Hans Journal of Computational Biology 计算生物学, 2021, 11(2), 21-29
Published Online June 2021 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2021.112003

Hans汉斯

新冠病毒DNA序列基于熵值的分布可视化

杨　宸，郑智捷

云南大学软件学院，云南　昆明
Email: 813109200@qq.com

收稿日期：2021年5月1日；录用日期：2021年6月2日；发布日期：2021年6月9日

摘　要

新型冠状病毒于2020年1月正式命名为2019-nCoV，时至今日该病毒的传播仍然没有得到良好的控制。病毒的DNA碱基序列在病毒的性状中起着决定性的作用，本文分析了四个国家(中国，美国，澳大利亚，德国)新冠病毒DNA之间的差异。选择各国新冠病毒DNA序列，以信息熵，相对熵，交叉熵的形式给出四国病毒之间差异的可视化分析。

关键词

新冠病毒，DNA序列，熵，可视化分析

Novel Coronavirus DNA Sequence Visualization Based on Entropy Distribution

Chen Yang, Jeffrey Zheng

School of Software, Yunnan University, Kunming Yunnan
Email: 813109200@qq.com

Received: May 1st, 2021; accepted: Jun. 2nd, 2021; published: Jun. 9th, 2021

Abstract

Novel coronavirus is officially named 2019-ncov in January 2020, and has not been well controlled until now. The DNA base sequence of avirus plays a decisive role in the character of avirus. This paper analyzes the differences between novel coronavirus DNA from four countries (China, America, Australia and Germany). The novel coronavirus DNA sequences from different countries were selected, and the visual analysis of the differences between four viruses was given in the form of

文章引用: 杨宸, 郑智捷. 新冠病毒 DNA 序列基于熵值的分布可视化[J]. 计算生物学, 2021, 11(2): 21-29.
DOI: 10.12677/hjcb.2021.112003

information entropy, relative entropy and cross entropy.

Keywords

Novel Coronavirus, DNA Sequence, Entropy, Visualization Analysis

1. 引言

2020 年 3 月底，新冠病毒在我国得到了基本的控制，但全球的疫情远远没有结束，社会上也出现了各种关于病毒来自哪个国家的讨论，这时候利用科学的方法来辨识病毒在微观上的差异也显得尤为重要。

DNA 的一级结构决定了基因的功能，欲想解释基因的生物学含义，首先必须知道其 DNA 顺序。

熵，定义为信息的期望值。信息熵常作为某个系统的信息含量的量化指标，所以信息熵会进一步用来作为系统方程优化的目标或者参数选择的判据。交叉熵是 Shannon 信息论中一个重要概念，可以用来度量两个概率分布的相似性，而且交叉熵在神经网络中又可以用作损失函数，所以在本文中交叉熵用来衡量不同 DNA 之间的相似性。相对熵在信息论中等价于两个概率分布的信息熵的差值，所以在本文中可以用于衡量两个 DNA 的差异；三种熵则为我们提供了一种新的研究方式，在基于生物学对 DNA 研究的方法上，合理利用三种熵值的公式来对不同的 DNA 进行测量技术，最后再将结果可视化。

2. 研究结构

2.1. 信息熵

什么是信息，信息一直是一个高度抽象的概念。在 1948 年香农提出了“信息熵”的概念，才得以将信息度量化。而信息熵一词则是其从热力学中借用而来的，Shannon 借鉴了热力学的概念，把信息中排除了冗余后的平均信息量称为“信息熵”，并给出了计算信息熵的数学表达式(见下式(1))。

$$H(U)=E\left[-\log p_i\right]=-\sum_{i=1}^{n}\log p_i \tag{1}$$

2.2. 相对熵

相对熵(relative entropy)，又被称为 Kullback-Leibler 散度(Kullback-Leibler divergence)或信息散度(information divergence)，是两个概率分布(probability distribution)间差异的非对称性度量[1]。在信息理论中，相对熵等价于两个概率分布的信息熵(Shannon entropy)的差值[2]。

在同样的字符集上，假设存在另一个概率分布 $Q(x)$，如果用概率分布 $P(x)$ 的最优编码，来符合分布 $P(x)$ 的字符编码那么表示这些字符就会比理想情况多用一些比特数。相对熵就是用来衡量这种情况下平均每个字符多用的比特数，因此可以用来衡量两个分布的距离(见下式(2))。

$$KL\left(P\|Q\right)=\sum_{x\in X}P(x)\log\frac{P(x)}{Q(x)} \tag{2}$$

2.3. 交叉熵

交叉熵(Cross Entropy)是 Shannon 信息论中一个重要概念，主要用于度量两个概率分布间的差异性信

息。[3]其公式定义如下(见下式(3))：

$$H(p,q)=\sum_{x}P(x)\log\left(\frac{1}{q(x)}\right) \tag{3}$$

2.4. 新冠病毒 DNA 碱基序列

本次测试使用的 DNA 来自于国家生物信息中心(CNCB)中公开的新型冠状病毒(2019nCoVR)。

2.5. 研究方法与研究模块

2.5.1. 参数

中国地区 DNA 序列，共计 29904 个碱基；

美国地区 DNA 序列，共计 29883 个碱基；

德国地区 DNA 序列，共计 29783 个碱基；

澳大利亚地区 DNA 序列，共计 29894 个碱基。

对于每个地区分别计算序列中 AG 碱基的信息熵。之后计算中国对美国，中国对德国，中国对澳大利亚的交叉熵与相对熵。

2.5.2. 计量模块

如图 1 所示，对于每个地区，计算出序列中“AG”“CT”“AC”“AT”“GC”“GT”六种组合序列所占比，该比值则可以作为熵值计算公式中 $P(x)$ 与 $Q(x)$ 来使用。

Figure 1. Measuring module
图 1. 计量模块

2.5.3. 处理模块

如图 2 所示，利用计量模块中得到频率计算熵值。得出的结果再进行下一步的可视化处理。

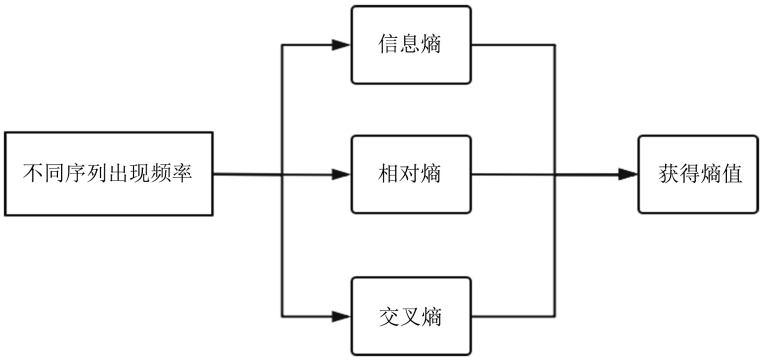

Figure 2. Processing module
图 2. 处理模块

2.5.4. 可视化模块

如图 3 所示，利用 Python 的可视化工具，对得到的熵值进行分析。

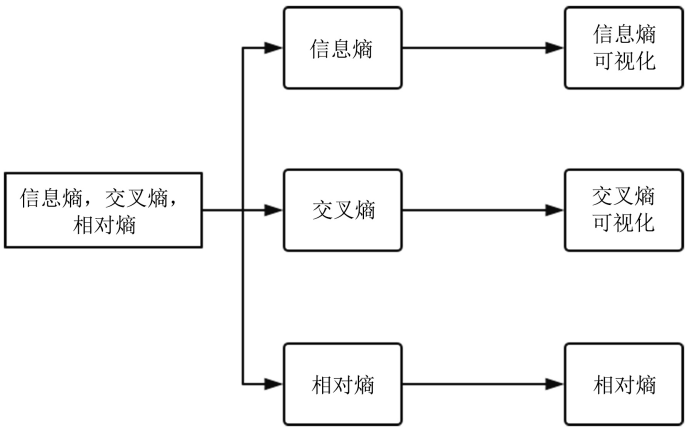

Figure 3. Visualization module
图 3. 可视化模块

3. 可视化结果分析

3.1. 四个地区，六种分布

如图 4 所示，可以看出四个地区的六种碱基序列分布基本呈现聚集的形式，符合不同地区之间新冠病毒的亲属关系，大致推断，每个地区的病毒并不是独立的；之后我们可以观察一下微观下分布的情况。

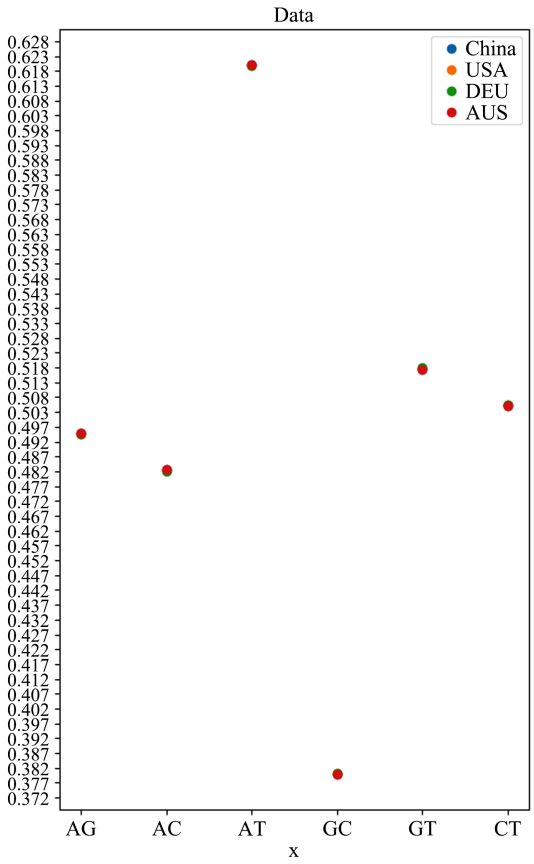

Figure 4. Six kinds of sequence frequency distribution
图 4. 六种序列频率分布

如图 5~10 所示，六种分布可以明显的看出，不同地区的新冠病毒基因序列是有这不小的差异，虽然差异在数量级上只有 0.001，但基于分析序列的所用的数据量庞大，可以知道其在宏观上是可以等价的，但是到了微观之处，该等价是不成立的。大概可以猜测六种分布的不同代表着新冠病毒在不同地区有着较为细微的差距。

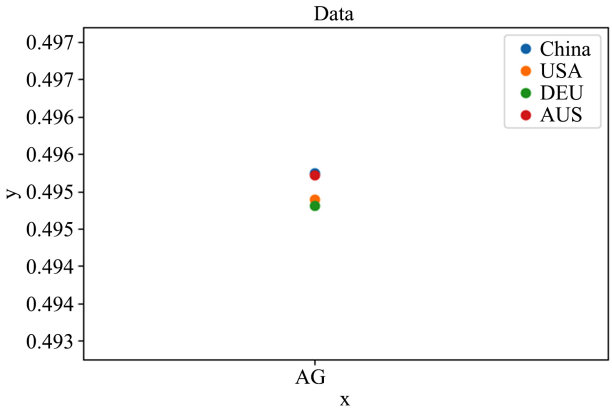

Figure 5. AG sequence frequency distribution
图 5. AG 序列频率分布

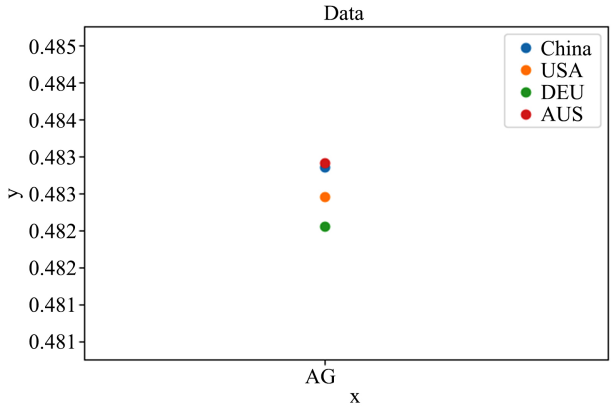

Figure 6. AC sequence frequency distribution
图 6. AC 序列频率分布

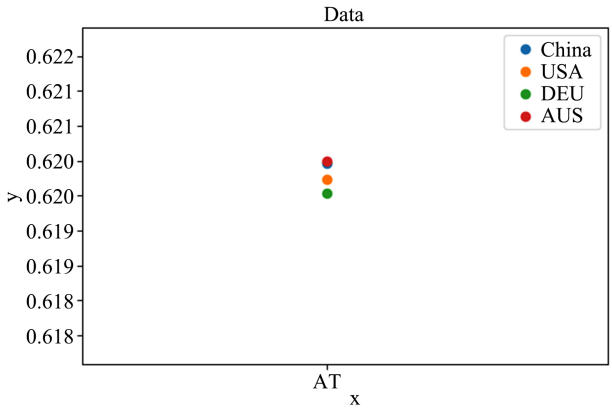

Figure 7. AT sequence frequency distribution
图 7. AT 序列频率分布

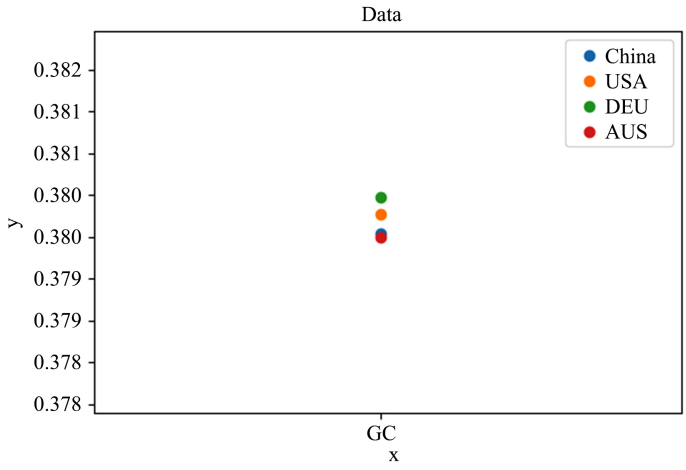

Figure 8. GC sequence frequency distribution
图 8. GC 序列频率分布

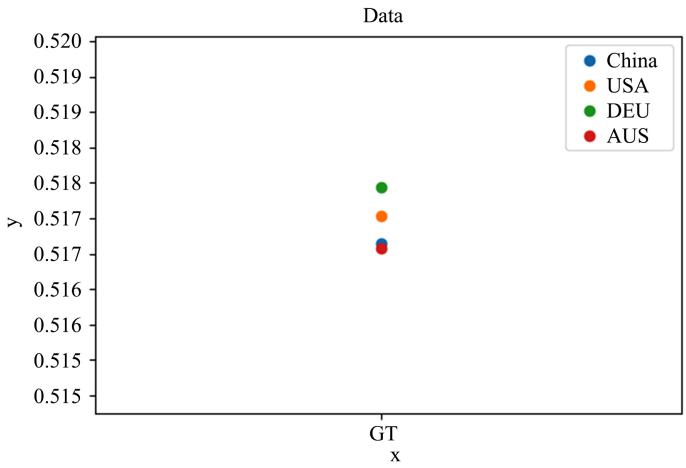

Figure 9. GT sequence frequency distribution
图 9. GT 序列频率分布

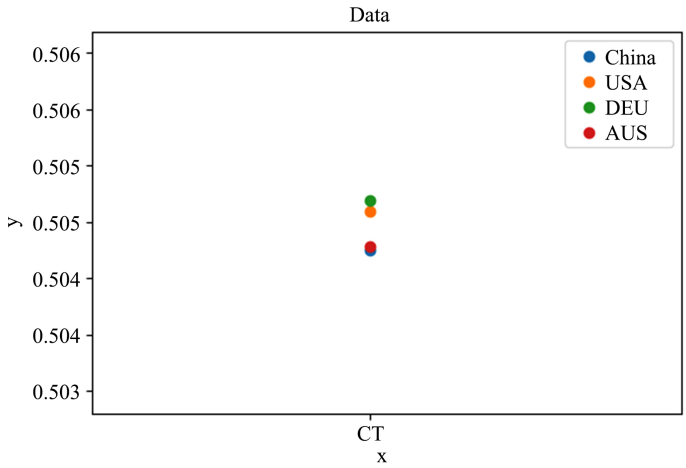

Figure 10. CT sequence frequency distribution
图 10. CT 序列频率分布

3.2. 基因序列的信息熵

公式

$$H(x)=-\sum_{i=1}^{n}p(x_i)\log p(x_i) \tag{4}$$

虽然从图 11 中可以看出各个地区病毒序列并没有太大的差别，但每个地区之间的病毒在微观上一定是存在一定差异的，从图 11 中根据信息熵的定义可以看出每个地区病毒所具有的“信息量”分布是不同的。从微观的角度可以看到四个地区之间的病毒是存在差异的。

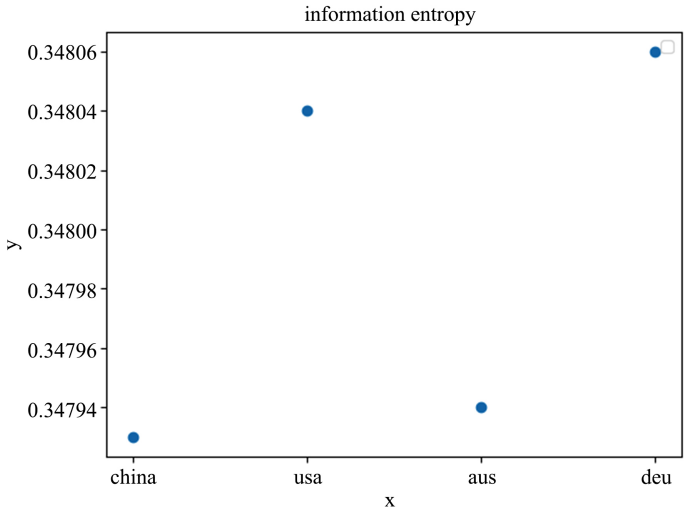

Figure 11. Information entropy of four regions
图 11. 四个地区信息熵值

3.3. 基因中 AG 分布的相对熵

对于同一个随机变量 AG 有两个单独的概率分布 P(AG)和 Q(AG)，相对熵来衡量这两个分布的差异。

公式

$$\begin{aligned}DKL(p\|q)&=-\int p(x)\ln q(x)-\left(-\int p(x)\ln p(x)\mathrm{d}x\right)\\&=\sum_{i=1}^{n}p(x_i)\log\left(\frac{p(x_i)}{q(x_i)}\right)\end{aligned} \tag{5}$$

如图 12 所示，相对熵用于衡量 AG 在不同国家的新冠序列中的分布差异，可以在图 12 中明显的看出，中国的新冠与美国和德国的新冠的基因序列中 AG 分布的差别不大，但中国的新冠与澳大利亚的新冠基因序列中 AG 分布与中国与美国，中国与德国之间的差距却非常明显，是否可以大致推断澳大利亚的新冠与中国，美国，德国之间存在某种差异。

3.4. 基因中 AG 分布的交叉熵

公式

$$DKL\_C=-\int p(x)\ln\left(q(x)\mathrm{d}x\right) \tag{6}$$

如图 13 所示交叉熵就是相对熵公式的前一半所得结果，此处去除的后一半是理论值，所以交叉熵更表现出真实性，而根据图 13 也可以看出来其结果与相对熵得出的结果一致。

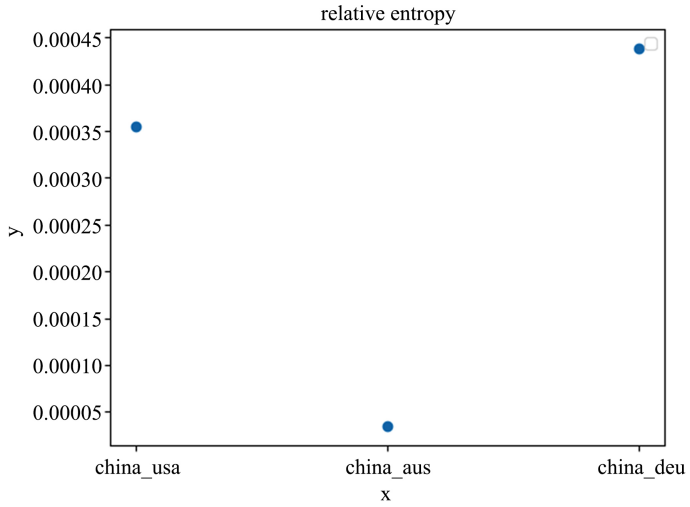

Figure 12. The relative entropy of the AG distribution
图 12. AG 分布的相对熵

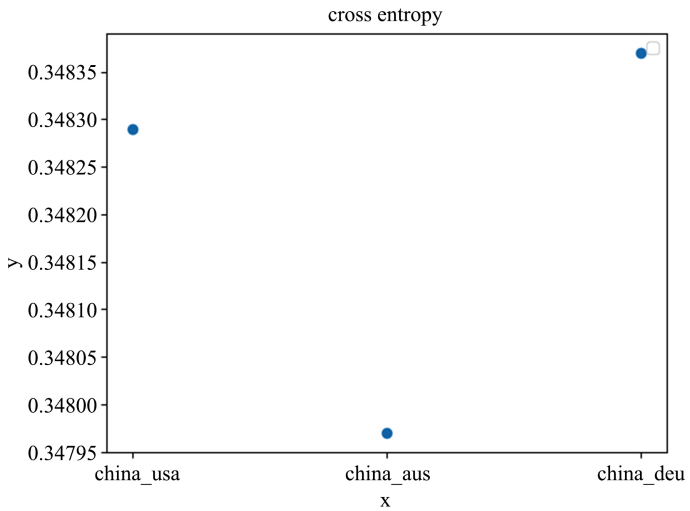

Figure 13. The cross entropy of the AG distribution
图 13. AG 分布的交叉熵

4. 总结

本文通过对四个地区的新型冠状病毒 DNA 序列信息熵，相对熵，交叉熵的计算，在实际过程中，可以主观的调节在对序列中不同碱基对的分组情况来得到结果，但最终的结果都是指向相同的结论，在此找出熵的最优值，来描述不同地区之间新冠病毒在微观上的差异，信息熵的引用与可视化的分析可以为之后关于基因的研究提供新的思路。

致　谢

感谢云南大学软件学院郑智捷教授的悉心指导，以及云南大学软件学院的大力支持。

参考文献

[1] Kullback, S. and Leibler, R.A. (1951) On Information and Sufficiency. *The Annals of Mathematical Statistics*, **22**,

79-86. https://doi.org/10.1214/aoms/1177729694

[2] Goodfellow, I., Bengio, Y. and Courville, A. (2016) Deep Learning (Vol. 1). MIT Press, Cambridge, 71-73.

[3] 百度百科. 交叉熵. https://baike.baidu.com/item/%E4%BA%A4%E5%8F%89%E7%86%B5/8983241?fr=aladdin