到参考度 *p* 的影响, 如果认为每个数据点都有可能作为聚类中心, 那么 *p* 就应取相同的值. 如果取输入的相似度的均值作为 *p* 的值, 得到聚类数量是中等的, 如果取最小值, 得到类数较少的聚类. 由于仿射聚类算法速度快, 精度高已经获得了广泛的应用[14,15]. 如下图所示, 算法主要分为两部分.

在第一步中, 首先将测序的到的DNA序列的连续 *l* 位序列的子串进行编码, 构建倒排索引. 此时能够高效的计算任意两个测序序列之间的包含的公共子串信息, 计算所有序列对之间的包含的公共子串信息仅为. 然后通过倒排索引, 基于DNA序列的公共子串的关系对 DNA 序列进行初步聚类, 并得到很多小类. 建立的倒排索引如图 4 所示, 对图中第两行进行解释: 即为编号为 1, 3, 10 和 11 的 DNA 串包含子串 AA.....AAAA , 串 1, DNA 串 57 包含子串 AA.....AAAT .

同时为了加速查找过程, 将这些子串使用整数进行编码, 将 A 编码为 00, T 编码为 01, G 编码为 10, C 编码为 11 此时将 AA.....AAAA 编码为 0, AA.....AAAT 编码为 1, AA...AAAG 编码为 2, AA....AAAC 编码为 3, 而 CC...CCCC 编码为 2^l-1. 注意只需要存储 DNA 序列编码中出现的编码数值, 而非所有的 2^l 个编码. 这样

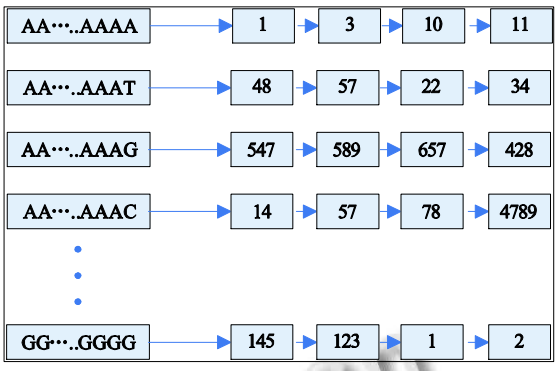

图 4　对子串进行编码, 并建立倒排索引

就可以将DNA子串编码得到一个整数, 并使用红黑树存储, 可以大大加速子串信息的定位和查找到包含该子串的 DNA 序列的信息. 假定对所有的 DNA 串进行编码后的到的所有互异的子串的数目为 *M*, 目标 DNA 串 *s* 长度为 *L*, 包含任意一个子串的 DNA 串的数目为 *K*, 可以在 $O(LK\log M)$ 的时间内统计出 DNA 串 *s* 与其他的所有 DNA 串包含的子串信息. 因为只需要查找到该串包含的编码数目, 随后再查找那些DNA序列包含这些编码, 既可以得到其他所有的NA串与串 *s* 的相似度信息. 示例红黑树如图 5 所示:

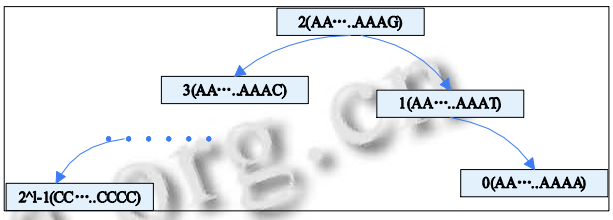

图 5　对子串的到的编码建立红黑树

在第二步中, 将第一步中得到的小类进行进一步处理, 首先剔除重复的连续位编码信息. 随后将的到的子类包含的连续 4 个碱基位点进行编码统计, 将每一个小类进行统计后得到 256 维的向量. 随后选取的到的向量间的距离来度量 DNA 序列间的距离并得到相似度矩阵, 最后使用仿射聚类传播算法对DNA序列进行聚类. 算法的概要执行流程如图 6.

3　实验及数据分析

3.1　实验平台

实验采用的是英特尔 4 核处理器, 它装配有英特尔 TM 4 核 3.1GHZ 处理器, 内存为 4G. 针对使用

MetaSim 模拟下一代测序技术产生的测序序列对算法进行了验证[16].

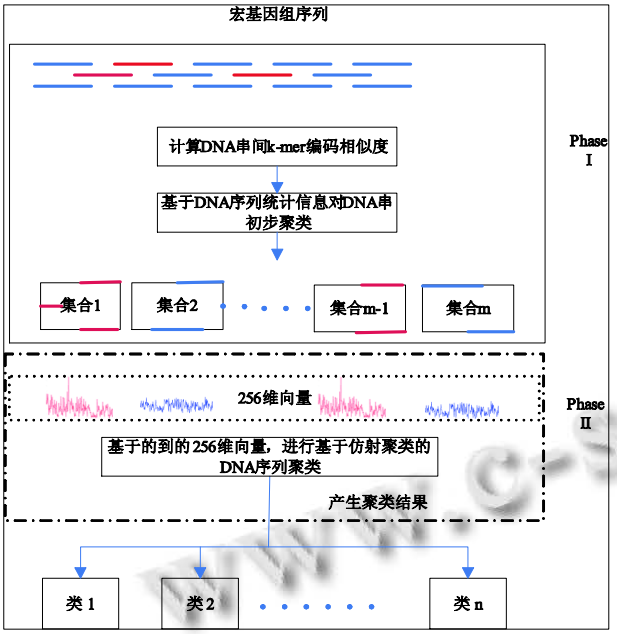

图 6 算法流程

MetaSim 是一个用来产生第二代测序数据集的软件，它可以根据已知生物序列，特定的测序方法，测序长度，错误率，来模拟产生对应的测序结果. 该文中用到的生物序列均下载自 NCBI 数据库，各个物种间的距离下载自 NCBI 物种距离数据库[17].

3.2 实验结果及数据分析

在使用 Metasim 仿真软件模拟中下一代测序技术时，设定相关参数为测序序列平均长度约为 75，序列的错误率为 1%，测序深度为 20. 而对测序物种选取了宏基因组中不同物种距离---种，属，类，纲(Species, Genus, Family, Order)的物种的 DNA 序列并进行聚类，并将使用基于仿射聚类算法的宏基因组序列物种聚类工具 MetaBinning 与基于图聚类的宏基因组序列物种聚类工具 TOSS、基于 k-means 的 MetaCluster 3.0 进行了对比.

由于算法是无监督的，在聚类数目与物种数目相同时，算法的错误率即为将 DNA 序列错误的聚类的比例占到所有序列的总数的比例. 而对于聚类数目与物种数目与物种数目不同时，也进行了分析. 当聚类数目大于物种数目时，算法误将一些物种的序列分为了多个类；而在聚类数目小于物种数目时，算法误将多个物种的 DNA 序列分到了一个类中. 由于聚类数的预知本身对聚类问题的研究就是十分困难的，在测试过程中，分别选取了简单的测试集和复杂的测试集. 统计聚类准确率时，如果一个类中包含的序列数目为 n 个，在这 n 条序列中来自物种 i 的序列最多，共有 m 个，我们就将该类定义为物种 i 的类，并且记 n-m 为聚类错误的序列的数目. 同理我们聚类的序列数目数目为 N，在聚类为 k 个类后，每个类中包含的聚类错误的序列的数目为 e_i，那么聚类准确率记为 $1-\sum_{i=1}^{k}(e_i)/N$. 测试结果如表 1 所示:

表 1 算法在不同物种距离下对模拟数据聚类精确度

ID	物种	物种距离	MetaBinning聚类准确率	TOSS聚类准确率	MetaCluster 3.0聚类准确率
1	Bacillus halodurans & Bacillus ubtilis	Specie	98.1%	98.7%	56.4%
2	Gluconobacter oxydans & Granulibacter ethesdensis	Genus	99.0%	96.5%	63.2%
3	Methanocaldococcus annaschii &Methanococcus maripaludis	Family	97%	99%	63.5%
4	Gluconobacter oxydans & Rhodospirillum rubrum	Order	99%	99%	78.7%
	Escherichia coli& Pseudomonas putida&Bacillus anthracis	Family & Order	94.3%	90%	71.6%

首先选取了宏基因组序列物种比较少的情况，仅包含 2～3 个物种，实验数据发现基于仿射聚类的算法可以有效的预测出物种的数目，并且可以得到较好的结果. 以上的测试数据表明，MetaCluster3.0 对于序列聚类时直接使用结构信息，仅适用于序列较长的情况，在序列较短的时候聚类效果较差. MetaBinning 对宏基因组序列物种聚类问题取得了很好的效果，与已有的 TOSS 基于图聚类的结果的聚类准确率类似. 并且随着物种距离正大，算法对序列物种聚类问题的计算结

果精确度越高, 这与宏基因组序列统计分析的到的微生物 DNA 序列间相似度随着物种距离增加的特性也相符.

随后选取了多个物种的情况(物种数目大于 4), 此时 MetaBinning 和 TOSS 均无法取得理想的聚类结果. 如上图中的测试数目 2 和 3 中的 DNA 序列使用 MetaBinning 和 TOSS 进行分析发现, MetaBinning 和 TOSS 均无法准确预测类的个数. 此时确定算法错误率对算法来说是不公平的, 因为决定聚类的个数, 对聚类问题来说本身就是一个十分复杂的问题, 而宏基因组物种分析对聚类数目精确性要求更高. 将会在下一步工作中, 将会进一步研究如何首先通过分析宏基因组测序序列特性来分析得到序列中包含的物种的个数来改进算法的设计.

同时选取了针对序列数目分别为 100000, 200000 及 400000 时, 对本文提出的聚类算法和已有的 TOSS 运行时间进行比对. 通过表 2 的时间对比可以发现 MetaBinning 的执行速度已有的基于图聚类的宏基因组序列物种聚类工具 TOSS 要快很多倍. 经过分析发现, TOSS 是基于图聚类算法的, 它首先计算任意两个串之间的相似度信息, 随后对的到的整个相似度矩阵进行聚类, 这样由于原始的聚类元素为所有 DNA 串的数目, 这使得聚类速度较慢. 而 MetaBinning 算法, 首先使用相似度信息将 DNA 串进行初步聚类, 得到了少量的小类. 随后使用仿射聚类算法对这些小类再进行聚类, 此时聚类的元素的数目远少于 TOSS, 故运行速度是 TOSS 的十倍以上. 从以上分析可以看出 MetaBinning 对宏基因组序列进行有效聚类, 而且算法快速稳健.

表 2 基于仿射聚类算法的宏基因组序列物种聚类和已有的基于图聚类算法的执行时间对比

序列数目	序列大小	Meta Binning 运行时间(s)	TOSS 运行时间(s)
100,000	15M	35	780
200,000	30M	65	1750
400,000	60M	152	3288

4 实验和讨论

随着下一代测序技术的不断普及和宏基因组技术的迅猛发展, 宏基因组序列分析已经成为了新的研究热点. 而宏基因组序列分析中将不同物种的DNA序列进行聚类的问题, 已经成为宏基因组学中亟待解决的问题. 为了提升宏基因组序列中物种聚类问题的准确率和性能, 本文首先从宏基因组序列和下一代测序技术的特性入手, 给出了序列的统计特性, 测序的 DNA 串间的公共子串, 然后再使引入仿射聚类算法到宏基因组序列物种聚类问题中. 选取了宏基因组测序序列模拟数据测试, 并和已有的宏基因组序列物种聚类工具进行对比, 验证了算法的效率及准确性.

由于算法在物种较多的情况下聚类效果有待改进, 因此在下一步中将会根据宏基因组序列信息来设计预测宏基因组序列中包含物种的数目, 改进宏基因组序列物种聚类的准确性. 同时也会考虑针对宏基因组研究问题中序列纠错及拼接问题进行进一步研究.

参考文献

1 Mardis ER. Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum, 2008, 9: 387–402.

2 Wendl M, Waterston R. Generalized gap model for bacterial artificial chromosome clone fingerprint mapping and shotgun sequencing. Genome Res, 2002, 12(1): 1943–1949.

3 Gill SR, Pop M, DeBoy RT. Metagenomic analysis of the human distal gut microbiome. Science, June 2006, 312: 1355–1359.

4 McHardy A, Marln H, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length dna fragments. Nauture Methods, 2007, 4(1): 63–72.

5 Sandberg R, Winberg G, Branden CI. Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. Genome Research, 2001, 11(8): 1404-1409.

6 Diaz N, Krause L, Goesmann A. TACOA-Taxonomic classil-cation of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics, 2009, 10(1): 56.

7 Wu YW, Ye YZ. A novel abundance-based algori- thm for binning metagenomic sequences using l-tuples. Proc. of the 14th annual international conference (RECOMB'10). Springer. 2010. 535–549.

8 Yang B. MetaCluster: unsupervised binning of environmental genomic fragments and taxonomic annotation. Proc. of the ACM Conference on Bioinformatics, Computational

(下转第 142 页)

法对运行时间和空间的要求较小, 完全满足用户对场景漫游实时性交互的要求. 将来的主要工作有: 将场景 LOD 判断及裂缝的消除等操作转移到 GPU 端进行; 研究动态场景的绘制技术, 用以表现动态变化的场景.

参考文献

1 Jin HL, Lu XP, Liu HJ. View dependent fast real-time generating algorithm for largescale terrain. Procedia Earth and Planetary Science, 2009, (1): 1147−1151.

2 Zhong DH, Liu J, Li MC. NURBS recon- struction of digital terrain for hydropowerengineering based on TIN model. Progress inNatural Science, 2008, (18): 1409−1415.

3 Kennelly PJ. Terrain maps displaying hill-shading with curvature. Geomorphology, 2008, 102(3): 567−577.

4 孔川,罗大庸刑用动态多分辨率 LOD 技术的地形简化研究. 计算机工程与应用,2010,46(27):32−41.

5 杨硕磊,郝爱民,王莉莉.运用矩阵结构的可并行地形层次细节算法.计算机辅助设计与图形学学报,2011,23(2):276−283.

6 Francisco RM, Zhang Q, Reid JF. Stereovision three-dimensional terrain maps for precision agriculture. Computers and Electronics in Agriculture, 2008, 60(2): 133−143.

7 张淑军,陈芳,周忠.基于四叉树剖分的 L0D 地形绘制算法. 系统仿真学报,2008,20(z1):25−32.

8 罗景馨,唐珊.基于改进四叉树分割和结点存储的 LOD 算法.计算机工程,2009,35(20):202−204.

9 李白云,赵春霞.GPU 实时构建四叉树的快速地形渲染算法. 计算机辅助设计与图形学学报,2010,22(12):2259−2264.

10 Chen H, Long AQ, Peng YH. Building panoramas from photographs taken with an uncalibrated hand-held camera. Chinese Journal of Computers, 2009, 32(2): 328−335.

11 郎兵.复杂场景中海量外存地形模型的实时绘制算法.系统仿真学报,2009,21(20):6510−6514.

12 李成名,王继周,马照事.数字城市三维地理空间框架原理与方法.北京:科学出版社,2008.

13 张和平.RFID 在供应链物流管理中的应用.中国市场, 2007,(41):104−105.

14 Genske DD, Heinrich K. A knowledge—based fuzzy expert system to analyze degraded terrain. Expert Systems with Applications, 2009, 36(2): 2459−2472.

(上接第 170 页)

Biology and Biomedicine(ACM-BCB). 2010. 170−179.

9 Leung HCM, Yiu SM, Yang B. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics, 2011, 27: 1489−1495.

10 Chatterji S, Yamazaki I, Bai Z. Compostbin: a DNA composition-based algorithm for binning environmental shotgun reads. Proc. of the 12th annual international conference (RECOMB'08). Springer. 2008. 17−28.

11 Tanaseichuk O, Borneman J, Jiang T. Separating metagenometic short reads into genomes via clustering. Proc. of the 11th Algorithms in Bioinformatics.(WABI). 2011. 298−313.

12 Zhou F. Barcodes for genomes and applications. BMC Bioinformatics, 2008, 9(1): 546.

13 Frey BJ, Dueck D. Clustering by passing messages between data points. Science, February 2007, 315: 972−976.

14 王开军,张军英,李丹.自适应仿射传播聚类.自动化学报, 2007,12(33):1242−1245.

15 许文竹,徐立鸿.基于仿射传播聚类的自适应关键帧提取. 计算机科学,2010,12(1):268−270.

16 Richter DC. MetaSim:a sequencing simulator for genomics and metagenomics, PLoS ONE,3,pp.e3373,2008.

17 ftp://ftp.ncbi.nih.gov/ ftp://www.ncbi.nlm.nih.gov/.