Open Journal of Natural Science 自然科学, 2021, 9(2), 281-290
Published Online March 2021 in Hans. http://www.hanspub.org/journal/ojns
https://doi.org/10.12677/ojns.2021.92031

Hans汉斯

基于前缀标识符及其位置的DNA序列比较

王　代，陆　超

辽宁师范大学，辽宁　大连
Email: 2568062365@qq.com

收稿日期：2021年2月24日；录用日期：2021年3月22日；发布日期：2021年3月29日

摘　要

分子序列比较是生物信息学中最基本、最主要的问题，DNA序列相似性分析是研究的重要的课题。非比对方法是研究序列比较的方法之一，它克服了比对方法的局限，其计算速度更快。本文从前缀标识符位置角度出发，利用信息熵，提出了序列分析的非比对方法。本文通过对生物序列构建前缀树，得到生物序列前缀标识符的基础上，以两两序列的共同前缀标识符为研究对象，提取它们在序列中位置信息，将它们的位置差的绝对值看成随机变量，利用信息熵，提出新的DNA序列相似性度量方法，建立有效的模型。将70个哺乳动物的线粒体DNA序列作为实验数据集，应用该模型得到的相似性距离构建生物进化树。该进化树的分类结果符合当前的生物学分类标准。

关键词

非比对方法，相似性度量，进化树，前缀标识符，信息熵

Comparison of DNA Sequences Based on Prefix Identifiers and Their Locations

Dai Wang, Chao Lu

Liaoning Normal University, Dalian Liaoning
Email: 2568062365@qq.com

Received: Feb. 24th, 2021; accepted: Mar. 22nd, 2021; published: Mar. 29th, 2021

Abstract

Comparison of molecular sequence is the most basic and important problem in bioinformatics. DNA sequence similarity analysis is an important research topic. Alignment-free method is one of the methods to study sequence comparison. It overcomes the limitation of alignment method and

文章引用: 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.
DOI: 10.12677/ojns.2021.92031

is faster than alignment method. In this paper, from the point of view of prefix identifier location, the alignment-free method of sequence analysis is proposed by using information entropy. Based on the prefix tree and the prefix identifier of biological sequences, the position information of pairwise sequences is extracted by using the common prefix identifiers of pairwise sequences. The absolute value of their position difference is regarded as random variable. Using information entropy, a new DNA sequence similarity measurement method is proposed and an effective model is established. Mitochondrial DNA sequences of 70 mammalian were used as experimental data sets. Construct the Phylogenetic tree based on the similarity distance obtained by the model. The classification results of Phylogenetic tree conform to the current biological classification.

Keywords

Alignment-Free Method, Similarity Measurement, Phylogenetic Tree, Prefix Identifier, Information Entropy

Copyright © 2021 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

1. 引言

近几十年来，随着人类基因组计划的实施和完成，分子生物学发展呈现出生物数据爆炸式增长的特点。面对呈指数增长趋势的海量数据，高效管理、准确解读、从而挖掘有用的生物信息是一项有意义的工作，同时也是生物、数学、计算机科学等多个领域专家学者面临的一大挑战[1]。本文将信息理论知识、前缀树知识与生物信息学知识相结合，提出一种新的序列相似性度量方法。

1973 年 Peter Weiner 提出了前缀树模型及其算法[2]。前缀树模型的提出启发很多在生物信息学领域的学者们，为非比对方法提供了新的研究手段。部分文献基于前缀树，利用序列最长公共子串，提出序列相似性度量模型[3]。有文献通过讨论 k 词在序列中的位置分布，来研究序列相似性[4]。本文将提出一种新的以 DNA 序列前缀标识符为研究对象的非比对方法。

本文将提出的新的 DNA 序列相似性度量模型在 70 条哺乳动物线粒体 DNA 序列上进行实验，根据生成的进化树，进行结果分析与讨论。将所得的结果首先与 NCBI 给出的生物学分类标准进行比较，证明本文提出的方法对该数据集具有正确的分类能力，其次与已发表文献结果进行比较，得出本结果在不同程度上由于前人结果。

2. 材料

本文中使用哺乳动物线粒体 DNA 序列作为数据集进行实验。本数据集中共有 70 条不同的哺乳动物线粒体 DNA 序列。序列长度最短为 16295，最长为 17447。并且 DNA 序列中不含有非 A、C、G、T 的字符。本文中的数据集均为从网站 National Center for Biotechnology Information“NCBI”(https://www.ncbi.nlm.nih.gov/)中 Genbank 数据库下载得到。具体信息见附录。

本数据集中序列来自于哺乳动物中原兽亚纲动物(Prototheria)、有袋类动物(Marsupialia)、有胎盘类动物(Placentalia)。其中有胎盘类哺乳动物的序列来自于非洲总兽目(Afrotheria)、贫齿目(Xenarthra)、灵长总目(Euarchontoglires)、劳亚兽总目(Laurasiatheria)四个超目。本数据集中有 17 个目，41 个科。

本文数据集与文献[5]有 63 个相同物种，有 3 条序列有相同的属，2 条序列有相同的科。

3. 方法

3.1. 前缀标识符

3.1.1. 线形序列前缀标识符定义

本文对前缀标识符的定义来源于文献[1]。对任意序列 S， $S=S(1)S(2)\cdots S(n)$，则 n 为序列 S 的长度，其中 $S(i)$ 称为序列 S 在位置 i 处的字符。将字符串 $S(i)S(i+1)S(i+2)\cdots S(j)$ 称为序列 S 中 i 处开始在 j 处结束的字符串。在序列 S 的结尾处添加一个结束字符，该字符必须与 S 中每一个字符都不相同，例如#。于是得到连接后的序列，记为 SS，则 $SS=S(1)S(2)\cdots S(n)\#$。若字符串 $S(i)S(i+1)\cdots S(j)$ 为位置 i 处的前缀标识符，则满足以下两个条件：(1) 字符串 $S(i)S(i+1)\cdots S(j)$ 序列 SS 中只出现一次(2) $S(i)S(i+1)\cdots S(j-1)$ 在序列 SS 中至少出现两次。例如，字符串 $S=acppb$，添加结束字符#，则 $SS=acppb\#$。那么 S 在位置 1 处的前缀标识符为 a，位置 3 处的前缀标识符为 pp，而不是 p，这是因为在 SS 中位置 3 和位置 4 处均出现字符 p，即不满足条件(1)，pp 在 SS 只出现一次，而 p 在 SS 中出现 2 次，满足条件(1)和(2)。

3.1.2. 环形序列前缀标识符定义

因为哺乳动物线粒体 DNA 序列是环形的，因此本文定义环形序列的前缀标识符，其定义如下：任意 DNA 序列 $S1$， $S1=S(1)S(2)\cdots S(n)$，若字符串 $S(i)S(i+1)\cdots S(j)$ 为位置 i 处的前缀标识符，则满足以下两个条件：(1) 在 $S(i)S(i+1)\cdots S(j)$ 序列 $S1$ 中只出现一次(2) $S(i)S(i+1)\cdots S(j-1)$ 在序列 $S1$ 中至少只出现两次。本文下述前缀标识符，均为对环形序列查找。

3.2. 信息熵理论

3.2.1. 信息熵的定义

信息熵是对信息的一种度量方法，是生物信息学研究中的一种重要工具[6] [7] [8]。在信息理论中，单符号的离散信息源是最简单的离散信源，一个信源所有可能发送的消息符号构成一个样本空间，用 $X=\{x_k\,|k=1,2,\cdots,n\}$ 表示信源 X 的样本集合，其中表示 x_k 信源所有发出的消息符号，用 $P_X=\{p_k\,|k=1,2,\cdots,n\}$ 表示概率集合，则 p_k 表示信源发出符号 x_k 的概率。信源 X 的概率空间表示为

$$\begin{bmatrix} X \\ P_X \end{bmatrix}=\begin{bmatrix} x_1 & x_2 & \cdots & x_n \\ p_1 & p_2 & \cdots & p_n \end{bmatrix}, 0\le p_i\le 1, \sum_{i=1}^{n} p_i=1 \tag{1}$$

信息熵是信源发出任何一个消息状态所具有的平均信息量[9]。对于离散无记忆的信源，它的熵定义为[9]

$$H(X)=-\sum_{i=1}^{n} p_i \log p_i \tag{2}$$

3.2.2. 信息熵熵的性质渐化性

信息熵满足如下条件，该性质称为信息熵的渐化性[10]。

$$H(X)=H(p_1,p_2,\cdots,p_n)=H(p_1+p_2,p_3,\cdots,p_n)+(p_1+p_2)H\left(\frac{p_1}{p_1+p_2},\frac{p_2}{p_1+p_2}\right) \tag{3}$$

$$0\le p(x_i)\le 1, k=1,2,\cdots,n, \sum_{i=1}^{n} p(x_n)=1, p(x_1)+p(x_2)>0$$

根据上面的等式，发现 $(p(x_1)+p(x_2))H\left(\frac{p(x_1)}{p(x_1)+p(x_2)},\frac{p(x_2)}{p(x_1)+p(x_2)}\right)$ 是非负数，因此等式右边的

第一项是小于等式左边，则意味着概率合并相加会导致熵变小。也就是说概率合并相加会导致概率分布越不均匀，从而熵变小。因此渐化性反映出概率分布越均匀，熵值越大。

3.3. 相似性度量

任意两条序列之间存在着差异，这些差异可以看作是在进化过程中所产生的突变。本文将两条序列的共同前缀标识符作为它们的共同特征，该标识符在两序列中的位置不同，由此可产生该标识符的位置差，进而有位置差的绝对值。本文认为位置差绝对值的情况越多，意味着进化过程中发生的突变越大，则两序列之间的差异性越大。也就是说位置差绝对值的概率分布越均匀，则两序列的距离越大。于是把该位置差的绝对值看成信源发出的消息符号，每一个绝对值都有相应的出现频率，将该频率看成概率，这些位置差绝对值的信息熵作为两序列之间的距离。

4. 模型

一、提取两序列的共同特征，即找到共同前缀标识符。

(1) 对于任意两条序列 $S1$、$S2$，分别查找所有位置处的前缀标识符，得到前缀标识符集合，分别记为 $I(S1)$、$I(S2)$；

(2) 根据前缀标识符集合 $I(S1)$、$I(S2)$，得到序列 $S1$、$S2$ 共同前缀标识符，记为 $e(i)$，并记录共同前缀标识符的个数，记为 m。其中，

$$m=\left|I(S1)\cap I(S2)\right|, e(i)\in I(S1)\cap I(S2), i=1,2,\cdots,m \tag{4}$$

二、计算共同前缀标识符在两序列中位置差的绝对值。

(1) 查找共同前缀标识符 $e(i)$ 在 $S1$ 和 $S2$ 中位置，分别记为 $pos^1(i)$、$pos^2(i)$，其中 $i=1,2,\cdots,m$；

(2) 计算出共同前缀标识符 $e(i)$ 对应的位置差绝对值，记为 $pos\_dis(i)$，

$$pos\_dis(i)=\left|pos^1(i)-pos^2(i)\right|, i=1,2,\cdots,m \tag{5}$$

则所有共同前缀标识符对应的位置差绝对值集合可表达为

$$POS\_DIS=\left\{pos\_dis(i)\middle|pos\_dis(i)=pos^1(i)-pos^2(i)\right|, i=1,2,\cdots,m\} \tag{6}$$

对于任意两个共同前缀标识符 $e(i)$，$e(j)$，它们在序列位置差绝对值 $pos\_dis(i)$、$pos\_dis(j)$ 可能相等，因此 POS_DIS 集合为多重集，其中 $i,j\in 1,2,\cdots,m$。

三、将上述位置差绝对值看成信源发出的符号，得到信源概率空间，计算信息熵，得到两序列间距离。

(1) 根据位置差绝对值对应的多重集 POS_DIS，找到位置差绝对值出现的所有情况，分别记为 $u\_pos\_dis(1),\cdots,u\_pos\_dis(h)$，则共有 h 中情况；

(2) 统计各类位置差绝对值出现的次数，记为 $n(1),n(2),\cdots,n(t)$ 则

$$n(1)+\cdots+n(h)=m \tag{7}$$

(3) 计算各类位置差绝对值出现的频率，记为 $f(t)$，则

$$f(t)=\frac{n(t)}{m}, t=1,2,\cdots,h \tag{8}$$

(4) 信源的概率空间为

$$\begin{bmatrix}X\\P_X\end{bmatrix}=\begin{bmatrix}u\_pos\_dis(1) & u\_pos\_dis(2) & \cdots & u\_pos\_dis(h)\\ f(1) & f(2) & \cdots & f(h)\end{bmatrix} \tag{9}$$

(5) 两序列的距离定义为

$$dis(S1,S2)=\sum_{t=1}^{h}-f(t)\log_2 f(t) \tag{10}$$

信息熵公式中对数的底可以有多种[10]，本文中计算两序列相似性时取以 2 为底的对数。

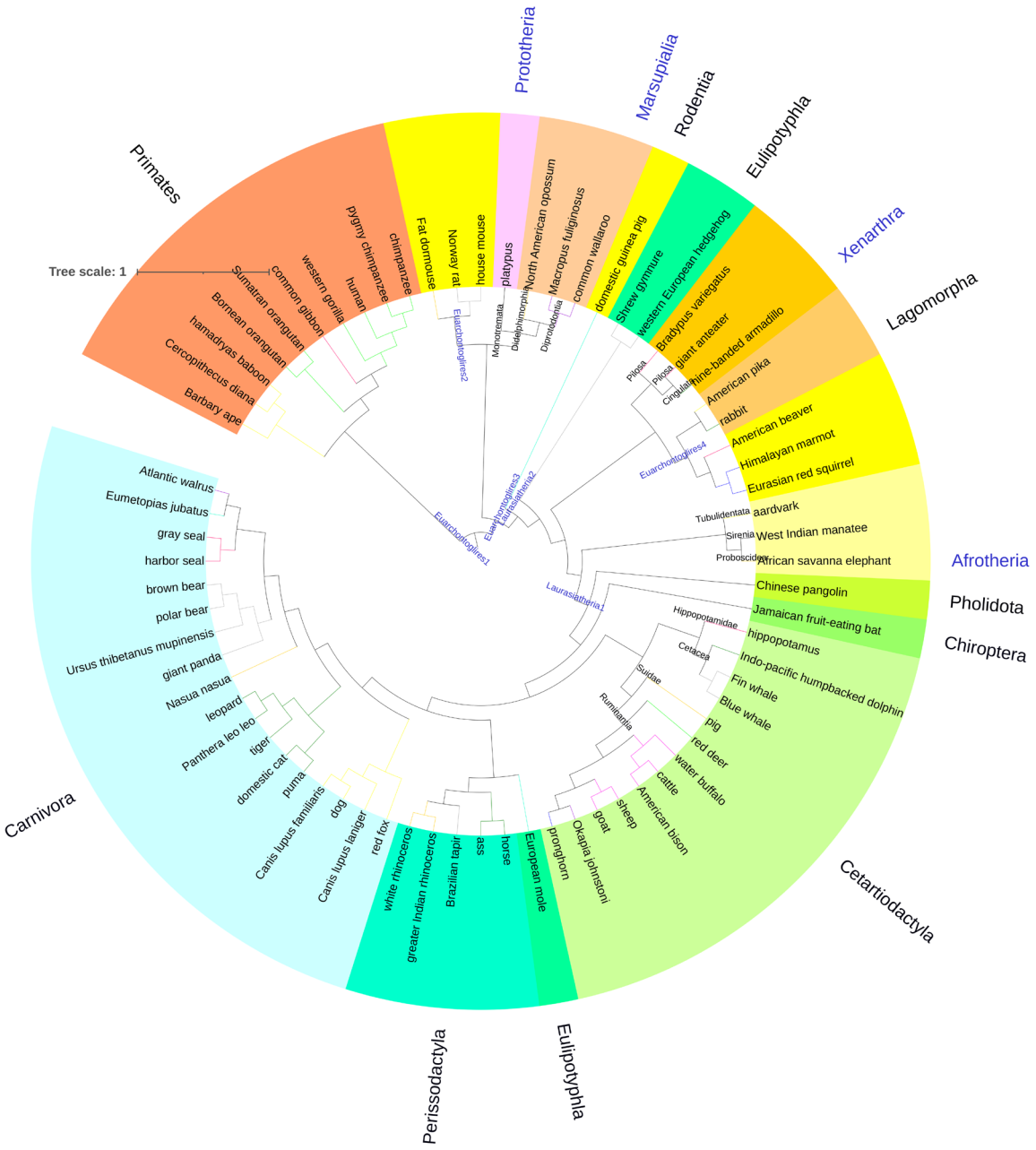

Figure 1. A phylogenetic tree of 70 mammalian mitochondrial DNA sequences by Methods of this article
图 1. 本文方法对 70 条哺乳动物线粒体 DNA 序列生成的系统发育树

在图 1 进化树中，蓝色字体表示哺乳动物的超目，黑色字体表示哺乳动物的各个目，进化树的分支代表科，同一目下不同颜色的分支表示不同的科。

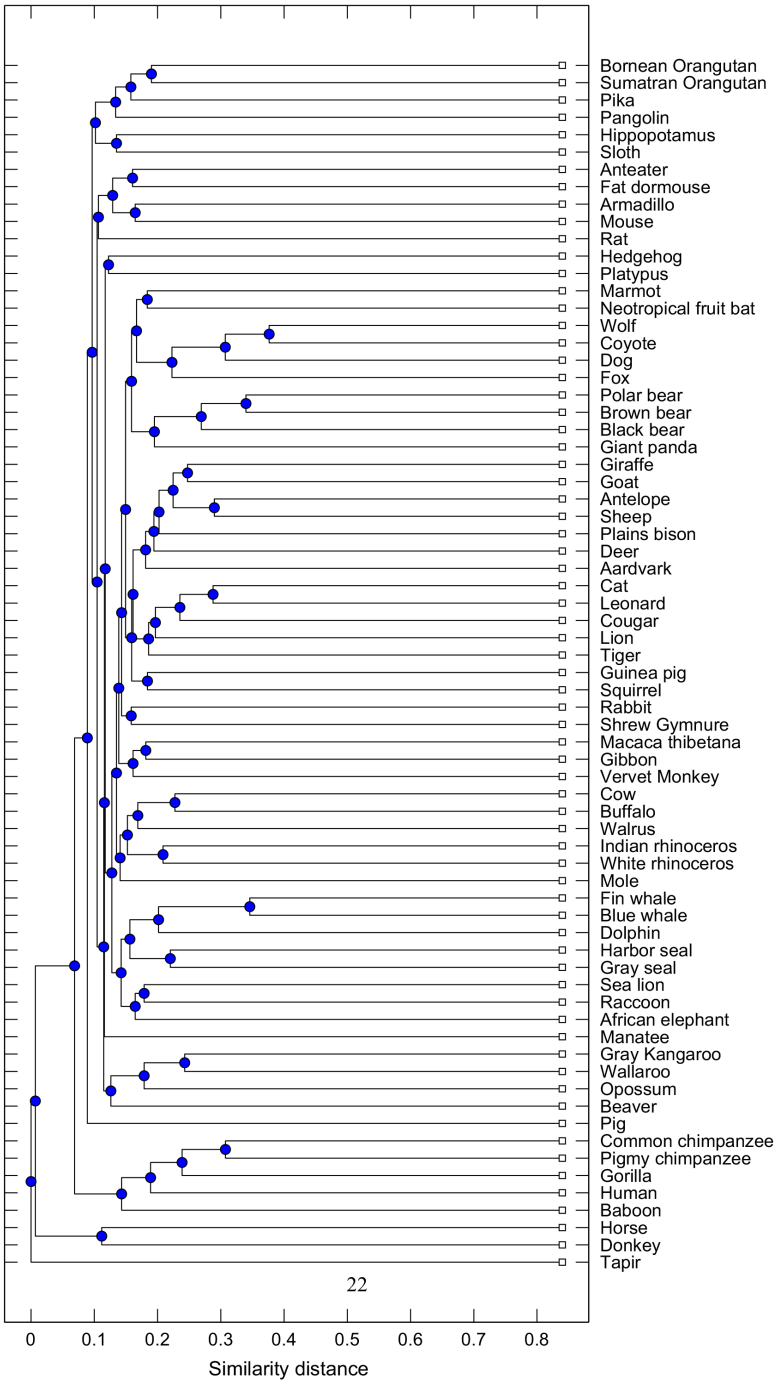

Figure 2. A phylogenetic tree of 70 mammalian mitochondrial genomes by The DFT distance of DNA sequences with the 2D numerical mapping [5]
图 2. 采用傅里叶分析方法对 70 条哺乳动物线粒体基因组生成的系统发育树[5]

5. 结果讨论与分析

本文运用 Matlab 软件编写代码并实现模型，运用邻接法(Neighbor-joining)构建进化树，运行时间为 90 秒。

根据生物学分类，哺乳动物分为原生哺乳动物 Prototheria、有袋哺乳动物 Marsupialia 和胎盘哺乳动物 Placentalia。其中胎盘哺乳动物又分为灵长总目 Euarchontoglires 和劳亚兽总目 Laurasiatheria。本文的进化树中将原生哺乳动物和有袋哺乳动物分别聚类到各自类别中，但是胎盘哺乳动物聚类欠佳。这是因为在图 1 进化树中，灵长总目分为四个分支，劳亚兽总目分成两个分支。在图中 2 进化树中，原生哺乳动物、有袋哺乳动物和胎盘哺乳动物均没有聚类清楚。

从超目的层次分析，本文的进化树将单孔目 Monotremata、有袋目 Marsupials、非洲兽总目 Afrotheria、贫齿目 Xenarthra 中动物分别聚在各自的超目中，劳亚兽总目 Laurasiatheria 中只有食虫目中两条序列没有正确聚类，分在劳亚兽总目之外，其余则聚类良好。灵长总目 Euarchontoglires 由灵长目、啮齿目和兔形目组成，本文中该总目动物聚类欠佳。在图 2 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当中。其中非洲兽总目中序列 African elephant 序列和 aardvark 序列与劳亚兽总目中的序列聚在一个节点，劳亚兽总目中 hippopotamus 序列与灵长总目中序列聚在一个节点，在贫齿目中 sloth 序列与灵长总目中序列聚在一个节点。

从目的层次分析，本文进化树啮齿目和贫齿目 Xenarthra 中的皮毛目 Pilosa 两个目的动物未聚到一起，其余各个目中动物聚类良好。其中鲸偶蹄目有鲸目 Cetacea、反刍亚目 Ruminantia、猪次目 Suina 和河马科 Hippopotamidae 构成，本文这四个目分别聚类良好。鲸偶蹄目在进化关系上也表现良好，与文献[11]一致。

在科的层次上分析，本文进化树灵长目中人科中分进来一个长臂猿科动物，鲸偶蹄目中牛科动物分开，其余各个科分别聚类良好。

综上所述，本文提出的序列相似性度量在哺乳动物数据上具有有效性。本文的进化树结果要优于傅里叶分析的进化树结果[5]。

6. 总结

本文以生物 DNA 序列以及它们的共同前缀标识符为研究对象，对共同前缀标识符在生物序列中的位置信息进行深入的分析研究，以信息熵作为研究手段，提出一个序列相似性度量方法，建立了一个高效的 DNA 序列比对模型。依据提出的模型，本文利用 70 条哺乳动物线粒体 DNA 序列构建生物进化树，与文献[5]提出的 DFT 方法得到的进化树进行比较，本文的方法则更有优势。

本文提出的方法优点有：(1) 算法的时间复杂度低。(2) 思路简单，易于理解。本文的方法也存在一些不足之处，需要改进完善。该方法只对两序列共同前缀标识符的位置信息进行研究，并没有对共同前缀标识符的长度、以及共同前缀标识符的个数信息研究。

参考文献

[1] 李霞, 雷建波, 等. 生物信息学[M]. 第 2 版. 北京: 人民卫生出版社, 2015: 1-8.

[2] 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

[3] Leimeister, C.-A. and Morgenstern, B. (2014) Kmacs: The k-Mismatch Average Common Substring Approach to Alignment-Free Sequence Comparison. *Bioinformatics*, **30**, 2000-2008. https://doi.org/10.1093/bioinformatics/btu331

[4] 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

[5] 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

[6] Vinga, S. (2013) Information Theory Applications for Biological Sequence Analysis. *Bioinformatics*, **15**, 1-14.

[7] Singh, K., Kumar, A. and Gupta, M.K. (2020) Modified k-String in Composition Vector Method for DNA Sequence Comparison Based on Maximum Entropy Principle. *Journal of Interdisciplinary Mathematics*, **23**, 31-41. https://doi.org/10.1080/09720502.2020.1721649

[8] Pinello, L., Lo Bosco, G. and Yuan, G.-C. (2013) Applications of Alignment-Free Methods in Epigenomics. *Bioinformatics*, **15**, 1-12.

[9] 詹青. 基于信息熵理论的基因组特性研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨工业大学, 2011.

[10] 吕峰, 王虹. 信息理论与编码[M]. 第 2 版. 北京: 人民邮电出版社, 2010: 20-100.

[11] Zurano, J.P., Magalhães, F.M., *et al*. (2019) Cetartiodactyla: Updating a Time-Calibrated Molecular Phylogeny. *Molecular Phylogenetics and Evolution*, **133**, 256-262. https://doi.org/10.1016/j.ympev.2018.12.015

附录

Table 1. Mitochondria DNA sequence information of 70 mammalian genomes
表 1. 70 条哺乳动物线粒体 DNA 序列信息

Accession	Common Name	Super Orders	Super Orders	Super Orders	Order	Family
Z29573	North American opossum	marsupialia			Didelphidae	Didelphidae
Y10524	common wallaroo	marsupialia			Diprotodontia	Macropodidae
NC_039717.1	Macropus fuliginosus	marsupialia			Diprotodontia	Macropodidae
AJ001588	rabbit	placentalia	Boreoeutheria	Euarchontoglires	Lagomorpha	Leporidae
AJ537415	American pika	placentalia	Boreoeutheria	Euarchontoglires	Lagomorpha	Ochotonidae
Y18001	hamadryas baboon	placentalia	Boreoeutheria	Euarchontoglires	Primates	Cercopithecidae
NC_002764	Barbary ape	placentalia	Boreoeutheria	Euarchontoglires	Primates	Cercopithecidae
NC_023963.1	Cercopithecus diana	placentalia	Boreoeutheria	Euarchontoglires	Primates	Cercopithecidae
D38115	Bornean orangutan	placentalia	Boreoeutheria	Euarchontoglires	Primates	Hominidae
D38113	chimpanzee	placentalia	Boreoeutheria	Euarchontoglires	Primates	Hominidae
V00662	human	placentalia	Boreoeutheria	Euarchontoglires	Primates	Hominidae
NC_002083	Sumatran orangutan	placentalia	Boreoeutheria	Euarchontoglires	Primates	Hominidae
D38114	western gorilla	placentalia	Boreoeutheria	Euarchontoglires	Primates	Hominidae
D38116	pygmy chimpanzee	placentalia	Boreoeutheria	Euarchontoglires	Primates	Hominidae
X99256	common gibbon	placentalia	Boreoeutheria	Euarchontoglires	Primates	Hylobatidae
NC_015108	American beaver	placentalia	Boreoeutheria	Euarchontoglires	Rodentia	Castoridae
AJ222767	domestic guinea pig	placentalia	Boreoeutheria	Euarchontoglires	Rodentia	Caviidae
AJ001562	Fat dormouse	placentalia	Boreoeutheria	Euarchontoglires	Rodentia	Gliridae
V00711	house mouse	placentalia	Boreoeutheria	Euarchontoglires	Rodentia	Muridae
X14848	Norway rat	placentalia	Boreoeutheria	Euarchontoglires	Rodentia	Muridae
AJ238588	Eurasian red squirrel	placentalia	Boreoeutheria	Euarchontoglires	Rodentia	Sciuridae
NC_018367	Himalayan marmot	placentalia	Boreoeutheria	Euarchontoglires	Rodentia	Sciuridae
NC_020679	pronghorn	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Antilocapridae
EU177871	American bison	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Bovidae
V00654	cattle	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Bovidae
AF533441	goat	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Bovidae
AF010406	sheep	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Bovidae
AY488491	water buffalo	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Bovidae
AB245427	red deer	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Cervidae
NC_020730.1	Okapia johnstoni	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Giraffidae
AJ010957	hippopotamus	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Hippopotamidae
AJ002189	pig	placentalia	Boreoeutheria	Laurasiatheria	Artiodactyla	Suidae
NC_002008.4	Canis lupus familiaris	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Canidae
U96639	dog	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Canidae
NC_011218.2	Canis lupus laniger	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Canidae
AM181037	red fox	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Canidae

Continued

U20753	domestic cat	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Felidae
EF551002	leopard	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Felidae
EF551003	tiger	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Felidae
KF907306	Panthera leo leo	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Felidae
NC_016470	puma	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Felidae
AJ428576	Atlantic walrus	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Odobenidae
NC_004030.2	Eumetopias jubatus	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Otariidae
X72004	gray seal	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Phocidae
X63726	harbor seal	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Phocidae
NC_020647.1	Nasua nasua	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Procyonidae
AF303110	brown bear	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Ursidae
EF212882	giant panda	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Ursidae
AF303111	polar bear	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Ursidae
DQ402478	Ursus thibetanus mupinensis	placentalia	Boreoeutheria	Laurasiatheria	Carnivora	Ursidae
X72204	Blue whale	placentalia	Boreoeutheria	Laurasiatheria	Cetacea	Balaenopteridae
X61145	Fin whale	placentalia	Boreoeutheria	Laurasiatheria	Cetacea	Balaenopteridae
EU557091	Indo-pacific humpbacked dolphin	placentalia	Boreoeutheria	Laurasiatheria	Cetacea	Delphinidae
AF061340	Jamaican fruit-eating bat	placentalia	Boreoeutheria	Laurasiatheria	Chiroptera	Phyllostomidae
X88898	western European hedgehog	placentalia	Boreoeutheria	Laurasiatheria	Eulipotyphla	Erinaceidae
NC_019626	Shrew gymnure	placentalia	Boreoeutheria	Laurasiatheria	Eulipotyphla	Erinaceidae
NC_002391	European mole	placentalia	Boreoeutheria	Laurasiatheria	Eulipotyphla	Talpidae
X97337	ass	placentalia	Boreoeutheria	Laurasiatheria	Perissodactyla	Equidae
X79547	horse	placentalia	Boreoeutheria	Laurasiatheria	Perissodactyla	Equidae
Y07726	white rhinoceros	placentalia	Boreoeutheria	Laurasiatheria	Perissodactyla	Rhinocerotidae
X97336	greater Indian rhinoceros	placentalia	Boreoeutheria	Laurasiatheria	Perissodactyla	Rhinocerotidae
AJ428947	Brazilian tapir	placentalia	Boreoeutheria	Laurasiatheria	Perissodactyla	Tapiridae
NC_016008	Chinese pangolin	placentalia	Boreoeutheria	Laurasiatheria	Pholidota	Manidae
AJ224821	African savanna elephant	placentalia		Afrotheria	Proboscidea	Elephantidae
NC_010302	West Indian manatee	placentalia		Afrotheria	Sirenia	Trichechidae
Y18475	aardvark	placentalia		Afrotheria	Tubulidentata	Orycteropodidae
Y11832	nine-banded armadillo	placentalia		Xenarthra	Cingulata	Dasypodidae
NC_028501.1	Bradypus variegatus	placentalia		Xenarthra	Pilosa	Bradypodidae
MH142215	giant anteater	placentalia		Xenarthra	Pilosa	Myrmecophagidae
X83427	platypus	Prototheria			Monotremata	Ornithorhynchidae