Hans Journal of Computational Biology 计算生物学, 2020, 10(3), 31-40
Published Online September 2020 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2020.103004

Hans汉斯

基于元胞自动机图的2019-nCoV溯源研究

薛广富，肖　绚*，许召春*

景德镇陶瓷大学生物大数据实验室，江西　景德镇
Email: jci_xgf@163.com, *jdzxiaoxuan@163.com, *jdzxuzhaochun@163.com

收稿日期：2020年8月6日；录用日期：2020年8月24日；发布日期：2020年8月31日

摘　要

病毒溯源是研究病毒人际传播规律和演化历史，发现病毒的传播机理，并寻找病毒源头是控制疫情的核心科研攻关环节。本文采用元胞自动机将病毒的基因组数据转换为图像，使用Canny边缘检测方法生成特征图像，采用SSIM值评估病毒序列特征图像的相似性，基于病毒序列图像相似性的聚类热图和A-T含量曲线，发现蝙蝠冠状病毒RaTG13和穿山甲冠状病毒在各个指标中都与2019-nCoV更为接近，进而推测出2019-nCoV可能来自蝙蝠，穿山甲可能是该病毒的中间宿主。

关键词

元胞自动机，基因序列可视化，2019-nCov，Canny边缘检测，结构相似性

2019-nCoV Traceability Research Based on Cellular Automata Diagram

Guangfu Xue, Xuan Xiao*, Zhaochu Xu*

Bio-Big Data Laboratory, Jingdezhen Ceramic University, Jingdezhen Jiangxi
Email: jci_xgf@163.com, *jdzxiaoxuan@163.com, *jdzxuzhaochun@163.com

Received: Aug. 6th, 2020; accepted: Aug. 24th, 2020; published: Aug. 31st, 2020

Abstract

Tracing the source of virus is the key link of scientific research to control the epidemic situation, which is to study the law of human transmission and evolution of virus, to understand the transmission mechanism of virus and to find the source of virus. This paper uses cellular automata to convert the genome data of the virus into feature images, and the Canny edge detection method was proposed to generate feature images, while SSIM was calculated to measure the structural similarity between the feature images of the two virus sequences. Through drawing the sequence image structure similarity clustering heat map and A-T content curve, it is concluded that bat coronavirus RaTG13 and pangolin

*通讯作者。

文章引用: 薛广富, 肖绚, 许召春. 基于元胞自动机图的 2019-nCoV 溯源研究[J]. 计算生物学, 2020, 10(3): 31-40.
DOI: 10.12677/hjcb.2020.103004

coronavirus are closer to 2019-nCoV in various indicators, and it is speculated that 2019-nCoV may come from bats, and pangolin may be the intermediate host of the virus.

Keywords

Cellular Automata, Gene Sequence Visualization, 2019-nCoV, Canny Edge Detection Method, Structural Similarity

1. 引言

新型冠状病毒的溯源问题仍然是一大挑战，已有研究团队在此方面进行了探索与研究。在发布2019-nCoV 基因组后不久，石正丽研究团队[1]发布了一种蝙蝠冠状病毒 RaTG13 的全基因组，RaTG13与 2019-nCoV 在全基因组水平上有 96%的同源性，这表明 2019-nCoV 很有可能来自蝙蝠。然而，由于这类蝙蝠与人类的直接接触非常罕见，类似于 SARS 和 MERS，2019-nCoV 似乎更有可能是从另一个中间宿主向人类的溢出，而不是直接来自蝙蝠。

此外，管轶等人[2]对华南地区马来亚穿山甲中 2019-nCoV 相关冠状病毒进行了鉴定，沈永义等人[3]对马来亚穿山甲中新冠类似病毒进行了分离与鉴定，陈金平[4]分析了穿山甲是否为新冠病毒的中间宿主，以及 Matthew C. Wong 等人[5]通过对冠状病毒进行重组，表明了新冠病毒或来源于穿山甲。张志刚等人[6]首次表述了 2019-nCoV 与穿山甲冠状病毒(Pangolin-CoV)、RaTG13、SARS 以及 MERS 的关系，研究表明穿山甲和蝙蝠一样，是 β 类冠状病毒的天然宿主，认为除了蝙蝠和穿山甲之外，2019-nCoV 还存在其它未知中间宿主。

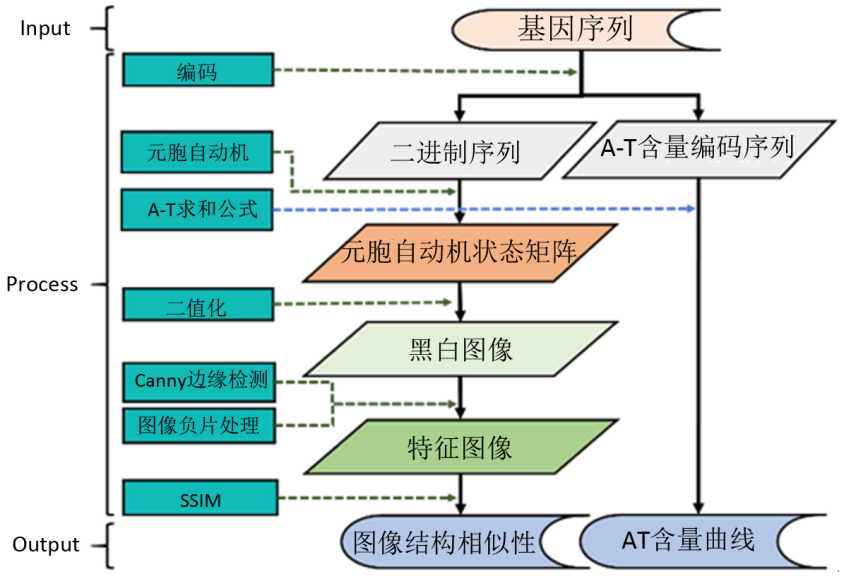

Figure 1. Data flow diagram of virus sequence visualization model
图 1. 病毒序列可视化模型数据流程图

本文使用元胞自动机对病毒的基因序列进行可视化处理(图 1)，可以将 2019-nCoV 与非 SARS 相关病毒序列准确区分，并分析了序列图像特征形成的原因，推测出 2019-nCoV 的中间宿主。

2. 数据与方法

2.1. 数据集

从 NCBI 网站下载了属于冠状病毒的 2019-nCoV 的全基因组序列，同时为了研究 2019-nCoV 的潜在中间宿主，还下载了其他病毒的全基因组序列进行对比分析。具体的病毒序列及其相关信息见表 1。

Table 1. List of virus sequence information of various species
表 1. 各物种病毒序列信息表

病毒名称	英文缩写	宿主	ACCESSION
2019 新型冠状病毒	2019-nCoV	人	MN908947
严重急性呼吸系统综合症冠状病毒	SARS-CoV	人	NC_004718
中东呼吸综合征冠状病毒	MERS-CoV	人	NC_019843
人冠状病毒 229E	HCoV-229E	人	NC_002645
人冠状病毒 NL63	HCoV-NL63	人	NC_005831
人冠状病毒 OC43	HCoV-OC43	人	NC_006213
人冠状病毒 HKU1	HCoV-HKU1	人	NC_006577
埃博拉病毒	Ebola	人	NC_002549
穿山甲冠状病毒	Pangolin-CoV	穿山甲	MT040336
蝙蝠冠状病毒 RaTG13	Bat-CoV RaTG13	蝙蝠	MN996532
菊头蝠 SARS 样冠状病毒 HKU3-2	Bat SARS HKU3-2	蝙蝠	DQ084199
扁颅蝠冠状病毒 HKU4-1	Bat-CoV HKU4-1	蝙蝠	EF065505
猪血凝性脑脊髓炎病毒	PHEV	猪	DQ011855
鹿冠状病毒	Water deer coronavirus	鹿	MG518518
牛冠状病毒	Bovine coronavirus	牛	NC_003045
骆驼冠状病毒 HKU23	Camel-CoV HKU23	骆驼	KT368891
马冠状病毒	Equine coronavirus	马	LC061272
鼠肝炎病毒	Murine hepatitis Virus	鼠	AY700211

2.2. 病毒序列可视化模型算法流程描述

元胞自动机[7] (Cellular Automata, CA)与和一些传统的方法相比，其可以较为容易地模拟仿真物种演化、晶格生长、流体形成和化学反应过程等难以解析表达的复杂现象[8]。

在元胞自动机中每个元胞的状态都受到其相邻元胞状态的影响。一维初等元胞自动机(ECA)是状态集中只有两个元素且邻居半径为 1 的一维元胞自动机。对于 ECA，邻居个数为 $N = 2$，因此确定下一状态的映射函数如下所示：

$$C_i^{t+1} = F\left(C_{i-1}^t, C_i^t, C_{i+1}^t\right) \tag{1}$$

F 是元胞自动机的演化规则[9]，C_i^{t+1} 表示元胞 i 在 $t + 1$ 时刻的状态。图 2 显示了一维 CA 的演变。

水平轴是空间，垂直轴是时间步长。行表示元胞空间在某一时间的整体状态，列表示同一个元胞在不同时间的状态。

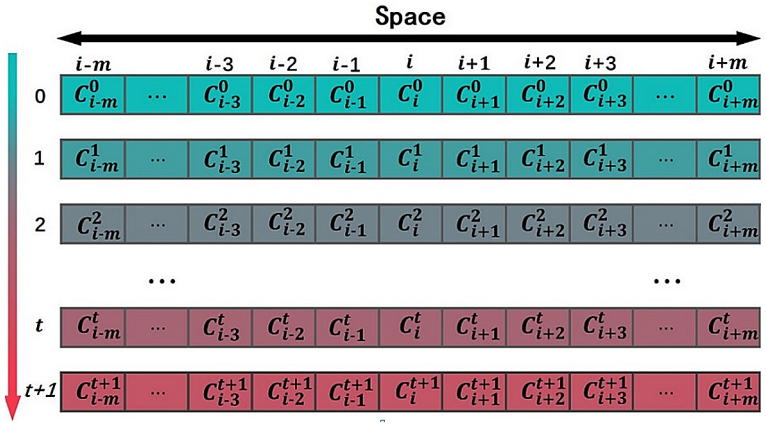

Figure 2. Evolution of one-dimensional CA
图 2. 一维 CA 的演变

公式(1)中演化规则含有中心元胞及其两个邻居，每个元胞都分别有两种状态，所以输入状态中一共有 $2^3=8$ 种组合方式：

$$\left[111\quad 110\quad 101\quad 100\quad 011\quad 010\quad 001\quad 000\right] \tag{2}$$

每一个输入条件都对应着两种输出状态 0 或 1，一共存在 $2^8=256$ 种状态组合。即一维初等元胞自动机总共存在 256 种演化规则。Wolfram 在研究这些元胞自动机的时候对他们进行了标号，即将每种输入条件的输出 0 或 1 排列看成一个 8 位 2 进制数。

例如对于 Wolfram 的 184 号规则，其输出状态可表示为：

$$F_{184}=(10111000)=184 \tag{3}$$

即 184 号规则的映射关系为：

$$\begin{array}{c} t \\ t+1 \end{array}\begin{bmatrix} 111 & 110 & 101 & 100 & 011 & 010 & 001 & 000 \\ \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\ \mathbf{1} & \mathbf{0} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix} \tag{4}$$

2.2.1. 基因编码

为了降低元胞自动机的计算复杂度，本文将 DNA 的四种碱基编码为 01 数字信号，每一种碱基都由两位二进制的形式表示。腺嘌呤(A)、鸟嘌呤(G)、胞嘧啶(C)和胸腺嘧啶(T)分别编码为“00”、“10”、“01”和“11”。通过以上编码，DNA 序列转化成数字序列，例如序列为“TTGCAGC”的基因片段由上述规则可以编码为“11111001001001”。这种编码方式具有可逆性，给定一个数字序列，也可以通过上述编码规则将其解码为 DNA 序列。

2.2.2. 元胞自动机状态矩阵生成

假设某 DNA 序列 S 的长度为 N，编码后可以得到长度 $2N$ 的数字序列，将此一维序列赋值为一维初等元胞自动机的初始元胞空间状态。在设置演化规则 F、时间(演化次数) T 后，本文采用循环边界条件对初始序列进行演化。

演化过程中，将各个元胞的状态存放在一个长度为 $2N$，高度为 T 的二维矩阵 Q，元胞状态的二维矩

阵 Q 的映射关系如下所示：

$$Q_{(i,t)} = C_i^t = F\left(C_{i-1}^{t-1}, C_i^{t-1}, C_{i+1}^{t-1}\right) \ (1 < i < 2N, 1 < t < T) \tag{5}$$

$$Q_{(1,t)} = C_1^t = F\left(C_{2N}^{t-1}, C_1^{t-1}, C_2^{t-1}\right) \ (1 < t < T) \tag{6}$$

$$Q_{(2N,t)} = C_{2N}^t = F\left(C_{2N-1}^{t-1}, C_{2N}^{t-1}, C_1^{t-1}\right) \ (1 < t < T) \tag{7}$$

$Q_{(i,t)}$ 表示第 i 行、第 t 列元素的值，C_i^t 表示在 t 时刻所处位置为 i 的元胞的状态。

2.2.3. 图像生成

元胞自动机图通常使用不同的颜色表示各个元胞的状态，在元胞状态为 0、1 的元胞自动机图中，常常使用黑白两种颜色来表示，即□ = 0，■ = 1。

在黑白图像中，灰度值为 0 表示白色，灰度值为 255 表示黑色。在将上述的元胞自动机状态矩阵 Q 可视化过程中，本文将该矩阵的元素进行二值化处理，从而使得生成的图像呈现出明显的纹理效果，具体公式如下所述：

$$Q'_{(ij)} = \begin{cases} 0, & Q_{(i,j)} = 0 \\ 255, & Q_{(i,j)} = 1 \end{cases} \ (1 < i < 2N, 1 < j < T) \tag{8}$$

通过上述操作可以生成一个尺寸为 $2N \times T$ 的黑白图像。

为了缓解的计算机的运行压力，本文中将所有生成的黑白图像进行缩放，其缩放后的图片尺寸为 10000 × 3000 (合计 30,000,000 个像素点)。

2.2.4. 图像特征处理

图像边缘是指某区域像素点的灰度值有阶跃变化或屋顶变化，其反映了图像中各个区域的灰度不连续性，边缘检测方法有助于突出图像的特征。Canny 边缘检测是 John Canny 在 1986 年提出的图像边缘检测算法，本文使用该算法提取病毒序列可视化后的图像特征，并使用图像中最大灰度值的像素点作为阈值 $\left(L = \max(r)\right)$，对所有的像素点进行负片处理：

$$T(r) = L - r \tag{9}$$

L 是待处理图片中最大灰度值的像素点，r 是待处理图片中的像素点灰度值，$T(r)$ 是已经过负片处理图像中像素点的灰度值。

如图 3 所示，图片 3A 为元胞自动机状态矩阵可视化的图像，图片 3B 为经过缩放后的图像，图片 3C 是经过 Canny 边缘检测后的特征图像，图片 3D 为特征图像经过负片处理后的特征图像。

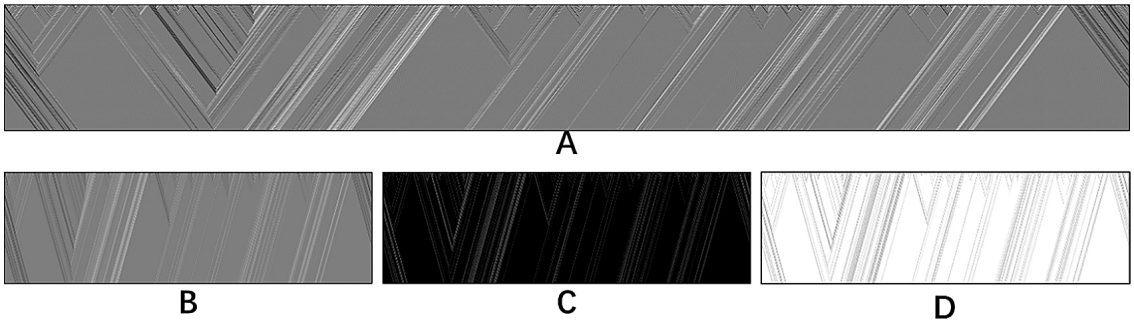

Figure 3. Feature images of the virus sequence after various stages of processing
图 3. 病毒序列经过各个阶段处理后的特征图像

2.3. 评估指标

2.3.1. SSIM 指标

结构相似性(Structural Similarity, SSIM)是由 Wang 等人提出并用于衡量两个图像之间相似性的方法，通过比较两图像的亮度、对比度和结构来衡量图像之间的相似性。为了比较各种病毒元胞自动机图的相似性，本文使用 SSIM 值描述病毒序列特征图像之间的相似性。

2.3.2. 基因序列 A-T 含量统计

DNA 序列中 A 与 T 之间含有 2 个氢键，C 与 G 之间含有 3 个氢键，这些氢键在 DNA 结构的稳定性中发挥着关键的作用。在 DNA 单链中，序列内部 A-T 和 C-G 之间的连接也影响着单链的稳定性和单链的结构。统计 A-T 和 C-G 在序列各个位置中的分布，可以得出不同病毒之间的基因序列的异同。

本文主要对序列中各个位置的 A-T 含量分布进行研究，DNA 序列的编码方式如下：

$$A=-1,T=1,C=0,G=0 \tag{10}$$

通过以上编码，DNA 序列便可以用“0，1，-1”的序列表示。通过求和函数可以将 DNA 序列转换成二维空间的一条曲线：

$$Sum(P,j)=\sum_{i=1}^{i=j}P_i,\quad j=1,2,\cdots,L \tag{11}$$

其中 P_i 表示处于第 i 个位置的碱基的编码，L 为序列 P 的长度。横坐标为序列碱基的位置，纵坐标表示各个位置的公式(11)的求和值，将各个位置所对应的值绘制成曲线，通过分析曲线的变化可以反映序列的特征。

3. 实验结果

本文应用以上的算法步骤，使用 184 号规则，演化次数为 5000，将病毒的基因序列转化为二维的图像，然后分析和比对各个病毒序列之间的图像特征。

3.1. 病毒基因序列的特征图像特点

2019-nCoV、RaTG13 和 pangolin-CoV 的特征图像，如图 4、图 5 和图 6 所示，其余病毒序列的特征图像已上传至 GitHub (https://github.com/MMCXXVI/Viral-sequence-visualization-model)。

通过对所有病毒的基因序列进行可视化处理，可以明显地发现所有冠状病毒的特征图像都有向左向右倾斜的条纹特征，但图片中向左倾斜“/”形条纹占大多数，且 SARS 相关病毒 2019-nCoV、SARS-CoV

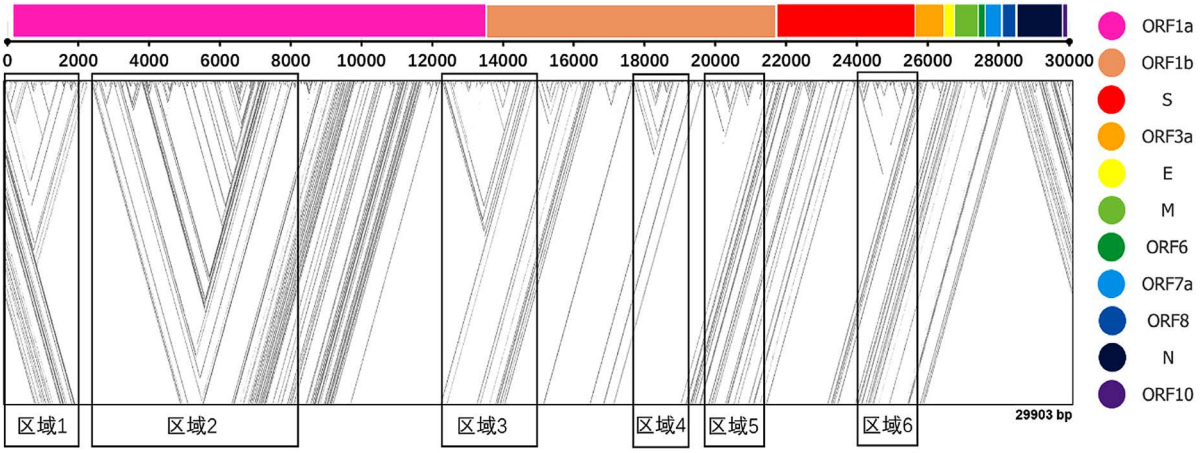

Figure 4. Sequence image of 2019-nCoV (MN908947)
图 4. 2019-nCoV (MN908947)病毒序列特征图像

中都含有 6 个明显的“V”字形交叉区域，且大多分布在“ORF1a”、“ORF1b”、“S”这三个基因区域，其中 2019-nCoV 特征图像的“V”字形交叉区域映射在序列中的位置大约为 1-2348bp、2394-7802bp、12251-14692、17610-18854、19787-20983bp、23878-25336bp。

在对蝙蝠 RaTG13 (MN996532) [1]和穿山甲 pangolin-CoV (MT040336)病毒序列的进行可视化处理后(图 5、图 6)，可以发现两者的特征图像都含有六个明显的“V”字形交叉区域，在与 2019-nCoV 的病毒序列特征图像进行比对后可以发现，三者的特征图像无论是在斜条纹的类型分布还是在“V”字形交叉区域的大小和位置分布上都极为相似。

同时，与冠状病毒特征图像有所不同的是，虽然埃博拉病毒的特征图像也有向左向右倾斜的条纹特征，但大多数为向右倾斜的“\”形条纹。

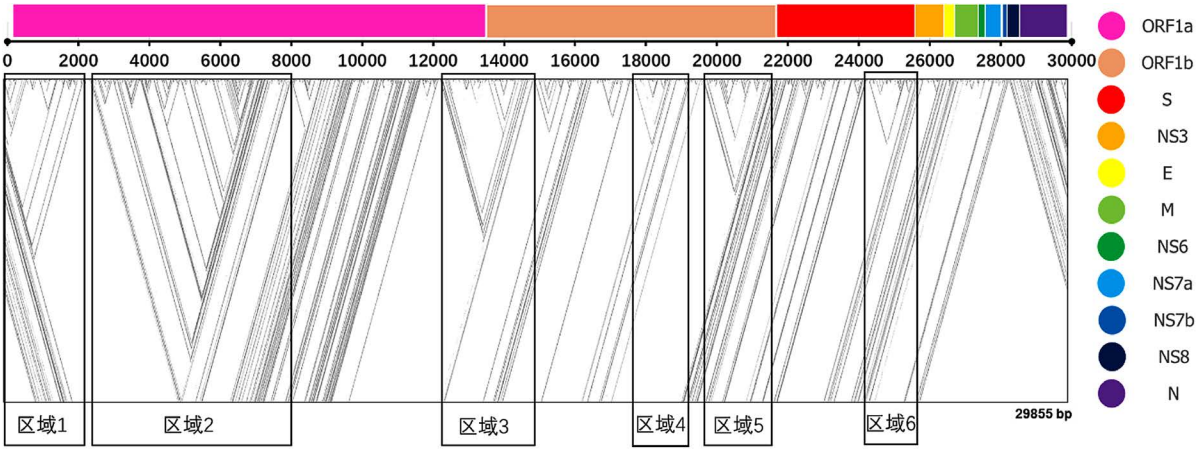

Figure 5. Sequence image of bat coronavirus RaTG13 (MN996532)
图 5. 蝙蝠冠状病毒 RaTG13 (MN996532)序列特征图像

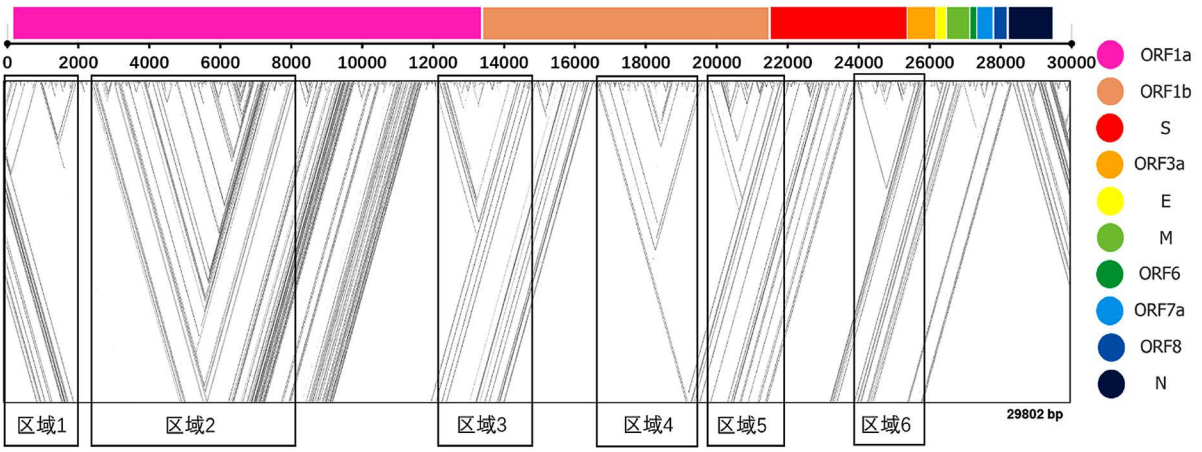

Figure 6. Sequence image of pangolin-CoV (MT040336)
图 6. 穿山甲冠状病毒 pangolin-CoV (MT040336)序列特征图像

3.2. 特征图像相似性分析

使用 SSIM 算法对数据集中的各个病毒特征图像之间的相似性进行计算，同时将结果绘制成聚类热图，如图 7 所示。

在感染人类的 8 种病毒序列中，只有 2019-nCoV 和 SARS 之间的图像相似性达到了 73.87%，其余病

毒的之间的特征图像相似性均小于 60%。

在分析感染人类的病毒和其他物种病毒之间的图像相似性后可以得知，与 2019-nCoV 特征图像相似性最高的病毒分别为蝙蝠冠状病毒 RaTG13 (77.12%)，穿山甲冠状病毒(73.36%)，蝙蝠 SARS 类冠状病毒 HKU3-2 (72.86%)；与 SARS 特征图像相似性最高的病毒分别为蝙蝠 SARS 类冠状病毒 HKU3-2 (74.97%)，蝙蝠冠状病毒 RaTG13 (74.22%)，穿山甲冠状病毒(72.67%)，其余病毒的之间的特征图像相似性均小于 65%。同时，蝙蝠冠状病毒 RaTG13、穿山甲冠状病毒与 2019-nCoV 之间的图像相似性均高于这两种冠状病毒与 SARS-CoV 之间的图像相似性。

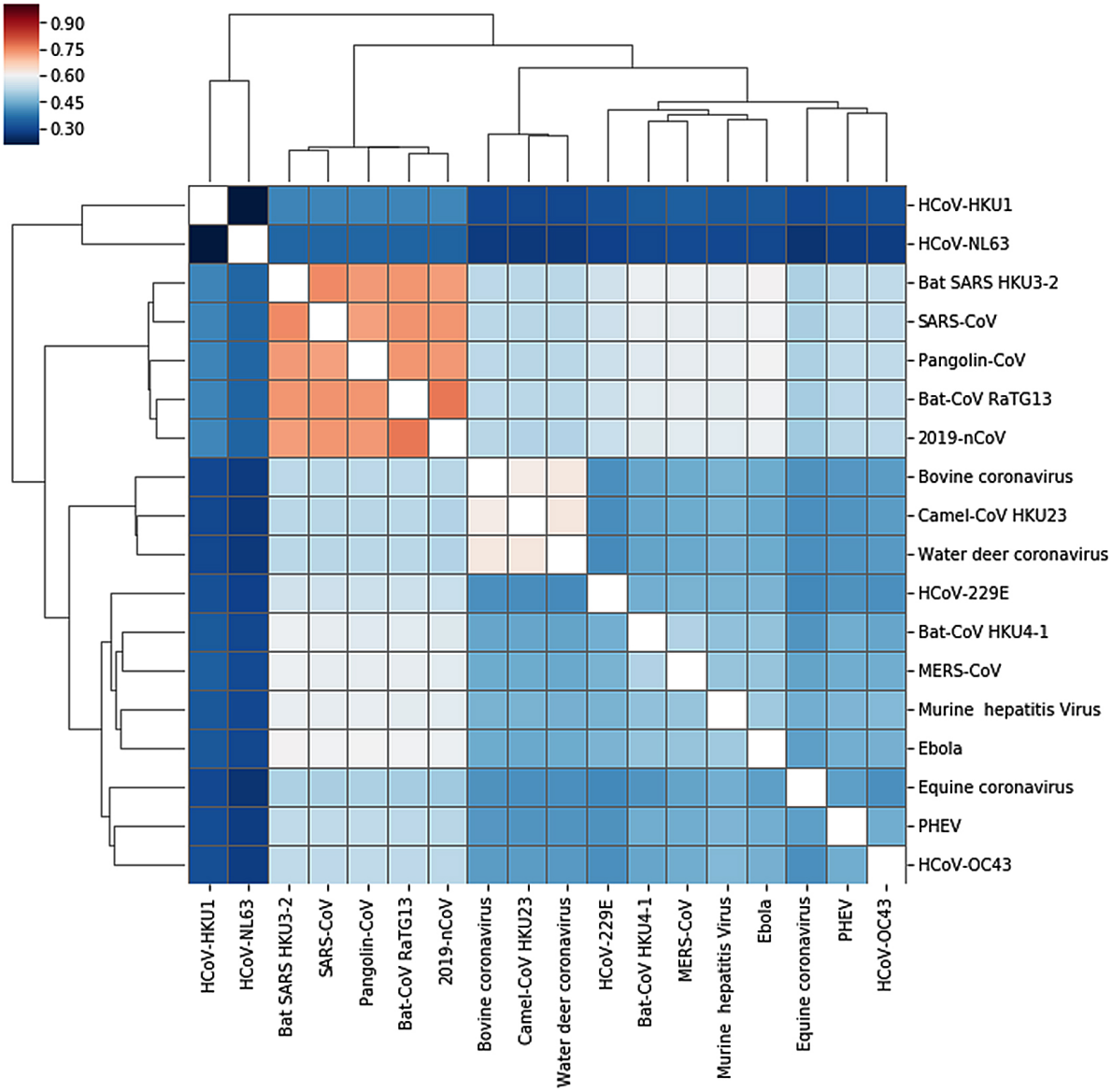

Figure 7. Clustering heat map of similarity between feature images of various virus sequences
图 7. 各个病毒序列特征图像之间相似性聚类热图

3.3. A-T 含量曲线

通过 1.3.2 所描述的方法对数据集中的病毒序列进行编码，然后计算得出的 A-T 含量曲线如图 8 所示。

在图 8A 中，曲线主要分为三种类型，一种是以 MERS 病毒为代表的呈平滑上升趋势的曲线，另一种是以 2019-nCoV 为代表的呈凹凸状上升的曲线，还有一种是以 Ebola 病毒为代表的一直呈下降趋势的曲线。

为了分析 2019-nCoV 的潜在宿主，本文对类型为凹凸状上升的曲线进行分析，其 A-T 含量曲线图像如图 8B 所示，该类型的曲线包含有 SARS-CoV、2019-nCoV、穿山甲冠状病毒、蝙蝠冠状病毒 RaTG13 和蝙蝠 SARS 类冠状病毒 HKU3-2。

在图 8B 中，可以明显看出有 6 个凹形区域，且五种病毒的 A-T 曲线变化趋势大致相同。本文以 2019-nCoV 特征图像中六个“V”字形交叉区域与该曲线进行比对，发现 A-T 含量曲线中的凹形区域与 2019-nCoV 特征图像中六个“V”字形交叉区域高度重合。通过公式 10 可以得知，曲线中出现凹形说明该区域的前半段以 A 为主，而后半段以 T 为主，所以在特征图像中出现“V”字形交叉区域。

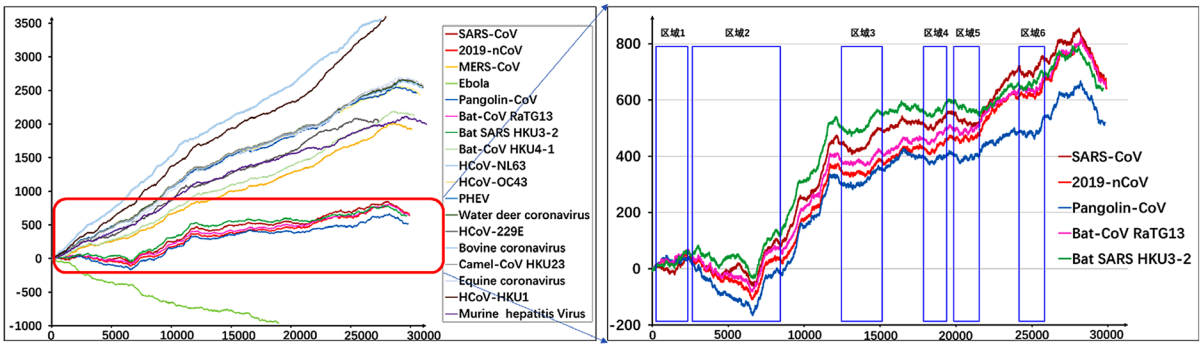

Figure 8. A: A-T content curve of each viral gene sequence; B: A-T content curve of ascending and descending; the blue box corresponds to the six “V” cross areas in the 2019-nCoV feature image; The X coordinate represents the position of base in the sequence (in bp), and the Y-coordinate represents the value calculated by Formula (11)

图 8. A：各个病毒基因序列的 A-T 含量曲线；B：凹凸状上升的 A-T 含量曲线，蓝色方框对应着 2019-nCoV 特征图像中六个“V”字形交叉区域；横坐标代表碱基在序列中的位置(单位为 bp)，纵坐标代表由公式(11)计算出的值

由此可以得出，在某个序列片段中，如果 A 的含量大于 T 的含量，会导致该片段所对应特征图像出现向左倾斜的“/”形条纹，反之则出现向右倾斜的“\”形条纹。且基因序列的特征图像中若要出现“V”字形交叉区域，该区域所处的序列片段中，前半段多以 A 为主，后半段多以 T 为主，且 T 与 A 的比例接近于 1:1。

同时，从图 8 可以看出，SARS 病毒和蝙蝠 SARS 类冠状病毒曲线之间的变化趋势更为接近；而蝙蝠冠状病毒 RaTG13、穿山甲冠状病毒的曲线与 2019-nCoV 更为接近。

4. 讨论与总结

本文从各个病毒序列的特征图像得出，蝙蝠冠状病毒 RaTG13、穿山甲冠状病毒和 2019-nCoV 的图像极为相似，都存在 6 个 V”字形交叉区域。再经过 SSIM 算法计算各个特征图像的相似性后得知，除 SARS 病毒之外，只有蝙蝠冠状病毒 RaTG13、穿山甲冠状病毒的特征图像与 2019-nCoV 序列的图像相似性均高于 70%。同时，通过 A-T 含量图可知，蝙蝠冠状病毒 RaTG13、穿山甲冠状病毒的曲线与 2019-nCoV 更为接近。

综上，我们推测 2019-nCoV 很有可能来自蝙蝠，且穿山甲很可能是该病毒的中间宿主。虽然本章节所提出的方法并不能直接证明 2019-nCoV 的来源与中间宿主，但若是通过对“V”字形交叉区域的序列片段进行分析，进一步了解产生该现象的机理，这将会有助于新冠肺炎的差异化的治疗和防控。

虽然现在没有证据能完全确定 2019-nCoV 的是经由穿山甲传播，但不可否认的是，穿山甲有很大可能是 2019-nCoV 的中间宿主。要回答 2019-nCoV 起源与中间宿主这一科学问题，不仅需要在生物信息学、免疫学、分子流行病学和病毒传播模式学等各个方面进行分析，同时还需要生物学家使用病毒组学、反转录聚合酶链反应以及酶联免疫吸附测定等方法进行验证，才能得出确定的答案。

基金项目

国家自然科学基金项目(31860312)；科技部政府间科技合作项目(国科外字[2018] 31 号)；江西省自然科学重点基金项目(20171ACB20023)；景德镇市科技计划项目(20192GYZD008-04)。

参考文献

[1] Zhou, P., Yang, X.-L., Wang, X.-G., *et al.* (2020) A Pneumonia Outbreak Associated with a New Coronavirus of Probable Bat Origin. *Nature*, **579**, 270-273.

[2] Lam, T.T., Jia, N., Zhang, Y., *et al.* (2020) Identifying SARS-CoV-2-Related Coronaviruses in *Malayan pangolins*. *Nature*, **583**, 282-285. https://doi.org/10.1038/s41586-020-2169-0

[3] Xiao, K., Zhai, J., Feng, Y., *et al.* (2020) Isolation and Characterization of 2019-nCoV-like Coronavirus from Malayan Pangolins. *BioRxiv*. https://doi.org/10.1101/2020.02.17.951335

[4] Liu, P., Jiang, J.-Z., Hua, Y., *et al.* (2020) Are Pangolins the Intermediate Host of the 2019 Novel Coronavirus (2019-nCoV)? *Biorxiv*. https://doi.org/10.1101/2020.02.18.954628

[5] Wong, M.C., Cregeen, S.J.J., Ajami, N.J., *et al.* (2020) Evidence of Recombination in Coronaviruses Implicating Pangolin Origins of nCoV-2019. *Biorxiv*. https://doi.org/10.1101/2020.02.07.939207

[6] Zhang, T., Wu, Q. and Zhang, Z. (2020) Probable Pangolin Origin of SARS-CoV-2 Associated with the COVID-19 Outbreak. *Current Biology*, **30**, 1246-1351. https://doi.org/10.1016/j.cub.2020.03.063

[7] Schwartz, J.T., Neumann, J.V. and Burks, A.W. (1967) Theory of Self-Reproducing Automata. *Quarterly Review of Biology*, **21**, 745. https://doi.org/10.2307/2005041

[8] 季海鹏. 基于元胞自动机法的 316LN 不锈钢动态再结晶组织预测[D]: [博士学位论文]. 秦皇岛: 燕山大学, 2013.

[9] Sirakoulis, G., Karafyllidis, I., Mizas, C., *et al.* (2003) A Cellular Automaton Model for the Study of DNA Sequence Evolution. *Computers in Biology and Medicine*, **33**, 439-53. https://doi.org/10.1016/S0010-4825(03)00017-9