

doi:10.3969/j.issn.1673-9833.2014.03.019

基于碱基间隔距离模型的多瘤病毒系统发育关系分析

周立前¹, 李 瑞¹, 温在义²

(1. 湖南工业大学 计算机与通信学院, 湖南 株洲 412007; 2. 湖南工业大学 理学院, 湖南 株洲 412007)

摘 要: DNA 序列的碱基间隔距离分析方法可以对完全基因组序列进行较好地分析, 但是对短基因序列分析的效果不佳。因此, 在碱基间隔距离的基础上, 提出了一种改进的 DNA 序列碱基间隔距离模型, 并结合欧式距离, 构建了 70 种多瘤病毒基因组的系统发育树。通过将所得系统发育树的拓扑结构与已有文献中的结果进行对比与分析, 发现所获得的结果同传统方法计算的结果基本一致, 验证了所提方法的有效性。

关键词: 完全基因组; 碱基间隔距离; 欧氏距离; 系统发育树

中图分类号: Q19

文献标志码: A

文章编号: 1673-9833(2014)03-0094-05

The Phylogenetic Analysis of Polyomavirus Based on the Inter-Nucleotide Distance Model

Zhou Liqian¹, Li Rui¹, Wen Zaiyi²

(1. School of Computer and Communication, Hunan University of Technology, Zhuzhou Hunan 412007, China;

2. School of Science, Hunan University of Technology, Zhuzhou Hunan 412007, China)

Abstract : The DNA sequence inter-nucleotide distance analysis method can better analyze the complete genome sequence, but it is not ideal for short genome sequence. Therefore based on inter-nucleotide distance, proposes an improved DNA sequence inter-nucleotide distance model, and combined with Euclidean distance, constructs phylogenetic tree of 70 kinds of polyoma virus genome. Through the comparison of topological structure of the obtained phylogenetic tree with results in the existing literature, finds that the obtained results are basically the same with the results computed by traditional method, and verifies the effectiveness of the proposed method.

Keywords : complete genomes; inter-nucleotide distances; Euclidean distance; phylogeny tree

1 研究背景

基因组系统发育分析是生物信息学中基因组数据处理的一个重要方面。通过对基因组系统发育的分析, 可以得出物种之间的亲缘关系与进化历程。

系统发育分析方法很多, 一般根据对碱基序列转换为数值序列方式的不同而将其分为序列比对方法和非序列比对方法。传统的序列比对方法主要有: 基因序列重排方法^[1]、编码蛋白的出现与不出现方法^[2]、基因组分与完全相似性方法^[3], 折叠的出现与直系

收稿日期: 2014-02-12

基金项目: 湖南省自然科学基金资助项目(13JJ3109), 湖南省教育厅基金资助重点项目(13A004)

作者简介: 周立前(1970-), 男, 湖南涟源人, 湖南工业大学教授, 博士, 主要从事生物信息学方面的教学与研究,

E-mail: zhouliq11@163.com

同源基因方法^[4]等。但是随着完全基因组数据的使用,这些序列比对方法对于大批量数据的比对不再适用^[5-6],其空间和时间复杂度一直没能达到令人满意的效果。因此,生物信息学专家们致力于非序列比对方面的研究,提出许多作为补充系统发育树构建方法的新方法,如信息论方法^[7-8]、主成分分析方法^[9]、奇异值分解方法^[5-6]、马尔科夫模型方法^[10-11]、分型方法^[12-15]、动力学语言方法^[16-17]等。

2005年,印度学者 A. S. S. Nair 等人基于序列中碱基的位置分布情况,提出了一种新的DNA序列数值化表示方法。该方法将DNA序列转换为一个与原序列等长的数字序列,数字序列中,每个数字表示两个相同的相邻碱基的位置差,这种位置差称为碱基间隔距离,因此,数值序列也称为碱基间隔距离序列^[18]。同时,他们通过傅里叶变换对碱基间隔序列进行了分析,得到了碱基间隔距离序列能很好地识别基因组序列中启动子区域的结论。然而, M. Akhtar 等人于2008年证实,采用碱基间隔距离序列分析序列外显子时,其精确度并不高^[19]。2009年,葡萄牙学者 V. Afreixo 等人对碱基间隔距离序列提出了一种新的分析方法,他们认为碱基间隔服从几何分布,通过计算几何分布的理论值与实际值的相对误差向量,可以获取DNA序列的生物信息,并且构建系统发育树^[20]。这种方法没有采用多重比对算法,可以方便地进行全基因组计算。2011年, V. Afreixo 等人又提出了以连续核苷酸距离描述DNA序列的方法^[21],即通过计算连续核苷酸的长度,构建出一个与原序列等长的数字序列,然后结合碱基间隔距离序列,对DNA序列进行数值化分析。但试验证实, V. Afreixo 等提出的方法^[20-21]只适用于对较长全基因组序列进行分析计算,而对于短基因组序列,如病毒序列等,则无法正确表示。

多瘤病毒 (*polyomavirus*) 是一种无包膜的双链DNA病毒,这类病毒会引发肿瘤,其中有些种类会感染人的呼吸系统,它们广泛地分布在脊椎动物中。多瘤病毒的系统发育分析是了解、分析多瘤病毒间的亲缘与进化关系的一种重要方式。通过对多瘤病毒亲缘关系的了解,不仅有助于多瘤病毒引发疾病的治疗,而且对疫苗药物的研制也有着重要的指导作用;同时,对于新的未知病毒性疾病的分析也有着重要的参考价值^[22]。

本文通过对碱基间隔距离的进一步分析研究,提出了一种改进的碱基间隔距离分析方法。即通过统计相邻的同碱基间其余3种碱基的分布,以更好地描述基因组序列,该法可包含更多DNA序列的生物

信息。同时,使用此方法并结合欧式距离方法,构建了70种多瘤病毒完全基因组的系统发育树,通过对比所获结果与已知系统发育关系判断所提方法的有效性,以期通过这种新方法的尝试为分析和处理分类与进化问题提供一种新工具。

2 数据与方法

2.1 基因组数据集

M. Pérez-Losada 等人^[22]用极大似然法与贝叶斯方法分析了72种多瘤病毒的进化关系,这些病毒包括9类哺乳动物病毒(共67种链)和2类鸟类病毒(共5种链)。因其中有2种病毒(NC_001536和K02737)现无法从GenBank中找到,所以本文主要分析余下的70种多瘤病毒基因组的系统发育关系,基因组的数据集与文献^[22]一致,数据来源于NCBI(National Center of Biotechnology Information)数据库(<http://www.ncbi.nlm.nih.gov/nucleotide>),其具体名称见图1。

2.2 方法

2.2.1 DNA序列的间隔数值表示

DNA序列的碱基间隔距离由DNA序列中邻近相同碱基的位置差构成。设由A, C, G, T碱基组成的集合为 $R, S=S_1, S_2, \dots, S_N$ 是长度为 N 的一条DNA序列。设 d^x 是碱基 $x(x \in R)$ 的间隔距离序列,若 k 是满足条件 $S_m=S_{m+k}$ 和 $m+k \leq N$ 的最小值,则 $d^x(m)=k$;否则 $d^x(m)=N-m$ (m 为碱基 x 所在DNA序列中的位置, N 为DNA序列长度)。例如,对于一条短DNA序列ACACGCTTGATACG:

$$d^A=(2, 8, 2, 2), \quad d^C=(2, 3, 7, 1),$$

$$d^G=(5, 5, 0), \quad d^T=(2, 1, 3, 3)。$$

式中, d^A, d^C, d^G, d^T 分别表示碱基A, C, G, T的间隔序列。由DNA序列可知,第1个碱基A的位置为1,第2个碱基A的位置为3,第3个碱基A的位置为11,第4个碱基A的位置为13。所以在 d^A 序列中,第1个间隔值为2,第2个间隔值为8,第3个间隔值为2,最后一个间隔值为位置与序列长度之差,为2。同样,可以计算得到其余3个碱基(C, G, T)的碱基间隔序列。如果将4种碱基的间隔距离序列按每个碱基的具体位置组合到一起,可以得到与原序列对应等长度的碱基间隔距离序列 d :

$$d=(2, 2, 8, 3, 5, 2, 7, 1, 3, 5, 2, 3, 2, 1, 0)。$$

如果每个碱基第1次出现的位置已知,如 k_0^A, k_0^C, k_0^G 和 k_0^T ,那么整条序列可被唯一地还原回来:

若 $[k_0^A, k_0^C, k_0^G, k_0^T]=[1, 2, 5, 6]$,则可还原碱基序列,为AC--GT-----。

根据碱基集合 R 的顺序依次恢复, 每组碱基间隔序列依次按照间隔距离恢复, 最后一个距离作为验证, 具体过程如下:

- $d_1=2$, 对应序列 ACA-GT-----;
- $d_2=8$, 对应序列 ACA-GT----A----;
- $d_3=2$, 对应序列 ACA-GT----A-A--;
- $d_4=2$, 对应序列 ACA-GT----A-A--;
- $d_5=2$, 对应序列 ACACGT----A-A--;
- $d_6=3$, 对应序列 ACACGTC--A-A--;
- $d_7=7$, 对应序列 ACACGTC--A-AC-;
- $d_8=1$, 对应序列 ACACGTC--A-AC-;
- $d_9=5$, 对应序列 ACACGTC--GA-AC-;
- $d_{10}=5$, 对应序列 ACACGTC--GA-ACG;
- $d_{11}=0$, 对应序列 ACACGTC--GA-ACG;
- $d_{12}=2$, 对应序列 ACACGTCT-GA-ACG;
- $d_{13}=1$, 对应序列 ACACGTCTTGA-ACG;
- $d_{14}=3$, 对应序列 ACACGTCTTGATACG;
- $d_{15}=3$, 对应序列 ACACGTCTTGATACG。

以上结果表明, 采用碱基间隔距离表示 DNA 序列的数值方法, 没有丢失 DNA 字符序列的生物信息, 可以用作 DNA 序列的计算。

2.2.2 DNA 序列的碱基间隔距离模型

V. Afreixo 等人^[20]研究了碱基间隔序列的几何分布特征, 认为碱基间隔距离符合几何分布中的 n 重伯努利实验模型, 碱基是独立同分布的随机变量。通过计算不同间隔距离 K 的理论值与实际观测值的相对误差, 构建一组 $1 \sim K_{\max}$ 长度的序列, 用于描述 DNA 序列特性。该方法对于碱基平均间隔距离大于 20 的 DNA 序列, 可以较好地进行描述, 但当 DNA 序列的碱基平均间隔距离小于 10 时 (如细小病毒全基因序列), 由于包含的信息较少, 不能很好地描述。

通过对 DNA 序列的间隔数值表示研究, 统计间隔距离间其余 3 种碱基的分布, 可以更好地描述序列的生物学特性, 特别是对于长度相对较短的基因序列。基于相同碱基间其余 3 种碱基的分布, 此处提出一种新的碱基间隔序列模型。对于不同间隔长度 K , 其间 3 种碱基存在 $N=3^{K-1}$ 种可能的组合, 将每一种组合数值化, 定义 R 集合中 A, C, G, T 碱基对应的值分别为 $\{1, 2, 3, 4\}$, 则每一种分布模式的各碱基数值之和为 $M(\alpha_1\alpha_2 \cdots \alpha_{K-1})$, 因此, 各分布的观察频率可以定义为:

$$p(\alpha_1\alpha_2 \cdots \alpha_{K-1}) = \frac{n(\alpha_1\alpha_2 \cdots \alpha_{K-1}) \cdot M(\alpha_1\alpha_2 \cdots \alpha_{K-1})}{\sum_{j=1}^N n_j(\alpha_1\alpha_2 \cdots \alpha_{K-1}) \cdot M_j(\alpha_1\alpha_2 \cdots \alpha_{K-1})}$$

式中: p 为观察频率; $n(\alpha_1\alpha_2 \cdots \alpha_{K-1})$ 是当碱基间隔序列中间隔距离为 K 时其余 3 种碱基字串为 $\alpha_1\alpha_2 \cdots \alpha_{K-1}$ 的出现次数。

上式表明, 观察频率 p 是一种分布模式的值占所有分布模式值的比例。由于 DNA 序列编码区由多段序列组成, 此处用 m 表示每个完全基因组中编码序列的个数。因此, 整个 DNA 编码区序列 K 距离的观察频率定义为:

$$p(\alpha_1\alpha_2 \cdots \alpha_{K-1}) = \frac{\sum_{i=1}^m n_i(s_1s_2 \cdots s_{K-1}) \cdot M(\alpha_1\alpha_2 \cdots \alpha_{K-1})}{\sum_{j=1}^N \sum_{i=1}^m n_{ji}(\alpha_1\alpha_2 \cdots \alpha_{K-1}) \cdot M_j(\alpha_1\alpha_2 \cdots \alpha_{K-1})}$$

式中 $n_i(s_1s_2 \cdots s_{K-1})$ 是一种当间隔距离为 K 时其余 3 种碱基字串为 $\alpha_1\alpha_2 \cdots \alpha_{K-1}$ 在一段 DNA 编码序列中出现的次数。统计各段编码序列该距离的字串出现的总次数, 通过计算该字串值占该距离所有字串总值的比例, 即可得出其观察频率。

对不同间隔距离 K , 分别以 $p(\alpha_1\alpha_2 \cdots \alpha_{K-1})$ 为元素, 构建一个长度为 $N=3^{K-1}$ 的组合向量:

$$P = (p_1, p_2, \dots, p_N)$$

依次计算碱基的间隔距离序列如 d^A 中不同长度距离 K ($K > 2$) 的组合向量, 并按顺序拼接起来, 得到组合向量 $Q^A = (P_{K=2}, P_{K=3}, \dots, P_{K=K})$ 。将 4 个碱基对应的组合向量 Q^X 按字母顺序依次排列, 即可得到新的描述模型向量, 即 $V_K = (Q^A, Q^C, Q^G, Q^T)$ 。

2.2.3 欧氏距离

欧氏距离是系统发育分析中的常用距离。设两个基因组 A 和 B 的所得向量分别为

$$P = (p_1, p_2, \dots, p_N), Q = (q_1, q_2, \dots, q_N)$$

则两个向量的欧氏距离定义为:

$$d(P, Q) = \sqrt{\sum_{j=1}^N (p_j - q_j)^2}$$

结合碱基间隔距离序列模型与欧氏距离计算方法, 分别计算完全基因组 DNA 序列与编码 DNA 序列物种间的距离, 可获得距离矩阵, 然后运用 Splits Tree V4.0 中的 NJ (neighbor-joining) 方法^[23]构建系统发育树。

3 结果与讨论

结合碱基间隔距离模型与欧氏距离, 计算了 70 种多瘤病毒完全的编码 DNA 序列。70 种多瘤病毒编码 DNA 序列的平均长度为 5 000 bp, 平均碱基间隔距离为 3.9, 具体的碱基分布情况如表 1 所示。

表1 70种多瘤病毒的编码DNA序列平均碱基间隔距离出现的频率分布

Table 1 Frequency distribution of the inter-nucleotide distances of 70 polyomavirus coding DNA sequences

距离	1	2	3	4	5	6	7	8	9	10
频率	0.26	0.18	0.13	0.10	0.07	0.06	0.04	0.03	0.02	0.007

由于多瘤病毒序列的长度较短, 碱基间隔距离大于10出现的频率均小于1%, 故不再统计, 仅计算碱基间隔距离模型 K 取值为2~9的所有系统发育树。通过将所得系统发育树与文献[22]中的树进行对比, 发现当最大距离 $K=7$ 时的树是最好的, 其拓扑结构与文献[22]中的树基本相同, 如图1所示。

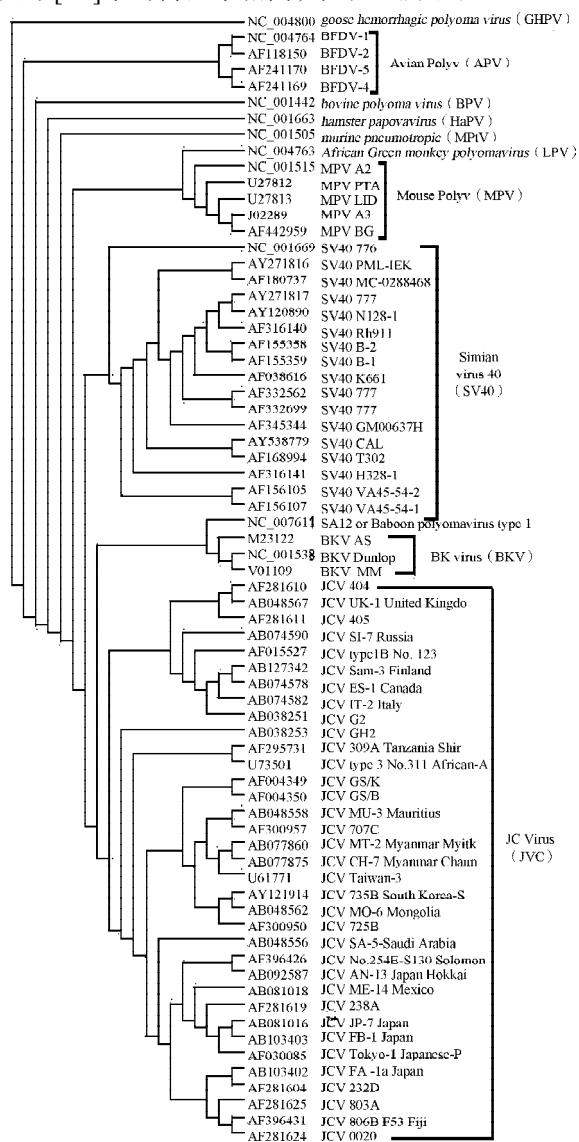


图1 基于碱基间隔距离模型结合欧氏距离方法构建的70种多瘤病毒系统发育树

Fig. 1 The phylogenetic tree of 70 polyomavirus based on the inter-nucleotide distance model combining with Euclidean distance method

由图1所示的系统发育树可见, 本实验的70种

多瘤病毒大体可分为两类: 哺乳类和鸟类多瘤病毒。4种鸟类多瘤病毒 (avian polyv, APV) 链聚集在一个分枝, 鹅出血性多瘤病毒 (*goose hemorrhagic polyoma virus*, GHPV) 靠近这个分枝, 而APV和GHPV与其它的哺乳动物多瘤病毒是分开的。换言之, 所得树中鸟类多瘤病毒 (包括APV与GHPV) 与哺乳动物多瘤病毒分离的现象, 支持文献[22]中多瘤病毒类分为两个子类的观点。这种分离与每个生物群体的不同滤过性病毒生命策略有关。在哺乳动物多瘤病毒类, 老鼠多瘤病毒 (mouse parvovirus, MPV)、猿猴病毒40 (simian virus 40, SV40)、BK病毒 (BK virus, BKV)、JC病毒 (JC virus, JCV) 分别聚集在所期望的不同分枝上。牛乳头病毒 (*bovine polyoma virus*, BPV)、仓鼠多瘤病毒 (*hamster papovavirus*, HaPV)、鼠科亲肺多瘤病毒 (*murine pneumotrophic*, MPtV) 和非洲绿猴多瘤病毒 (*African Green monkey polyomavirus*, LPV) 4种多瘤病毒相邻, 其结构与文献[22]和文献[17]中的有些不一致, 说明这4种病毒的亲缘关系不太确定, 但又有相近的亲缘关系。BKV链聚成一个分枝, 此分枝与SA12组合形成一个有亲缘关系的大分枝; 同时, 这个大分枝与JCV分枝组合在一起, 形成一个有亲缘关系的更大分枝: 这也与文献[22]和文献[17]中的结果一致。

从以上分析可以看出, 所得结果基本上能反映多瘤病毒和它们的宿主之间存在一些系统发育的共同关系。

4 结语

在本研究中, 运用基于碱基间隔距离模型的方法, 并结合欧氏距离方法对多瘤病毒序列进行建模处理, 构建并分析了70种多瘤病毒的系统发育树, 通过与已有文献结果对比, 可知所得树的拓扑结构与当前已知的用传统方法计算的树的结构基本一致, 从而进一步验证了以前方法的可靠性。

对70种多瘤病毒进行系统发育关系分析, 既可以根据多瘤病毒的特点预测出一些新的多瘤病毒, 也能够更全面地学习多瘤病毒的进化关系。所构建的方法通过构建简单的数学模型实现, 与V. Afreixo等人^[20]所提出的碱基间隔距离模型方法相比, 本文所提方法可以处理长度较短的生物序列, 能较好地建立短碱基序列的数学模型; 与传统的序列比对方法相比, 其思路更简单直接, 速度更快。且所提方法不需要复杂的去噪步骤, 结果显示, 碱基间隔距离模型方法能实现这种功能。

可见,通过这种新方法的尝试,能提供一个分析和处理分类与进化问题的新工具。

参考文献:

- [1] Sankoff D, Leaduc G, Antoine N, et al. Gene Order Comparisons for Phylogenetic Inference: Evolution of the Mitochondrial Genome[J]. Proceedings of the National Academy of Sciences, 1992, 89(14): 6575-6579.
- [2] Gibbon S T F, House C H. Whole Genome-Based Phylogenetic Analysis of Free-Living Microorganisms[J]. Nucleic Acids Research, 1999, 27(21): 4218-4222.
- [3] Tekaia F, Lazcano A, Dujon B. The Genome Tree as Revealed from Whole Proteome Comparisons[J]. Genome Research, 1999, 9(6): 550-557.
- [4] Lin J, Gerstein M. Whole Genome Tree Based on the Occurrence of Folds and Orthologs: Implications for Comparing Genomes at Different Levels[J]. Genome Research, 2000, 10(6): 808-818.
- [5] Stuart G W, Moffet K, Baker S. Integrated Gene Species Phylogenies from Unaligned Whole Genome Protein Sequences[J]. Bioinformatics, 2002, 18(1): 100-108.
- [6] Stuart G W, Moffett K, Leader J J. A Comprehensive Vertebrate Phylogeny Using Vector Representations of Protein Sequences from Whole Genomes[J]. Molecular Biology and Evolution, 2002, 19(4): 554-562.
- [7] Li Ming, Badgeb J H, Chen Xin, et al. An Information-Based Sequence Distance and Its Application to Whole Mitochondrial Genome Phylogeny[J]. Bioinformatics, 2001, 17(2): 149-154.
- [8] 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.
- [9] Edwards S V, Fertil B, Giron A, et al. A Genomic Schism in Birds Revealed by Phylogenetic Analysis of DNA Strings [J]. Systems Biology, 2002, 51(4): 599-613.
- [10] Qi Ji, Luo Hong, Hao Bailin. CVTree: A Phylogenetic-Tree Reconstruction Tool Based on Whole Genomes[J]. Nucleic Acids Research, 2004, 32(2): 45-47.
- [11] Qi Ji, Wang Bin, Hao Bailin. Whole Proteome Prokaryote Phylogeny Without Sequence Alignment: A K-String Composition Approach[J]. Journal Molecular Evolution, 2004, 58(1): 1-11.
- [12] Yu Z G, Anh V, Lau K S. Multifractal and Correlation Analysis of Protein Sequences from Complete Genomes[J]. Physical Review E, 2003, 68(2): 021913.
- [13] Yu Z G, Anh V, Lau K S. Chaos Game Representation of Protein Sequences Based on the Detailed HP Model and Their Multifractal and Correlation Analyses[J]. Journal of Theoretical Biology, 2004, 226(3): 341-348.
- [14] Yu Z G, Anh V, Lau K S, et al. The Genomic Tree of Living Organisms Based on a Fractal Model[J]. Physics Letters A, 2003, 317(3): 293-302.
- [15] Chu K H, Qi J, Yu Z G, et al. Origin and Phylogeny of Chloroplasts: A Simple Correlation Analysis of Complete Genomes[J]. Molecular Biology and Evolution, 2004, 21(1): 200-206.
- [16] Yu Z G, Zhou L Q, 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.
- [17] Yu Z G, Zhou L Q, Chu K H, et al. Phylogenetic Analysis of Polyomaviruses Based on Their Complete Genomes[C]//The 4th International Conference on Natural Computation. Jinan: Conference Publication, 2008: 80-84.
- [18] Nair A S S, Mahalashmi T. Visualization of Genomic Data Using Inter-Nucleotide Distance Signals[C]//Processing of IEEE Genomic Signal Processing. Bucharest: Conference Publication, 2005: 11-13.
- [19] Akhtar M, Epps J, Ambikairajah E. Signal Processing in Sequence Analysis: Advances in Eukaryotic Gene Prediction[J]. IEEE Journal of Selected Topics in Signal Processing, 2008, 2(3): 310-321.
- [20] Afreixo V, Bastos C A C, Pinho A, et al. Genome Analysis with Inter-Nucleotide Distance[J]. Bioinformatics, 2009, 25(23): 3064-3070.
- [21] Afreixo V, Bastos C A C, Pinho A, et al. Genome Analysis with Distance to the Nearest Dissimilar Nucleotide[J]. Journal of Theoretical Biology, 2011, 275(1): 52-58.
- [22] Pérez-Losada M, Christensen R G, McClellan D A, et al. Comparing Phylogenetic Codivergence Between Polyomaviruses and Their Hosts[J]. Journal of Virology, 2006, 80(12): 5663-5669.
- [23] 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.

(责任编辑:廖友媛)

