畜牧兽医学报 2018,49(9):1861-1869
Acta Veterinaria et Zootechnica Sinica
doi: 10.11843/j.issn.0366-6964.2018.09.007

基于高速双倍自助法对蒙古马系统发育树的研究

任爱珍[1],白东义[2],赵一萍[2],侯 娜[1],芒 来[2*]
(1. 内蒙古农业大学理学院,呼和浩特 010018; 2. 内蒙古农业大学动物科学学院
内蒙古自治区蒙古马遗传资源保护及马产业工程实验室,呼和浩特 010018)

摘 要: 旨在利用高速双倍自助法研究蒙古马系统发育树的问题。本研究选用 30 匹蒙古马的 mtDNA D-loop 区碱基序列,其中包括 5 个蒙古马类群(巴尔虎马、三河马、乌审马、乌珠穆沁马、锡尼河马)各 6 匹,对其所有 105 个可能系统树(普氏野马作为外群,6 匹)进行分析,利用生物信息学软件 Mega6 和 PAMAL4.9,以最大似然法估计蒙古马的最大似然系统树。最后利用生物信息软件 CONSEL、R3.0 中的 SDBP 包等计算 105 个候补系统树的基于高速双倍自助法的可靠度(speedy double bootstrap *P*-value,SDBP)。结果表明,SDBP 值最大的蒙古马系统树共有 3 个分支,巴尔虎马与乌审马合并为一个分支,具有较近的遗传关系;乌珠穆沁马与锡尼河马具有较近的遗传关系;三河马独立于其他 4 种蒙古马,构成一个系统发育分支。SDBP 值最大的蒙古马系统树与三河马和其他 4 类蒙古马具有较远的预测血缘关系是一致的,同时此结果的系统树也具有最大似然值。本研究结果还表明,采用最大似然法结合 SDBP 值分析蒙古马的系统拓扑关系是较最大似然法结合渐近无偏(approximately unbiased,AU) 值以及结合自助(bootstrap *P*-value,BP)值分析蒙古马系统拓扑关系更有效。同时也比非加权组平均法(unweighted pair-group method with arithmetic means,UPGMA) 有效,而与邻接法(neighbor joining,NJ)得到了相近的结论。上述结果提示,我们的研究结果使蒙古马的进化理论奠定在更坚实的遗传学基础上,可推动蒙古马的进化理论进一步发展与完善。

关键词: 蒙古马;普氏野马;高速双倍自助法;系统进化树
中图分类号:S821.2 **文献标志码:**A **文章编号:** 0366-6964(2018)09-1861-09

Phylogenetic Tree Analysis of Mongolian Horses Using the Speedy Double Bootstrap Method

REN Ai-zhen[1], BAI Dong-yi[2], ZHAO Yi-ping[2], HOU Na[1], DUGARJAVIIN Mang-lai[2*]
(1. *College of Science, Inner Mongolia Agricultural University, Hohhot* 010018, *China*;
2. *Inner Mongolia Autonomous Region Mongolian Horse Genetic Resources Protection and Horse Industry Engineering Laboratory, College of Animal Science, Inner Mongolia Agricultural University, Hohhot* 010018, *China*)

Abstract: The aim of this study was to analyze the phylogenetic tree of Mongolian horses using the speedy double bootstrap method. The mitochondrial DNA D-loop base sequences of 30 modern horses were used, which included 5 Mongolian horse subspecies (Baerhu, Sanhe, Wushen, Ujimqin and Xinihe horses) 6 for each subspecies and 6 Przewalski's horse as the outgroup, and all 105 possible phylogenetic trees were analyzed. Mega6 and PAMAL4.9 were used, and maximum likelihood phylogenetic tree of Mongolian horse were estimated by maximum likelihood

收稿日期:2018-01-22
基金项目:内蒙古农业大学博士科研启动基金(BJ2014-14);内蒙古自治区自然科学基金(2017ZD06;2018MS01006);内蒙古自治区科学技术厅项目(201603002);国家自然科学基金(31472070)
作者简介:任爱珍(1974-),女,蒙古族,内蒙古呼和浩特人,博士,副教授,主要从事统计模型信赖度研究及其应用,Tel:0471-4306727,E-mail:1620594116@qq.com
*** 通信作者:**芒 来,博导,教授,主要从事马属动物遗传育种与繁殖研究,E-mail: dmanglai@163.com

method. Finally, the 3rd order accurate *P*-value (speedy double bootstrap *P*-value, SDBP) of the 105 phylogenetic trees was calculated with the CONSEL, SDBP packages in R3.0. The phylogenetic tree of Mongolian horse with the largest SDBP value had 3 branches; Baerhu and Wushen horses were clustered together, Ujimqin and Xinihe horses were clustered together, and Sanhe horse formed an independent branch which was sister to all other Mongolian subspecies. The Mongolian horse phylogenetic tree with the largest SDBP value was consistent with the predicted relationship between Sanhe and the other 4 Mongolian horse subspecies, and this topological tree also had the highest likelihood value. The results of this study also showed that it was more effective to analyze the topological relationships of Mongolian horses using the maximum likelihood method combined with the SDBP value than using the maximum likelihood method combined with the approximately unbiased (AU) value and the bootstrap *P*-value (BP) value. Additionally, this approach was more effective than the unweighted pair-group method with arithmetic means (UPGMA), and had similar results to the neighbor joining (NJ) method. The results of this study provided robust genetic data which would produce a more complete picture of Mongolian horse evolution.

Key words: Mongolian horses; Przewalski's horse; speedy double bootstrap method; phylogenetic tree

各种分子系统进化树的构建方法只给出了一个真实进化树的点估计,需要对其可靠性给出概率描述[1]。目前提出的常用的基于评价最大似然树的可靠性方法有:自助法、下平-长谷川法(shimodaira-hasegawa method,SH)、多尺度自助法(multiscale bootstrap method, MBP)、双倍自助法(double bootstrap method,DBP)、高速双倍自助法(speedy double boostrap method,SDBM)[2-5],其中利用自助法计算出的 *P*-值用自助 *P*-值(bootstrap *P*-value, BP)表示。但是杨子恒[6-7]在其有关分子进化树的综述和著作中总结并指出,理论研究和计算机模拟比较都表明,BP 与它所估计的"重建的树是真实树的概率"间存在着较大的差距,利用 BP 进行检验在统计上是保守的,即不易得出"估计的系统树受资料显著支持"的结论[8-9]。因此,提出了几种复杂的自助法以得到更高精度的研究结果,其中之一是多尺度自助法,Shimodaira[4]用渐近无偏(approximately unbiased,AU)表示其计算的 *P*-值。这个方法具有 3 次精度,计算量是 O(MB),其中 M 是重抽样的自助样本尺度的个数(重抽样的自助样本尺度是原样本容量 n 除以重抽样本容量 m 的比值),B 是重新采样的样本个数。这个方法已成功地在许多实际问题中得到应用[4]。然而,问题是它的计算量比较大。因此,研究者提出了高速双倍自助法[5],用高速双倍自助 *P*-值(speedy double boostrap *P*-value,SDBP)表示其 *P*-值。这个方法克服了计算 AU 速度慢和因外插估计出的数值不稳定的缺陷,解决了为获得 3 次精度 *P*-值所需计算量大的难题,且也是具有 3 次精度而计算量仅为 O(B)。在哺乳动物系统树的应用研究表明,3 次精度 *P*-值中 SDBP 计算速度是最快的[5]。

国外关于马系统发育的研究较早,但是对于蒙古马系统全面的研究比较少[10-14]。国内有关马和蒙古马的进化与遗传多态性的研究比较多[15-21]。2006 年李金莲[19]对内蒙古锡尼河马、巴尔虎马、乌审马、乌珠穆沁马 4 个类型的蒙古马及引入品种纯血马、培育品种三河马和地方品种广西矮马共 309 个样品的 mtDNA D-loop 序列利用非加权组平均法(unweighted pair-group method with arithmetic means, UPGMA)进行聚类分析得出,乌珠穆沁马和乌审马亲缘关系最近,锡尼河马和巴尔虎马亲缘关系较近,与二者次近的是三河马,它们与广西矮马和纯血马亲缘关系都较远。2010 年王小斌等[21]对蒙古马与世界范围内的家马、古马、普氏野马的 mtDNA D-loop 序列进行系统发育分析,并根据邻接法(neighbor joining,NJ)分析结果,共分出 A、B、C、D、E、F、G 7 个分支,其中蒙古马分在 A、B、C、D、E、F 6 个分支中。2013 年陈建兴等[20]针对亚欧马、蒙古马、普氏野马的 *ND*4 基因序列进行了遗传多样性分析,并根据 NJ 树分析结果得出,普氏野马和蒙古家马聚在一起,用数据证明了普氏野马和蒙古家马具有较近的亲缘关系。我国对于蒙古马系统发

育分析研究有一定的成果,但是建树方法多采用UPGMA 法和 NJ 法。杨子恒[6]指出,由于对绝大多数资料来讲分子钟的假定都不能成立,所以 UPGMA 法在使用中存在缺陷。另外,杨子恒[6]还指出,最大似然法由于计算量庞大,仅在少量研究中进行了比较,不过这些研究都表明其效率在现有方法中较高,比使用相同模型的 NJ 法或最小二乘距离矩阵法效率要高[22-24]。因此可以利用最大似然法估计蒙古马系统进化树,对估计出的系统进化树进行可靠性评价,因此,可靠性检验方法也需要进一步完善,推断出更精确的进化关系。

本研究以最大似然法估计蒙古马的系统发育树,然后对估计出的系统树利用高速双倍自助法计算其可靠度 SDBP 值,并给出了 SDBP 值最大的系统发育树。这将为蒙古马的适应性进化以及基因结构和功能的研究奠定基础。

1　材料与方法

1.1　试验材料

本研究从美国国家生物技术信息中心(National center of biotechnology information,NCBI)网站中获取了 30 匹中国蒙古家马与 6 匹普氏野马的 mtDNA D-loop 全序列(GenBank 登录号见表 1),这些序列包含巴尔虎马、三河马、乌审马、乌珠穆沁马以及锡尼河马 5 个类群。从每个类群中选取了 6 个样本。此外,普氏野马(P)有 6 个样本,共计 36 匹马。

表 1　蒙古马 D-loop 区碱基序列的 GenBank 登录号

Table 1　GenBank accession numbers of Mongolian horse D-loop sequences

品种 Breed	数量/匹 Quantity	GenBank 登录号 GenBank accession number
巴尔虎马 Baerhu	6	JN790867～ JN790872
三河马 Sanhe	6	JN790879～JN790884
乌审马 Wushen	6	JN790885～JN790890
乌珠穆沁马 Ujimqin	6	JN790897～JN790902
锡尼河马 Xinihe	6	JN790891～JN790896
普氏野马 Przewalski's	6	DQ017347、DQ017373、GU565340、GU565529、JN398402、JN395403

1.2　最大似然系统发育树的计算

设有一组生物种类数为 S,每种生物 DNA 的长度为 n 的碱基序列见表 2,为了一般性和简单明了,利用了虚拟的 DNA 碱基序列。在统计学中,记 S 种生物 DNA 碱基序列为矩阵 $X_{s\times n}=(x_1,\cdots,x_n)$,其中 $x_i(i=1,\cdots,n)$ 表示表 2 中 DNA 碱基序列的第 i 位点。S 种(不包含外群的种数)生物可有 $K=(2S-3)!!$ 棵候补无根系统树(本研究所建的树都是无根树),每棵系统树记为 $T_k(k=1,\cdots,K)$,第 k 棵系统树的概率函数:

$$f_k(x;\theta_k) \tag{1}$$

其中,k 表示第 k 棵系统树,θ_k 表示第 k 棵系统树的概率函数参向量,其包含碱基置换模型中的参数和树状拓扑中的若干个树枝长度参数,x 为 DNA 碱基序列中的一个位点所表示的碱基列。

表 2　S 种生物 DNA 碱基序列

Table 2　DNA base sequences of S species

生物的序号 Number of species	DNA 碱基序列 DNA base sequence									
	1	2	3	4	5	6	7	8	…	n
1	T	C	A	C	C	T	A	T		G
2	T	C	A	C	T	T	A	T	…	G
3	T	C	A	T	C	T	A	T	…	G
…	…	…	…	…	…	…	…	…	…	…
S	T	C	A	C	C	T	A	T	…	G

表中列的"1,2,3,…,S"表示生物的序号;行的"1,2,3,…,n"表示 DNA 碱基序列的位点序号

The "1, 2, 3,…, S" in the columns denote the serial number of species; the "1, 2, 3,…, n" in the rows denote the serial number of sites

根据以往的研究报道设定碱基置换模型中的参数，所以碱基置换模型是已知的，只有树枝长度是未知。本研究中的 S 种生物指的是 S=5 个类群的蒙古马，而 $K=(2\cdot5-3)!!=105$ 是 5 个类群蒙古马的所有候补树，5 个类群蒙古马的 DNA 碱基序列矩阵为 $X_{5\times309}$，其中位点总数 $n=309$（以下计算步骤中都设成 $S=5, K=105, X_{5\times309}$）。求解最大似然系统发育树有如下 4 个步骤。

步骤 1：计算每棵系统发育树的最大似然值。在实际的计算中先利用 Mega6 对表 2 的 DNA 碱基序列和外群的 DNA 碱基序列一同进行多重比对，比对后去掉有空位的列，然后得到处理后的 DNA 碱基序列，建一个文件夹作为输入文件。S 种（不包含外群的种数）生物的（2S-3）!! 棵不同树形拓扑按 Newick 格式全部手写出来，建一个文件夹作为另一个输入文件。利用软件 PAML4.9 以最大似然估计法计算每棵系统树的对数似然函数：

$$l_k=l_k(\theta_k;X_{s\times n})=\sum_{i=1}^{n}\log f_k(x_i;\theta_k) \tag{2}$$

其中，k 表示第 k 棵系统树，θ_k 表示树枝长度向量，x_i 表示 DNA 碱基序列中的第 i 位碱基列，$X_{s\times n}$ 表示 S 种生物 DNA 碱基序列矩阵，$f_k(x;\theta_k)$ 是第 k 棵系统树的概率函数。利用软件 PAML4.9 计算出的对数似然函数可估计出树枝长度向量 θ_k 的最大似然估计 $\hat{\theta}_k$：

$$\hat{\theta}_k=\arg\max_{\theta_k\in\Theta_k}\sum_{i=1}^{n}\log f_k(x_i;\theta_k),k=1,\cdots,K \tag{3}$$

其中，k、θ_k、x_i 所表示的含义与公式（2）相同，Θ_k 表示第 k 棵树的树枝长度向量 θ_k 所在的向量空间，K 表示所有可能的候补树。把 $\hat{\theta}_k$ 代入 $l_k=l_k(\theta_k;X_{s\times n})=\sum_{i=1}^{n}\log f_k(x_i;\theta_k)$ 中得到每棵系统树的对数似然估计值 $\hat{l}_k=\hat{l}_k(\hat{\theta}_k;X_{s\times n})$，简记为 $\hat{l}_k$。

步骤 2：建立最大似然矩阵。利用软件 PAML4.9 计算每棵系统树对于每个位点的对数似然估计值，按行的形式摆放，得到最大对数似然矩阵：

$$\begin{bmatrix}\log f_1(x_1;\hat{\theta}_1) & \cdots & \log f_1(x_n;\hat{\theta}_1)\\ \vdots & & \vdots\\ \log f_K(x_1;\hat{\theta}_K) & \cdots & \log f_K(x_n;\hat{\theta}_K)\end{bmatrix} \tag{4}$$

简记为

$$\begin{bmatrix}\hat{l}_{11} & \cdots & \hat{l}_{1n}\\ \vdots & & \vdots\\ \hat{l}_{K1} & \cdots & \hat{l}_{Kn}\end{bmatrix}$$

其中，$\hat{l}_{kj}=\log f_k(x_j;\hat{\theta}_k),k=1,\cdots,K;j=1,\cdots,n$ 是第 k 棵树的第 j 位点序列的最大对数似然估计值。

步骤 3：计算最大对数似然向量。利用软件 CONSEL 将公式（4）中每行数据相加，可得最大对数似然向量：

$$\hat{l}=\begin{bmatrix}\hat{l}_1\\ \vdots\\ \hat{l}_K\end{bmatrix} \tag{5}$$

其中，分量 $\hat{l}_k,k=1,\cdots,K$ 是公式（3）中的 $\hat{\theta}_k$ 代入 $l_k=l_k(\theta_k;X_{s\times n})=\sum_{i=1}^{n}\log f_k(x_i;\theta_k)$ 中得到的 $\hat{l}_k$。然后记 $\hat{l}_{\hat{k}}=\max\{\hat{l}_1,\hat{l}_2,\cdots,\hat{l}_K\}$ 为所有系统树最大对数似然值 $\hat{l}_k,k=1,\cdots,K$ 中的最大数值。

步骤 4：估计出最大似然树。由 $\hat{l}_k=\max\{\hat{l}_1,\hat{l}_2,\cdots,\hat{l}_K\}$ 得到最大似然树编号 $\hat{k}=\arg\max\{\hat{l}_1,\hat{l}_2,\cdots,\hat{l}_K\}$。

1.3 计算系统发育树的可靠度 SDBP 值

对 BP 以及 AU 的计算步骤这里省略掉。系统树可靠性评价首先要提出如下的假设，这个假设是属于多元假设检验问题。

$H_1:\mu_1=\max_{i=1,\cdots,K}\mu_i$，$H_1$ 表示第 1 棵系统树为真实系统树；

…

$H_K:\mu_K=\max_{i=1,\cdots,K}\mu_i$，$H_K$ 表示第 K 棵系统树为真实系统树。

其中，$\mu_K=E_q[l_k(\hat{\theta}_k;X'_{S\times n})],k=1,\cdots,K$ 表示第 k 棵系统树对数似然值的期望值，这里 $X'_{S\times n}$ 设为一组未知的 DNA 碱基序列；$\mu_k=E_q[l_k(\hat{\theta}_k;X'_{S\times n})]$ 中的 q 表示 $q(x)$，$q(x)$ 是每个位点碱基列 x 的真实分布，$l_k(\hat{\theta}_k;X'_{S\times n})$ 表示第 k 棵系统树的对数似然值，它与利用公式（3）估计出的 $\hat{l}_k=\hat{l}_k(\hat{\theta}_k;X_{s\times n})$ 是不同的未知对数似然值，$\hat{\theta}_k$ 是公式（3）中的已知向量。由于碱基位点序列 x 的真实分布 $q(x)$ 是未知的，所以 $l_k(\hat{\theta}_k;X'_{S\times n})$ 与 $\mu_k=E_q[l_k(\hat{\theta}_k;X'_{S\times n})]$，$k=1,\cdots,K$ 也都是未知的。以 H_1 为例说明计算 SDBP 的值，计算其他假设的 SDBP 值的步骤类似。

第 1 棵系统树为真实系统树的假设检验：

$H_1:\mu_1=\max\limits_{i=1,\cdots,K}\mu_i$ 为原假设,即认为第 1 棵系统树为真实系统树,否则为备择假设。

原假设 H_1 的概率值 SDBP 的计算步骤:

(1)计算每棵系统树的最大对数似然值 $\hat{l}_k,k=2,\cdots,K$ 与第 1 棵系统树 $\hat{l}_1$ 的差值,然后取最大值,即符号距离 d:

$$d=\max_{j=2,\cdots,K}\hat{l}_j-\hat{l}_1 \tag{6}$$

(2)重抽样

重抽样的样本取自 K 元正态总体 $N_K(\hat{\mu},\sum)$,即 $l^*\sim N_K(\hat{\mu},\sum)$。其中 $\hat{\mu}=(\hat{\mu},\cdots,\hat{\mu}_K)$ 表示公式(5)中的最大对数似然向量 $\hat{l}$ 在假设 H_1 所表示的领域 G 的边界上的射影,$\hat{\mu}=(\hat{\mu}_1,\cdots,\hat{\mu}_K)$ 的计算可用 PAVA (pool adjacent violators algorithm method)[25-26] 来计算:

$$\hat{\mu}_1=\frac{\sum_{j\in W}\hat{l}_j}{\sharp W},\hat{\mu}_j=\min(\hat{\mu}_1,\hat{\mu}_j),j=2,\cdots,K \tag{7}$$

其中,集合 W 是集合 $\{1,\cdots,K\}$ 的子集,并且包含元素 1,符号 $\sharp W$ 表示集合 W 所含元素的个数。另一方面,最大对数似然向量 $\hat{l}$ 的协方差矩阵 $\sum=(\sigma_{ij})$ 的元素 σ_{ij},可按如下的公式计算[27]:

$$V(\hat{l}_i)=n\times\frac{1}{n-1}\sum_{h=1}^{n}[\log f_i(x_h;\hat{\theta}_i)-\frac{1}{n}\sum_{h=1}^{n}\log f_i(x_h;\hat{\theta}_i)]^2 \tag{8}$$

$$Cov(\hat{l}_i,\hat{l}_j)=n\times\frac{1}{n-1}\sum_{h=1}^{n}\{[\log f_i(x_h;\hat{\theta}_i)-\frac{1}{n}\sum_{h=1}^{n}f_i(x_h;\hat{\theta}_i)]\times[\log f_j(x_h;\hat{\theta}_j)-\frac{1}{n}\sum_{h=1}^{n}\log f_j(x_h;\hat{\theta}_j)]\} \tag{9}$$

从总体 $N_K(\hat{\mu},\sum)$ 产生 B(一般设为1 000或10 000)个自助样本向量:

$$l_1^*=\begin{pmatrix}l_{11}^*\\ \vdots\\ l_{K1}^*\end{pmatrix},\cdots,l_B^*=\begin{pmatrix}l_{1B}^*\\ \vdots\\ l_{KB}^*\end{pmatrix} \tag{10}$$

(3)计算 SDBP 值

对每个自助样本 $l_b^*,b=1,\cdots,B$ 按公式(6)计算符号距离 $d_b^*,b=1,\cdots,B$,最终得到 SDBP 计算公式:

$$SDBP=\frac{\sharp\{d_b^*>d\}}{B},b=1,\cdots,B \tag{11}$$

其中 d 是公式(6)表示的量。

2　结　果

2.1　蒙古马系统进化树可靠度分析

将获取的 36 匹马的 D-loop 碱基全序列导入 Mega6 软件中,对所有马匹 D-loop 区片段序列进行多重序列比对,比对后把空位的位点全部删除,结果每匹马的碱基长度为 309 bp。

2.1.1　PAML4.9 软件输出最大似然进化树

将多重比对之后的碱基序列和所有的 105 棵候补树组成的两个文件夹输入 PAML4.9 软件中的 baseml 程序,并选择 REV 碱基替换模型分析。运行程序 baseml 后可得 105 棵树每个位点的对数似然值,即公式(4)。这些对数似然值可得出 105 棵树中的最大似然树的序号。

2.1.2　CONSEL、R3.0 中 SDBP 包给出各系统树的可靠度　　将所得 105 棵树的每个位点的对数似然值,利用 CONSEL 软件,计算各个系统树的 AU 值和 BP 值。最后利用 R3.0 中 SDBP 包计算各个树的 SDBP 值,可得 SDBP 值最大的蒙古马系统发育进化树。

表 3 展示了 SDBP 值大于等于 0.05 的 60 棵系统进化树的 SDBP 值,旁边的数值是 AU 值和 BP 值。表 3 中"排序"表示按每棵系统进化树的最大对数似然值降序生成的树的排序,"树形"表示 PAML4.9 软件中输入 105 棵系统进化树建立的树文件中系统树的序号。综合考察结果,按系统树的 SDBP 值分析系统树,序号为 100 的系统树的 SDBP 值最大为 0.980。而按系统树的 AU 值和 BP 值分析进化树,序号为 66 和 19 的可靠度最大,分别为 0.717 和 0.215。采用最大似然法结合 SDBP 值分析蒙古马的系统拓扑关系是较最大似然法结合渐近无偏(approximately unbiased,AU)值以及结合自助(bootstrap P-value,BP)值分析蒙古马系统拓扑关系更有效。将 t100 号、t66 号以及 t19 号系统树的拓扑结构画成二分系统树的树形,得到图 1、图 2、图 3。

表 3 30 匹蒙古马的 60 棵候选树 SDBP、AU 以及 BP 值的计算结果

Table 3 The results of SDBP, AU and BP values of 60 candidate phylogenetic trees for 30 Mongolian horses

排序 Rank	树形 Tree	SDBP 值 SDBP value	AU 值 AU value	BP 值 BP value	排序 Rank	树形 Tree	SDBP 值 SDBP value	AU 值 AU value	BP 值 BP value
1	100	0.980	0.697	0.026	31	4	0.440	0.012	0.000
2	66	0.933	0.717	0.096	32	79	0.348	0.093	0.000
3	19	0.751	0.563	0.215	33	70	0.351	0.097	0.000
4	89	0.777	0.570	0.161	34	102	0.345	0.098	0.000
5	30	0.769	0.499	0.101	35	41	0.353	0.095	0.000
6	78	0.776	0.497	0.101	36	9	0.339	0.081	0.000
7	43	0.778	0.498	0.101	37	53	0.352	0.096	0.000
8	103	0.768	0.366	0.003	38	5	0.348	0.059	0.000
9	37	0.771	0.496	0.101	39	63	0.327	0.000	0.101
10	24	0.780	0.160	0.005	40	42	0.329	0.000	0.005
11	49	0.774	0.496	0.101	41	35	0.329	0.000	0.101
12	99	0.649	0.446	0.210	42	52	0.337	0.000	0.210
13	73	0.649	0.447	0.209	43	50	0.339	0.000	0.209
14	82	0.596	0.374	0.148	44	58	0.337	0.000	0.148
15	77	0.678	0.154	0.003	45	93	0.340	0.000	0.003
16	16	0.539	0.144	0.009	46	6	0.331	0.000	0.009
17	55	0.558	0.234	0.050	47	21	0.335	0.000	0.050
18	32	0.559	0.238	0.051	48	83	0.336	0.000	0.051
19	74	0.565	0.237	0.051	49	17	0.326	0.000	0.051
20	76	0.559	0.237	0.051	50	22	0.336	0.000	0.051
21	13	0.564	0.199	0.002	51	20	0.333	0.000	0.002
22	44	0.557	0.128	0.001	52	69	0.307	0.000	0.001
23	94	0.558	0.134	0.000	53	60	0.300	0.000	0.000
24	25	0.582	0.226	0.000	54	23	0.296	0.000	0.000
25	26	0.577	0.221	0.000	55	87	0.306	0.033	0.000
26	92	0.573	0.002	0.000	56	98	0.313	0.031	0.000
27	34	0.576	0.002	0.000	57	72	0.312	0.032	0.000
28	95	0.584	0.232	0.000	58	10	0.316	0.032	0.000
29	7	0.580	0.231	0.000	59	14	0.309	0.032	0.000
30	62	0.583	0.235	0.000	60	12	0.312	0.032	0.000

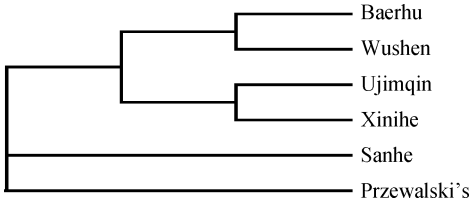

图 1 SDBP 值最大系统树 t100 拓扑结构

Fig. 1 The topology of tree t100 with the highest SDBP (speedy double bootstrap *P*-value) value

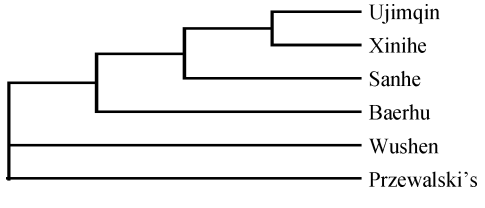

图 2 AU 值最大系统树 t66 拓扑结构

Fig. 2 The topology of tree t66 with the highest AU (approximately unbiased) value

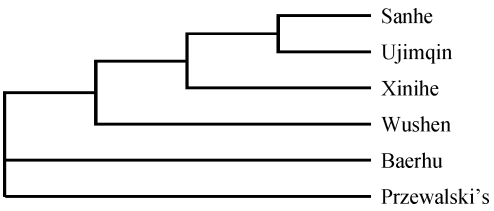

图 3　BP 值最大系统树 t19 拓扑结构
Fig. 3　The topology of tree t19 with the highest BP (bootstrap *P*-value) value

SDBP 值最大的蒙古马系统树(图 1)共有 3 个分支,巴尔虎马与乌审马合并为一个分支,具有较近的遗传关系;乌珠穆沁马与锡尼河马具有较近的遗传关系;三河马与其他 4 种蒙古马独立地分出来构成一个系统发育分支。SDBP 值最大的蒙古马系统树与三河马和其他 4 类蒙古马具有较远的预测血缘关系是一致的,同时此结果的系统树也具有最大似然值。

2.2　蒙古马 NJ 树与 UPGMA 树分析

将多重比对之后的 36 个碱基序列输入 Mega6 软件中,利用 Mega6 软件工具栏中 Data 下的 Setup/Select Taxa and Groups 标签对 36 个碱基序列按表 1 分组,结果得到带有 6 组分组信息的 36 个碱基序列并保存。然后对带有分组信息的 36 个碱基序列,利用工具栏中 Data 下的 Setup/Compute Between and Groups Means 标签计算 6 组间的距离。其次对组间距离的结果利用工具栏中 Phylogeny 下的 Construct Phylogeny 标签采用 NJ 方法和 UPGMA 方法分别建立系统树。最后得到蒙古马的 NJ 系统树和 UPGMA 系统树,把它们画成二分系统树的树形,得到图 4 和图 5。结合 SDBP 值分析蒙古马的系统拓扑关系较非加权组平均法(unweighted pair-group method with arithmetic means, UPGMA) 有效,而与邻接法(neighbor joining, NJ)得到了相近的结论。

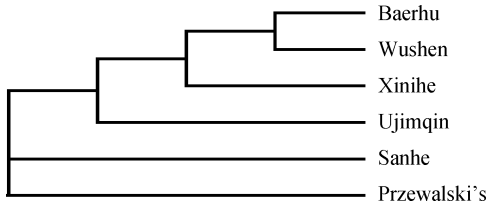

图 4　5 个蒙古马类群的 NJ 树拓扑结构
Fig. 4　The topology of NJ trees of 5 Mongolian horse subspecies

这次的试验对进化时间没有做分析,只是对拓扑结构做的分析。所有 105 棵树都是以普氏野马为外群的无根树,所以图 1～图 5 中的进化树是没有时间标尺和时间方向的进化系统树。

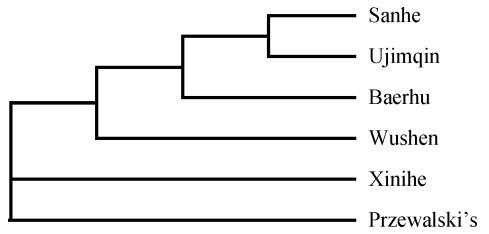

图 5　5 个蒙古马类群的 UPGMA 树拓扑结构
Fig. 5　The topology of UPGMA tree of 5 Mongolian horse subspecies

3　讨　论

从统计学假设检验理论来判断,每棵系统发育树的 SDBP 值相当于原假设 H_k 的 P- 值,$k=1,\cdots,105$。所以如果以显著性水平检验各原假设的话,各原假设的 SDBP 值大于或等于 0.05 的 60 个系统树都不被拒绝。如何从这 60 个系统树中找出真实的系统树是分子进化学中的一个重要问题。一般采用生物学等其他信息在选择过的范围里再选择。在 SDBP 值大于或等于 0.05 的 60 个系统树都不被拒绝的情况下,我们认为结合一个或几个种群马与其他种群马的遗传关系的已知信息,从 60 棵系统树中判断出一个符合这个已知信息的系统树作为蒙古马的系统进化树在逻辑与统计学方法上都是正确的。所以根据已有研究报道,三河马的血缘主要是在后贝加尔马和其杂种马的基础上,混入呼伦贝尔草原蒙古马的血液,经几十年精心培育而成[28]。根据三河马这样的血缘关系,预测三河马应该从系统进化发育树首先分离出来独立作为一个分支,而其他4 种蒙古马种群作为另一个分支。从 t100、t66 以及 t19 3 个系统树看符合这一条件的只有 t100,而 t100 是最大似然树,其 SDBP 值也是最大的。所以,笔者认为 t100 比较真实地反映了蒙古马的系统进化关系。对比 3 次精度 AU 值最大的系统树 t66 与SDBP值最大的 t100,它们都有以乌珠穆沁马与锡尼河马为叶子组成的一个独立小分支,而在 BP 值最大的 t19 里没有这样的独立小分支。

对比 2006 年李金莲[19]用 UPGMA 的结果与本研究中 3 次精度 SDBP 值最大的 t100,它们没有共同的由两个叶子组成的独立小分支,在拓扑结构上

也不一样。2006 年李金莲[19]用 UPGMA 的结果与 3 次精度的 AU 值最大的系统树 t66 以及 BP 值最大的 t19 相比,无论在拓扑结构还是独立分支上也没有相似之处。由于 2010 年王小斌等[21]是把蒙古马与世界范围内的家马、古马、普氏野马一起进行系统发育分析,即没有进行分类分析,所以本研究结果与王小斌等[21]的 NJ 树结果没有可比性,陈建兴等[20]的结果也存在同样的问题,因而也无可比性。对比本研究用 NJ 方法与 UPGMA 分析出的结果(图 4 与图 5)可知,蒙古马的 NJ 树与 SDBP 值最大的 t100 树,它们的分支在系统树中的顺序完全一样,只是系统树拓扑结构上有些差异。而蒙古马的 UPGMA 树拓扑关系与 BP 值最大的 t19 树非常接近。这说明在每匹马序列长度比较少的情况下 NJ 树的结果要比 UPGMA 树的结果更接近 SDBP 值最大的系统树。综上所述,本研究通过利用 30 匹蒙古马线粒体 DNA 中的 D-loop 高变区序列对 5 类蒙古马的 105 个候补系统发育树进行分析,结果表明,采用最大似然法结合 SDBP 值分析生物的系统拓扑关系是较最大似然法结合 AU 值以及结合 BP 值分析生物的系统拓扑关系更有效的方法。同时也比 UPGMA 法有效,而与 NJ 法得到的结论相似。

另外,当种与种之间分类数或种内部的分类数超过 6 类(包括外群),例如是 7 类时,利用最大似然法估计最大似然树时,所有候补树的个数为 945 个。由于树个数的增加,导致最大似然法估计的计算量也增加。所以当分类数超过 6 类时基于最大似然法的 AU 和 SDBP 方法的计算量也很大,这也是 AU 和 SDBP 方法的限制所在。而 BP 方法除了在最大似然法的应用外,也可以在 NJ 法以及 UPGMA 法中使用。另一方面,SDBP 方法中的 $\hat{\mu}=(\hat{\mu}_1,\cdots,\hat{\mu}_K)$是最大对数似然向量 $\hat{l}$在假设 H_1 所表示的领域 G 的边界上的射影,$\hat{\mu}=(\hat{\mu}_1,\cdots,\hat{\mu}_K)$的计算在其他系统树构树方法如 NJ 法与 UPGMA 中很难估计。在这些情况下,更适合用 AU 或 BP 计算系统树的可靠度。

4 结 论

本研究利用 30 匹蒙古马的线粒体 DNA 中的 D-loop 高变区序列对蒙古马的 105 个候补系统发育树进行了分析,采用最大似然法,并利用生物信息软件 Mega6、PAMAL4.9 等估计出蒙古马的最大似然树。最后利用 CONSEL 和 R3.0 中的 SDBP 软件包等计算了 105 个候补系统发育树的 SDBP 值。结合三河马的特殊血缘关系分析出:除去普氏野马共有 3 个分支,巴尔虎马与乌审马合并为一个分支,具有较近的遗传关系;乌珠穆沁马与锡尼河马具有较近的遗传关系;三河马与其他 4 种蒙古马独立地分出来构成一个系统发育分支。通过对蒙古马的分析表明,采用最大似然法结合 SDBP 值分析生物的系统拓扑关系是较最大似然法结合 AU 值以及结合 BP 值分析生物的系统拓扑关系更有效的方法。同时也比 UPGMA 法有效,而与 NJ 法得到相似的结论。

参考文献(References):

[1] 吕宝忠.分子进化树的构建[J].动物学研究,1993,14(2):186-193.
LÜ B Z. The construction of molecular evolutionary trees[J]. *Zoological Research*,1993,14(2):186-193. (in Chinese)

[2] FELSENSTEIN J. Confidence limits on phylogenies: an approach using the bootstrap[J]. *Evolution*,1985,39(4):783-791.

[3] SHIMODAIRA H, HASEGAWA M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference[J]. *Mol Biol Evol*,1999,16(8):1114-1116.

[4] SHIMODAIRA H. An approximately unbiased test of phylogenetic tree selection[J]. *Syst Biol*,2002,51(3):492-508.

[5] REN A Z,ISHIDA T,AKIYAMA Y. Assessing statistical reliability of phylogenetic trees via a speedy double bootstrap method[J]. *Mol Phylogenet Evol*,2013,67(2):429-435.

[6] 杨子恒.分子进化树的统计推断[J].遗传,1995,17(S1):92-96.
YANG Z H. Statistical estimation of molecular evolutionary trees[J]. *Hereditas (Beijing)*,1995,17(S1):92-96. (in Chinese)

[7] 杨子恒.计算分子进化[M].钟扬,译.上海:复旦大学出版社,2008:207-215.
YANG Z H. Computational molecular evolution[M]. ZHONG Y, trans. Shanghai: Fudan University Press,2008:207-215. (in Chinese)

[8] ZHARKIKH A, LI W H. Statistical properties of bootstrap estimation of phylogenetic variability from nucleotide sequences. II. Four taxa without a molecular clock[J]. *J Mol Evol*,1992,35(4):356-366.

[9] HILLIS D M, BULL J J. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis[J]. *Syst Biol*,1993,42(2):

182-192.

[10] XU X F, ÁRNASON Ú. The complete mitochondrial DNA sequence of the horse, *Equus caballus*: extensive heteroplasmy of the control region[J]. *Gene*, 1994, 148(2):357-362.

[11] ISHIDA N, OYUNSUREN T, MASHIMA S, et al. Mitochondrial DNA sequences of various species of the genus *Equus* with special reference to the phylogenetic relationship between Przewalskii's wild horse and domestic horse[J]. *J Mol Evol*, 1995, 41(2): 180-188.

[12] KIM K I, YANG Y H, LEE S S, et al. Phylogenetic relationships of Cheju horses to other horse breeds as determined by mtDNA D-loop sequence polymorphism[J]. *Anim Genet*, 1999, 30(2):102-108.

[13] PETERSENJ L, MICKELSON J R, RENDAHL A K, et al. Genome-wide analysis reveals selection for important traits in domestic horse breeds[J]. *PLoS Genet*, 2013, 9(1): e1003211.

[14] SCHUBERT M, JÓNSSON H, CHANG D, et al. Prehistoric genomes reveal the genetic foundation and cost of horse domestication[J]. *Proc Natl Acad Sci U S A*, 2014, 111(52): E5661-E5669.

[15] 李金莲,芒　来,石有斐. 利用微卫星标记对蒙古马和纯血马遗传多样性的研究[J]. 畜牧兽医学报, 2005, 36(1):6-9.
LI J L, MANG L, SHI Y F. Evaluation of genetic diversity of Mongolian horse and thoroughbred horse using microsatellite markers[J]. *Acta Veterinaria et Zootechnica Sinica*, 2005, 36(1):6-9. (in Chinese)

[16] 任秀娟,赵一萍,萨如拉,等. 马属动物进化特征概述[J]. 畜牧兽医学报, 2017, 48(3):385-392.
REN X J, ZHAO Y P, SARULA, et al. The review for the characteristics of the equus evolution[J]. *Acta Veterinaria et Zootechnica Sinica*, 2017, 48(3):385-392. (in Chinese)

[17] 赵启南,芒　来,白东义,等. 蒙古马高负荷运动训练前后转录组差异表达分析[J]. 畜牧兽医学报, 2017, 48(6):1007-1016.
ZHAO Q N, MANG L, BAI D Y, et al. Mongolian horses transcriptome differential expression analysis before and after a high load exercise training[J]. *Acta Veterinaria et Zootechnica Sinica*, 2017, 48(6):1007-1016. (in Chinese)

[18] 芒　来,杨　虹. 蒙古马遗传多样性研究进展[J]. 遗传, 2008, 30(3):269-276.
MANG L, YANG H. Progress in the study of genetic diversity of Mongolian horse[J]. *Hereditas (Beijing)*, 2008, 30(3):269-276. (in Chinese)

[19] 李金莲. 中国蒙古马遗传多样性和分子系统进化研究[D]. 呼和浩特:内蒙古农业大学, 2006.
LI J L. Study on genetic diversity and molecular phylogeny evolution in Chinese Mongolian horse[D]. Hohhot: Inner Mongolia Agricultural University, 2006. (in Chinese)

[20] 陈建兴,孙玉江,乌尼尔夫,等. 基于 ND4 基因分析家马的母系起源[J]. 黑龙江畜牧兽医, 2013(3): 8-10, 14.
CHEN J X, SUN Y J, WUNIERFU, 等. Analysis of the maternal origins of domestic horses based on ND4 gene[J]. *Heilongjiang Animal Science and Veterinary Medicine*, 2013(3):8-10, 14. (in Chinese)

[21] 王小斌,秦　芳,张云生,等. 蒙古马 mtDNA D-loop 区遗传多样性与多重母系起源[J]. 西北农林科技大学学报:自然科学版, 2010, 38(3):47-51.
WANG X B, QIN F, ZHANG Y S, et al. Mitochondrial DNA D-loop genetic diversity and multiple maternal origins in Mongolian horses[J]. *Journal of Northwest A & F University: Natural Science Edition*, 2010, 38(3):47-51. (in Chinese)

[22] FUKAM-KOBAYASHI K, TATENO Y. Robustness of maximum likelihood tree estimation against different patterns of base substitutions[J]. *J Mol Evol*, 1991, 32(1):79-91.

[23] HASEGAWA M, KISHINO H, SAITOU N. On the maximum likelihood method in molecular phylogenetics[J]. *J Mol Evol*, 1991, 32(5):443-445.

[24] YANG Z H. JSTOR: systematic biology[J]. *Syst Biol*, 1994, 43(3):329-342.

[25] AYER M, BRUNK H D, EWING G M, et al. An empirical distribution function for sampling with incomplete information[J]. *Ann Math Stat*, 1955, 26(4): 641-647.

[26] ZHAO H B. Comparing several treatments with a control[J]. *J Stat Plan Inference*, 2007, 137(9): 2996-3006.

[27] KISHINO H, MIYATA T, HASEGAWA M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts[J]. *J Mol Evol*, 1990, 31(2): 151-160.

[28] 陈建兴. 中国蒙古马的遗传多样性与系统发育及起源研究[D]. 呼和浩特:内蒙古农业大学, 2012.
CHEN J X. Investigation of the genetic diversity, phylogeny and origins of Chinese Mongolian horses[D]. Hohhot: Inner Mongolia Agricultural University, 2012. (in Chinese)

(编辑　郭云雁)