Hans Journal of Computational Biology 计算生物学, 2021, 11(1), 12-19
Published Online March 2021 in Hans. http://www.hanspub.org/journal/hjcb
https://doi.org/10.12677/hjcb.2021.111002

Hans汉斯

# 新冠病毒基因序列S蛋白信息熵可视化分布

**吴　可，张月晴，黄嘉政，董芯宇，张舒智，郑智捷**

云南大学软件学院，云南　昆明
Email: 1099161824@qq.com

收稿日期：2021年2月11日；录用日期：2021年3月11日；发布日期：2021年3月23日

## 摘　要

**新冠肺炎(COVID-19)在全球范围爆发，至今仍未得到有效控制。新冠病毒(SARS-CoV-2)表面的刺突蛋白(spike protein, S)在病毒传播中起着十分重要的作用，针对它的分析在疾病预防与免疫中具有重要的应用价值。本文分析了新型冠状病毒基因序列的碱基分布及S蛋白基因的突变情况。针对相关新冠病毒基因序列进行多种可视化处理及分析，选择多条S蛋白基因序列，运用BLAST以及MEGA6软件进行信息比对、对齐，再进行信息熵的计算、展示可视化分布及相关分析。结果显示，新冠病毒基因碱基的整体分布具有对称性，由于选择的S蛋白数量不大变异量较小，其信息熵可视化分布呈现的特征聚点数目也较少。**

## 关键词

**COVID-19，SARS-CoV-2，S蛋白，信息熵，可视化分布**

# Visual Distribution of Information Entropy on SARS-CoV-2 Spike Protein

**Ke Wu, Yueqing Zhang, Jiazheng Huang, Xinyu Dong, Shuzhi Zhang, Jeffrey Zheng**

School of Software, Yunnan University, Kunming Yunnan
Email: 1099161824@qq.com

Received: Feb. 11th, 2021; accepted: Mar. 11th, 2021; published: Mar. 23rd, 2021

## Abstract

**At present, the COVID-19 is breaking out on a global scale, and it has not been effectively controlled. Because the surface spike protein of SARS-CoV-2 genomes plays an important role in the**

文章引用: 吴可, 张月晴, 黄嘉政, 董芯宇, 张舒智, 郑智捷. 新冠病毒基因序列 S 蛋白信息熵可视化分布[J]. 计算生物学, 2021, 11(1): 12-19. DOI: 10.12677/hjcb.2021.111002

**spread of the virus, it provides valuable information for fighting COVID-19 and vaccine practices. This paper analyzed the base distribution of SARS-CoV-2 genomes and the mutation of S protein gene. It made visualization to analyze the relevant gene sequences. Multiple S protein gene sequences are selected, then the BLAST and MEGA6 are applied to compare and align them. Then S proteins are calculated their information entropy, and made visualization of their entropy distributions. The visual results show that the base distributions of SARS-CoV-2 genomes have symmetrical properties. Due to smaller number of S proteins selected, there are only a limited number of clustering on their distributions of information entropy.**

## Keywords

**COVID-19, SARS-CoV-2, S Protein, Information Entropy, Visual Distribution**

Copyright © 2021 by author(s) and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

## 1. 引言

2020 年 12 月中旬，英国首次报告了传染性极强的新冠(SARS-CoV-2)病毒突变株：B.1.1.7 毒株，这一毒株很快成为伦敦地区主要毒株。

S 蛋白(spike protein，刺突蛋白)在冠状病毒进入宿主细胞时，主要由跨膜刺突(S)糖蛋白进行介导并参与入胞过程，在新型冠状病毒的传播过程中发挥着重要的作用[1]。在研究新冠病毒基因序列时，S 蛋白表位也一直是抗体研发的研究对象，它具有一定的代表性和特征性[2]，所以主要以分析 S 蛋白为出发点来进行进一步的研究[3]。而 B.1.1.7 毒株在 S 蛋白受体结合结构域(RBD) 501 位置出现突变，这让我们将目光放到了 B.1.1.7 毒株的 S 蛋白上。

目前，基因研究的常规方式是通过基因特征提取、基因序列定位等方式来找到关键位点进行研究[4]。

信息熵[1]的概念自 1948 年被香农提出后，随着科技的发展，已突破香农信息论的范围，在生物医学领域被研究应用并取得成果，成为现代生物医学领域中的一种新思路、新方法。因此本文将信息熵引入生物信息学领域，根据信息熵的概念和香农公式[2]，结合生物学领域的计算和分析要求，对需要的生物信息进行熵值计算和分析，之后将计算出的熵值在二维空间进行投影，以达到可视化分析的目的[3] [5] [6] [7]。

本文将在第二章中介绍系统架构，使用 S 蛋白熵值可视化将 S 蛋白的基因序列进行处理、计算、投影成二维可视化图示，在第三章中以可视化的结果来对复杂的 S 蛋白进行展示。

## 2. 系统架构

### 2.1. 参数解释

$m$：DNA 序列各分组长度(这里我们选取了 $m=80$ 进行可视化处理分析)；

$M$：每条序列组数，$M=\frac{ALL}{m}$ ($ALL$ 为每条序列总碱基数)；

$P_{Ni}$：每组分组对应排列 $N(i=1,2,3,\cdots,M$；N = A、C、G、T、AC、AG、AT、CG、CT、GT、ACG、ACT、AGT、CGT、ACGT，顺序无关)的出现概率，$P_{Ni}=\frac{SUM(N)}{m}$；

[1]信息熵：香农从热力学中借用过来，以描述信源的不确定度。

$E_N$：每条序列各排列分组的信息熵($N$ = A、C、G、T、AC、AG、AT、CG、CT、GT、ACG、ACT、AGT、CGT、ACGT)，计算公式为

$$E_N = -\sum_{i=1}^{M} P_{Ni} \log_2 P_{Ni}$$

$(E_N, E_N)$：每条序列信息熵映射生成图像上的点，如$(E_A, E_C)$即表示该条序列碱基 $A$ 与 $C$ 的信息熵在图像上的映射。

## 2.2. 架构

系统整体架构如图 1 所示，分为输入、计数、处理、投影、输出五个模块，其中核心功能模块有三个，分别是处理、测量、投影模块。


**Figure 1.** Architecture
**图 1.** 架构

计数模块的功能是将新冠病毒基因序列 S 蛋白以 m 个碱基为一段，进行四种碱基(A，T，C，G)的数量统计；在处理模块中需要对各段信息熵进行计算，再累加各段信息熵得到 S 蛋白的熵值；在投影模块中，可以选择查看单张信息熵分布图或 225 张信息熵分布图的全排列。

## 2.3. 计数模块

计数模块是针对输入的编码 S 蛋白的基因序列进行处理，下载的编码 S 蛋白的基因序列自动以 80 个碱基为一段分好，不需要手动分段，因此直接导入计数程序即可计算出每一段中四种碱基(A，T，C，G)的数量，架构图如图 2 所示。

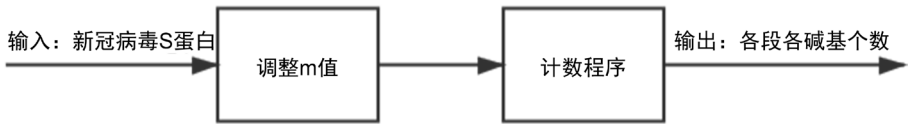


**Figure 2.** Counting module
**图 2.** 计数模块

## 2.4. 处理模块

处理模块主要是根据香农公式对计数模块输出的结果进行处理，分别计算出整条编码 S 蛋白基因的 15 种碱基组合($N$ = A、C、G、T、AC、AG、AT、CG、CT、GT、ACG、ACT、AGT、CGT、ACGT)的熵值[5] [6]，架构如图 3 所示。

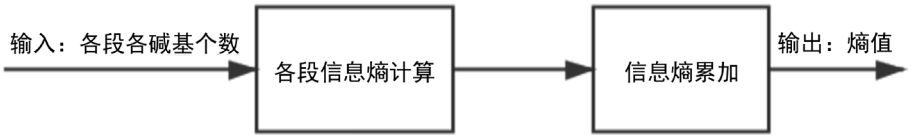


**Figure 3.** Processing module
**图 3.** 处理模块

### 2.5. 投影模块

投影模块可以将除含空集外的 15 个碱基组合中的任意两个的信息熵值作为 X、Y 轴，生成相关的信息熵分布图像，并使用不同颜色加以区分，也可以直接生成 225 张散点图，以 N 中排列顺序输出到同一张大图中，架构如图 4 所示。

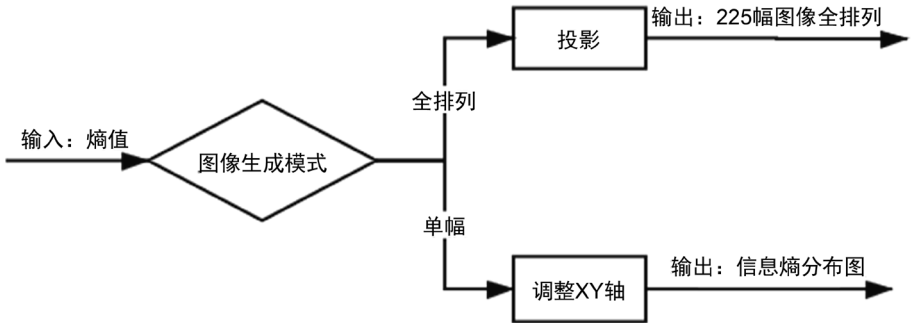


**Figure 4.** Projection module
**图 4.** 投影模块

## 3. 过程与结果分析

### 3.1. 序列选择

本文意图通过研究新冠病毒变异毒株 S 蛋白与初期毒株的 S 蛋白信息熵的差异，找出毒株突变的生物学原理与生物学表达，为生物信息、生命科学等方面提供一定的研究基础，因此我们选择了 B.1.1.7 毒株中的四条序列和原始序列，提取 S 蛋白段进行尝试。

### 3.2. 序列比对

选取初期序列(NC_045512.2)作为基准序列和 B.1.1.7 毒株棘突蛋白的四条序列(三条来自苏格兰的序列 CVR5974、CVR6031、CVR6032，一条来自英格兰的序列 204590575)，使用对齐工具 MEGA6 对齐五条序列，根据 S 蛋白的前后碱基排列特征截取五条序列的 S 蛋白段。

对病毒基因序列的 S 蛋白段进行碱基对比分析，如图 5 所示。

查询序列统计信息

| Sequence ID | Length(bp) | GC% | Total N | BLASTN Hit |
|---|---|---|---|---|
| NC_045512.2_21563-25384 | 3822 | 37.31% | 0 | Yes |
| hCoV-19_Scotland_CVR5974_2020 | 3819 | 37.29% | 3 | Yes |
| hCoV-19_Scotland_CVR6031_2020 | 3819 | 37.29% | 1 | Yes |
| hCoV-19_Scotland_CVR6032_2020 | 3819 | 37.29% | 1 | Yes |
| hCoV-19_England_204590575_2020 | 3822 | 36.39% | 98 | Yes |

**Figure 5.** Sequence information
**图 5.** 序列信息

从图中信息可以看出自英格兰的病毒序列包含较多 N 碱基，取自苏格兰的三条病毒序列相似度较高，变异毒株的 S 蛋白与初期序列均有差异。

### 3.3. 可视化分析

如图 6 为各碱基组合($N$ = A、C、G、T、AC、AG、AT、CG、CT、GT、ACG、ACT、AGT、CGT、ACGT)对应的信息熵全映射。从图中我们清晰直观地可以看出，五条序列的碱基信息熵区间均在 3 到 4 之间，成聚集的情况，而且均存在差别，这表明了突变的发生。

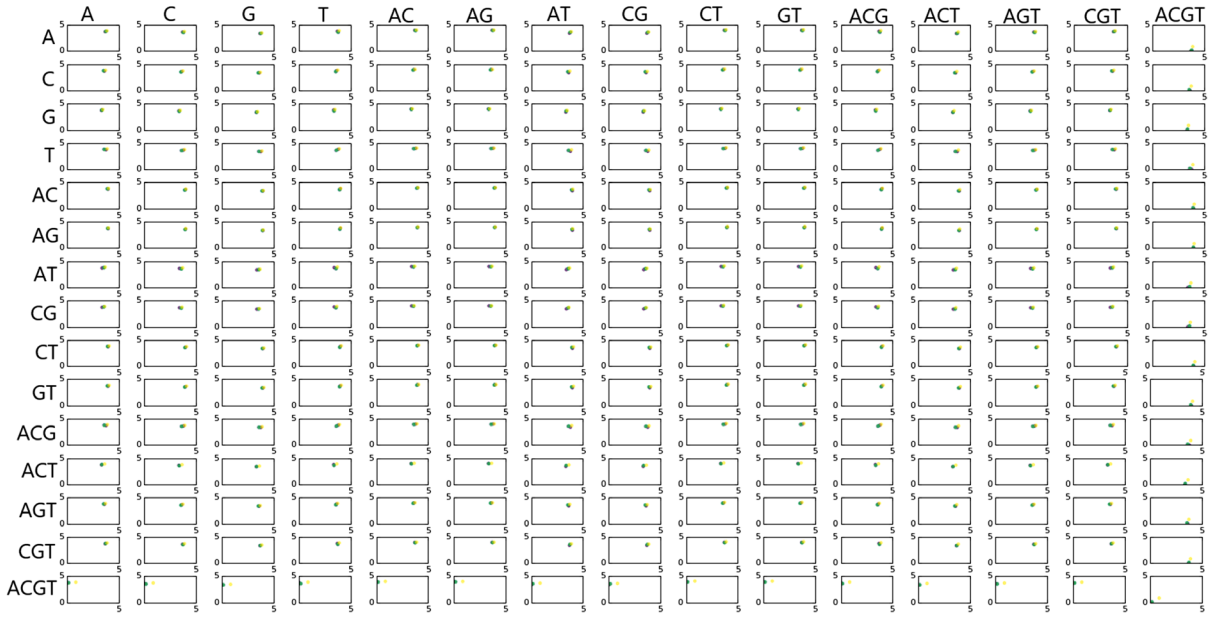


**Figure 6.** Total mapping of base pairs
**图 6.** 碱基组合的全映射

如图 7 所示，可以选取任意目标碱基组合放大，对图像映射进行更有效的处理分析。

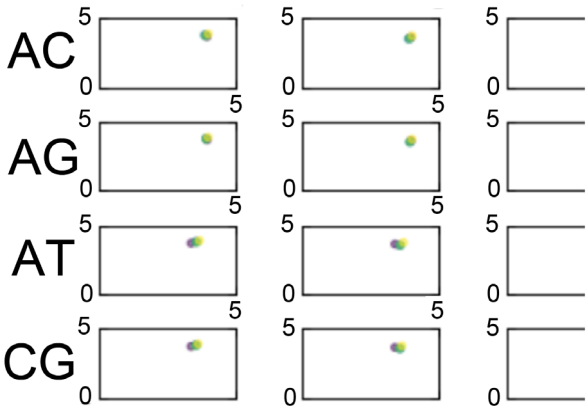


**Figure 7.** Zooming in on target pairs mapping
**图 7.** 放大目标组合映射

此处我们选取了 A、C、AT、CG 组合进行细致的规律可视化分析，从图 8~11 中看出，来自苏格兰的三条序列 CVR5974、CVR6031、CVR6032 的信息熵映射呈集聚之势，可以推测三者之间的亲属关系，且相对于另外的两条序列的映射，在表现上具有不稳定性。而参考序列 NC_045512.2 与英格兰的序列 204590575 在趋势上呈现不变性，但明显可以区别参考序列、苏格兰三条序列与英格兰一条序列之间的

关系。突变的一部分特性在图像上得以体现。

S 蛋白个别位点的变化就有可能会造成整个病毒的变异，在病毒功能性中起到关键作用。信息熵分布图的差异性就取决于样本的差异性，相同的 S 蛋白样本只会产生一个特征点。因此通过信息熵分布图像，可以批量导入 S 蛋白段后放大观察，选取有差异的基因序列继续细化观察，通过全排列的图像可以基本发现出现差异的碱基变化，从而获取到一些特征信息，再配合其他工具细化分析。

同时，信息熵可视化作为一种分析方法，可以扩展应用到一切病毒的研究分析中，包括整段的基因序列分析，有着广泛的应用前景。

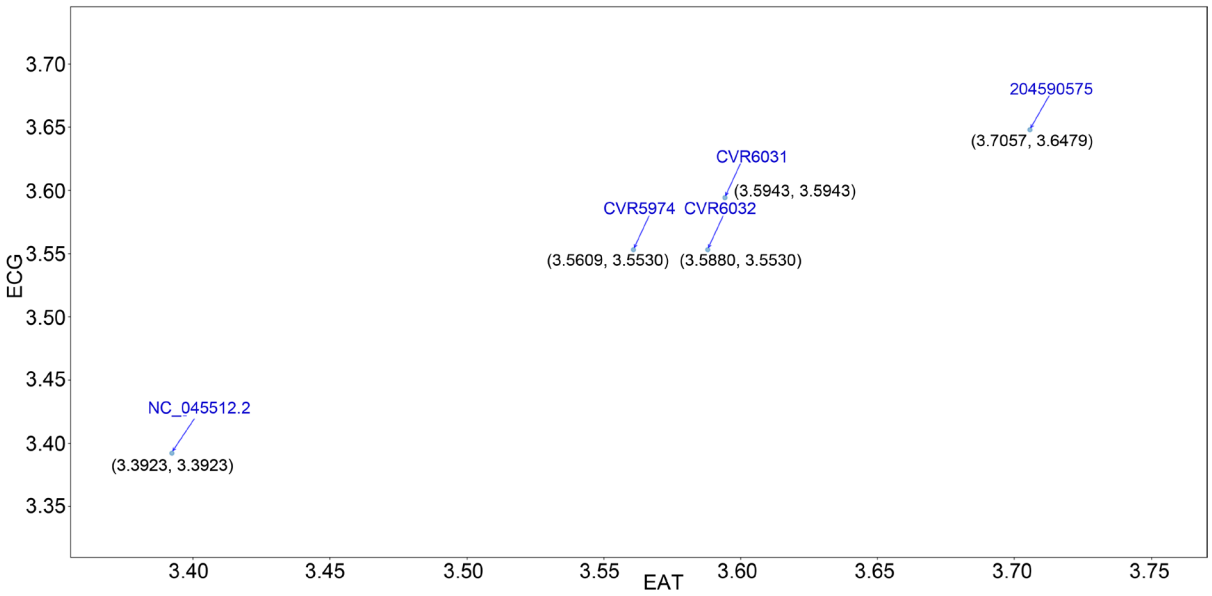


**Figure 8.** EA-EC mapping
**图 8.** EA-EC 映射

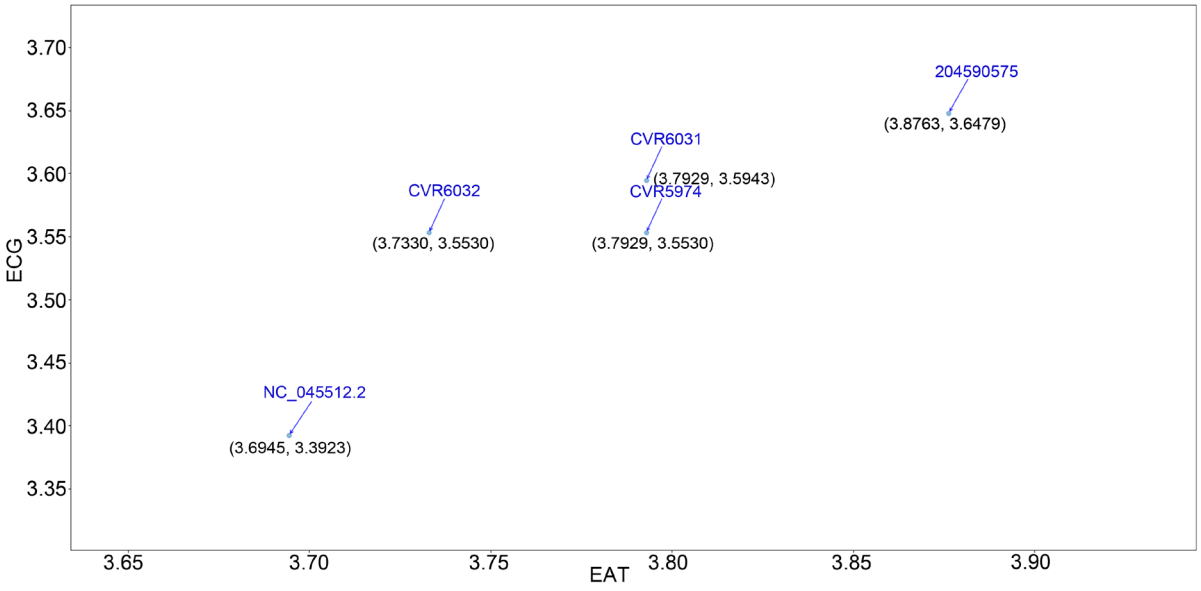


**Figure 9.** EA-ECG mapping
**图 9.** EA-ECG 映射

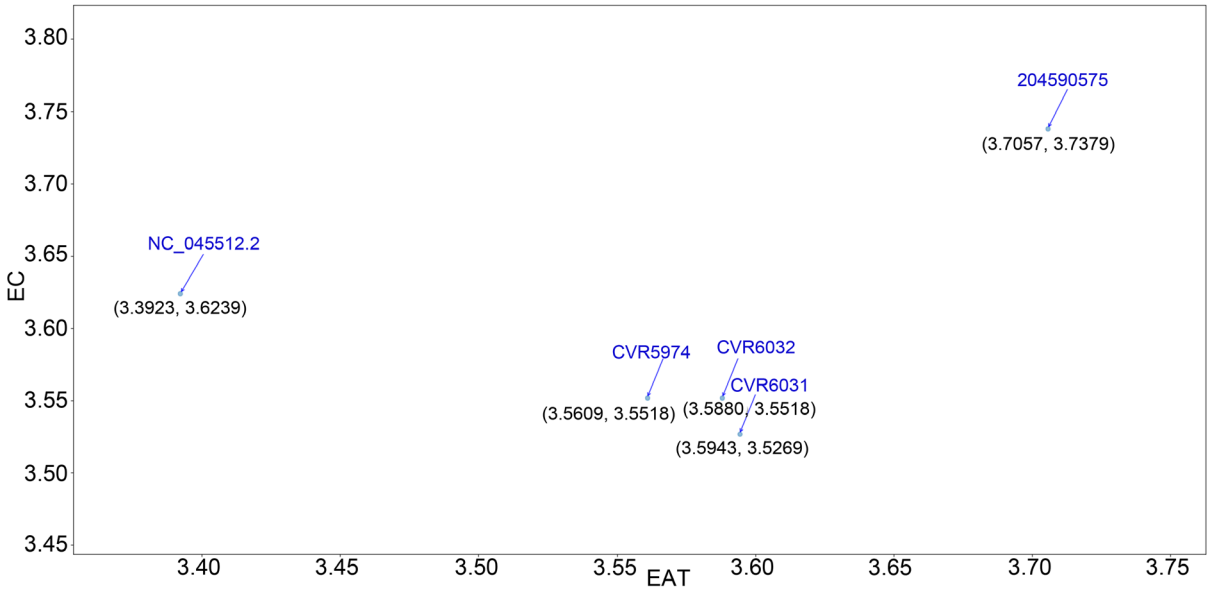


**Figure 10.** EAT-EC mapping
**图 10.** EAT-EC 映射

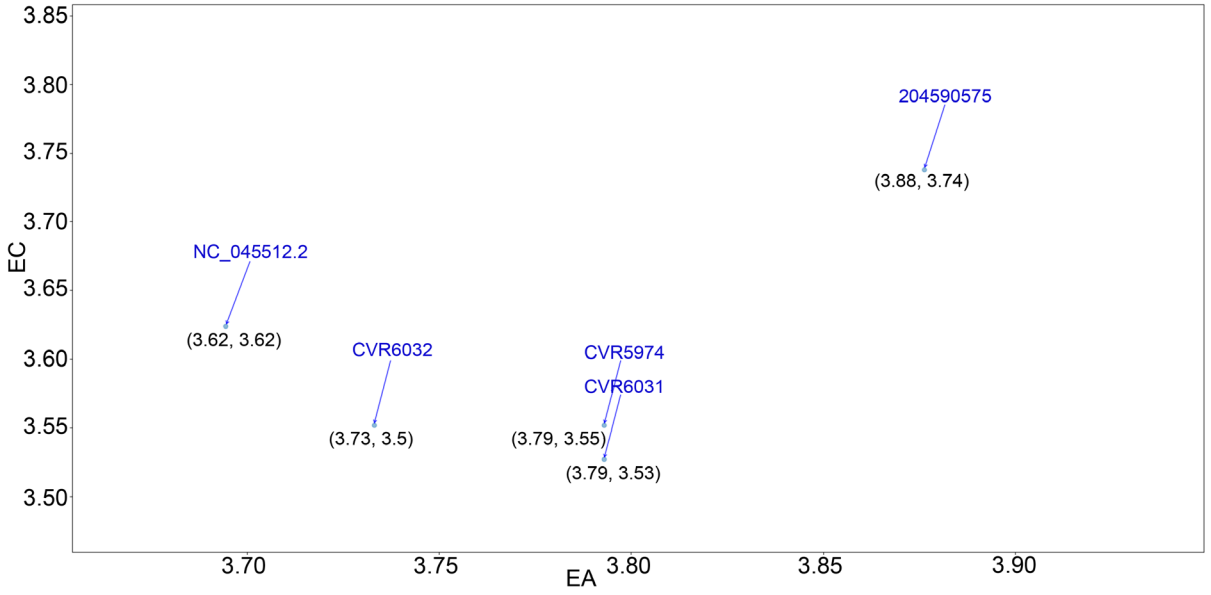


**Figure 11.** EAT-ECG mapping
**图 11.** EAT-ECG 映射

## 4. 总结

本文通过对新型冠状病毒变异毒株和原始病毒进行信息熵计算及投影，在实现过程中可以根据需求调整每组碱基数量，将数据转换为更加直观的彩色散点图，根据需要放大局部并进行分析。相比于传统的生物研究方法，信息熵投影可以提高数据分析效率，为生物信息、计算生物等方面研究提供了新思路和研究基础。

## 致　谢

感谢郑智捷教授的悉心指导，感谢云南大学软件学院对本项目的支持。

## 基金项目

国家自然科学基金项目 62041213。

## 参考文献

[1] 许湘, 李鹏, 魏香. 基于新型冠状病毒S蛋白结构及入胞机制的抗体与药物研发[J]. 中国生物化学与分子生物学报, 2021, 37(1): 1-10.

[2] 徐进, 冯宝龙, 王清艳, 梁瑾, 王靖飞. 氨基酸序列集熵值计算工具实现及应用[J]. 生命科学, 2008(3): 415-420.

[3] Ke, Z.L., Oton, J., Qu, K., *et al.* (2020) Structures and Distributions of SARS-CoV-2 Spike Proteins on Intact Virions. *Nature*, **588**, 498-502. https://doi.org/10.1038/s41586-020-2665-2

[4] 曲恒熠, 袁鑫, 马林威, 郑智捷. 烟草 DNA 序列分组测量可视化分析[J]. 计算生物学, 2019, 9(2): 14-21.

[5] Qiao, M., Liu, R.Y., Wang, Z.H., Li, X.M. and Zheng, J. (2021) Visualizations of Topologic Entropy on SARS-CoV-2 Genomes in Multiple Regions. *EC Neurology*, **S1**, 86-93. https://doi.org/10.21203/rs.3.rs-65305/v2

[6] Zhou, Y. and Zheng, J. (2021) Visualizations of Combinatorial Entropy Index on Whole SARS-CoV-2 Genomes. *EC Neurology*, **S1**, 101-109.

[7] Wu, R.X., Qiao, M. and Zheng, J. (2021) 2D Visual Analysis of SARS-CoV-2. *EC Neurology*, **S1**, 110-116. https://doi.org/10.21203/rs.3.rs-68275/v2