doi:10.3969/j.issn.1673-9833.2012.02.003

一种构建细小病毒进化树的加权距离方法

周立前[1]，杨碧波[2]

（1.湖南工业大学 计算机与通信学院，湖南 株洲 412007；2.湘潭大学 数学与计算科学学院，湖南 湘潭 411105）

摘 要：应用对数关联距离与互信息距离加权的方法，对完全基因组DNA序列和蛋白质序列构建了30种细小病毒系统发育树。构建的发育树均将30种细小病毒分成细小病毒亚科和浓核病毒亚科两个大的分枝，其结构与国际病毒学分类委员会第八版报道的结果及已有文献的结果基本一致。且基于蛋白质序列构建的系统发育树比基于完全基因组DNA序列构建的要好。

关键词：完全基因组；系统发育树；组合向量；对数关联距离；互信息距离；加权距离

中图分类号：Q19 **文献标志码**：A **文章编号**：1673-9833(2012)02-0010-06

A Weighted Distance Method to Construct Parvovirus Phylogenetic Tree

Zhou Liqian[1]，Yang Bibo[2]

（1. School of Computer and Communication，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. School of Mathematics and Computational Science，Xiangtan University，Xiangtan Hunan 411105，China）

Abstract：The phylogenetic trees of 30 parvoviruses are reconstructed by the weighted methods of log-correlation distance and mutual information distance. The trees are made of two branches of parvovirinae and densovirinae, and the structures are mainly consistent with the eighth report of the International Committee on Taxonomy of Viruses and results of the existing literature. The tree based on the protein sequences construction is better than that based on entire genome DNA sequences construction.

Keywords：complete genomes；phylogenetic tree；composition vector；log-correlation distance；mutual information distance；weighted distance

0 引言

病毒是比细菌还小、没有细胞结构、只能在细胞中增殖的一类微生物[1]。人们通常用形态学特征（包括衣壳大小、形状、结构等）和物理化学特性及抗原性等为特征来区分病毒[1]。对病毒基因组测序后，许多病毒的科、属之间的进化关系可通过单个基因或基因家族来分析。国际病毒分类委员会（International Committee on Taxonomy of Vviruses，ICTV）每5年报道一次病毒的分类标准。

细小病毒于1970年被确定为一个家族，包括所有含线状、自吸和长度在5 kB以内的单链DNA（single-stranded DNA，ss DNA）基因组的无包膜病毒[2-3]。细小病毒首先由K. I. Berns[4]根据基因组组织、病毒复制和宿主范围等进行分类，也有人提出应该根据序列同源性数目和基因组组织来进行分

收稿日期：2012-01-05
基金项目：湖南省国际合作基金资助项目（2011WK3032）
作者简介：周立前（1970-），男，湖南涟源人，湖南工业大学副教授，博士，主要研究方向为生物信息学，
E-mail：zhoulq11@163.com

类[5]。根据 ICTV 第八版[6]的报道，细小病毒家族分为细小病毒亚科（Parvovirinae）和浓核病毒亚科（Densovirinae），细小病毒亚科中的病毒主要感染脊椎动物，通过脊椎动物细胞培养，并且能频繁地与其它病毒交叉传染，而浓核病毒亚科中的病毒主要感染节肢动物和其它的无脊椎动物[7−8]。由于浓核病毒对宿主感染的致命性，因而浓核病毒被建议作为病虫害防治的代表[8−9]。近年来，J. R. Kerr 等人对人类与啮齿类动物细小病毒及其宿主的一致性与相似性进行了研究[10]。

分析物种间系统发育关系的传统方法大多是基于序列比对的方法，但是考虑到基因组中包含成百万直至上亿个碱基字符，逐个字符的比较已变得不现实[11]。因此，迄今已提出了许多构建物种系统发育树的非序列比较方法，如：信息理论方法[12−14]、主成分分析法[15]、奇异值分解法[11,16−17]、分形分析法[18−20]、Markov 模型法[1,21−22]、动力学语言法[23−24]、对数关联距离及 Fourier 变换加 Kullback-Leibler 散度距离法[25]等。在本课题前期的研究中，用对数关联距离法[25]与互信息方法[14]，构建了基于线粒体完全基因组的 64 种脊椎动物系统发育树，树的结构由哺乳动物、鱼类和初龙次亚纲 3 部分组成，获得的结果与已知的脊椎动物系统发育关系完全一致。本文拟采用对数关联距离与互信息距离加权的方法，基于完全基因组重构 30 种细小病毒的系统发育关系。

1　数据与方法

1.1　基因组数据集

从文献[8]的表 1 和文献[26]的表 3 中选取 30 种细小病毒作为研究对象，包括 20 种细小病毒亚科病毒和 10 种浓核病毒亚科病毒。细小病毒亚科病毒包括，阿留申水貂病毒属：aleutian mink disease virus（AMDV, NC_001662）；牛犬细小病毒属：minute virus of canines（MVC, NC_004442）；依赖病毒属：adeno-associated virus 1（AAV1, NC_002077），adeno-associated virus 2（AAV2, NC_001401），adeno-associated virus 3（AAV3, NC_001729），adenoassociated virus 4（AAV4, NC_001829），adeno-associated virus 5（AAV5, NC_006152），adeno-associated virus 7（AAV7, NC_006260），adeno-associated virus 8（AAV8, NC_006261），avian adeno-associated virus ATCC VR-865（AAAVa, NC_004828），avian adeno-associated virus strain DA-1（AAAVd, NC_006263），bovine adeno-associated virus（BAAV, NC_005889），bovine parvovirus-2（BPV, NC_006259），goose parvovirus（GPV, NC_001701）和 muscovy duck parvovirus（MDPV, NC_006147）；红细胞病毒属：B19 virus（B19V, NC_000883）；细小病毒属：canine parvovirus（CPV, NC_001539），LuIII parvovirus（LuIIIV,NC_004713），mouse parvovirus 3（MPV, NC_008185）和 porcine parvovirus（PPV, NC_001718）。浓核病毒亚科病毒包括，短浓核病毒属：aedes albopictus densovirus（AalDNV, NC_004285）；浓核病毒属：acheta domesticus densovirus（AdDNV, NC_004290），diatraea saccharalis densovirus（DsDNV, NC_001899），galleria mellonella densovirus（GmDNV, NC_004286），junonia coenia densovirus（JcDNV, NC_004284）和 mythimna loreyi densovirus (MIDNV, NC_005341)；艾特拉浓核病毒属：bombyx mori densovirus 1（BmDNV1, NC_003346），bombyx mori densovirus 5（BmDNV5, NC_004287）和 casphalia extranea densovirus（CeDNV, NC_004288）；烟色大蠊浓核病毒属：periplaneta fuliginosa densovirus（PfDNV, NC_000936）。其中 AAV7，AAV8，AAAVa，BPV，MPV，AdDNV 和 CeDNV 在 ICTV 第八版的报道中没有准确分类。

1.2　方法

本文中分析完全基因组的 2 种数据：完全基因组 DNA 序列（含编码区和非编码区）和蛋白质序列。一个 DNA 或蛋白质序列分别由 4 个不同的碱基或 20 个不同的氨基酸组成，每个编码序列都可根据遗传密码子[27]转换成相应的蛋白质序列。

1.2.1　组合向量

首先，把含 4 种碱基的 DNA 序列和含 20 种氨基酸的蛋白质序列看成符号序列，然后考虑长度为 K 的子串（称为 $K-$ 串），总共存在 $N=4^K$（对 DNA 序列）和 $N=20^K$（对蛋白质序列）种可能的 $K-$ 串类型。假设符号序列长度为 L，用长度为 K 的窗口在这个符号序列中移动，每次移动一个位置，由此来计算每种 $K-$ 串类型在这个长度为 L 的串中出现的频率。一个 $K-$ 串类型的观察频率定义为

$$p(\alpha_1\alpha_2\cdots\alpha_K)=\frac{n(\alpha_1\alpha_2\cdots\alpha_K)}{L-K+1},$$

式中 $n(\alpha_1\alpha_2\cdots\alpha_K)$ 是子串 $\alpha_1\alpha_2\cdots\alpha_K$ 在这个长度为 L 的串中出现的次数。

对于编码 DNA 和蛋白质序列，用 m 表示每个完全基因组中编码 DNA 或蛋白质序列的个数，那么 $K-$ 串的总观察频率定义为

$$\frac{\sum_{j=1}^{m} n_j(\alpha_1\alpha_2\cdots\alpha_K)}{\sum_{j=1}^{m}(L_j-K+1)},$$

式中：

$n_j(\alpha_1\alpha_2\cdots\alpha_K)$表示 K- 串$\alpha_1\alpha_2\cdots\alpha_K$在第$j$ 个编码 DNA 或蛋白质序列中出现的次数；

L_j表示完全基因组中第j个编码 DNA或蛋白质的长度。

对所有可能的 K- 串$\alpha_1\alpha_2\cdots\alpha_K$，用$p(\alpha_1\alpha_2\cdots\alpha_K)$作为元素构建一个基因组的组合向量。为进一步简化符号，用p_i表示第 i 个 K- 串类型（i = 1, 2, …, N，这 N 个子串按固定的字母表顺序排列）的观察频率。由此，对一个基因组构建了一个组合向量

$$(p_1p_2\cdots p_N)。$$

1.2.2 对数关联距离

对于基因组 A，用上面的方法可获得组合向量

$$\boldsymbol{P}=(p_1p_2\cdots p_N);$$

同样，对基因组 B 可获得组合向量

$$\boldsymbol{Q}=(q_1q_2\cdots q_N)。$$

先定义向量 $\boldsymbol{P}$, $\boldsymbol{Q}$ 间的夹角的余弦为

$$\cos\theta=\frac{\sum_{j=1}^{N}p_jq_j}{\sqrt{\sum_{j=1}^{N}p_j^2}\cdot\sqrt{\sum_{j=1}^{N}q_j^2}};$$

再定义两个向量 $\boldsymbol{P}$, $\boldsymbol{Q}$ 的距离为

$$d(\boldsymbol{P},\boldsymbol{Q})=-\log[(1+\cos\theta)/2]。$$

G. W. Stuart 等人[16-17]在 SVD（singuar value decomposition）中应用了这个对数关联距离。

1.2.3 互信息方法

设 X 和 Y 是两个离散时间变量，X取值 x_i（i = 1, 2, …, n）的概率为$p(x_i)$，Y取值y_j（j = 1, 2, …, m）的概率为$p(y_i)$，两个变量的联合概率为$p(x_i,y_j)$，它们的熵定义如下：

$$H(X)=-\sum_{i=1}^{n}p(x_i)\log p(x_i),$$

$$H(Y)=-\sum_{j=1}^{m}p(y_j)\log p(y_j),$$

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{m}p(x_i,y_j)\log p(x_i,y_j)。$$

条件熵的定义为[28]

$$H(X|Y)\equiv H(X,Y)-H(Y)。$$

根据文献[28]，有

$$H(X)\geqslant H(X|Y)\geqslant 0。$$

进而 X 和 Y 的平均互信息定义为

$$I(X,Y)\equiv H(X)+H(Y)-H(X,Y)=$$
$$H(X)-H(X|Y)=$$
$$H(Y)-H(Y|X)。$$

最后定义基于 X 和 Y 之间的互信息距离为

$$d(X,Y)=1-I(X,Y)/H(X,Y)。$$

1.2.4 加权距离

对于两个基因组 A 和 B，对应的组分矢量分别为

$$\boldsymbol{P}=(p_1p_2\cdots p_N),$$
$$\boldsymbol{Q}=(q_1q_2\cdots q_N)。$$

设对数关联距离为 $d_1(\boldsymbol{P},\boldsymbol{Q})$，互信息距离为 $d_2(\boldsymbol{P},\boldsymbol{Q})$，加权距离定义为

$$d(\boldsymbol{P},\boldsymbol{Q})=\alpha d_1(\boldsymbol{P},\boldsymbol{Q})+\beta d_2(\boldsymbol{P},\boldsymbol{Q}),$$

式中$\alpha+\beta=1$。

在具体的计算中，α和β的取值见表 1。

表 1 α与β取值表

Table 1 Values of α and β

α	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
β	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1

根据加权距离公式计算所研究的两种数据（完全基因组 DNA 序列和蛋白质序列）不同物种间的距离，获得相应的距离矩阵，然后用邻接方法[29]以及软件 Splits Tree4 V4.10[30]构建系统发育树。

2 结果与讨论

对完全基因组 DNA 序列，分别构建了 K 取 8~13，α和β取表 1 中各组值时的系统发育树；对蛋白质编码氨基酸序列，分别构建 K 取 3~6，α和β取表 1 中各组值时的系统发育树。将所有构建出的系统发育树，与 ICTV 第八版报告[6]、ICTV 第七版报告[7]，以及笔者以前的研究结果[31]进行比较，结果表明：对于完全基因组 DNA 序列，当 K=11, α=0.8, β=0.2 时，构建出的系统发育树最好（见图 1）；对于蛋白质编码氨基酸序列，当 K=6, α =0.1, β=0.9 时，构建出的系统发育树最好（见图 2）。

从图 1 和图 2 看出，两图中大部分病毒都归入其相应的科、属，与 ICTV 第八版和文献[31]的结果基本一致，但有 3 个病毒在两个图中与文献[31]的结果不一样。

一是属于红细胞病毒属的 B19V 病毒，在本文的图 1 中表现出与浓核病毒亚科的病毒有更近的亲缘关系；在图 2 中，B19V 虽被划分在细小病毒亚科，但被

划入依赖病毒属，这一方面体现了 B19V 病毒进化关系可以有更深入的探讨空间，另一方面也说明了本文的方法还不够好，在文献[31]中用互信息距离方法作进化分析时，B19V 病毒也被划入依赖病毒属。

二是属于浓核病毒属的美洲蟋蟀浓核病毒（AdDNV）在图 1 中并没有划入浓核病毒属，在文献[31]中，AdDNV 病毒也是排列在最靠近其他病毒属的位置，值得说明的是在 ICTV 第八版报告[6]中仍然没有将 AdDNV 病毒给出准确的分类，这说明 AdDNV 病毒的具体分类还值得进一步研究。

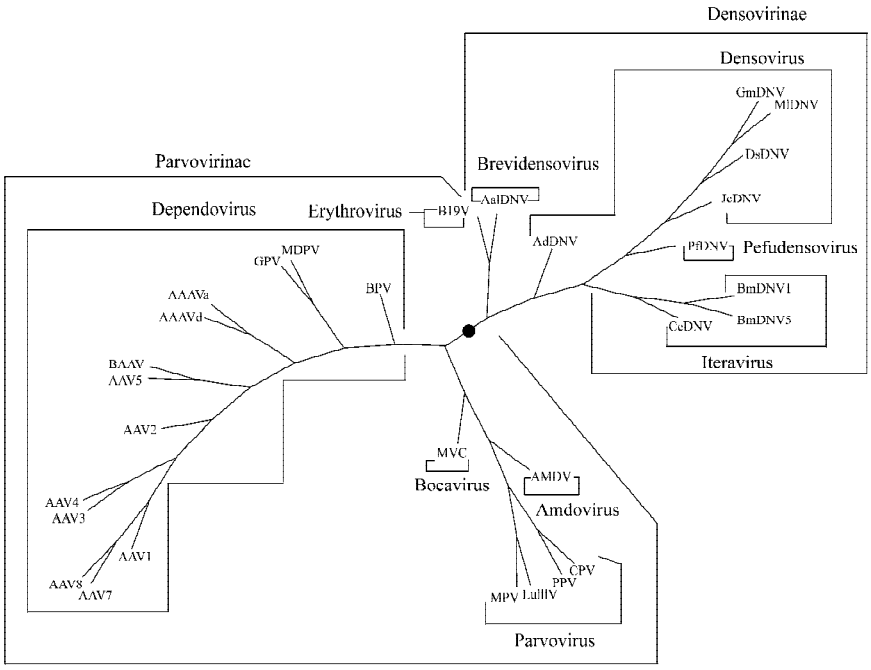

图 1　基于对数关联距离和互信息距离加权对完全基因组 DNA 序列构建的 30 种细小病毒系统发育树

Fig. 1　The phylogenetic tree of 30 parvovirus based on entire genome DNA sequences construction using the weighted methods of log-correlation distance and mutual information distance

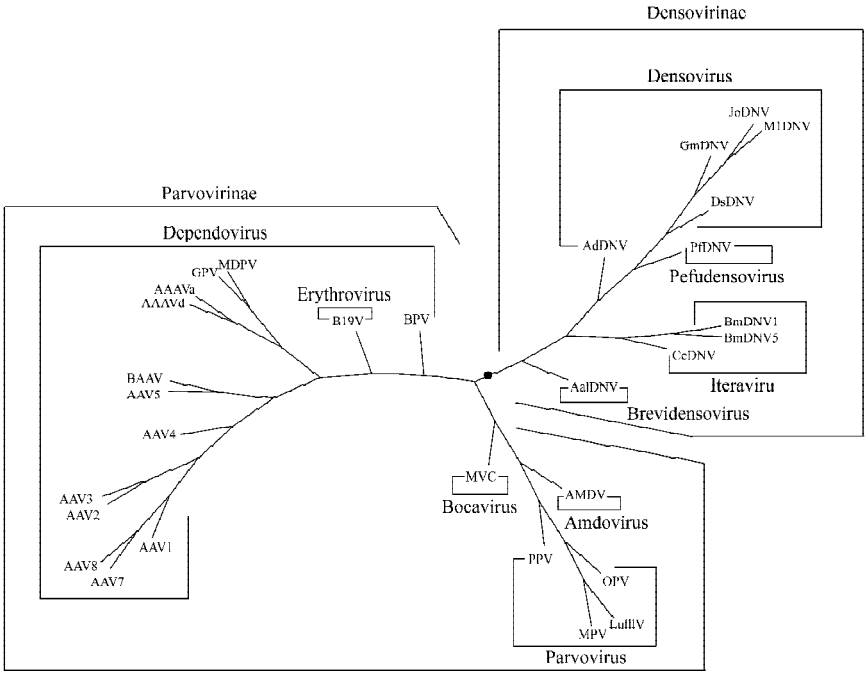

图2　基于对数关联距离和互信息距离加权对蛋白质序列构建的30种细小病毒系统发育树

Fig. 2　The phylogenetic tree of 30 parvovirus based on the protein sequences construction using the weighted methods of log-correlation distance and mutual information distance

三是属于烟色大蠊浓核病毒属的烟色大蠊浓核病毒（PfDNV）被划入浓核病毒属，在文献[31]中也出现了同样的情况，通过查找相关文献，发现L. Li等人[32]的研究认为将PfDNV病毒划为浓核病毒属更合适，在ICTV第七版[7]也曾将PfDNV病毒划为浓核病毒属，但在ICTV第八版[6]又将其划为烟色大蠊浓核病毒属，说明在此处还存在研究探讨的空间。

将图1和2的结果分别与文献[31]的结果进行比较可知：对于AdDNV病毒，图1中位置稍有不同，图2中结果一致；对于B19V病毒，图1中结果偏移较大，图2中其位置仅与BPV病毒换了位置。其余病毒在树中的位置与文献[31]中的结果基本一致。因此，图2的结构比图1更好。

7种病毒：AAV7，AAV8，AAAVa，BPV，MPV，AdDNV和CeDNV在ICTV第八版报道[6]中没有准确分类，除了AdDNV在上面已经讨论外，其它6种病毒均与文献[31]获得的结果相一致，这种情况再一次为这6个病毒的分类提供了依据。

3 结语

本研究中，采用对数关联距离与互信息距离进行加权的方式，研究了30种细小病毒的系统发育关系，获得的结果与ICTV第八版的报道以及相关文献中的研究结果基本一致。加权距离的方法，思想简单，计算速度快，不需要序列比对，对大批量数据的处理比较方便。因此，希望通过对这种新方法的尝试，能为分析和处理病毒分类与进化问题提供一个新的工具。

参考文献：

[1] Gao L，Qi J. Whole Genome Molecular Phylogeny of Large dsDNA Viruses Using Composition Vector Method[J]. BMC Evol. Biol.，2007，7：1-7.

[2] Chapman M S，Rossmann M G. Structure, Sequence, and Function Correlations among Parvoviruses[J]. Virology，1993，194(2)：491-508.

[3] Tattersall P，Cotmore S F. The Parvoviruses[M]. London：Hodder Arnold，2005：407-438.

[4] Berns K I. Parvoviridae: The Viruses and Their Replication [M]. 3rd ed. Philadelphia：Lippincott-Raven，1996：2173-2197.

[5] Tijssen P. Molecular and Structural Basis of The Evolution of Parvovirus Tropism[J]. Acta Veterinaria Hungarica，1999，47(3)：379-394.

[6] Fauquet C M，Mayo M A，Maniloff J，et al. Virus Taxonomy：Eighth Report of The International Committee on Taxonomy of Viruses[R]. [S. l.]：Academic Press，2005.

[7] Van Regenmortel M H V，Fauquet C M，Bishop D H L，et al. Virus Taxonomy: Seventh Report of The International Committee on Taxonomy of Viruses[R]. [S. l.]：Academic Press，2000.

[8] Kerr J R. The Parvoviridae: An Emerging Virus Family[J]. Infect. Dis. Rev.，2000，2(3)：99-109.

[9] Belloncik S. Potential Use of Densonucleosis Viruses as Biological Control Agents of Insect Pests[M]. Boca Raton：CRC Press，1988：285-289.

[10] Kerr J R，Boschetti N. Short Regions of Sequence Identity between the Genomes of Human and Rodent Parvoviruses and Their Respective Hosts Occur within Host Genes for Cytoskeleton, Cell Adhesion and Wnt Signaling[J]. J. Gen. Virol.，2006，87(12)：3567-3575.

[11] Stuart G W，Moffet K，Baker S. Integrated Gene Species Phylogenies from Unaligned Whole Genome Protein Sequences[J]. Bioinformatics，2002，18(1)：100-108.

[12] Li M，Badger J H，Chen X，et al. An Information-Based Sequence Distance and Its Application to Whole Mitochondrial Genome Phylogeny[J]. Bioinformatics，2001，17(2)：149-154.

[13] Yu Zuguo，Jiang Po. Distance, Correlation and Mutual Information among Portraits of Organisms Based on Complete Genomes[J]. Physics Letters A，2001, 286(1)：34-46.

[14] Yu Zuguo, Mao Zhi，Zhou Liqian，et al. A Mutual Information Based Sequence Distance for Vertebrate Phylogeny Using Complete Mitochondrial Genomes[C]// The 3rd International Conference on Natural Computation. Haikou：Conference Publications，2007，2：253-257.

[15] Edwards S V，Fertil B，Giron A，et al. A Genomic Schism in Birds Revealed by Phylogenetic Analysis of DNA Strings [J]. System Biology，2002，51(4)：599-613.

[16] Stuart G W，Berry M W. An SVD-Based Comparison of Nine Whole Eukaryotic Genomes Supports a Coelomate Rather Than Ecdysozoan Lineage[J]. BMC Bioinformatics，2004，5：204.

[17] Stuart G W，Moffet K，Leader J J. A Comprehensive Vertebrate Phylogeny Using Vector Representations of Protein Sequences from Whole Genomes[J]. Mol. Biol. Evol.，2002，19(4)：554-562.

[18] Yu Zuguo, Anh V，Lau K S. Multifractal and Correlation Analysis of Protein Sequences from Complete Genome[J]. Phys. Rev. E，2003，68(2)：021913.

[19] Yu Zuguo，Anh V，Lau K S. Chaos Game Representation, and Multifractal and Correlation Analysis of Protein Sequences from Complete Genome Based on Detailed HP Model[J]. Journal of Theoretical Biology，2004，226(3)：

341-348.

[20] Yu Zuguo，Anh V，Lau K S et al. The Phylogenetic Analysis of Prokaryotes Based on A Fractal Model of the Complete Genomes[J]. Phys. Lett. A，2003，317：293-302.

[21] Qi J，Luo H，Hao B L. CVTree: A Phylogenetic Tree Reconstruction Tool Based on Whole Genomes[J]. Nucleic Acids Research，2004，32(2)：45-47.

[22] Qi J，Wang B， Hao B L. Whole Proteome Prokaryote Phylogeny without Sequence Alignment: A K-String Composition Approach[J]. Journal of Molecular Evolution，2004，58(1)：1-11.

[23] Yu Zuguo，Zhou Liqian，Anh V，et al. Phylogeny of Prokaryotes and Chloroplasts Revealed by a Simple Composition Approach on All Protein Sequences from Whole Genome without Sequence Alignment[J]. Journal of Molecular Evolution，2005，60(4)：538-545.

[24] Yu Zuguo，Zhou Liqian，Chu K H，et al. Phylogenetic Analysis of Polyomaviruses Based on Their Complete Genomes[C]//The 4th International Conference on Natural Computation. Ji'nan：Conference Publications，2008，5：80-84.

[25] Zhou Liqian，Yu Zuguo，Anh V，et al. Logcorrelation Distance and Fourier Transform with Kullback-Leibler Divergence Distance for Construction of Vertebrate Phylogeny Using Complete Mitochondrial Genomes[C]// The 3rd International Conference on Natural Computation. Haikou：Conference Publications，2007，2：304-308.

[26] Farkas S L，Benkö M，Elo P，et al. Genomic and Phylogenetic Analyses of an Adenovirus Isolated from a Corn Snake (Elaphe Guttata) Imply Common Origin with the Members of the Proposed New Genus Atadenovirus[J]. J. Gen. Virol.，2002，83(10)：2403-2410.

[27] Brown T A. Genetics[M]. 3rd ed. London：Chapman，1998.

[28] Gray R M. Entropy and Information Theory[M]. New York：Springer Verlag，1990.

[29] Saitou N，Nei M. The Neighbor-Joining Method: A New Method for Reconstructing Phylogenetic Trees[J]. Mol. Biol. Evol.，1987，4(4)：406-425.

[30] Huson D H，Bryant D. Application of Phylogenetic Networks in Evolutionary Studies[J]. Mol. Biol. Evol.，2006，23(2)：254-267.

[31] Zhou Liqian，Bai Jumin，Yang Bibo. The Methods of Log-Correlation Distance and Mutual Information for Constructing Genome Phylogenetic Tree of Parvoviruses [C]//2010 3rd International Conference on Biomedical Engineering and Informatics(BMEI 2010). Xiangtan：Conference Publications，2010，6：2610-2613.

[32] Li L，Guo H，Zhang J，et al. Studies on Reclassifying of Periplaneta Fuliginosa Densovirus：in Chinese[J]. Virologica Sinica，2003，18(5)：486-491.

（责任编辑：邓光辉）