基于文本挖掘筛选COVID-19症状相关核心基因及分析潜在药物

李俊玲[1,3],何太平[2*],王晓辉[3*],杨峥嵘[3] (1.广东医科大学公共卫生学院,广东东莞 523808; 2.广东医科大学公共卫生学院,广东湛江 524023; 3.深圳市疾病预防控制中心,广东深圳 508055)

摘 要: 目的 筛选新型冠状病毒肺炎(COVID-19)症状相关的核心基因及信号通路,进而分析潜在的靶向药物。**方法** 从文本挖掘数据库中获取与COVID-19主要症状相关的7个基因列表,取交集得到新的基因集。R软件进行基因功能注释和通路富集分析,Cytoscape软件可视化STRING数据库中的蛋白互作网络信息,并按分值筛选出核心基因,再通过DGIDB数据库分析基因与药物之间的作用关系,确定潜在的靶向药物。**结果** 通过文本挖掘到97个与COVID-19症状相关的基因,其涉及的生物学过程主要为:Stat5蛋白酪氨酸磷酸化的正调控以及激活T细胞增殖的正向调控等;富集到的通路分别是细胞因子及其受体的相互作用、JAK-STAT信号通路和疟疾相关免疫反应等。最终通过基因-药物间的相互作用分析得到了9个密切相关的核心基因(即KIT、ACE、ESR1、TNF、VEGFA、IL1B、IL6、IL5、TGFB1),进而筛选到22个潜在药物。**结论** 该研究通过大数据挖掘以及生物信息学分析方法对相关的基因进行功能注释以及通路分析,并通过基因药物之间的相互作用,筛选出一些与COVID-19症状相关的潜在药物,为COVID-19的对症治疗提供科学依据。

关键词: COVID-19; 文本挖掘; 核心基因; 潜在药物
中图分类号: R181.8 **文献标志码:** A **文章编号:** 2096-3610(2021)03-0259-05

Screening of core genes related to COVID-19 symptoms and analysis on potential drugs based on text mining

LI Jun-ling[1,3], HE Tai-ping[2*], WANG Xiao-hui[3*], YANG Zheng-rong[3] (1.School of Public Health, Guangdong Medical University, Dongguan 523808, China; 2.School of Public Health, Guangdong Medical University, Zhanjiang 524023, China; 3.Shenzhen Center for Disease Control and Prevention, Shenzhen 508055, China)

Abstract: Objective To screen out core genes and signaling pathway related to COVID-19 symptoms and further analyze potential targeted drugs. Methods Seven gene lists related to main COVID-19 symptoms were obtained from the text mining database and a new gene set is obtained through intersection. Gene function annotation and pathway enrichment analysis was conducted with R software. Cytoscape was used to visualize the information of protein interaction network in the STRING database, and the core genes were screened out according to the value. Then, the drug-gene interaction was analyzed with the DGIDB database to determine the potential targeted drugs. Results Through text mining, 97 genes related to COVID-19 symptoms were identified, and the main biological process involved was positive regulation of tyrosine phosphorylation of Stat5 protein and positive regulation of activation of T cell proliferation. The enriched pathways were the interaction of cytokines and their receptors, JAK-STAT signaling pathway and Malaria-associated immune response. Through the drug-gene interaction analysis, 9 core genes closely related (KIT, ACE, ESR1, TNF, VEGFA, IL1B, IL6, IL5 and TGFB1) were obtained and then 24 related drugs were screened out. Conclusion In this study, functional annotation and pathway analysis is performed for related genes with big data mining and bioinformatics analysis and some potential drugs related to COVID-19 symptoms were screened out, providing scientific evidence for symptomatic treatment of COVID-19.

Key words: COVID-19; text-mining; core genes; potential drugs

基金项目: 深圳市科技创新委基础研究(No. JCYJ20180508152244835),深圳市医学重点学科建设基金(No. SZXK064)
收稿日期: 2020-10-02; **修订日期:** 2020-12-04

作者简介: 李俊玲(1994–),女,在读硕士研究生
通信作者: 何太平(1970–),男,硕士,教授,E-mail:htp@gdmu.edu.cn
王晓辉(1971–),男,博士,主任技师,E-mail:wangxh@szcdc.net

SARS-CoV-2是于2020年1月通过基因测序确定的一种新型冠状病毒，该病毒引起新型冠状病毒肺炎（COVID-19），简称“新冠肺炎”，并造成了世界范围的广泛流行[1]。SARS-CoV-2的持续传播，对国际公共卫生构成了巨大威胁[2]，全球科学家都在积极探索新冠肺炎治疗和预防的方法，其中药物研究是一个热点。我们通过大数据的挖掘以及生物信息学的分析筛选出新冠肺炎症状相关的核心基因，并鉴定出潜在对症治疗药物，希望能为新冠肺炎药物的研发提供一些帮助。

1 材料和方法

1.1 文本挖掘确定基因列表

查阅文献确定了临床诊断中与COVID-19密切相关的7个主要临床症状，分别是“发烧”“咳嗽”“重症肺炎”“呼吸困难”“ 呼吸窘迫”“乏力”“肌肉酸痛”，随后通过文本挖掘数据库（http://pubmed2ensembl.ls.manchester.ac.uk/）搜索这7个关键词，得到7组与这些关键词相关的基因列表，以供下一步分析。

1.2 基因功能注释和通路富集分析

基因注释功能分析主要包括基因的生物学过程、分子功能和细胞组分这三个方面。通路富集分析能够从分子水平确定基因参与最主要的信号转导途径。这两种分析方法主要是基于超几何分布检验，通过*P*值大小筛选出这些基因主要显著富集在何种生物学过程、分子功能、细胞组分及通路。利用R软件中的clusterProfiler包对基因进行功能注释和通路富集的统计分析和可视化[3]。

1.3 蛋白互作网络分析

用STRING（https://string-db.org/）数据库来确定蛋白之间相互作用的信息，用Cytoscape软件来可视化STRING数据库中导出的蛋白互作数据信息，再利用互作网络中分值由大到小的顺序确定核心基因[4]。

1.4 药物基因相互作用分析

DGIDB数据库是探索药物与基因之间的相互关系的分析平台，将上述分析后产生的核心基因作为靶向基因并使用DGIDB数据库来筛选出潜在治疗药物，并最终用R软件中的ggalluvival包来可视化药物-基因-通路之间的相互关系[5]。

2 结果

2.1 确定基因列表

通过文本挖掘数据库对7个COVID-19主要症状进行检索，得到7组基因集，涉及的基因数量分别为：768、563、456、627、879、1 148、809个，对这7组基因集取交集，得到包含97个基因的基因列表（图1）。

2.2 基因功能注释和富集通路分析

使用R软件对97个基因进行基因功能注释和通路富集分析。基因功能注释显示（如图2，仅显示*P*值最小的前6个条目）这些基因的生物学过程主要集中在Stat5蛋白酪氨酸磷酸化的正调控（*P*=3.86E-11）、活化T细胞增殖的正调控（*P*=7.35E-06）等96个功能簇；这些基因的细胞组分构成主要是细胞外隙（*P*=1.79E-18）、胞外区（*P*=1.00E-12）和质膜外侧（*P*=7.64E-09）等21个功能簇；这些基因的分子功能方面主要集中在细胞因子受体结合（*P*=6.15E-12）、生长因子活性（*P*=6.14E-11）等35个功能簇。

富集通路分析显示这些基因一共富集到47条通路，同样选择6条*P*值最小的通路（如图3），分别是细胞因子与细胞因子受体的相互作用（*P*=5.00E-12）、JAK-STAT信号通路（*P*=6.93E-10）、疟疾相关免疫反应（*P*=1.35E-09）、炎症性肠病（IBD）（*P*=5.93E-09）、T细胞受体信号通路（*P*=6.18E-07）、同种异体移植排斥（*P*=1.34E-06）以及PI3K-Akt信号通路（*P*=2.57E-06）。

2.3 构建互作网络并选取核心基因

利用String数据库对97个基因进行相互作用分析，构建出了一个拥有85个节点，917条边的蛋白互作网络图（如图4）。利用Cytoscape软件筛选出85个

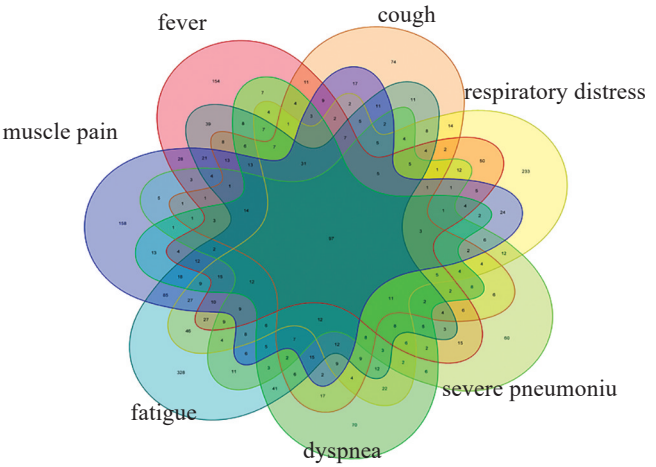

图1 与COVID-19症状密切相关的基因列表

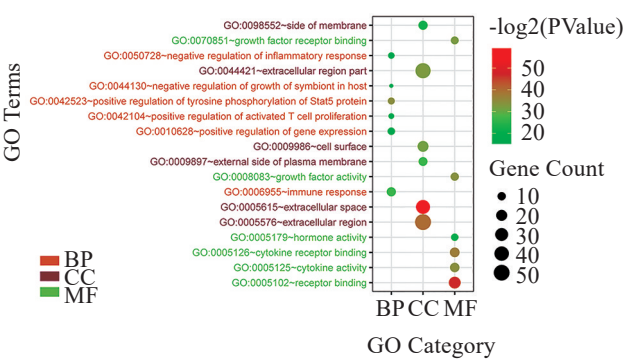

图2 与COVID-19症状相关的基因功能注释气泡图

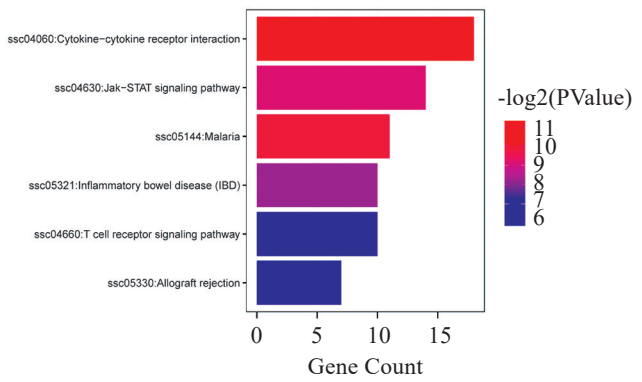

图3　通路富集分析柱状图

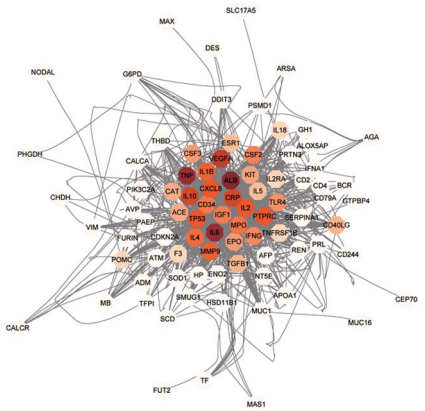

图4　基因相互作用网络图

基因节点中与其他基因之间相互联系最为密切的，分值最高的前30个基因（如图4中颜色较深部分），为下一步数据分析做准备。

2.4　药物基因相互作用分析

我们用上述筛选到的30个基因进一步分析药物与基因之间的相互作用，最终筛选到9个核心基因，涉及到22个潜在药物，可能对COVID具有一定的对症治疗作用（如表1）。

3　讨论

COVID-19作为一种新发的传染病，至今尚无确定的特效药物。对于大多数轻症及无症状感染者，此疾病具有一定的自愈能力或者不需要特别的治疗，但是对于一些危重患者来说，情况却很严峻。有研究表明，即使在积极的治疗情况下，进入ICU的COVID-19重症患者28 d内病死率高达61.5%，其中有47.0%的重症死亡患者并无基础疾病[6]。因此，迫切需要筛选有效的COVID-19药物来降低病死率。文本挖掘等生物信息学分析工具为药物筛选提供了一条快速通道，通过KEGG通路富集分析，按照*P*值由小到大的顺序鉴定出3条与COVID-19症状密切相关的通路；进一步通过蛋白互作分析以及基因与药物间的相互

表1　基于核心基因治疗COVID-19的潜在药物汇总表

编号	药物	基因	药物基因相互作用	评分	批准	临床应用范围
1	伊马替尼	KIT	抑制剂	98	是	慢性髓性白血病和恶性胃肠道间质肿瘤
2	舒尼替尼	KIT	抑制剂	46	是	抗肿瘤，降血糖
3	达沙替尼	KIT	拮抗剂/抑制剂	45	是	抗肿瘤
4	索拉非尼	KIT	拮抗剂/抑制剂	35	是	抗肿瘤
5	卡托普利	ACE	抑制剂	27	是	降压药
6	西拉普利	ACE	抑制剂	11	是	原发性高血压，肾性高血压，心力衰竭
7	赖诺普利	ACE	抑制剂	10	是	高血压，充血性心力衰竭
8	他莫昔芬	ESR1	激动剂	17	是	抗肿瘤，激素，免疫调节剂
9	托瑞米芬	ESR1	调制器	15	是	转移性乳腺癌
10	乙炔雌二醇	ESR1	激动剂	14	是	前列腺癌，避孕，更年期综合征
11	乙烯雌酚	ESR1	激动剂	13	是	激素替代剂，抗肿瘤药物
12	英夫利昔单抗	TNF	抑制剂	17	是	抗肿瘤药物，类风湿关节炎
13	伊那西普	TNF	抑制剂，抗体	12	是	抗风湿
14	阿达木单抗	TNF	抗体，抑制剂	12	是	类风湿关节炎，强直性脊柱炎
15	沙利度胺	TNF	抑制剂	11	是	II型麻风病，盘状红斑狼疮
16	兰尼单抗	VEGFA	抑制剂	14	是	糖尿病黄斑水肿，抗肿瘤
17	贝伐单抗	VEGFA	抑制剂，抗体	10	是	各种转移性癌症
18	阿柏西普	VEGFA	抑制剂，抗体	7	是	转移结肠癌，视网膜血管性疾病
19	康纳单抗	IL1B	抑制剂，抗体	7	是	系统性幼年性特发性关节炎，周期性发热综合征
20	司妥昔单抗	IL6	拮抗剂，抑制剂	4	是	淋巴瘤
21	美泊利单抗	IL5	拮抗剂，抑制剂	5	是	嗜酸性粒细胞性哮喘和变应性肉芽肿性血管炎
22	透明质酸酶	TGFB1	抑制剂	5	是	降低细胞间质的黏性，药物渗透剂

作用，筛选出9个与COVID-19症状密切相关的基因。

KEGG通路富集分析结果表明，与COVID-19症状高度相关的通路与细胞因子及其受体间的相互作用相关。多项临床实验表明，在SARS-CoV2的感染过程中，COVID-19患者的淋巴细胞和NK细胞计数显著降低，细胞因子水平却显著升高[7]，出现“细胞因子风暴”，使宿主免疫反应过度，造成急性肺部损伤、多器官衰竭以及不良的预后等严重的后果[8]。但细胞因子发挥其生物学功能是需要通过与靶细胞表面的相应受体结合才能将信号转导到细胞内部，因此，细胞因子与其受体的相互作用是重要的治疗靶点。细胞因子及其受体互作网络极其复杂，需要分析处于核心的细胞因子。在严重的SARS-CoV-2感染病例中，IL-6水平显著升高，是最常被检测出来并被报道[9-11]，而IL-6受体与IL-6结合进一步促进IL-6的生物学作用，加剧“细胞因子风暴”的进程。我们鉴定到的药物司妥昔单抗可以有效地阻断两者的结合，避免激活信号传导通路[12]，可能是COVID-19严重感染病例的有效治疗手段。

其次是JAK-STAT信号通路，在炎症反应时，细胞因子与其受体相互作用增强，进一步激活JAK，发生JAK的自磷酸化以及STATs的二聚化，随后二聚化后的STATs进入到细胞核中参与细胞的免疫调节等生物学过程，进一步促进“细胞因子风暴”[13]。因此，通过JAK抑制剂治疗由SARS-CoV-2引起的“细胞因子风暴”可能是一种有效策略。经过检索文献发现鲁索替尼作为JAK的抑制剂，相比其他药物耐受性较好并且在老年人群体中也适用，可能对COVID-19患者出现的免疫反应过度症状有比较好的效果[14]。

另一条与COVID-19症状高度相关的通路是疟疾相关免疫反应信号通路，疟原虫感染及其治疗药物均有其特点，最古老的治疗药物为氯喹，后来逐步改进到磷酸氯喹、羟氯喹等衍生药物。在武汉、荆州、广州、上海、北京、重庆、宁波等多家医院进行的试点实验表明磷酸氯喹可以有效地抑制肺炎的恶化，缩短COVID-19的病程[15]。国家卫建委发部的《新型冠状病毒肺炎诊疗方案（试行第八版中）》也指出磷酸氯喹可以继续试用，在临床应用上进一步评价它的疗效。综合分析来看，磷酸氯喹可以有效地调节与COVID-19相关的病理学通路。

在进一步的基因与药物相互作用分析中，我们发现KIT基因所筛选到的靶向药物最多且评分最高。KIT基因是一种III类酪氨酸酶受体，它的表达异常可能会使宿主细胞发生多种肿瘤[16-18]。我们经过一系列的生物信息学分析发现KIT与COVID-19具有很强的相关性。伊马替尼是我们鉴定到评分最高的KIT抑制剂，它在针对严重性呼吸窘迫综合征以及中东呼吸综合征冠状病毒的体外实验中显示出具有抗病毒活性[19]。今年6月份，国外相关临床研究显示，一位38岁的确诊女性病例在经过羟氯喹和利托那韦的双重治疗后病情却再次复发之际，改用伊马替尼进行治疗后情况好转并顺利出院[20]。这提示伊马替尼治疗COVID-19具有继续进行临床研究的价值。

基因筛选评分其次的是ACE基因，它与ACE2基因是肾素-血管紧张素系统（RAS）中的两个不可或缺的调制器，两者之间相互保持平衡对于维持RAS的稳定具有重要的作用，预计可以有效地降低COVID-19的死亡率和发病率。ACE2由于与冠状病毒表面的胞膜蛋白有很好的亲和力使得使其消耗而表达水平下降[21-22]，此时ACE和ACE2之间的表达不平衡促使血管紧张素II的水平不受限制，增加血管的通透性并引起血管收缩，从而导致急性肺损伤并促进纤维化。一项包含1 128名COVID-19伴高血压患者的多中心回顾性研究显示，使用ACE抑制剂的住院患者比不使用的死亡风险降低[23]，我们研究中鉴定到的一些ACE抑制剂，如卡托普利等药物可能在一定程度上具有治疗和预防由COVID-19引起的急性肺损伤的问题。

基因筛选评分第三是ESR1基因，是一种介导雌激素发挥生物学效应的配体依赖转录因子。COVID-19流行病学资料显示，不同性别、年龄群体对于新冠病毒所表现出的炎症反应是不同的。有研究表明，除了个体差异，雌激素可能是造成这一差别的重要原因[24]。雌激素可以调节中性粒细胞、巨噬细胞等免疫细胞的发育，使B细胞介导的适应性免疫产生特异性抗体[25]，抑制NF-kB通路介导的炎症反应，可能降低肺部损伤。这个机理在动物实验中已得到证实[26]，国外已有学者提出雌激素可能会降低COVID-19的死亡率[27-28]，因此对老年女性新冠患者尝试用外源雌激素治疗的方法可能更具有实际意义。

除了上述描述的评分较高的3个基因之外，还有其他一些与新冠症状密切相关的细胞因子及受体，包括TNF、VEGFA、IL-1B、IL5、TGFB1，同时也确定了一些基因对应的靶向药物。综上，我们利用生物信息学的分析方法筛选出一些COVID-19症状相关的核心基因，之后通过基因功能富集分析和通路富集分析对这些基因的生物学功能、涉及的信号通路做了进一步的分析，同时分析出与核心基因相互作用的潜在药物。希望此类研究能为COVID-19的预防和治疗提供一定的方向指引。

参考文献：

[1] 吴家园，赖天文，刘华锋，等.广东省湛江市新型冠状病毒肺炎流行趋势的初步预测[J].广东医科大学学报，2020，38（2）：148-152.

[2] RUBIN E J, BADEN L R, MORRISSEY S. Audio interview: Caring for patients with Covid-19 [J]. N Engl J Med, 2020, 38 (16): e50.

[3] YU G, WANG L G, HAN Y, et al. Cluster profiler: An R package for comparing biological themes among gene clusters [J]. Omics, 2012, 16(5):284-287.

[4] DONCHEVA N T, MORRIS J H, GORODKIN J, et al. Cytoscape String App: Network analysis and visualization of proteomics data [J]. J Proteome Res, 2018, 18(2):623-632.

[5] COTTO K C, WAGNER A H, FENG Y Y, et al. DGIdb 3.0: A redesign and expansion of the drug – gene interaction database [J]. Nucleic Acids Res, 2018, 46(D1):D1068-D1073.

[6] YANG X, YU Y, XU J, et al. Clinical course and outcomes of critically ill patients with SARS-CoV-2 pneumonia in Wuhan, China: A single-centered, retrospective, observational study [J]. Lancet Respir Med, 2020, 8(5):475-481.

[7] ZHANG X, ZHANG Y, QIAO W, et al. Baricitinib, a drug with potential effect to prevent SARS-COV-2 from entering target cells and control cytokine storm induced by COVID-19 [J]. Int Immunopharmacol, 2020, 86:106749.

[8] HUANG C, WANG Y, LI X, et al. Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China [J]. Lancet, 2020, 395(10223):497-506.

[9] CHEN G, WU D, GUO W, et al. Clinical and immunological features of severe and moderate coronavirus disease 2019 [J]. J Clin Investig, 2020, 130(5):2620-2629.

[10] RUAN Q, YANG K, WANG W, et al. Clinical predictors of mortality due to COVID-19 based on an analysis of data of 150 patients from Wuhan, China [J]. Intensive Care Med, 2020, 46(5):846-848.

[11] GAO Y, LI T, HAN M, et al. Diagnostic utility of clinical laboratory data determinations for patients with the severe COVID-19 [J]. J Med Virol, 2020, 92(7):791-796.

[12] CRISAFULLI S, ISGRò V, LA CORTE L, et al. Potential role of anti-interleukin (IL)-6 drugs in the treatment of COVID-19: rationale, clinical evidence and risks [J]. BioDrugs, 2020, 34(4):415-422.

[13] SEIF F, AAZAMI H, KHOSHMIRSAFA M, et al. JAK inhibition as a new treatment strategy for patients with COVID-19 [J]. Int Arch Allergy Immunol, 2020, 181(6):467-475.

[14] MARKHAM A, KEAM S J. Peficitinib: First global approval [J]. Drugs, 2019, 79(8):887-891.

[15] GAO J, TIAN Z, YANG X. Breakthrough: Chloroquine phosphate has shown apparent efficacy in treatment of COVID-19 associated pneumonia in clinical studies [J]. Biosci Trends, 2020, 14(1):72-73.

[16] COMODO-NAVARRO A N, FERNANDES M, BARCELOS D, et al. Intratumor heterogeneity of KIT gene mutations in acral lentiginous melanoma [J]. Am J Dermatopath, 2020, 42(4):265-271.

[17] LEBEDEV T D, VAGAPOVA E R, POPENKO V I, et al. Two receptors, two isoforms, two cancers: Comprehensive analysis of KIT and TrkA expression in neuroblastoma and acute Myeloid leukemia [J]. Front Oncol, 2019, 9:1046.

[18] OHSHIMA K, FUJIYA K, NAGASHIMA T, et al. Driver gene alterations and activated signaling pathways toward malignant progression of gastrointestinal stromal tumors [J]. Cancer Sci, 2019, 110(12):3821-3833.

[19] COLEMAN C M, SISK J M, MINGO R M, et al. Abelson kinase inhibitors are potent inhibitors of severe acute respiratory syndrome coronavirus and middle east respiratory syndrome coronavirus fusion [J]. J Virol, 2016, 90(19): 8924-8933.

[20] MORALES-ORTEGA A, BERNAL-BELLO D, LLARENA-BARROSO C, et al. Imatinib for COVID-19: A case report [J]. Clin Immunol, 2020, 218:108518.

[21] HOFFMANN M, KLEINE-WEBER H, SCHROEDER S, et al. SARS-CoV-2 Cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor [J]. Cell, 2020, 181(2):271-280.

[22] WALLS A C, PARK Y J, TORTORICI M A, et al. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein [J]. Cell, 2020, 181(2):281-292.

[23] ZHANG P, ZHU L, CAI J, et al. Association of inpatient use of angiotensin-converting enzyme inhibitors and angiotensin II receptor blockers with mortality among patients with hypertension hospitalized with COVID-19 [J]. Circ Res, 2020, 126(12):1671-1681.

[24] ZENG F, DAI C, CAI P, et al. A comparison study of SARS-CoV-2 IgG antibody between male and female COVID-19 patients: A possible reason underlying different outcome between sex [J]. J Med Virol, 2020, 92(10):2050-2054.

[25] CUTOLO M, SMITH V, PAOLINO S. Understanding immune effects of oestrogens to explain the reduced morbidity and mortality in female versus male COVID-19 patients. Comparisons with autoimmunity and vaccination [J]. Clin Exp Rheumatol, 2020, 38:383- 386.

[26] CHANNAPPANAVAR R, FETT C, MACK M, et al. Sex-based differences in susceptibility to severe acute respiratory syndrome coronavirus infection [J]. J Immunol, 2017, 198 (10):4046-4053.

[27] LU R, ZHAO X, LI J, et al. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding [J]. Lancet, 2020, 395 (10224):565-574.

[28] SUBA Z. Prevention and therapy of COVID-19 via exogenous estrogen treatment for both male and female patients: Prevention and therapy of COVID-19 [J]. J Pharm Pharm Sci, 2020, 23(1):75-85.