Advances in Applied Mathematics 应用数学进展, 2021, 10(6), 1937-1944
Published Online June 2021 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2021.106204

Hans汉斯

基于前缀位置及允许错配的DNA序列进化分析

高胜男

辽宁师范大学，辽宁 大连
Email: 820272439@qq.com

收稿日期：2021年5月10日；录用日期：2021年5月29日；发布日期：2021年6月16日

摘 要

对DNA序列相似性的比对分析是生物信息学中的重要问题。由于多序列比对(MSA)方法耗时较长，因此非比对方法的应用变得流行起来。DNA序列中的基因突变对序列比对分析的影响是不可忽视的，突变的存在使本应该被匹配的位置丢失。本文在构建环形前缀树以及获得前缀集的基础上，考虑完全匹配与错配两个匹配法则，提取序列中最佳的匹配位置，创建了新的关于位置差的非比对方法，对多序列进行成对比对，并运用邻接(Neighbor-Joining)法构建进化树，从而得到有效的进化关系。

关键词

非比对方法，前缀集，错配，进化树

DNA Sequence Evolution Analysis Based on Prefix Position and Allowable Mismatch

Shengnan Gao

Liaoning Normal University, Dalian Liaoning
Email: 820272439@qq.com

Received: May 10th, 2021; accepted: May 29th, 2021; published: Jun. 16th, 2021

Abstract

The comparison and analysis of DNA sequence similarity is an important issue in bioinformatics. Because the multiple sequence alignment (MSA) method takes a long time, the application of non-alignment methods has become popular. The influence of genetic mutations in DNA sequences on sequence comparison analysis cannot be ignored. The existence of mutations makes the positions that should have been matched lose. Based on the construction of the ring prefix tree and the acquisition of the prefix set, this paper considers the two matching rules of perfect match and mismatch, extracts the best matching position in the sequence, and creates a new non-alignment me-

文章引用: 高胜男. 基于前缀位置及允许错配的 DNA 序列进化分析[J]. 应用数学进展, 2021, 10(6): 1937-1944.
DOI: 10.12677/aam.2021.106204

thod for position difference, which is used for multiple sequences. Perform pairwise comparisons and use the Neighbor-Joining method to construct evolutionary trees to obtain effective evolutionary relationships.

Keywords

Alignment-Free Method, Prefix Set, Mismatch, Phylogenetic Tree

1. 引言

生物信息学作为一门前沿交叉学科，它不仅推动了分子生物学的发展，而且它使得数学与生物学在这个框架中实现了新的结合。随着生物信息学的崛起，遗传序列的系统发育分析变得重要起来，它成为了研究所有类型生物之间进化的必要条件，与此同时了解群体之间的自然关系也非常重要，例如目、科以及属等[1]。系统发育分析的方法通常依赖于多序列比对，由于生物序列的数量急剧增加，传统的序列比对方法变得不可行，大量基于遗传序列的数字表征的非比对方法被提出，用来补偿传统的基于比对方法的低效性[2]。

目前在生物序列中的比对中，主要的研究对象是 DNA 序列与蛋白质序列。在基因组中，遗传信息存储在 DNA 序列中,同时在 DNA 序列中，突变在分子水平以或多或少随机的方式发生，从而选择决定了进化的方向[3]。突变的方式可能是点突变、插入、缺失或者重排，有些非比对方法对于数据缺失会更敏感，可能会造成错误的匹配，因此突变的存在不允许被忽视，本文将突变考虑到 DNA 序列的两两比对中。

无论是对 DNA 序列还是蛋白质序列研究，都提出过用 K-词的数量与分布来度量相似性，但对于如何选择使用 K-词的长度没有一个标准，在文献[4]中闻佳等人提出复合 K-词联合来唯一地量化每个蛋白质序列，同理这种想法在 DNA 序列研究中也适用，也就是利用前缀集来唯一刻画每个 DNA 序列。

1973 年 Peter Weiner 提出了前缀树模型及其算法[5]。每一个前缀都是唯一出现在序列中的 K-词，以单拷贝序列的形式存在于序列中，是整个基因组中独有的，代表了生物序列中独特的特征，因此具有很大的研究价值。有文献利用信息熵，基于共同前缀的位置差提出新的 DNA 序列相似性度量方法[6]，也有文献利用欧氏距离，针对前缀集提出基于共同前缀的位置差法(CPPA 算法) [7]，本文在 CPPA 算法基础上加入错配法则，也就是考虑突变存在，创建了新的模型。通过对 70 条哺乳动物线粒体 DNA 序列进行实验，在完全匹配与错配法则的双重查找后得到成对序列最佳匹配的所有位置，再进行位置差计算，最后构建系统发育树，得到优于上述两个文献的结果。

2. 材料

本文应用到的实验数据集来自 NCBI (National Center for Biotechnology Information)的 Genebank 数据库，数据集与文献[6]中所用相同，均为 70 条哺乳动物的线粒体完整基因，经过处理以后的 70 条 DNA 序列中，不含有非 A、C、G、T 的字符。

数据集中序列由原兽亚纲类(Prototheria)、有袋类(Marsupialia)、有胎盘类(Placentalia)三类构成，其中原兽亚纲类序列只来自 1 个目，有袋类序列来自 2 个目，有胎盘类序列来自非洲兽总目(Afrotheria)中的 3 个目、贫齿目(Xenarthra)中的 2 个目、灵长总目(Euarchontoglires)中的 3 个目和劳亚兽总目(Laurasiatheria)

中的 6 个目，总共由 17 个目构成，其中又包含了 41 个科，因此聚类相对困难。

3. 方法

3.1. 环形前缀标识的定义

对于一条生物序列 S，其位置 i 处长度为 m 的前缀标识需满足以下两个条件：

1) 从生物序列的第 i 个位置开始，到第 $i+m-1$ 个位置结束的字符串在序列中是唯一出现的，并且是唯一出现的字符串中长度最小的。

2) 此字符串去掉最后一个字符后，在序列中至少出现两次[8]。

3.2. 环形前缀集的定义

令 $S=(s_1,s_2,s_3,\cdots,s_N)$ 是一条长度为 N 的核苷酸序列，其中 $s_i\in\{A,C,G,T\}$，$i=1,2,3,\cdots,N$。将序列 S 的前 n 个碱基拼接在序列的末端，构成长度为 $N+m$ 的新核苷酸序列 $S_1=(s_1,s_2,\cdots,s_N,s_{N+1},\cdots,s_{N+n})$，对于新序列 S_1 的第 i 个位置，都存在唯一的前缀 $\text{prefix\_string}(i)$，由 $\text{prefix\_string}(i)$，$1\le i\le n$，构成的集合 $I(S)$ 称为序列 S 的环形前缀集。

3.3. 完全匹配与错配法则

由于存在基因突变的可能，本文在 CPPA 算法的前缀树模型基础上进行改进，主要在获取两条序列共同前缀的过程中遵循完全匹配与错配两种法则，其中完全匹配则认为只有每个位置的字符都相同的前缀才是匹配的，例如序列 S_1 前缀集中的 ACTT 与序列 S_2 前缀集中的 ACTT 是匹配的，也就是 CPPA 算法中的匹配方法。

本文加入的允许错配则认为序列 S_1 前缀集中的 ACTT 与序列 S_2 前缀集中的 ACTTA，ACTTC，ACTTG，ACTTT 均是可以进行匹配的，最后根据它们的位置信息，在 ACTTA，ACTTC，ACTTG，ACTTT 中选取与 S_1 中 ACTT 的最佳匹配。错配原则是在完全匹配原则进行之后，在剩余前缀集中进行的。

3.4. 模型介绍

第一步：利用环形前缀树法构建 n 条序列的环形前缀集 $I(S_i)$，$i=1,2,\cdots,n$，接下来成对处理序列 S_i 与 S_j，其中 $i,j=1,2,\cdots,n$。

第二步：根据完全匹配原则，获取 S_i 与 S_j 序列的共同前缀集 $e_{ij}=I(S_i)\cap I(S_j)$，同时记录共同前缀的个数 n，并且记录共前缀集 e_{ij} 在序列 S_i 与序列 S_j 中的位置(与 ACCP 法中位置的处理方式相同，本模型中获取的位置信息也需标准化)，分别记为 $P_i=(P_i^1,P_i^2,\cdots,P_i^n)$，$P_j=(P_j^1,P_j^2,\cdots,P_j^n)$。对应的位置差为：

$$d_k=\min\left(\left|P_i^k-P_j^k\right|,1-\left|P_i^k-P_j^k\right|\right),\ k=1,2,\cdots,n。\tag{1}$$

第三步：获取 S_i 与 S_j 序列的剩余前缀集 $I'(S_i)$ 与 $I'(S_j)$，并且记录剩余前缀的位置集 P_i'，P_j'，其中

$$I'(S_i)=I(S_i)-e_{ij},\ I'(S_j)=I(S_j)-e_{ij}。$$

第四步：将剩余前缀集 $I'(S_i)$ 与 $I'(S_j)$ 中的前缀均去掉最后一个字符，获得新的字符串集合 $I''(S_i)$ 与 $I''(S_j)$，利用错配原则，获取交集 $e_{ij}^1=I'(S_i)\cap I''(S_j)$，以及交集 $e_{ij}^2=I''(S_j)\cap I''(S_i)$，同时记录交集的总个数 m，在记录两个交集中的元素在 S_i 与 S_j 的标准化位置时，会出现交集中的某个元素在 S_i 与 S_j 出现的位置不是一对一，而是一对多的情况，这样会出现多个位置差，只需选择最佳的位置差，处理如下：

假设交集中的元素 a 在 S_i 中出现的标准化位置为 L_1，在 S_j 中出现的标准化位置为 L_1'，L_2'，L_3'，那么对于元素 a 的位置差为：

$$d=\min\left(\left|L_1-L_i'\right|,1-\left|L_1-L_i'\right|\right),\ \text{其中}i=1,2,3 \tag{2}$$

将第二步与第四步所获得的位置差合并成一维向量：

$$D=\left(d_1,d_2,\cdots,d_n,d_{n+1},\cdots,d_{n+m}\right) \tag{3}$$

第五步：大多数非比对方法都是仅基于计数方法的，即仅考虑公共邻接单词的频率，而无需对其在序列中的位置进行任何校正[9]，本文不仅考虑计数方法，而且还要考虑其在序列中的位置。计数方法被频繁使用是因为计数方法比匹配长度方法在较高的差异情况下更准确[10]，因此在考虑错配时加入的参数一为交集的总个数 $m+n$，同时基于共同前缀的位置差来计算进化距离，序列间的进化距离越小，序列就越相近，因此加入参数二为序列对的最小长度 $\min\left(len1,len2\right)$，最终定义相似性距离为：

$$\text{distance}\left(S_i,S_j\right)=\frac{\min\left(len1,len2\right)\times\text{sum}\left(D\right)}{\left(m+n\right)^2},\ \text{其中}len1,\ len2\text{分别是}S_i\text{与}S_j\text{的长度} \tag{4}$$

4. 模型结果与分析结论

4.1. 模型结果

通过用 MATLAB 软件编写程序实现了本文的模型，运用邻接法构建了进化树，并且用 iTOL 网站进行画树，得到了更直观的关于 70 条哺乳动物 DNA 序列的系统发育树(见图 1)。为了体现本实验方法的有效性，同时也用 MATLAB 软件实现了 ACCP 法模型，构建了关于 70 条哺乳动物 DNA 序列的系统发育树(见图 2)，还与基于共同前缀的位置差法(见图 3)以及傅里叶分析方法[11] (见图 4)的系统发育树的结果都进行比对分析。

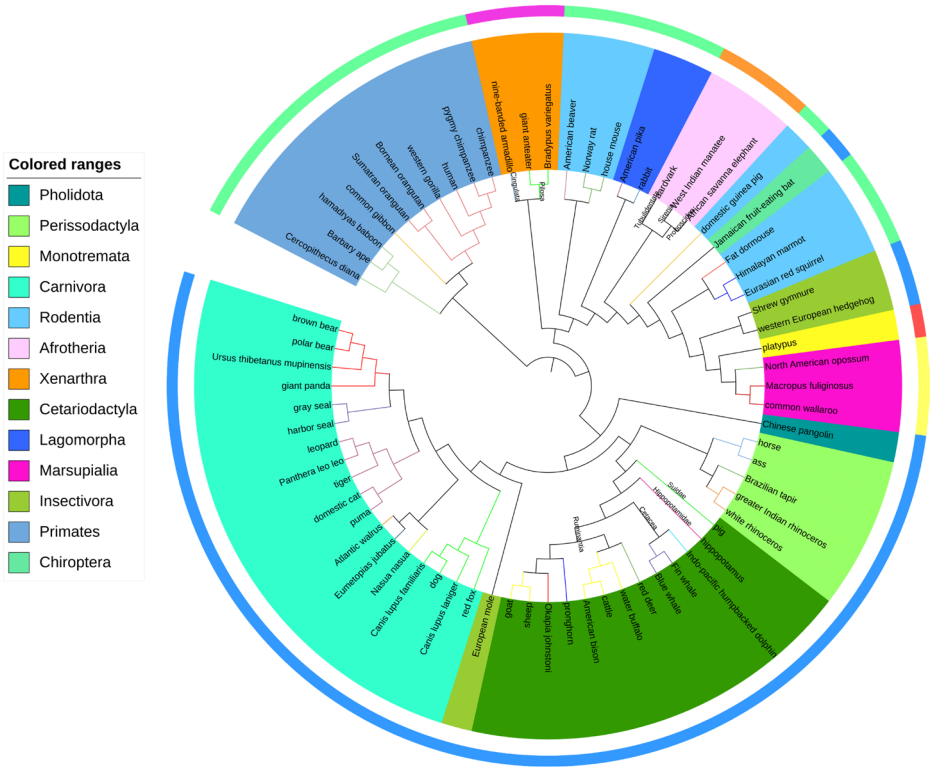

Figure 1. Based on this experimental method, a phylogenetic tree of 70 mammalian mitochondrial DNA sequences was established
图 1. 基于本实验方法建立 70 个哺乳动物线粒体 DNA 序列的系统发育树

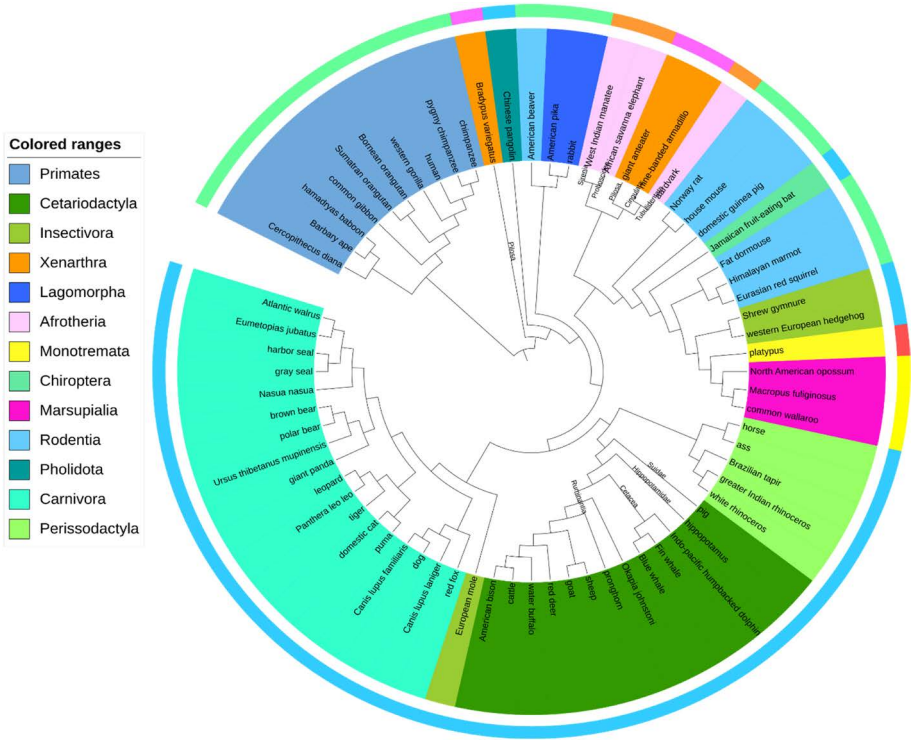

Figure 2. Establishment of a phylogenetic tree of 70 mammalian mitochondrial DNA sequences based on the ACCP method
图 2. 基于 ACCP 方法建立 70 个哺乳动物线粒体 DNA 序列的系统发育树

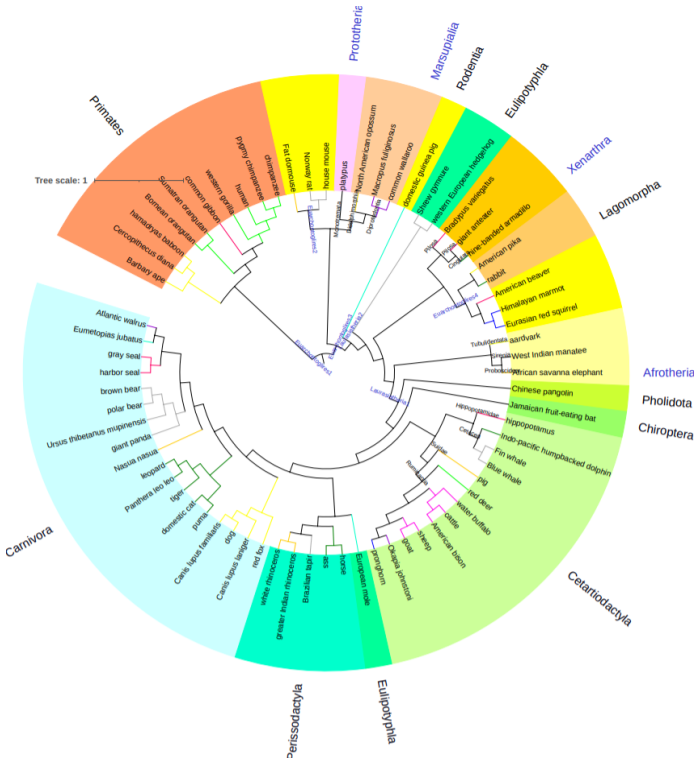

Figure 3. A phylogenetic tree of 70 mammalian mitochondrial DNA sequences by the position difference method based on the common prefix [6]
图 3. 基于共同前缀的位置差法建立 70 个哺乳动物线粒体 DNA 序列的系统发育树[6]

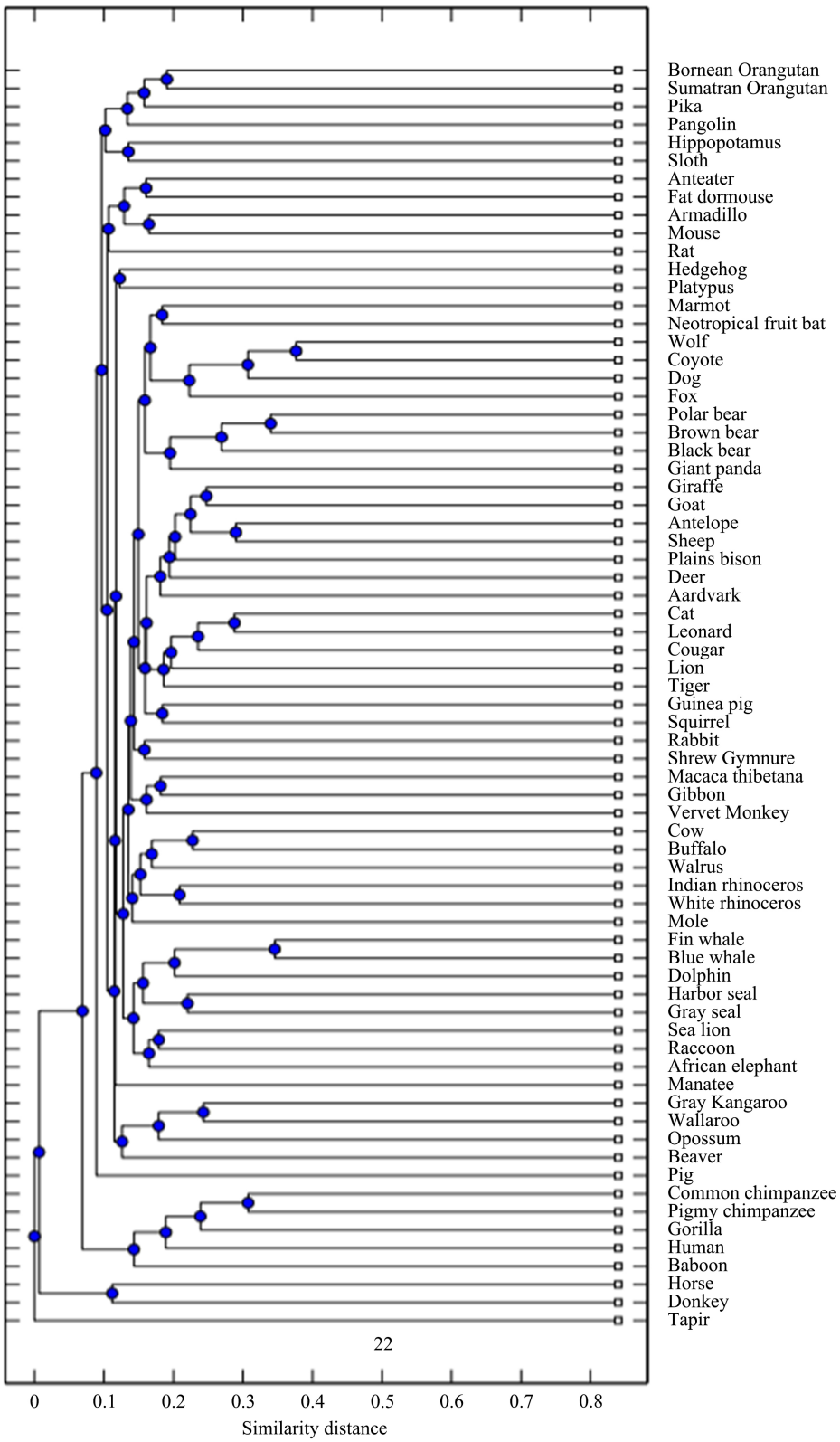

Figure 4. A phylogenetic tree of 70 mammalian mitochondrial genomes by The DFT distance of DNA sequences with the 2D numerical mapping [11]

图 4. 采用傅里叶分析方法对 70 条哺乳动物线粒体 DNA 基因组生成的系统发育树[11]

4.2. 结果分析

在超目的角度进行分析，图 1 与图 3 的进化树中单孔目(Monotremata)、有袋目(Marsupials)、非洲兽总目(Afrotheria)以及贫齿目(Xenarthra)均正确聚类，而图 2 中的非洲兽总目(Afrotheria)和贫齿目(Xenarthra)发生了错误聚类，均被分成两部分，由此可看出本文的非比对方法已经优于 ACCP 法[7]。在图 4 进化树中，非洲兽总目、贫齿目、劳亚兽总目和灵长总目均没有将序列聚类到各自的类别，聚类结果最差。

在目与科的角度分析，图 3 中啮齿目(Rodentia)被分成 4 部分，并且所涉及的 4 个亚目也未能正确聚类，贫齿目(Xenarthra)中的皮毛目(Pilosa)下的两个目的动物未聚到一起，而图 1 中的啮齿目(Rodentia)分成 3 部分，并且亚目下所有的科均聚类正确，同时贫齿目(Xenarthra)中的皮毛目(Pilosa)与有带下目(Cingulata)均聚类正确。图 3 中灵长目(Primates)的人科(Hominidae)中分进来一个长臂猿科(Hylobatidae)动物，而图 1 中灵长目(Primates)的所有科均聚类正确，由此可看出本文的非比对方法也优于基于共同前缀的位置差法[6]。

可以直观看出，无论是在超目，还是目与科的层面分析，本实验结果都是优于傅里叶分析方法，本实验方法不足的是未能解决劳亚兽总目(Laurasiatheria)与灵长总目(Euarchontoglires)的聚类错误问题。

4.3. 结论

通过结果分析可以清楚地知道，把共同前缀标识符作为研究对象，用位置差作为度量因素是有效的方法，将突变考虑进去以后的进化分析模型则是更优的，得到的进化树也符合生物进化关系。但是本实验方法仍然存在不足，比如耗时略长，并且没有考虑到共同前缀标识符的长度以及出现的概率，对于距离的度量仍可以考虑其他因素，因此还有很大的改进空间。

致　谢

感谢两位师姐对我的帮助，不仅给我提供了许多参考资料，而且对于软件的应用问题给我耐心讲解，在此由衷感谢。

参考文献

[1] Tian, K., Zhao, X., Yau, S.S.-T. (2018) Convex Hull Analysis of Evolutionary and Phylogenetic Relationships between Biological Groups. *Journal of Theoretical Biology*, **456**, 34-40. https://doi.org/10.1016/j.jtbi.2018.07.035

[2] Wen, J., Chan, R.H.F., Yau, S.-C., He, R.L. and Yau, S.S.T. (2014) *K*-mer Natural Vector and Its Application to the Phylogenetic Analysis of Genetic Sequences. *Gene*, **546**, 25-34. https://doi.org/10.1016/j.gene.2014.05.043

[3] Ji, Q., Bin, W. and Bai-IinHao (2004) Whole Proteome Prokaryote Phylogeny without Sequence Alignment: A K-String Composition Approach. *Journal of Molecular Evolution*, **58**, 1-11. https://doi.org/10.1007/s00239-003-2493-7

[4] Zhang, Y.Y., Wen, J. and Yau, S.S.-T. (2018) Phylogenetic Analysis of Protein Sequences Based on a Novel *K*-mer Natural Vector Method. *Genomics*, **111**, 1298-1305. https://doi.org/10.1016/j.ygeno.2018.08.010

[5] Weiner, P. (1973) Linear Pattern Matching Algorithms. 14*th Annual Symposium on Switching and Automata Theory* (*Swat* 1973), USA, 15-17 October 1973, 1-11. https://doi.org/10.1109/SWAT.1973.13

[6] 王代, 陆超. 基于前缀标识符及其位置的 DNA 序列比较[J]. 自然科学, 2021, 9(2): 281-290. https://doi.org/10.12677/OJNS.2021.92031

[7] 陆超, 王代. 基于共同前缀位置的哺乳动物 mtDNA 序列系统发育分析[J]. 自然科学, 2021, 9(2): 272-280. https://doi.org/10.12677/OJNS.2021.92030

[8] 张欣. 基于后缀树的 DNA 序列进化树构建研究[D]: [硕士学位论文]. 大连: 辽宁师范大学, 2019.

[9] Amiri, S., and Dinov, I.D. (2016) Comparison of Genomic Data via Statistical Distribution. *Journal of Theoretical Biology*, **407**, 318-327. https://doi.org/10.1016/j.jtbi.2016.07.032

[10] Bernard, G., Chan, C.X., Chan, Y.-B., Chua, X.-Y., Cong, Y.N., Hogan, J.M., Maetschke, S.R. and Ragan, M.A.

(2017) Alignment-Free Inference of Hierarchical and Reticulate Phylogenomic Relationships. *Briefings in Bioinformatics*, **20**, 426-435. https://doi.org/10.1093/bib/bbx067

[11] Yin, C.C. and Yau, S.S.-T. (2015) An Improved Model for Whole Genome Phylogenetic Analysis by Fourier Transform. *Journal of Theoretical Biology*, **382**, 99-110. https://doi.org/10.1016/j.jtbi.2015.06.033