计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(9):163−169 [doi: 10.15888/j.cnki.csa.006545]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 16S rRNA 序列物种鉴定的改进向量空间模型算法①

祝　斌[1,2], 亓合媛[3], 马俊才[1,3]

[1](中国科学院 计算机网络信息中心, 北京 100190)
[2](中国科学院大学, 北京 100049)
[3](中国科学院 微生物研究所, 北京 100101)
通讯作者: 祝　斌, E-mail: zhubin@im.ac.cn

摘　要: 在物种鉴定领域中, 权威方法是基于 BLAST 的序列比对算法, 然而该算法出现计算量过于庞大, 运算效率低以及资源消耗较高等问题. 为解决以上问题, 本文借鉴经典文献中的 K-String 组份向量方法, 对向量空间模型作出改进, 将其应用于基于 16S rRNA 序列的物种鉴定领域, 并在巴拿赫空间的理论体系下, 对改进向量空间模型算法中的遗传距离公式进行等价替换, 给出不同范数背景下对应的遗传距离公式, 供科研人员参考. 本文从计算效率和物种鉴定效果两个方面来判断改进算法的性能, 最终得到如下结论: 欧几里得空间下的内积范数从计算效率上较经典的 blast 算法具有显著优势, 而其分类效果在检出率这一方面, 达到了比对结果的一致性.

关键词: 16S rRNA 基因序列; 改进向量空间模型算法; 非序列对比; 物种鉴定; 分类

引用格式: 祝斌,亓合媛,马俊才.基于 16S rRNA 序列物种鉴定的改进向量空间模型算法.计算机系统应用,2018,27(9):163–169. http://www.c-s-a.org.cn/1003-3254/6545.html

Improved VSM Algorithm in Species Identification Based on 16S rRNA Gene Sequences

ZHU Bin[1,2], QI He-Yuan[3], MA Jun-Cai[1,3]

[1](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)
[2](University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China)

Abstract: In the field of species identification, the traditional algorithm is based on the BLAST method, which is regarded as the authoritative method, but the method has a series of problems such as complex calculating process, time-consuming, as well as space-consuming. In this study, we propose an improved VSM algorithm based on K-String compositional vector method, and give the alternative norm-format formula in calculating the genetic distance between species in the Banach space for the reference of other scientific researchers. In this study, the computational efficiency and the result of the species identification are the two aspects to determine the properties of the improved method. The conclusion is that the calculating time of improved VSM algorithm based on 2-norm has decreased obviously than that of the BLAST algorithm, in addition, the result of classification demonstrates good consistence and convergence with the comparison result in terms of detection rate.

Key words: 16S rRNA gene sequence; improved VSM algorithm; non-sequence alignment; species identification; species classification

在过去的几十年中, 随着生物学数据的大量累积, 以及计算机技术、数学和生物学交叉学科的崛起, 21 世纪已经进入了云计算和大数据时代. 云计算时代也为基因序列比对能够在较短时间完成提供了坚实的

① 基金项目: 国家高技术研究发展计划 (863 计划)(2014AA021501)
Foundation item: National High-Tech R & D Program of China (863 Program)(2014AA021501)
收稿时间: 2018-02-01; 修改时间: 2018-02-28; 采用时间: 2018-03-08; csa 在线出版时间: 2018-08-16

基础. 物种鉴定是用来描述物种间近缘关系和进化层次的非常有用的一种工具. 最初, 物种鉴定常常基于单个基因序列或是很少的几个基因序列进行对比, 这种方法虽然简单易行, 但是由于横向基因转移 (Horizontal Gene Transfer, HGT)、并系同源基因 (Paralog) 以及物种进化差异等因素的出现, 这种方法受到了质疑.

基因测序方法在一定程度上解决了单基因序列比对出现的问题, 保证了系统发育树的合理性. 但是, 与此同时, 随着序列数据的增加, 计算时间呈指数式增长 (如图 1 所示), 因此, 计算效率成为了亟待解决的问题.

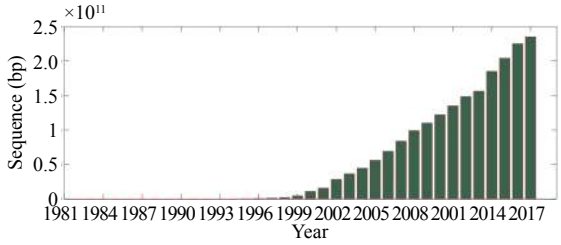

图 1 Genbank 数据库序列数目随年份走势图

近期, 关于物种鉴定的方法逐步地出现了全基因组序列对比. 在此期间, 众多学者提出: 比较两个完整的基因组意义并不大, 原因在于: 每个物种都有自己特定的基因含量和基因顺序, 此外基因组的数量是不同的; 另外, 微生物的基因数据也是需要人工处理操作的, 不同的实验室处理数据的不同会造成结果在一定程度上的差异, 进而使得结果不具备更完整的说服力, 非序列比对方法便应运而生.

非序列比对方法在计算效率上明显优于前者, 操作也较简单, 运算效率较高. 近几年, 关于非序列对比的方法也不断更新. 目前比较常见的方法有: K 串组份向量方法, (0, l) 序列法, DNA Walk, 压缩矩阵法, 表示法, CGR 方法, Nandy 二维图形[1]等. 近年的新兴算法在物种鉴定方面上的应用上不够广泛, 在所有的非序列比对算法中, 使用最为广泛且传统的算法为基于 TF-IDF 检索技术的向量空间模型 (Vector Space Model, VSM) 算法, 然而其物种鉴定的分类效果得不到保证[2], 原因在于该算法没有借鉴到微生物的背景, 因此无法消除在基因突变和物种进化的背景下, 基因序列的噪音影响. 因此, 在确定非序列对比算法具备了提高运算效率的优点, 以及向量空间模型算法在众多经典和最新的文献[3]中使用较为广泛的特点之后, 为此, 本文以如何改进向量空间模型算法, 进一步达到提高运算效率和保证分类效果质量两方面为主要目的.

在众多生物系统发育相关性水平指标中, 16S rRNA 基因序列具有如下特征:

1) 普遍存在于一切细胞内;

2) 机体生理功能稳定且重要;

3) 在微生物中含量高, 且容易提取;

4) 编码基因比较稳定;

5) 序列相对保守;

6) 相对分子量适中;

7) 基因序列长度适中;

8) 既含有高度保守的序列区域, 又含有高度变化的序列区域.

基于以上的各个特点, 16S rRNA 基因序列具备最佳的鉴定特征, 是本文改进向量空间模型算法的应用数据, 可以为物种鉴定打下坚实的基础.

综上, 本文以 16S rRNA 基因序列为应用对象, 使用改进向量空间模型算法为核心, 以达到快速分类和保证分类质量的研究目的.

1 背景及相关工作

分子生物系统发展史的出现以及基因测序方面的进步, 大大加深了人们对物种进化的理解. 因此, 物种分类和鉴定在分子水平上的进步已经为微生物的分类提供了一个具有实用价值的工具.

目前分子系统发展史有两大重要研究成果: 一是线粒体和叶绿体之间具有内生共体特性, 二是目前为止, 生物可划分为古生菌, 细菌, 和真核生物三个生物领域. 然而, 随着完整的微生物基因组数据的逐步添加, 实验结果逐渐地对公众预期提出了质疑[4], 在这一争议过程中, 仍有几个实验试图从完整的基因组中推断出原核生物发展史. 以上实验使用的方法包括利用基因含量[5], 直系同源基因簇的存在/缺失值比例[6], 父系树[7], 保存基因对[8]等方法. 然而这些方法最终都依赖于序列比对这一传统思路, 到目前为止, 还没有一种能够被广泛接受且用于从完整基因组数据中推断出系统发育树的方法.

此后, 逐渐出现了非序列比对的方法[9], 计算效率和结果都得到了广泛的认可, 因此成为了除 BLAST 算法以外物种分类与鉴定方面不可或缺的方法. 而向量空间模型 (VSM) 算法在众多前沿文献中使用的频率

较高, 由此可见, 目前向量空间模型算法是非序列比对算法中构建系统发育树的主流算法. 因此, 对其算法的改进具有重大意义.

根据相关文献[10]的说明, 截至目前, 使用 16S rRNA 基因序列对物种进行鉴定和分类的项目有: 美国的 Greengenes, RDP 核糖体数据库, 以及韩国的 EzTaxon. 以上项目的核心基础仍是利用 BLAST 局部比对算法进行快速分类, 输出初始排名结果, 随后使用双序列全局比对, 给出在参考样本数据库中与待测序列最为接近的排名序列, 以此作为参考, 对样本序列进行鉴定和分类.

根据前面的分析, 我们发现, 用于物种鉴定的主流算法仍是基于 BLAST 的序列比对算法, 然而由于该算法出现计算量过于庞大, 运算效率低以及资源消耗较高等问题, 使用 VSM 方法能够有效地解决上述问题.

VSM 算法的运算效率相比于 BLAST 算法更优, 此特点解决了 BLAST 算法的核心问题, 但该算法的不足之处在于其分类效果远远没有主流 BLAST 鉴定算法更为优越. 因此, 对 VSM 算法的改进就具有了现实意义, 而改进的 VSM 算法可以作为物种鉴定的另一种有效工具方便科研人员参考和使用.

此外, 经典文献[11] 提到的 K-String 组份向量算法在病毒[12], 原核生物[13–17], 真菌[18], 叶绿体序列[19]以及人体的肠道元基因组[20]有了成功的应用.

综上所述, 本文旨在对常用的 VSM 算法进行改进, 将该改进 VSM 算法应用于基于 16S rRNA 序列的物种鉴定领域, 达到运算效率和分类质量两方面的提高效果. 本文后续的内容逻辑为: 在第 2 节介绍两种 VSM 模型算法以及两种算法的区别, 一种是基于 TF-IDF 检索技术的 VSM 模型算法, 另一种是借鉴经典文献[1]后的改进 VSM 模型算法. 此外, 本文还给出了改进 VSM 模型算法中遗传距离在巴拿赫空间下的等价替代公式, 并给出了相关说明; 同时, 第 2 节给出本文为测试改进 VSM 算法运算效率和分类排名质量两方面效果所使用的数据集来源, 以及对应的运算时间和排名效果结果汇总及相应分析; 第 3 节是对接下来研究工作的讨论与未来展望.

2 VSM 算法与改进 VSM 算法

2.1 VSM 算法原理[21,22]介绍

本文将以 16S rRNA 基因序列分析为研究背景, 介绍 VSM 在该背景下的操作流程.

一个物种 16S rRNA 基因序列文本, 其碱基只有 AGCT 四种, 将碱基序列划分为不同的 K 子串, 那么此排列方式就有 4^K 种可能, 通过计算词频和逆文档频率, 最终得到该 16S rRNA 序列文本对应的权重向量, 维数为 1×4^K.

假设, $d_1=\{t_1,t_2,\cdots,t_n\}$, $d_2=\{t'_1,t'_2,\cdots,t'_n\}$分别代表两条 16S 序列对应的权重向量, $\cos(d_1, d_2)$ 表示该两条序列的相似度量值, 如图 2 所示.

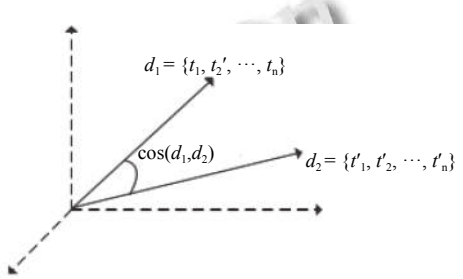

图 2 序列相似度

图 2 中, 每一项的权重都由词频和逆文档频率综合表示.

假设有 N 条样本序列, 记为$D=\{d_1,d_2,\cdots,d_N\}$, 词频 f_{ij} 表示 K 串词项 w_i 在序列 d_j 中出现的次数, n_i 为含文本 w_i 的数量, 逆文本频率计算公式为:

$$f_i^{-1}=\log\frac{N}{n_i} \tag{1}$$

其中, 序列 d_j 中词项 w_i 的 TF-IDF 权重公式为:

$$t_{ij}=f_{ij}\times\log\frac{N}{n_i} \tag{2}$$

最后, 样本序列相似度量值计算公式为:

$$\cos(d_i,d_j)=\frac{\sum_{k=1}^{4^K}t_{ik}\times t_{jk}}{\sqrt{\sum_{k=1}^{4^K}t^2{}_{ik}\times\sum_{k=1}^{4^K}t^2{}_{jk}}} \tag{3}$$

2.2 改进 VSM 算法原理介绍

该算法涉及 6 个步骤, 将分别作出说明.

第一步. 计算 K 串词项出现的频率.

长度为 L 的 16S rRNA 序列$\alpha_1\alpha_2\cdots\alpha_L$, 选取长度为 $K(K<L)$ 的子串在该序列上沿着一个方向逐个碱基进行移动, 定义这个含有 K 个核苷酸的短串序列为 $\alpha_1\alpha_2\cdots\alpha_K$, 则该串在整个 DNA 序列中出现的频率就等于频数 (记为$f(\alpha_1\alpha_2\cdots\alpha_k)$) 比上总数 ($L$–$K$+1), 公式如下:

$$p(\alpha_1\alpha_2\cdots\alpha_k)=\frac{f(\alpha_1\alpha_2\cdots\alpha_k)}{L-K+1} \tag{4}$$

第二步. 计算随机突变背景下的噪音频率.

随机突变在分子水平上或多或少以随机的方式发生, 而基因重组的选择决定了进化的方向. 以上因素导致了一些 K 串词项产成了一定的随机性. 为了还原该 K 串词项的原始频率, 本文需要对此噪音进行刻画, 根据最大熵原理的推导过程, 我们得到了噪音频率公式:

$$p^0(\alpha_1\alpha_2\cdots\alpha_k)=\frac{p(\alpha_1\alpha_2\cdots\alpha_{k-1})p(\alpha_2\alpha_3\cdots\alpha_k)}{p(\alpha_2\alpha_3\cdots\alpha_{k-1})} \tag{5}$$

第三步. 计算修正后 K 串词项频率.

综合前两步, 本文给出了修正后频率计算公式:

$$a(a_1.a_2,\cdots,a_k)=\begin{cases}\dfrac{p(\alpha_1\alpha_2\cdots\alpha_k)-p^0(\alpha_1\alpha_2\cdots\alpha_k)}{p^0(\alpha_1\alpha_2\cdots\alpha_k)} & p^0(\alpha_1\alpha_2\cdots\alpha_k)\neq 0\\ 0 & p^0(\alpha_1\alpha_2\cdots\alpha_k)=0\end{cases} \tag{6}$$

第四步. 计算每一个 16S rRNA 序列修正后的特征向量.

将每一个可能的序列子串$\alpha_1\alpha_2\cdots\alpha_K$的频率作为一个物种的特征向量的元素. 为了进一步简化这一个定义, 我们定义a_i为所有排列好的 K 子串中第 i 种子串类型对应特征向量中的第 i 个分量. 这里 i 从 1 到 $4K$ 循环. 因此, 我们可以得出对于 16S rRNA 序列 A 的特征向量:

$$A=(a_1,a_2,\cdots,a_{4K}) \tag{7}$$

以此类推, 对于物种 B, 我们仍有特征向量:

$$B=(b_1,b_2,\cdots,b_{4K}) \tag{8}$$

第五步. 计算各序列间的遗传距离.

这里同样以序列 A, B 为例, 两序列间的遗传距离使用传统的夹角余弦进行表示, 公式如下:

$$C(A,B)=\frac{\sum_{i=1}^{4^K}(a_i\times b_i)}{\left(\sum_{i=1}^{4^K}a_i^2\times\sum_{i=1}^{4^K}b_i^2\right)^{\frac{1}{2}}} \tag{9}$$

由公式 (9) 知, 夹角余弦数值的取值范围为[–1, 1]. 若将夹角余弦记做物种间的遗传距离, 则有: 两物种特征向量对应的遗传距离越大, 说明两个物种之间的相关性越强; 反之, 遗传距离越小, 说明两物种之间的相关性越弱. 为了符合直观, 表达相关性强, 对应遗传距离小; 相关性弱, 则遗传距离大的说法, 本文对此距离公式进行标准化修正, 公式为:

$$D(A,B)=\frac{1-C(A,B)}{2} \tag{10}$$

第六步. 计算待测样本与参考序列库之间的遗传距离, 从小到大进行排序, 输出前十名相关性最强的序列及其遗传信息, 以辅助科研人员参考和进行物种分类和鉴定工作.

2.3 巴拿赫空间下等价替换的遗传距离公式

遗传距离的定义是计算分子生物学中一个重要环节. 该距离的定义需要满足以下三个条件 (记 D(x, y) 为两个物种间的遗传距离):

非负性: D(x, y)≥0, D(x, y)=0 等价于 x=y;

对称性: D(y, x)=D(x, y);

三角形不等式: 任意三个物种 z, x, y, 距离恒满足: D(z, y)+D(y, x)≥D(x, y).

显然, 在本文的第 2.2 节中的公式 (10) 符合遗传距离的定义. 这里值得一提的是, 夹角余弦公式使用的是内积空间下的 2-范数. 根据向量范数的等价性定理:

设$||x||_s$, $||x||_t$ 为 R^n 上向量的任意两种范数, 则存在常数 c_1, c_2>0,使得对一切 $x\in R^n$, 有 $c_1||x||_s\leq||x||_t\leq c_2||x||_s$.

以及极化恒等式: 实线性空间上的内积和范数有以下关系:

$$<x,y>=\frac{1}{4}(||x+y||^2-||x-y||^2) \tag{11}$$

综合上述内容, 本文将公式 (8) 中内积范数进行重新定义, 给出在 1-范数和无穷范数下的计算公式, 公式如下:

$$C(A,B)=\frac{\frac{1}{4}(||A+B||_1{}^2-||A-B||_1{}^2)}{||A||_1||B||_1} \tag{12}$$

$$C(A,B)=\frac{\frac{1}{4}(||A+B||_\infty{}^2-||A-B||_\infty{}^2)}{||A||_\infty||B||_\infty} \tag{13}$$

2.4 改进 VSM 模型算法[23,24]运算效率和排名结果分析

2.4.1 待测样本与测试数据集

本文所使用的 16S rRNA 样本序列数据, 来源于 (863 计划, 课题编号: 2014AA021501) 中通过质检工具 pipeline 筛选整理出的高质量 16S rRNA 基因序列参考数据库.

这里, 本文从参考数据库中随机选取了 8000 条样本序列. 其中, 将前 6000 条为参考序列样本库, 剩余的

2000 条作为待测样本进行测试.

为了简化名称, 这里依次定义序列编号为 G_1, G_2,…, G_{6000}, G_{6001},…, G_{8000}. 其中, G_1, G_2,…, G_{6000} 为参考数据库, G_{6001},…, G_{8000} 为待测样本.

2.4.2 改进 VSM 算法运算效率与 blast 运算效率结果

结合第 2.2 节所述, 可以发现, 本文所选取的 6000 条参考样本序列文本可以通过改进向量空间模型算法进行计算, 得出对应的 6000 个特征向量, 是本实验的预处理阶段. 因此, 以上 6000 个特征向量的运算时间完全不需要计入该算法的运算时间, 这也是该算法提高运算效率的一大优势.

这里, 本文首先按照第 2.2 节所述的操作步骤逐一进行: (这里以 K=4 为例)

第一步. 对前 6000 条 16S rRNA 参考样本序列 G_1, G_2,…, G_{6000}, 逐一带入公式 (6), 计算出每一个序列文本对应的修正频率特征向量 A_1,A_2,…, A_{6000}.

说明. 此阶段为数据预处理阶段, 不占用算法计算时间; 其中每一个特征向量的维数为: 1×4^4 即 1×256.

第二步. i=6001, 对待测样本 G_i 计算出对应的修正频率特征向量 A_i.

第三步. 计算 G_i 与 G_1, G_2,…, G_{6000} 序列之间的遗传距离 $C(G_i, G_1)$,$C(G_i, G_2)$,…, $C(G_i, G_{6000})$, 依次记为遗传距离 d_1,d_2,…, d_{6000}.

第四步. 对上述 6000 个遗传距离 d_1,d_2,…, d_{6000} 按照递增的顺序进行排序, 输出相关性较高的前十条序列排名结果作为物种鉴定的参考初步排名结果.

第五步: i=i+1, 直至计算至最后一个待测样本 G_{8000}.

说明 1. 其中第 2, 3, 4 步为一个待测样本与 6000 条参考样本序列 G_1, G_2,…, G_{6000} 的整个运算过程, 其花费时间也为该改进向量空间模型算法的单个样本进行排名输出的运算时间.

说明 2. 本文将 i 从 6001 依次逐个循环至 8000 进行操作运算, 消耗的总时间除以 2000, 记为改进 VSM 模型算法的测试时间.

紧接着, 本文使用 BLAST 本地构建 6000 条参考数据库, 使用 blastn 程序对 2000 条待测样本进行逐一运算, 输出结果, 其花费的时间除以 2000, 同样记为 blast 算法测试时间.

说明 1. 这里使用 blastn 命令:

blastn -query 6001.fa -db Sequence6000 -evalue 1e-5 -out blast6001.xls -outfmt 6 -num_alignments 10 -num_threads 1

说明 2. 以上参数中, -query6001.fa 为待测样本 G_{6001} 的 fa 格式文件, -db Sequence6000 表示 6000 条参考样本的本地化数据库, -evalue 1e–5 表示控制误差, -outfmt 6 表示输出文件排版格式按照格式 6 进行输出, -num_alignments 10 表示输出排名前 10 的序列结果, -num_threads 1 表示单线程.

说明 3. 本文使用改进 VSM 算法, 使用的是 c 程序, 改进 VSM 算法和 blastn 算法均在 Ubuntu 12.04.4 LTS 同一个操作环境下运行.

最后, 综合上述两项内容的操作, 本文给出了改进 VSM 模型算法和 BLAST 算法运行时间.

表 1 改进 VSM 算法与 BLAST 算法运行效率 (单位: ms)

K	4	5	6	7	8
改进 VSM	0.035	0.117	0.447	1.183	4.535
BLAST	237	237	237	237	237

备注：数值表示: 平均一条待测样本与 6000 条序列遗传距离计算过程所消耗的时间

2.4.3 改进 VSM 算法运算效率与 BLAST 算法排名结果

本文选择输出前 10 名用于比较两种算法的鉴定效果, 原因在于: 本文使用的参考数据集序列数量为随机抽样后的 6000 条序列, 序列数量相对较小; 且在物种鉴定领域中, 一般输出 BLAST 相似度 98% 以上的排名结果, 这里使用 BLAST 输出序列相似度 98% 以上的序列数均小于 10 条,因此选择前 10 名作为评价的参考标准.

按照第 2.2 节的操作进行, 本文得出了对应的排名结果, 这里以待测样本 G_{6001} 的排名结果为例, 如表 2 所示.

表 2 改进 VSM 算法与 BLAST 算法排名结果

排名	K=4	K=5	K=6	K=7	K=8	BLAST
1	G_{5971}	G_{5997}	G_{5997}	G_{5997}	G_{5997}	G_{5997}
2	G_{4653}	G_{4101}	G_{4101}	G_{4101}	G_{4101}	G_{5994}
3	G_{5994}	G_{4100}	G_{4100}	G_{4100}	G_{5994}	G_{6000}
4	G_{4101}	G_{4556}	G_{5994}	G_{3791}	G_{4100}	G_{3792}
5	G_{4100}	G_{3783}	G_{4354}	G_{3792}	G_{6000}	G_{3791}
6	G_{4154}	G_{5995}	G_{4102}	G_{6000}	G_{4102}	G_{4102}
7	G_{4158}	G_{5999}	G_{6000}	G_{5995}	G_{5999}	G_{4101}
8	G_{767}	G_{5994}	G_{5995}	G_{4154}	G_{5995}	G_{5999}
9	G_{766}	G_{3784}	G_{4154}	G_{4158}	G_{3482}	G_{5995}
10	G_{768}	G_{1611}	G_{4158}	G_{5999}	G_{4556}	G_{4100}

本文 K=8 时, 将改进 VSM 算法输出排名与 blast

排名结果重复率进行统计, 最终得出: 所有的 2000 个待测 16S rRNA 样本序列, 通过使用改进 VSM 算法输出的前十名排名结果, 其检出率已达到 98.0%.

此外, 若将输出前 10 名序列信息, 改为输出前 50, 或前 100 名, 我们发现检出率和 K 相关, 随着 K 越大, 算法的检出率相对会越优; 且当 K=10, 输出前 100 名序列信息时, 检出率达到了 97.6%, 证明了该算法的收敛性.

2.4.4 改进 VSM 算法与 blast 算法对比综合分析

根据表 2, 我们可以看到: 随着 K 由 4 逐步递增至 K=8, 其输出的排名结果检出率由 30% 上升至 90%; 此外, G_{6000} 的排名也逐渐靠前, 以及排名第一的 G_{5997} 和 BLAST 的第 1 名结果吻合.

根据表 1, 我们可以看出, 随着 K 的增大, 运算时间也成约 4 倍递增, 然而当 K=8 时, BLAST 运算时间约为改进 VSM 模型算法的 50 倍.

综合以上运算效率和排名结果两方面的分析, 我们可以得出改进 VSM 算法维持了其计算效率的优越性, 并改进了排名结果, 提高了检出率.

3 展望

本文提出的改进 VSM 算法, 是将经典文献中的 K 串组份向量空间模型算法应用于微生物 16S rRNA 序列的物种鉴定中, 并对遗传距离公式进行改进, 以期克服传统 VSM 模型算法在物种鉴定方面上的不足, 进一步提高物种鉴定的检出率, 最终保证物种鉴定的质量效果.

后续的研究工作还包括: 改进 VSM 模型算法多线程模板设置, 进一步提升该算法的运算效率. 初步设置思路为: 将参考数据集划分成多个模块, 然后将待测样本分别与各个模块进行比对, 输出各自的遗传距离向量, 接着将各个向量汇集成一个完整的向量, 最终对该向量进行排序输出最终结果.

参考文献

1 冯思玲. 系统发育树构建方法研究. 信息技术, 2009, (6): 38–40. [doi: 10.3969/j.issn.1671-3176.2009.06.018]

2 张会敏, 冯友军. 一株野生细菌的 16Sr DNA 序列分析与系统发育树的构建. 生物信息学, 2005, 3(1): 1–4. [doi: 10.3969/j.issn.1672-5565.2005.01.001]

3 王章群, 解增言, 蔡应繁, 等. 系统发育基因组学研究进展. 遗传, 2014, 36(7): 669–678.

4 李强, 左光宏, 郝柏林. 从完全基因组出发建立原核生物亲缘关系和分类系统时遇到的数学问题. 中国科学: 物理学 力学天文学, 2014, (12): 7.

5 何亮, 谢小军, 王冲, 等. 16S rRNA 序列同源性分析结合系统发育树构建鉴定 6 株生殖道乳杆菌. 中华全科医学, 2013, 11(4): 617–618.

6 王颜颜, 夏茂宁, 欧维正, 等. 16S rRNA 和 secA1 基因构建临床诺卡菌的系统发育树比较. 贵阳医学院学报, 2017, 42(4): 409–415.

7 Liu J, Wang H, Yang H, *et al*. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. Nucleic Acids Research, 2013, 41(1): e3. [doi: 10.1093/nar/gks828]

8 Chu KH, Li CP, Qi J. Ribosomal RNA as molecular barcodes: A simple correlation analysis without sequence alignment. Bioinformatics, 2006, 22(14): 1690–1701. [doi: 10.1093/bioinformatics/btl146]

9 Chu KH, Xu M, Li CP. Rapid DNA barcoding analysis of large datasets using the composition vector method. BMC Bioinformatics, 2009, 10(S14): S8.

10 任清福, 孙清岚, 马俊才. 基于 16S rRNA 基因序列分析的物种辅助分类研究与实现. 科研信息化技术与应用, 2015, 6(5): 48–57.

11 Qi J, Wang B, Hao BI. Whole proteome prokaryote phylogeny without sequence alignment: A K-string composition approach. Journal of Molecular Evolution, 2004, 58(1): 1–11. [doi: 10.1007/s00239-003-2493-7]

12 Sinclair L, Osman OA, Bertilsson S, *et al*. Microbial community composition and diversity via 16S rRNA gene amplicons: Evaluating the illumina platform. Plos One, 2015, 10(2): e0116955. [doi: 10.1371/journal.pone.0116955]

13 Sun Y, Cai Y, Huse SM, *et al*. A large-scale benchmark study of existing algorithms for taxonomy-independent microbial community analysis. Briefings in Bioinformatics, 2012, 13(1): 107–121. [doi: 10.1093/bib/bbr009]

14 Thomas T, Gilbert J, Meyer F. Metagenomics–a guide from sampling to data analysis. Microbial Informatics and Experimentation, 2012, 2(1): 3. [doi: 10.1186/2042-5783-2-3]

15 Hao BL, Gao L. Prokaryotic branch of the tree of life: A composition vector approach. International Journal of Systematic and Evolutionary Microbiology, 2008, 46: 258–262.

16 Cui H, Zhang X. Alignment-free supervised classification of metagenomes by recursive SVM. BMC Genomics, 2013,

14(1): 1–12. [doi: 10.1186/1471-2164-14-1]

17 Daniel MD, Price MN, Julia G, *et al*. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME Journal, 2012, 6(3): 610–618. [doi: 10.1038/ismej.2011.139]

18 Wang H, Xu Z, Gao L, *et al*. A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evolutionary Biology, 2009, 9: 1471–2148.

19 Chu KH, Qi J, Yu ZG, *et al*. Origin and phylogeny of chloroplasts revealed by a simple correlation analysis of complete genomes. Molecular Biology and Evolution, 2004, 21: 200–206.

20 Liu J, Wang H, Yang H, *et al*. Composition-based classification of short metagenomic sequences elucidates the landscapes of taxonomic and functional enrichment of microorganisms. Nucleic Acids Research, 2013, 41: 1–10. [doi: 10.1093/nar/gks1039]

21 徐浩广, 王宁, 刘佳明, 等. 基于自然语言检索的综合相似度计算算法. 计算机系统应用, 2017, 26(6): 170–175. [doi: 10.15888/j.cnki.csa.005815]

22 Carrera-Trejo JV, Sidorov G, Miranda-Jiménez S, *et al*. Latent dirichlet allocation complement in the vector space model for multi-label text classification. International Journal of Combinatorial Optimization Problems and Informatics, 2015, 6(1): 7–19.

23 Daniel MD, Price MN, Julia G, *et al*. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. ISME Journal, 2012, 6(3): 610–618. [doi: 10.1038/ismej.2011.139]

24 Grossi De SMF, Guimaraes LM, Batista JAN, *et al*. Compositions and methods for modifying gene expression using the promoter of ubiquitin conjugating protein coding gene of soybean plants. US, US9012720. 2015.