doi:10.3969/j.issn.1672-5565.2013.02.04

基于 k-mer 组分信息的系统发生树构建方法

刘红梅[1],刘国庆[2,3]*

(1. 乌海市人民医院呼吸科,内蒙古 乌海 016000; 2. 内蒙古科技大学数理与生物工程学院,内蒙古 包头 014010;
3. 内蒙古科技大学生物工程与技术研究所,内蒙古 包头 014010)

摘 要:随着越来越多基因组的测序完成,基于全基因组的非比对的系统发生分析已成为研究热点。不同的生物物种或个体基因组之间的核酸组分不完全相同。遗传语言-DNA 序列的信息很大程度上反映在其 k-mer 频数中。基于基因组序列 k-mer 频数的系统发生树则从新的角度为我们提供物种之间的亲缘关系。本文定义基于 k-mer 频数的信息参数,并用它表征基因组序列,计算不同基因组之间信息参数的距离,用邻接法对84个病毒构建了系统发生树,发现构建的系统发生树很大程度上与已有的系统发生树相吻合。

关键词:系统发生树;k-mer 频数;距离矩阵

中图分类号:Q61 **文献标识码**:A **文章编号**:1672-5565(2013)-02-100-05

A method for constructing phylogenetic tree based on k-mer information

LIU Hong-mei[1], LIU Guo-qing[2,3]*

(1. *Department of respiration, People's Hospital, Wuhai* 016000, *China*; 2. *School of Mathematics, Physics and Biological Engineering, Inner Mongolia University of Science and Technology, Baotou* 014010, *China*; 3. *The Institute of Bioengineering and Technology, Inner Mongolia University of Science and Technology, Baotou* 014010, *China*)

Abstract: With the success in the sequencing of complete genomes, the phylogenetics analysis by alignment-free methods based on complete genomes has been a hot topic. The nucleotide composition is different across species or populations. The information in the genetic language-DNA can be reflected largely in its k-mer frequencies. The phylogenetic tree based on k-mer frequencies would provide us the evolutionary relation among organisms from a novel perspective. In this study, the genomes of 84 large viruses are characterized by an information parameter, which is defined based on k-mer frequencies in the sequences; then the distances among the virus genomes are calculated and a phylogenetic tree is constructed for the viruses by using neighbor-joining method. The obtained phylogenetic tree is largely in agreement with the others' tree.

Keywords: Phylogenetic Tree; k-mer Frequency; Distance Matrix

1 引言

系统发生学(phylogenetics)是从生物大分子的信息确定不同生物之间亲缘关系的学科。系统发生树的构建方法有两大类:基于序列比对(alignment)的方法和基于全基因组的非比对方法。随着基因系统发生学研究快速的发展,基于比对的系统发生分析也逐渐暴露出它的局限性,如基于比对的方法无法规避单基因或基因家族给进化研究带来的不确定性。随着越来越多的基因组测序完成,基于全基因组的非比对的系统发生分析已成为研究热点[1]。不同的生物物种或个体基因组之间的核酸组分不完全相同。遗传语言-DNA 序列的信息很大程度上反映在其 k-mer 频数中。基于基因组序列 k-mer 频数的系统发生树从新的角度为我们提供物种之间的亲

收稿日期:2011-09-01;**修回日期**:2012-10-03.

基金项目:国家自然科学基金(61102162)、内蒙古自治区高等学校科学研究项目(NJ10098)和内蒙古科技大学创新基金(2009NC005)资助。

作者简介:刘红梅,女,内蒙古通辽人,乌海市人民医院呼吸科医生。

*通讯作者:刘国庆,副教授,硕士生导师,Tel: 0472-5954358, E-mail: gqliu1010@163.com.

缘关系。

本文定义基于 k-mer 频数的信息参数,用基于 256 种 4-mer 的信息参数表征基因组序列,计算信息参数在不同基因组之间的距离,用邻接法对 84 个病毒构建了系统发生树,通过对比分析证实了这种系统发生树的可靠性。

2　材料与方法

2.1　材料

从 NCBI(ftp://ftp.ncbi.nih.gov/genomes)下载了 328 个病毒基因组全序列。为避免统计序列中 k-mer 频数时序列长度引起的涨落,从 328 个病毒基因组中进一步筛选出基因组长度大于 10kb 的 DNA 病毒序列。为了进化树的比较起见,我们从大于 10kb 的 DNA 病毒序列中挑选出 84 条在文献[2]中提到的序列,其详细信息见表 1。

表 1　84 个双链 DNA 病毒的名称、缩写和 NCBI 收录号等信息

Table 1　84 dsDNA virus names, abbreviations, and NCBI accession numbers

Species	Abbr.	Accession	Family	Genus
Bovine adenovirus D	BAdV－4	NC_002685.2	Adenoviridae	Atadenovirus
Ovine adenovirus D	OAdV－D	NC_004037.1	Adenoviridae	Atadenovirus
Duck adenovirus A	DAdV－A	NC_001813.1	Adenoviridae	Atadenovirus
Fowl adenovirus A	FAdV－A	NC_001720.1	Adenoviridae	Aviadenovirus
Fowl adenovirus D	FAdV－D	NC_000899.1	Adenoviridae	Aviadenovirus
Bovine adenovirus B	BAdV－B	NC_001876.1	Adenoviridae	Mastadenovirus
Canine adenovirus	CAdV	NC_001734.1	Adenoviridae	Mastadenovirus
Human adenovirus A	HAdV－A	NC_001460.1	Adenoviridae	Mastadenovirus
Human adenovirus C	HAdV－C	NC_001405.1	Adenoviridae	Mastadenovirus
Human adenovirus D	HAdV－D	NC_002067.1	Adenoviridae	Mastadenovirus
Human adenovirus E	HAdV－E	NC_003266.2	Adenoviridae	Mastadenovirus
Murine adenovirus A	MAdV－A	NC_000942.1	Adenoviridae	Mastadenovirus
Ovine adenovirus A	OAdV－A	NC_002513.1	Adenoviridae	Mastadenovirus
Porcine adenovirus C	PAdV－C	NC_002702.1	Adenoviridae	Mastadenovirus
Bovine adenovirus A	BAdV－A	NC_006324.1	Adenoviridae	Mastadenovirus
Human adenovirus F	HAdV－F	NC_001454.1	Adenoviridae	Mastadenovirus
Porcine adenovirus A	PAdV－A	NC_005869.1	Adenoviridae	Mastadenovirus
Tree shrew adenovirus	TSAdV	NC_004453.1	Adenoviridae	Mastadenovirus
Turkey adenovirus A	TAdV－A	NC_001958.1	Adenoviridae	Siadenovirus
African swine fever virus	ASFV	NC_001659.1	Asfarviridae	Asfivirus
Adoxophyes orana granulovirus	AdorGV	NC_005038.1	Baculoviridae	Granulovirus
Agrotis segetum granulovirus	AsGV	NC_005839.2	Baculoviridae	Granulovirus
Cryptophlebia leucotreta granulovirus	CrleGV	NC_005068.1	Baculoviridae	Granulovirus
Cydia pomonella granulovirus	CpGV	NC_002816.1	Baculoviridae	Granulovirus
Phthorimaea operculella granulovirus	PhopGV	NC_004062.1	Baculoviridae	Granulovirus
Plutella xylostella granulovirus	PlxyGV	NC_002593.1	Baculoviridae	Granulovirus
Xestia c－nigrum granulovirus	XecnGV	NC_002331.1	Baculoviridae	Granulovirus
Autographa californica nucleopolyhedrovirus	AcMNPV	NC_001623.1	Baculoviridae	Nucleopolyhedrovirus
Bombyx mori nucleopolyhedrovirus	BmNPV	NC_001962.1	Baculoviridae	Nucleopolyhedrovirus
Helicoverpa armigera nucleopolyhedrovirus G4	HearNPVG4	NC_002654.1	Baculoviridae	Nucleopolyhedrovirus
Gallid herpesvirus 1	GaHV－1	NC_006623.1	Herpesviridae	Iltovirus
Gallid herpesvirus 2	GaHV－2	NC_002229.2	Herpesviridae	Mardivirus
Gallid herpesvirus 3	GaHV－3	NC_002577.1	Herpesviridae	Mardivirus
Meleagrid herpesvirus 1	MeHV－1	NC_002641.1	Herpesviridae	Mardivirus
Human herpesvirus 1	HHV－1	NC_001806.1	Herpesviridae	Simplexvirus
Human herpesvirus 2	HHV－2	NC_001798.1	Herpesviridae	Simplexvirus
Cercopithecine herpesvirus 2	CeHV－2	NC_006560.1	Herpesviridae	Simplexvirus
Bovine herpesvirus 1	BoHV－1	NC_001847.1	Herpesviridae	Varicellovirus
Bovine herpesvirus 5	BoHV－5	NC_005261.1	Herpesviridae	Varicellovirus
Cercopithecine herpesvirus 9	CHV－7	NC_002686.1	Herpesviridae	Varicellovirus
Equid herpesvirus 1	EHV－1	NC_001491.2	Herpesviridae	Varicellovirus

(续表 1)

Species	Abbr.	Accession	Family	Genus
Equid herpesvirus 4	EHV - 4	NC_001844.1	Herpesviridae	Varicellovirus
Suid herpesvirus 1	SuHV - 1	NC_006151.1	Herpesviridae	Varicellovirus
Murid herpesvirus 1	MuHV - 1	NC_004065.1	Herpesviridae	Muromegalovirus
Murid herpesvirus 2	MuHV - 2	NC_002512.2	Herpesviridae	Muromegalovirus
Human herpesvirus 6B	HHV - 6B	NC_000898.1	Herpesviridae	Roseolovirus
Human herpesvirus 7	HHV - 7	NC_001716.2	Herpesviridae	Roseolovirus
Tupaiid herpesvirus 1	TuHV - 1	NC_002794.1	Herpesviridae	unclassified Betaherpesvirinae
Callitrichine herpesvirus 3	CalHV - 3	NC_004367.1	Herpesviridae	Lymphocryptovirus
Human herpesvirus 4	HHV - 4	NC_001345.1	Herpesviridae	Lymphocryptovirus
Alcelaphine herpesvirus 1	AlHV - 1	NC_002531.1	Herpesviridae	Rhadinovirus
Bovine herpesvirus 4	BoHV - 4	NC_002665.1	Herpesviridae	Rhadinovirus
Equid herpesvirus 2	EHV - 2	NC_001650.1	Herpesviridae	Rhadinovirus
Human herpesvirus 8	HHV - 8	NC_003409.1	Herpesviridae	Rhadinovirus
Murid herpesvirus 4	MuHV - 4	NC_001826.1	Herpesviridae	Rhadinovirus
Saimiriine herpesvirus 2	SaHV - 2	NC_001350.1	Herpesviridae	Rhadinovirus
Ictalurid herpesvirus 1	IcHV - 1	NC_001493.1	Herpesviridae	Ictalurivirus
Ostreid herpesvirus 1	OsHV - 1	NC_005881.1	Herpesviridae	Unassigned Herpesviridae
Psittacid herpesvirus 1	PsHV - 1	NC_005264.1	Herpesviridae	Unassigned Herpesviridae
Ateline herpesvirus 3	AtHV - 3	NC_001987.1	Herpesviridae	unclassified Herpesviridae
Invertebrate iridescent virus 6	IIV - 6	NC_003038.1	Iridoviridae	Iridovirus
Lymphocystis disease virus - isolate China	LCDV - IC	NC_005902.1	Iridoviridae	Lymphocystivirus
Lymphocystis disease virus 1	LCDV - 1	NC_001824.1	Iridoviridae	Lymphocystivirus
Infectious spleen and kidney necrosis virus	ISaKNV	NC_003494.1	Iridoviridae	Megalocytivirus
Frog virus 3	FV - 3	NC_005946.1	Iridoviridae	Ranavirus
Singapore grouper iridovirus	SiGV	NC_006549.1	Iridoviridae	Ranavirus
Shrimp white spot syndrome virus	WSSV	NC_003225.1	Nimaviridae	Whispovirus
Paramecium bursaria Chlorella virus 1	PBCV - 1	NC_000852.3	Phycodnaviridae	Chlorovirus
Canarypox virus	CNPV	NC_005309.1	Poxviridae	Avipoxvirus
Fowlpox virus	FWPV	NC_002188.1	Poxviridae	Avipoxvirus
Myxoma virus	MYXV	NC_001132.2	Poxviridae	Leporipoxvirus
Rabbit fibroma virus	SFV	NC_001266.1	Poxviridae	Leporipoxvirus
Camelpox virus	CMLV	NC_003391.1	Poxviridae	Orthopoxvirus
Cowpox virus	CPXV	NC_003663.2	Poxviridae	Orthopoxvirus
Ectromelia virus	ECTV	NC_004105.1	Poxviridae	Orthopoxvirus
Vaccinia virus	VACV	NC_001559.1	Poxviridae	Orthopoxvirus
Variola virus	VARV	NC_001611.1	Poxviridae	Orthopoxvirus
Bovine papular stomatitis virus	BPSV	NC_005337.1	Poxviridae	Parapoxvirus
Orf virus	ORFV	NC_005336.1	Poxviridae	Parapoxvirus
Swinepox virus	SWPV	NC_003389.1	Poxviridae	Suipoxvirus
Yaba monkey tumor virus	YMTV	NC_005179.1	Poxviridae	Yatapoxvirus
Yaba - like disease virus	YDV	NC_002642.1	Poxviridae	Yatapoxvirus
Melanoplus sanguinipes entomopoxvirus	MSEV	NC_001993.1	Poxviridae	Betaentomopoxvirus
Heliothis zea virus 1	HZV - 1	NC_004156.1		unclassified dsDNA viruses

2.2 方法

2.2.1 相对偏离度

如果把长度为 k 的核苷酸片段看作是一种“字”(k-mer),那么 k-mer 的频数就是长度为 k 的窗口在核苷酸序列上顺次移动时出现的次数。当 k 较大时, k-mer 的频数分布,构成了基因组的一个“等价表示”,即从 k-mer 的频数分布可以唯一地确定基因组序列[3-4]。组成 DNA 序列的核苷酸有 4 种:A(腺嘌呤),G(鸟嘌呤),T(胸腺嘧啶)和 C(胞嘧啶)。研究表明,相比其它 k-mer,4-mer 频数分布最能表征物种基因组[5]。

定义基于 4-mer 的关联信息指标。对于一条 DNA 序列,度量其 256 种 4-mer 使用频数的总体偏好性的指标(也称作相对偏离度)定义为:

$$RD_{ijkl} = \frac{p_{ijkl} - p_i p_j p_k p_l}{p_i p_j p_k p_l} \quad (1)$$

其中四联体 $ijkl$ 中核苷酸 i、j、k、l 的出现频率分别为 p_i、p_j、p_k 和 p_l,四联体 $ijkl$ 在序列中单碱基步长的扫描中出现的频率为 p_{ijkl}。相对偏离度 RD_{ijkl} 表征序列中每一种四联体的出现频率相对于独立序列的偏离程度,其中 $p_{ijkl} = p_i p_j p_k p_l$ 时 $RD_{ijkl} = 0$,表示这段序列中该四联体的使用次数没有偏好性,即序列中这种四联体的出现频率与独立序列(即单碱基组分约束下的随机组合序列)中出现的频率是相同的。可根据 RD_{ijkl} 值大于还是小于零来判断某一特定 4-mer 在序列中过多或过少出现。

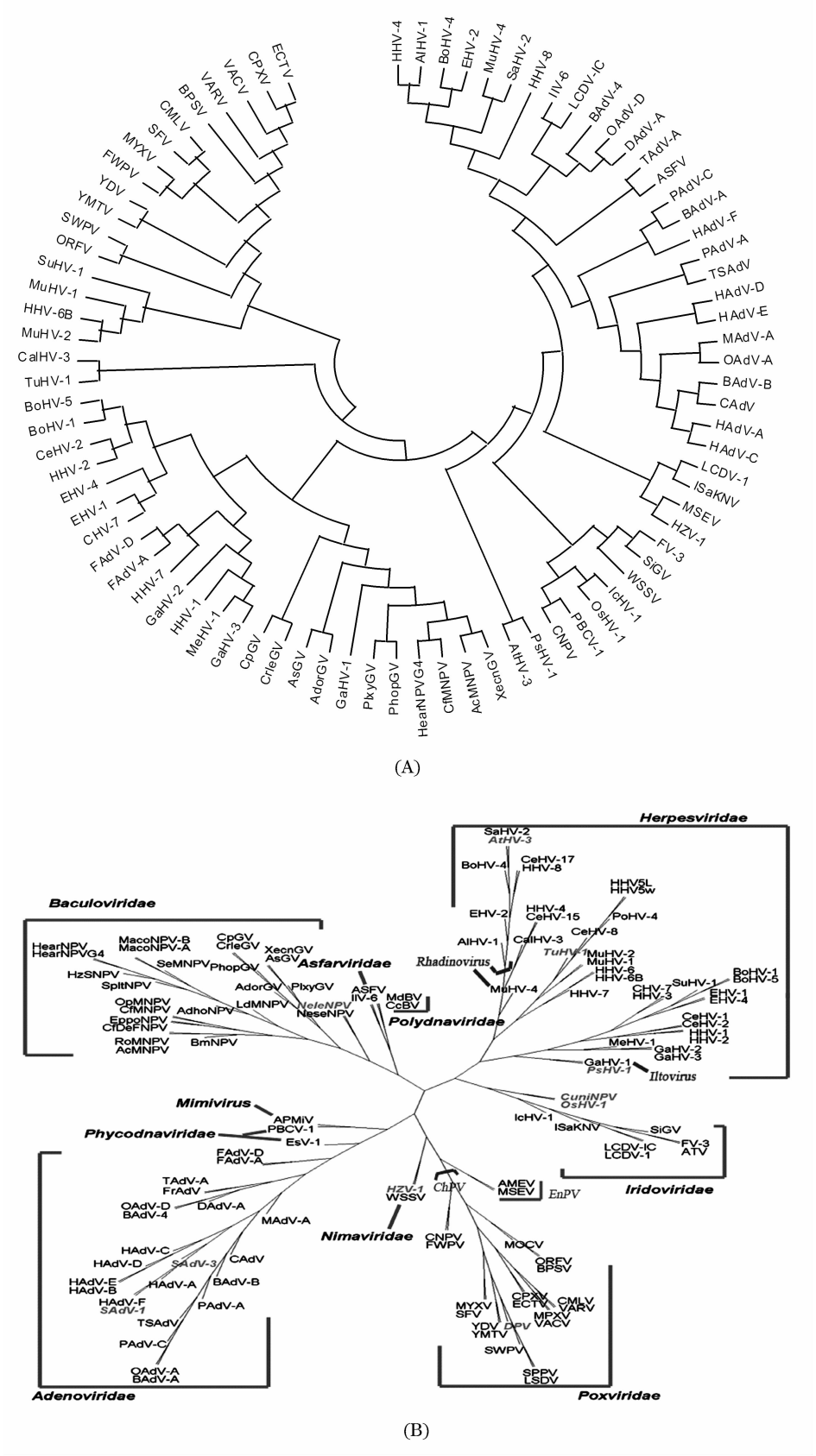

图 1　系统发生树比较

(A)用本文方法构建的系统发生树(B)文献[2]中的系统发生树

Fig. 1　The comparison of phylogenetic trees.

(A) a phylogenetic tree based on the method in this study　(B) the phylogenetic tree in the reference[2]

2.2.2 距离矩阵

每条序列我们都可以用一个由 4-mer 的相对偏离度组成的 256 维向量来表示。任意两条序列 A,B 之间的相关性则用相应的两个向量之间的夹角的余玄函数表示[6]:

$$C(A,B) = \frac{\sum_{i=1}^{N} a_i \times b_i}{\left[\sum_{i=1}^{N} a_i^2 \times \sum_{i=1}^{N} b_i^2\right]^{\frac{1}{2}}} \quad (2)$$

其中 a_i 表示 A 向量的第 i 个元素,b_i 表示 B 向量的第 i 个元素。两条序列 A 和 B 之间的距离定义为:

$$D(A,B) = \frac{1 - C(A,B)}{2} \quad (3)$$

由于相关性 C(A, B)在 -1 和 1 之间变化,距离 D(A, B)归一化在区间(0,1)内。每两个序列之间有一个距离,则对所有序列而言,构成一个距离矩阵。

2.2.3 邻接法

本文采用的建系统发生树的方法是邻接法(neighbor-joining method)。邻接法是一种应用最广的合并算法,最早由 Saitou 和 Nei 提出[7],尽管邻接法通常无法找到精确的最小进化树,只能找到近似的最小进化树,但是它的计算速度快,准确率较高,因此被广泛应用于系统发育分析中。邻接法不需要关于分子钟的假设,不考虑任何优化标准,基本思想是进行类的合并时,不仅要求待合并的类是相近的,而且要求待合并的类远离其他的类,从而通过对完全没有解析出的星型进化树进行分解,不断改善星型进化树。

3 结果与讨论

系统发生树是描述生物形成或进化顺序的拓扑结构。构建系统发生树的方法有基于比对的方法和基于全基因组的非比对方法。本文中使用的方法属于基于全基因组的非比对方法。我们用邻接法,结合 k-mer 信息参数,对 84 条较长的病毒序列构建了一幅系统发生树,并通过与其它可靠性较高的系统发生树相比较(图 1)[2],评估了此树的可靠性。

对比分析两个进化树中各病毒的位置相关性和拓扑结构,发现本文所建的系统发生树中绝大部分病毒的聚类情况和文献[2]中的一致。但也有几个不同的,它们分别是 IIV-6 和 LCDV-IC 没有聚在一起;SiGV、WSSV、FV-3 脱离所属进化枝;FAdV-D,FAdV-A 和 HHV-7 脱离所属进化枝。

本文用基于 k-mer 频数的信息参数表征基因组,并用基因组全序列的 4-mer 组分偏好性作为主要信息对 84 个病毒重构系统发生树,并将其与已有树相比较,分析其可信度。结果发现,本文构建的系统发生树中绝大部分病毒的聚类规则与已有系统发生树相一致。这说明本文提出的构建的系统发生树的基于全基因组的非比对方法能够真实地反映物种之间的亲缘关系。我们还发现,少数几个病毒的系统发生关系在两个树之间表现出差异。这两种构建进化树的方法区别在于:一个是用基于 k-1 阶的马尔科夫模型扣除了背景突变压力对 k-mer 频数的影响,而另一个是从实际 k-mer 频数中扣除了组分约束随机组合理论频数以便使相对偏离度指标反映 k-mer 中的碱基关联信息。

参考文献(References)

[1] 高扬. 碱基关联矩阵法在 DNA 病毒亲缘关系研究中的应用[D]. 呼和浩特:内蒙古大学,2011.

[2] Gao L, Qi J. Whole genome molecular phylogeny of large dsDNA viruses using composition vector method [J]. BMC Evolutionary Biology, 2007, 7:41.

[3] Xie H, Hao B. Visualization of K - tuple distribution in prokaryote complete genomes and their randomized counterparts[A]. CSB2002 Proceedings (C). Los Alamitos, California: IEEE Computer Society, 2002. 31-42.

[4] 罗辽复. DNA 序列信息内容的普适关系[J]. 合肥学院学报, 2005, 15(1):1-7.

[5] Zhou F, Olman V, Xu Y. Barcodes for genomes and applications [J]. BMC Bioinformatics, 2008, 9:546.

[6] Qi J, Wang B, Hao BL. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach [J]. J Mol Evol, 2004, 58(1):1-11.

[7] Saitou N, Nei M. The neighbor - joining method: a new method for reconstructing phylogenetic trees [J]. Molecular Biology and Evolution, 1987, 4(4): 406-425.