

新型冠状病毒中SARS-CoV-2/Human/CHN/Beijing_BJ0613-12-HJ/2020毒株的分析

胡义锋^{1*}, 夏鹏亮^{2#}, 马东方¹

¹长江大学农学院, 湿地生态与农业利用教育部工程研究中心/主要粮食作物产业化湖北省协同创新中心, 湖北 荆州

²湖北省烟草公司恩施市分公司, 湖北 恩施

收稿日期: 2021年9月8日; 录用日期: 2021年9月23日; 发布日期: 2021年9月29日

摘要

生物信息学分析的方法被广泛用于多个物种的多个基因家族的分析。随着新冠病毒(SARS-CoV-2)的蔓延, 导致感染肺炎的人数也越在与日俱增, 对于已知的SARS-CoV-2序列的鉴定相继被报道。对于具体的SARS-CoV-2序列的分析可能有助于帮助分析其他新增SARS-CoV-2序列, 为新增的新型冠状病毒的传播途径、分子诊断和临床治疗提供基础的信息。本研究选取北京新增的病毒株名为SARS-CoV-2/human/CHN/Beijing_BJ0613-12-HJ/2020的SARS-CoV-2, 采用生物信息学方法对系统发育、蛋白质特征进行分析, 并预测了它们的蛋白模型。以期望为抗病毒药物的研发提供有效的依据或提供新的思路。

关键词

SARS-CoV-2, 生物信息学, 系统发育分析, 毒株

Analysis of SARS-CoV-2/Human/CHN/Beijing_BJ0613-12-HJ/2020 Strain among COVID-19

Yifeng Hu^{1*}, Pengliang Xia^{2#}, Dongfang Ma¹

¹Engineering Research Center of Ecology and Agricultural Use of Wetland/Hubei Collaborative Innovation Center for Main Grain Industry, Ministry of Education, College of Agriculture, Yangtze University, Jingzhou Hubei

²Enshi Tobacco Company of Hubei Province, Enshi Hubei

Received: Sep. 8th, 2021; accepted: Sep. 23rd, 2021; published: Sep. 29th, 2021

*第一作者。

#此作者和第一作者同等贡献。

文章引用: 胡义锋, 夏鹏亮, 马东方. 新型冠状病毒中 SARS-CoV-2/Human/CHN/Beijing_BJ0613-12-HJ/2020 毒株的分析[J]. 生物医学, 2021, 11(4): 220-226. DOI: 10.12677/HJBM.2021.114028

Abstract

Bioinformatics analysis methods were used widely in the analysis of multiple gene families in different species. The number of people infected with pneumonia is increasing day by day because of the spread of the COVID-19 (SARS-CoV-2). Fortunately, the analysis of those reported and identified known SARS-CoV-2 sequences may help analyze other newly similar, and provide underlying information for the transmission route, molecular diagnosis, and clinical treatment of the novel coronavirus SARS. In this study, the newly-added virus strain (SARS-CoV-2/human/CHN/Beijing_BJ0613-12-HJ/2020, SARS-CoV-2 in Beijing) was selected and used the bioinformatics method to analyze the phylogeny and protein characteristics. Those protein models were analyzed and predicted for the expectation to provide a basis or new ideas for the development of antiviral drugs.

Keywords

SARS-CoV-2, Bioinformatics, Phylogenetic Analysis, Strain

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

2019 冠状病毒病(Corona Virus Disease 2019, COVID-19)与严重急性呼吸综合征(severe acute respiratory syndrome, SARS)和中东呼吸综合征(Middle East respiratory syndrome, MERS)同属冠状病毒属所致疾病, 其临床表现从轻微的流感样症状到严重危及生命的重症肺炎[1] [2] [3]。2019 年 12 月有新型冠状病毒(SARS-CoV-2)感染的新型冠状病毒肺炎(COVID-19)对全球公共卫生造成了巨大的威胁[4] [5]。2020 年 1 月 12 日, 世界卫生组织正式将其命名为 2019-nCoV [6]。新型冠状病毒的传染性强, 有症状和无症状感染者均可传播, 现在研发有效的治疗方法迫在眉睫。从 2019 年 12 月新型冠状病毒(SARS-CoV-2)大面积爆发后, 我国投入了大量的财力物力人力, 给我国经济造成了巨大损失, 同时也严重危害了人类的健康。与此同时, 疫情迅速蔓延, 造成人员恐慌、严重的社会危害, 引起了中国政府以及世界有关组织的密切关注[7]。2019-nCoV 的出现提醒了我们建立系统的冠状病毒监测网络的重要性。也给突发疫情防控和应急科研提出了新的挑战。

新型冠状病毒属于 β 属冠状病毒, 其遗传物质是单条正义 RNA 链。目前, SARS-CoV-2 基因组序列已经绘制完成(NCBI BioProject: PRJNA485481) [8]。2019 新型冠状病毒信息库 (<https://ngdc.cnpc.ac.cn/ncov/>)也在不断更新数据。有报道基于生物信息学分析方法已经将部分病毒基因组数据构建了系统发育分析树, 但是关于具体的 SARS-CoV-2 的序列分析还未被报道。我们的研究从生物信息学角度出发对 SARS-CoV-2 的部分序列进行分析, 以期可以为研究 SARS-CoV-2 的分子诊断, 筛选抗病毒药物靶点提供有效依据, 并且期望可以为抗 SARS-CoV-2 的疫苗或药物提供新的思路。

2. 材料与方法

2.1. 序列的搜索与鉴定

本文的研究对象是病毒株名为 SARS-CoV-2/human/CHN/Beijing_BJ0613-12-HJ/2020 的 SARS-CoV-2。从新型冠状病毒信息库(<https://ngdc.cnpc.ac.cn/ncov/>)下载了 SARS-CoV-2 的 12 个蛋白序列, 并同时下载

了蝙蝠冠状病毒和 SARS 的参考序列, 表 1 列出了 3 种病毒的名称和对应的序列号。通过上述下载的数据, 利用软件 Mega7.0 的最大自然法推断系统发育关系, 构建系统发育树, 并通过 ITOL (<http://itol.embl.de>) 美化系统发育树。

2.2. 蛋白质特征分析

使用蛋白质分析工具 ExPASy Server10 (SIB Bioinformatics Resource Portal, <https://prosite.expasy.org/PS50011>) 预测 SARS-CoV-2 蛋白的特征, 包括蛋白质长度, 分子量(MW), 等电点(pI), 稳定性和亲水性的平均值(GRAVY) [9]。

2.3. 蛋白模型的预测

使用 SWISS-MODEL server (<https://www.swissmodel.expasy.org/>) 对 12 个 SARS-CoV-2 蛋白序列进行三维建模[10]。在模板的基础上, 预测的模型试探性地最大化了测试序列的比对范围、百分比识别率和可信度。

3. 研究结果

3.1. 系统发育树

系统发育树分析表明 3 种病毒的蛋白序列可以分为 a, b, c 三组(见表 1 和图 1)。a 组中有 6 个 SARS 序列, 2 个蝙蝠冠状病毒序列, 2 个 SARS-CoV-2 序列。b 组中有 17 个 SARS 序列, 1 个蝙蝠冠状病毒序列, 3 个 SARS-CoV-2 序列。c 组中有 15 个 SARS 序列, 5 个蝙蝠冠状病毒序列, 7 个 SARS-CoV-2 序列。结果显示, 和蝙蝠冠状病毒比起来, 本文中我们研究的 SARS-CoV-2 与 SARS 的亲缘关系可能更接近。

Table 1. Virus name and corresponding serial number

表 1. 病毒名称及对应的序列号

病毒名称	序列号	分组	序列号	分组	病毒名称	序列号	分组
SARS	YP_009725309.1	a	YP_009742608.1	b	SARS-CoV-2	GWHPANPA000001	a
	YP_009725310.1	a	YP_009725305.1	b		GWHPANPA000003	a
	YP_009724389.1	a	YP_009725306.1	b		GWHPANPA000002	b
	YP_009724390.1	a	YP_009742610.1	b		GWHPANPA000005	b
	YP_009725302.1	a	YP_009742611.1	b		GWHPANPA000008	b
	YP_009742613.1	a	YP_009742612.1	b		GWHPANPA000004	c
	YP_009725307.1	c	YP_009742615.1	b		GWHPANPA000006	c
	YP_009725308.1	c	YP_009742616.1	b		GWHPANPA000007	c
	YP_009742609.1	c	YP_009742617.1	b		GWHPANPA000009	c
	YP_009742614.1	c	YP_009725295.1	b		GWHPANPA000010	c
	YP_009725311.1	c	YP_009725297.1	b		GWHPANPA000011	c
	YP_009725312.1	c	YP_009724392.1	b		GWHPANPA000012	c
	YP_009725255.1	c	YP_009725299.1	b	YP_009824990.1	a	
	YP_009725318.1	c	YP_009725300.1	b	YP_009824994.1	a	
	YP_009724391.1	c	YP_009725301.1	b	YP_009824988.1	b	
	YP_009725298.1	c	YP_009724395.1	b	YP_009824989.1	c	
	YP_009724393.1	c	YP_009725304.1	b	YP_009824991.1	c	
	YP_009724394.1	c			YP_009824992.1	c	
	YP_009724396.1	c			YP_009824993.1	c	
	YP_009724397.2	c			YP_009824995.1	c	
YP_009725303.1	c						

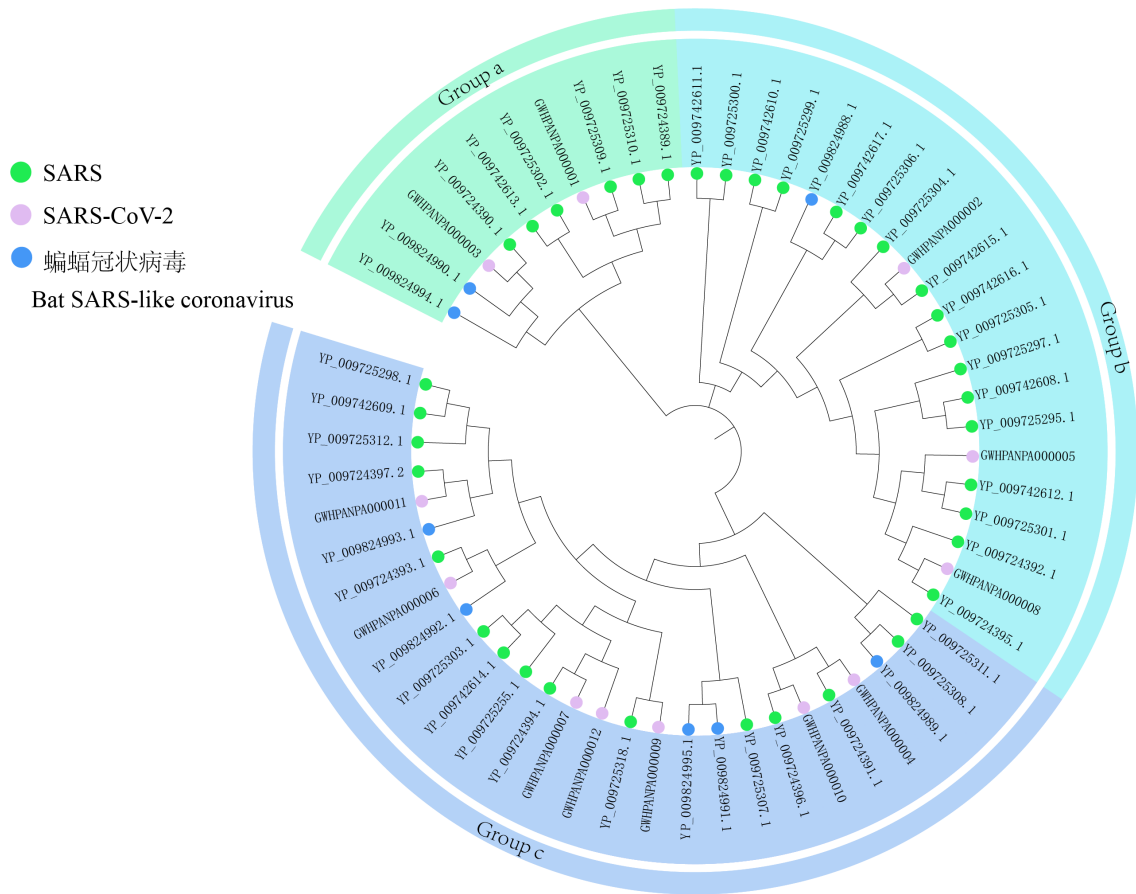


Figure 1. Phylogenetic tree of SARS, SARS-CoV-2, Bat SARS-like coronavirus
图 1. SARS、SARS-CoV-2、蝙蝠冠状病毒的系统发育树

3.2. 蛋白质特征分析

预测结果显示, SARS-CoV-2 蛋白的氨基酸数量从 38~7096 aa 不等, 平均值是 1179.08 aa。分子量在 4449.23~794,063.8 kDa 之间, 平均值为 131,664.31 kDa。等电点从 4.17 到 10.09 变化, 平均值是 6.90, 其中有 7 个 SARS-CoV-2 的等电点小于 7, 被预测为酸性蛋白质, 剩余的被预测为碱性蛋白质, 其中序列号 GWHPANPA000011 的等电点是 10.09, 是所有蛋白质序列中等电点最大的蛋白质, 根据等电点数值的变化, 我们做了折线图, 可以明确的看出等电点的变化(见图 2)。蛋白质不稳定指数分布范围在 16.06~55.81 之间, 平均值为 38.35, 表明大多数 SARS-CoV-2 是稳定蛋白, 但是 GWHPANPA000008, GWHPANPA000009, GWHPANPA000010, GWHPANPA000011 被预测是不稳定的。SARS-CoV-2 蛋白质的亲水性(GRAVY)在 0.98~1.449 之间变化, 平均值为 0.30, 表明它们大部分是亲水性蛋白质, 但是 GWHPANPA000005, GWHPANPA000009 被预测不是亲水性蛋白质(见表 2)。

Table 2. Protein characteristics table
表 2. 蛋白质特征表

序列号	氨基酸数量 (aa)	分子量	pI	不稳定指数	脂肪指数	亲水性平均值 (GRAVY)
GWHPANPA000001	7096	794,063.8	6.32	33.27	稳定	-0.069
GWHPANPA000002	4405	489,978.88	6.04	34.9	稳定	-0.023

Continued

GWHPANPA000003	1273	141,120.43	6.32	32.86	稳定	84.67	-0.077
GWHPANPA000004	275	31,122.94	5.55	32.96	稳定	103.42	0.275
GWHPANPA000005	75	8365.04	8.57	38.68	稳定	144	1.128
GWHPANPA000006	222	25,146.62	9.51	39.14	稳定	120.86	0.446
GWHPANPA000007	61	7272.54	4.6	31.16	稳定	130.98	0.233
GWHPANPA000008	121	13,744.17	8.23	48.66	不稳定	100.74	0.318
GWHPANPA000009	43	5180.27	4.17	50.96	不稳定	156.51	1.449
GWHPANPA000010	121	13,831.01	5.42	45.79	不稳定	97.36	0.219
GWHPANPA000011	419	45,696.82	10.09	55.81	不稳定	52.53	-0.98
GWHPANPA000012	38	4449.23	7.93	16.06	稳定	107.63	0.637
平均值	1179.08	131,664.31	6.9	38.35	稳定	106.22	0.297

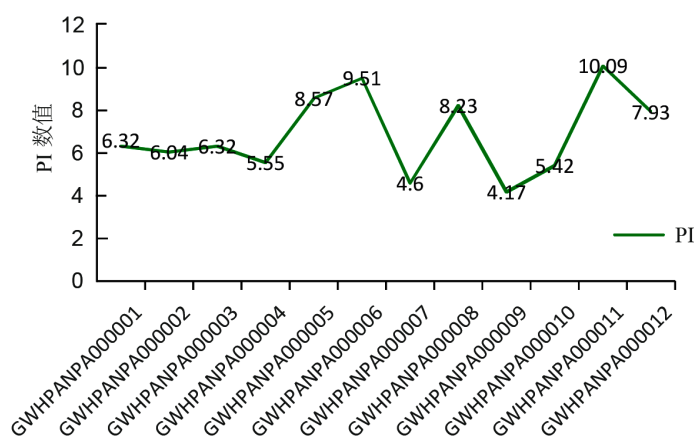


Figure 2. Line diagram of predicted isoelectric points of 12 SARS-CoV-2 protein sequences

图 2. 预测的 SARS-CoV-2 的 12 个蛋白序列的等电点的折线图

3.3. 蛋白模型的预测

使用 SWISS-MODEL server (<https://www.swissmodel.expasy.org/>)对 12 个 SARS-CoV-2 蛋白序列进行三维建模[10]。在模板的基础上,预测的模型试探性地最大化了测试序列的比对范围、百分比识别率和可信度。3D 模型结果显示,预测的 SARS-CoV-2 蛋白质的三级结构各不相同,其中 GWHPANPA000006, GWHPANPA000008, GWHPANPA000010 不含 α 螺旋,剩余 9 个序列都含有 α 螺旋,它在蛋白质结构中起着重要作用(见图 3)。对 GWHPANPA000006 的模型预测描述说明是 SARS-CoV-2 ORF3a 的 Cryo-EM 结构, GWHPANPA000008 的模型描述说明是 SARS 冠状病毒 ORF7a 编码的 X4 蛋白的溶液结构, GWHPANPA000010 的模型描述说明是 SARS 冠状病毒 ORF7a 附件蛋白的结构。

4. 讨论

对于 SARS-CoV-2 的传染性强,传染范围广,各个国家都在加紧研发有效的药物来控制新冠肺炎的蔓延。目前,已有大量的基于全基因组测序和临床防治的 SARS-CoV-2 研究[1] [5] [11],但是仍然缺乏强有效的疫苗和抗病毒药物。随着 SARS-CoV-2 的传染性增强,感染人数的增多,病毒也在不断变异,因此从生物进化学、遗传学、基因组学的角度深入认识和了解 SARS-CoV-2 是必不可少的。

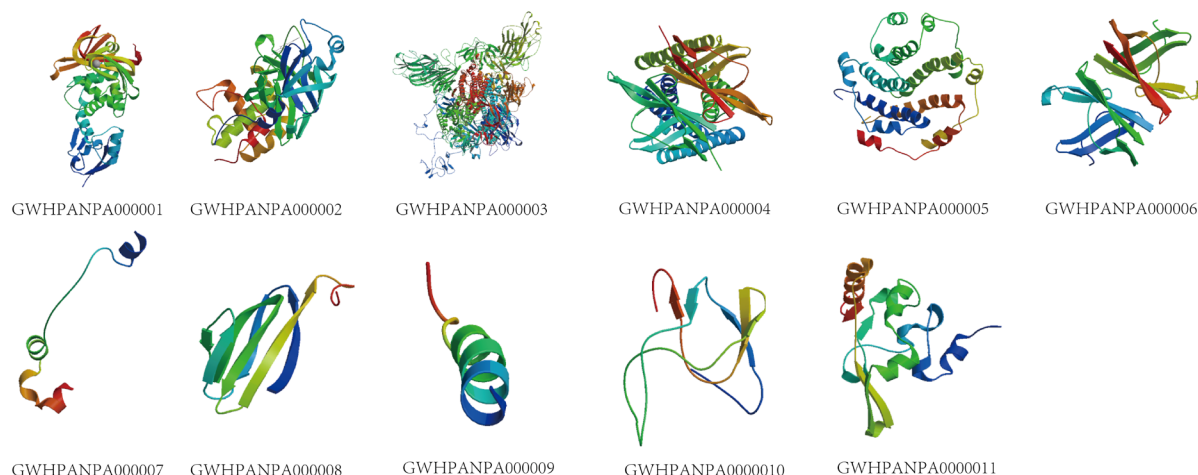


Figure 3. The predicted tertiary structure of the protein

图 3. 预测的蛋白质三级结构

在我们的研究中通过多序列比对，蛋白质特征的分析 and 预测蛋白质特征等生物学方法分析了 SARS-CoV-2/human/CHN/Beijing_BJ0613-12-HJ/2020 的序列，结果显示和蝙蝠冠状病毒比起来，本文中我们研究的 SARS-CoV-2 与 SARS 的亲缘关系可能更接近。蛋白质的氨基酸数量、分子量、等电点等特征变化很大。预测蛋白质模型结果显示有一部分含有 α 螺旋，另外几个的模型结构简单，同时预测这几个模型时对它们的模型描述说明都与它们的 ORF 有关。这种简单和复杂的蛋白结构之间可能存在功能上的联系或者蛋白的相互作用。有趣的是我们并未预测到 GWHPANPA000012 的模型，可能是因为它的结构过于简单或者复杂。从图中我们可以看到，预测的蛋白模型有复杂结构(例如：GWHPANPA000003)也有简单结构(例如：GWHPANPA000007)。说明这些蛋白质之间可能存在功能联系或者相互作用。

依据 SARS-CoV-2 复制的过程，目前可检测病毒的技术或者方法有定量 PCR、多重 PCR、反转录等，主要针对病毒正链单股 RNA、抗体等[12] [13]。本文中我们的分析期望可以了解 SARS-CoV-2 的根源有所贡献，并且可以为精准快速地进行核酸分子的诊断、有效的预防、和临床的防治提供新的思路。同时期望为抗 SARS-CoV-2 药物的研发提供有效的依据，也可以为其它流行性病毒诊断防控提供新的思路和方法。

致 谢

感谢实验室的师兄师姐教会我们做生物信息学分析，感谢马东方老师对本文的指导。

参考文献

- [1] Wu, F., Zhao, S., Yu, B., Chen, Y.-M., Wang, W., Song, Z.-G., *et al.* (2020) A New Coronavirus Associated with Human Respiratory Disease in China. *Nature*, **579**, 265-269. <https://doi.org/10.1038/s41586-020-2008-3>
- [2] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., *et al.* (2020) A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, **382**, 727-733. <https://doi.org/10.1056/NEJMoa2001017>
- [3] Rampal, L. and Liew, B.S. (2020) Coronavirus Disease (COVID-19) Pandemic. *The Medical Journal of Malaysia*, **75**, 95-97.
- [4] Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., *et al.* (2020) Early Transmission Dynamics in Wuhan, China, of Novel Coronavirus-Infected Pneumonia. *New England Journal of Medicine*, **382**, 1199-1207. <https://doi.org/10.1056/NEJMoa2001316>
- [5] Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., *et al.* (2020) Clinical Features of Patients Infected with 2019 Novel Corona-Virus in Wuhan, China. *Lancet*, **395**, 497-506. [https://doi.org/10.1016/S0140-6736\(20\)30183-5](https://doi.org/10.1016/S0140-6736(20)30183-5)

- [6] 中国发展网. 武汉市卫健委通报: 1月13日无新增新型冠状病毒感染的肺炎病例[EB/OL]. <http://www.chinadevelopment.com.cn/>, 2021-01-13.
- [7] 康国锋, 李铜, 陈郁, 刘鑫源, 罗勇军. 2019新型冠状病毒(2019-nCoV)致病机制及防控措施分析[J]. 解放军预防医学杂志, 2020, 38(1): 107-110.
- [8] 熊子军, 张喆, 王雅琦, 朱德芹, 郭佑民. 新型冠状病毒的基因组变异与分子诊断[J]. 西安交通大学学报(医学版), 2020, 41(4): 473-478.
- [9] Li, R., An, J., You, C., Shu, J., Wang, X.-F. and Hao, Y.-J. (2018) Identification and Expression of the CEP Gene Family in Apple (*Malus domestica*). *Journal of Integrative Agriculture*, **17**, 348-358. [https://doi.org/10.1016/S2095-3119\(17\)61653-8](https://doi.org/10.1016/S2095-3119(17)61653-8)
- [10] Lovell, S.C., Davis, I.W., Rd, A.W., de Bakker, P.I.W., Word, J.M., Prisant, M.G., *et al.* (2003) Structure Validation by Calpha Geometry: Phi, Psi and Cbeta Deviation. *Proteins*, **50**, 437-450. <https://doi.org/10.1002/prot.10286>
- [11] Zhou, P., Yang, X., Wang, X., Hu, B., Zhang, L., Zhang, W., *et al.* (2020) A Pneumonia Outbrake Associated with a New Coronavirus of Probable Bat Origin. *Nature*, **579**, 270-273. <https://doi.org/10.1038/s41586-020-2012-7>
- [12] Corman Victor, M., Landt, O., Kaiser, M., Molenkamp, R., Meijer, A., Chu, D.K.W., *et al.* (2020) Detection of 2019 Novel Coronavirus (2019-nCoV). By Real-Time RT-PCR. *Eurosurveillance*, **25**, Article ID: 2000045. <https://doi.org/10.2807/1560-7917.ES.2020.25.3.2000045>
- [13] Wang, M., Cao, R., Zhang, L., Yang, X., Liu, J., Xu, M., *et al.* (2020) Remdesivir and Chloroquine Effectively Inhibit There Cently Emerged Novel Coronavirus (2019-nCoV) *in Vitro*. *Cell Research*, **30**, 269-271. <https://doi.org/10.1038/s41422-020-0282-0>