Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Computers in Biology and Medicine 145 (2022) 105455

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Conserved protein targets for developing pan-coronavirus drugs based on sequence and 3D structure similarity analyses

Minfei Ma [a,b,1], Yanqing Yang [a,b,1], Leyun Wu [a,b], Liping Zhou [a,b], Yulong Shi [a,b], Jiaxin Han [a,c], Zhijian Xu [a,b,*], Weiliang Zhu [a,b,**]

[a] CAS Key Laboratory of Receptor Research, Stake Key Laboratory of Drug Research, Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China
[b] School of Pharmacy, University of Chinese Academy of Sciences, Beijing 100049, China
[c] School of Chinese Materia Medica, Nanjing University of Chinese Medicine, Nanjing, 210046, China

ARTICLE INFO

Keywords:
Pharmaceutical informatics
Drug design
Target discovery
Medicinal chemistry
Broad-spectrum drugs

ABSTRACT

There are 7 known human pathogenic coronaviruses, which are HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1, MERS-CoV, SARS-CoV and SARS-CoV-2. While SARS-CoV-2 is currently caused a severe epidemic, experts believe that new pathogenic coronavirus would emerge in the future. Therefore, developing broad-spectrum anti-coronavirus drugs is of great significance. In this study, we performed protein sequence and three-dimensional structure analyses for all the 20 virus-encoded proteins across all the 7 coronaviruses, with the purpose to identify highly conserved proteins and binding sites for developing pan-coronavirus drugs. We found that nsp5, nsp10, nsp12, nsp13, nsp14, and nsp16 are highly conserved both in protein sequences (with average identity percentage higher than 52%, average amino acid conservation scores higher than 5.2) and binding pockets (with average amino acid conservation scores higher than 5.8). We also performed the similarity comparison between these 6 proteins and all the human proteins, and found that all the 6 proteins have similarity less than 25%, indicating that the drugs targeting the 6 proteins should have little interference of human protein function. Accordingly, we suggest that nsp5, nsp10, nsp12, nsp13, nsp14, and nsp16 are potential targets for pan-coronavirus drug development.

1. Introduction

Coronavirus, with spikes resembled on the virus surface like a crown, have attracted a great deal of world's attention due to the severe impact on human health and global economies. Among various coronaviruses, severe acute respiratory syndrome coronavirus (SARS-CoV, year 2003), middle east respiratory syndrome coronavirus (MERS-CoV, year 2012) and severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2, year 2019) are known as highly transmissible pathogens that cause significant human morbidity and mortality [1,2]. A recent research suggests that human exposure to and spillover of SARS-related coronaviruses may be substantially underestimated, and the researchers estimated that around 400,000 people are infected with SARS-related coronaviruses annually in South and Southeast Asia [3]. It is reasonable to deduce that the harm of the coronavirus to human may not stop at the outbreak of COVID-19. In addition, the multiple mutations in SARS-CoV-2 Omicron variant may affect the current therapy [4]. Although some drugs for the treatment of COVID-19 can be found in Therapeutic Target Database (http://db.idrblab.net/ttd/) to be approved or out of the clinical trial, it's still a crucial task to find broad-spectrum treatments against the coronaviruses that have emerged or that may appear in the future [5]. Therefore, identifying potential proteins for pan-coronavirus drugs is of great importance.

Coronaviruses, which mainly encode 20 proteins, are enveloped viruses containing a positive-sense and single-stranded RNA genome [6–8]. The genome organization for a coronavirus is

* Corresponding author. CAS Key Laboratory of Receptor Research, Stake Key Laboratory of Drug Research, Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China.
** Corresponding author. CAS Key Laboratory of Receptor Research, Stake Key Laboratory of Drug Research, Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China.
E-mail addresses: zjxu@simm.ac.cn (Z. Xu), wlzhu@simm.ac.cn (W. Zhu).
[1] These two authors contribute equally to the work.

https://doi.org/10.1016/j.compbiomed.2022.105455
Received 7 February 2022; Received in revised form 20 March 2022; Accepted 23 March 2022
Available online 26 March 2022

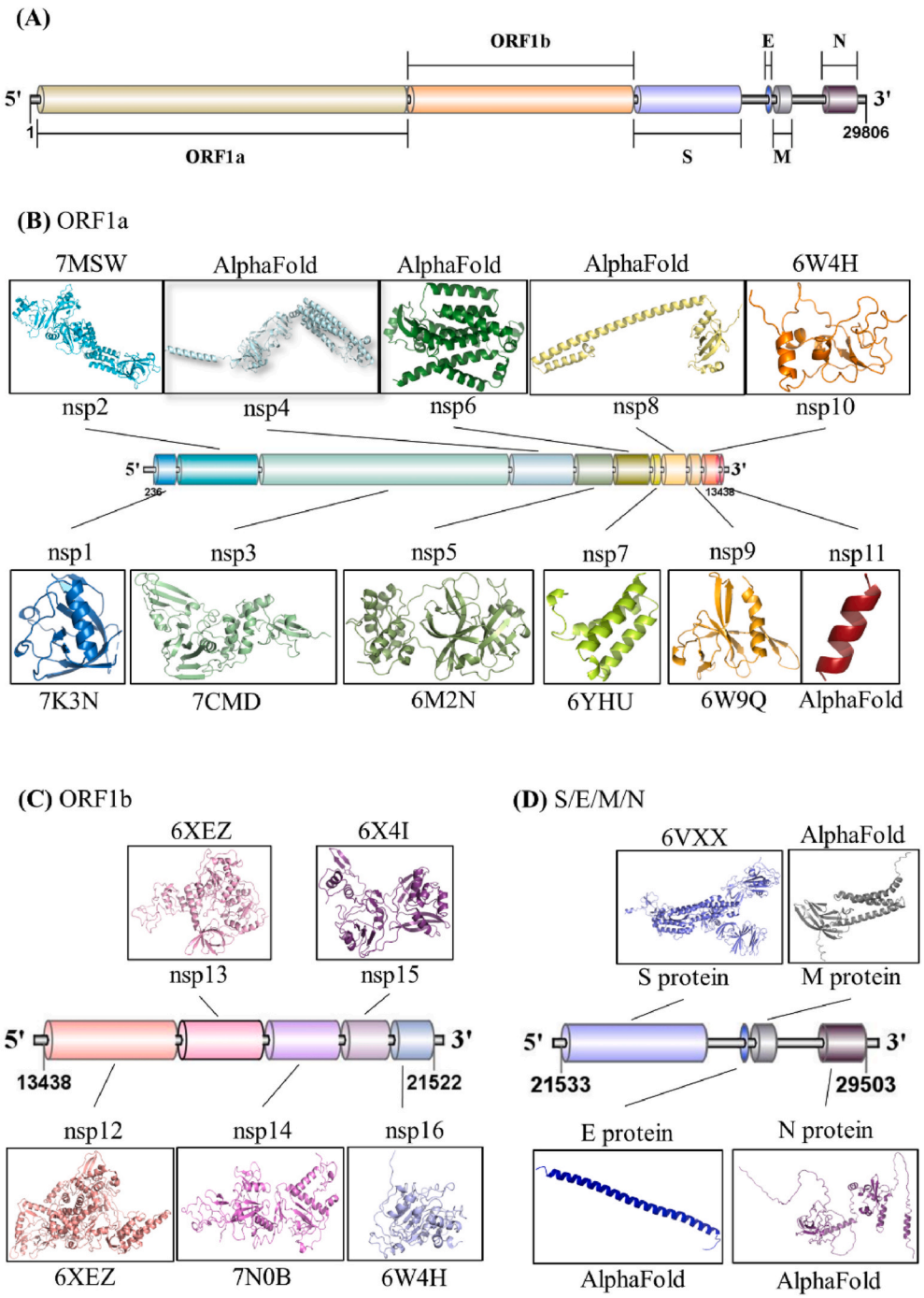

Fig. 1. Genomic sequence and protein structures of SARS-CoV-2. (A) The genome organization for SARS-CoV-2 full sequence (GenBank: OL518896.1). (B) Genomic organization of SARS-CoV-2 ORF1a and structures of ORF1a protein products (nsp1-nsp11). (C) Genomic organization of SARS-CoV-2 ORF1b and structures of ORF1b protein products (nsp12-nsp16). (D) Structures of SARS-CoV-2 structural proteins. Proteins with experimentally determined structures are marked with their PDB ID, proteins without experimental structures are predicted by AlphaFold v2.0.

5′-leader-UTR-replicase(ORF1ab)-Spike (S)-Envelope (E)- Membrane (M)-Nucleocapsid (N)-3′UTR-poly A [9]. The open reading frames 1 ab occupy the first two thirds of the genome and is translated into 2 polyproteins pp1a and pp1ab. The pp1a and pp1ab are cleaved into nsp1 to nsp16 [10,11]. These non-structural proteins play significant roles in the regulation of viral RNA replication and transcription [12,13]. The later reading frames encode four structural proteins: spike (S) protein, envelop (E) protein, membrane (M) protein, and nucleocapsid (N) protein (Fig. 1) [14]. These structural proteins are vital for viral assembly and release of virus-like particles (VLPs) by transfected cells [12,15,16].

Human coronaviruses were first identified in the 1965 [17,18]. The seven coronaviruses that cause disease in humans include 4 common human coronaviruses (HCoVs), namely HCoV-229E, HCoV-OC43, HCoV-NL63, HCoV-HKU1, that are circulating globally in the human population, and 3 coronaviruses (SARS-CoV, MERS-CoV, SARS-CoV-2) which have caused major outbreaks of deadly pneumonia in the 21st century [19–21].

In this study, we collected the sequence and structural information of 20 proteins encoded by the 7 known pathogenic human coronaviruses, including 4 structural proteins (spike protein, envelop protein, membrane protein, nucleocapsid protein) and 16 non-structural proteins (nsp1-nsp16). After protein multiple sequence alignment (MSA), we evaluated the conservation of the 20 proteins among the human coronaviruses. Subsequently, we focused on the potential ligand-binding pockets of proteins, analyzed the pockets conservation and the characteristic laws of amino acids in the pockets, as well as carried out molecular docking to explore the rationality of the pocket conservative. Our results will be helpful for looking for broad-spectrum drugs to fight the

Table 1
Information of the non-structural and structural proteins of the 7 human pathogenic coronaviruses.[a]

Protein	Alternative name	ORF	Length range (a.a.)	HCoV-229E		HCoV-OC43		SARS-CoV		HCoV-NL63		HCoV-HKU1		MERS-CoV		SARS-CoV-2	
				UniProt ID	PDB ID	UniProt ID	PDB ID	UniProt ID	PDB ID	UniProt ID	PDB ID	UniProt ID	PDB ID	UniProt ID	PDB ID	UniProt ID	PDB ID
Non-structural proteins																	
nsp1	Leader protein	ORF1a	110–246	P0C6X1		P0C6X6		P0C6X7	2HSX	P0C6X5		P0C6X4		K9N7C7		P0DTD1	7K3N
nsp2		ORF1a	587–788	P0C6X1		P0C6X6		P0C6X7		P0C6X5		P0C6X4		K9N7C7		P0DTD1	7MSW
nsp3	Papain-like proteinase	ORF1a	1979–1564	P0C6X1		P0C6X6		P0C6X7	5TL7	P0C6X5		P0C6X4		K9N7C7	4RNA	P0DTD1	7CMD
nsp4		ORF1a	477–508	P0C6X1		P0C6X6		P0C6X7		P0C6X5		P0C6X4		K9N7C7		P0DTD1	
nsp5	3C-like proteinase	ORF1a	302–311	P0C6X1	2ZU2	P0C6X6		P0C6X7	2ZU5	P0C6X5	7E6R	P0C6X4	3D23	K9N7C7	4RSP	P0DTD1	6M2N
nsp6		ORF1a	279–294	P0C6X1		P0C6X6		P0C6X7		P0C6X5		P0C6X4		K9N7C7		P0DTD1	
nsp7		ORF1a	83–92	P0C6X1		P0C6X6		P0C6X7	2KYS	P0C6X5		P0C6X4		K9N7C7		P0DTD1	6YHU
nsp8		ORF1a	194–201	P0C6X1		P0C6X6		P0C6X7		P0C6X5		P0C6X4		K9N7C7		P0DTD1	
nsp9		ORF1a	109–114	P0C6X1	2J97	P0C6X6		P0C6X7	1UW7	P0C6X5		P0C6X4		K9N7C7		P0DTD1	6W9Q
nsp10	Growth factor-like peptide	ORF1a	135–141	P0C6X1		P0C6X6		P0C6X7	3R24	P0C6X5		P0C6X4		K9N7C7	5YN5	P0DTD1	6W4H
nsp11		ORF1a	13–17	P0C6U2		P0C6U7		P0C6U8		P0C6U6		P0C6U5		K9N638		P0DTD1	
nsp12	RNA-directed RNA polymerase	ORF1b	927–947	P0C6X1		P0C6X6		P0C6X7	6NUR	P0C6X5		P0C6X4		K9N7C7		P0DTD1	6XEZ
nsp13	Helicase	ORF1b	597–611	P0C6X1		P0C6X6		P0C6X7	6JYT	P0C6X5		P0C6X4		K9N7C7	5WWP	P0DTD1	6XEZ
nsp14	Proofreading exoribonuclease	ORF1b	518–535	P0C6X1		P0C6X6		P0C6X7	5C8S	P0C6X5		P0C6X4		K9N7C7		P0DTD1	7N0B
nsp15	Uridylate-specific endoribonuclease	ORF1b	343–375	P0C6X1	4S1T	P0C6X6		P0C6X7	2H85	P0C6X5		P0C6X4		K9N7C7	5YVD	P0DTD1	6X4I
nsp16	2′-O-methyltransferase	ORF1b	298–303	P0C6X1		P0C6X6	7NH7	P0C6X7	3R24	P0C6X5		P0C6X4		K9N7C7	5YN5	P0DTD1	6W4H
Structural proteins																	
S protein		ORF2	1173–1356	P15423	7CYD	P36334	6OHW	P0DTC2	6ACD	Q6Q1S2	7KIP	Q0ZME7	5I08	K9N5Q8	5 × 5F	A0A679G9E9	6VXX
E protein		ORF4	67–84	S5YAG7		Q4VID3		P59637		Q5SBN7		Q5MQC8		A0A166ZLT5		A0A6C0QFP9	
M protein		ORF5	219–230	P15422		Q4VID2		P59596		Q6Q1R9		Q5MQC7		K9N7A1		A0A6V7AL93	
N protein		ORF9a	377–448	P15130		P33469		P59595		Q6Q1R8		Q5MQC6		K9N4V7		A0A6C0T6Z7	

[a] The underlined PDB ID indicates that the structure has resolved ligand structure.

Table 2
The conservation ranking result of the 20 proteins among the 7 human coronaviruses.[a]

Protein Identity Percentages				Amino Acid Conservation Score			
Order	Protein	Average	Variance	Order	Protein	Average	Variance
1	nsp13	67.443%	0.012	**1**	nsp13	5.800	10.483
2	nsp12	66.107%	0.012	**2**	nsp16	5.648	10.235
3	nsp16	63.623%	0.011	**3**	nsp12	5.579	9.471
4	nsp14	59.489%	0.013	**4**	nsp14	5.440	8.706
5	nsp10	57.715%	0.015	**5**	nsp11	5.385	6.391
6	nsp5	52.181%	0.019	**6**	nsp5	5.288	8.368
7	nsp9	51.273%	0.020	**7**	nsp15	5.277	8.628
8	nsp15	51.086%	0.015	**8**	nsp7	5.217	7.134
9	nsp7	50.476%	0.028	**9**	nsp10	5.209	8.554
10	nsp11	49.625%	0.031	**10**	nsp9	5.168	8.140
11	nsp8	41.812%	0.045	**11**	M protein	5.090	6.055
12	M protein	40.653%	0.026	**12**	nsp8	5.076	5.545
13	nsp4	39.339%	0.025	**13**	nsp4	5.074	5.545
14	N protein	36.773%	0.023	**14**	nsp6	5.024	4.348
15	S protein	34.602%	0.028	**15**	N protein	5.022	4.863
16	nsp6	34.503%	0.027	**16**	nsp1	5.022	2.822
17	nsp3	31.174%	0.020	**17**	nsp3	5.017	3.365
18	E protein	29.884%	0.031	**18**	E protein	5.013	3.154
19	nsp1	23.391%	0.031	**19**	S protein	5.009	4.322
20	nsp2	22.641%	0.017	**20**	nsp2	4.970	2.358

[a] The ranking result based on the average of sequence identity percentages is displayed on the left, while the ranking result based on the average of each amino acid's conservation score is displayed on the right.

currently known human coronavirus as well as coronavirus that could emerge in the future.

2. Methods

2.1. Data collection

Sequence information of the 4 structural proteins and 16 non-structural proteins were collected from Universal Protein Resource (UniProt) for each of the 7 known human coronaviruses. Subsequently, the experimentally determined structure of the 140 proteins were searched in the Protein Data Bank (PDB). Proteins with resolved structures were downloaded from PDB and processed into the form of monomer, while those without resolved structures were predicted by using AlphaFold v2.0 [22].

2.2. Protein multiple sequence alignment and conservative analysis

MSA was performed by Clustal W via the webserver Clustal Omega (https://www.ebi.ac.uk/Tools/msa/clustalo/) [23,24]. The results of MSA then were used to perform conservative analysis by the ConSurf Server (https://consurf.tau.ac.il/) [25]. The rate of evolution at each site was calculated using the empirical Bayesian, and the continuous conservation scores were then divided into nine levels (from grade 1 for most variable positions to grade 9 for the most conserved positions) of discrete scales in ConSurf [26]. The discrete scales were projected onto the protein sequences and structures of SARS-CoV-2 for visualization. In order to compare the conservative property of the 20 coronavirus proteins of the coronaviruses, two evaluation methods were used to rank protein conservation: one is the average value of sequence similarity percentages between the same proteins of different viruses, and the other is the average value of each amino acid's conservation scores that calculated by the ConSurf Server.

2.3. Protein pocket generation

All the potential ligand binding pockets of the 140 proteins were predicted by D3Pockets (http://www.d3pharma.com/D3Pocket/index.php) [27]. The pockets for further analysis in this study were selected based on the following criteria: the pockets with endogenous or reported ligand, the pockets with druggability score of 1 or ranked in top by D3Pockets. That is, if there are ligands in the experimental determined structures, select the pockets where the ligands are located. For proteins those without reported ligand, the pockets predicted to be druggable by D3Pockets are chosen preferentially. If the proteins do not meet the above two criteria, the first one ranked by D3Pockets was selected.

2.4. Pockets conservative assessment and molecular docking verification

Pocket conservation was assessed by the conservation scores that calculated using the ConSurf Server for each amino acid in the protein pockets. To verify the rationality of the predicted pocket conservative results, we carried out molecular docking for the SARS-CoV-2 proteins with ligands in their resolved structures (nsp3, nsp5, nsp12, nsp13 and nsp16) by the docking program Glide SP of Schrödinger Release 2020 [28–30].

2.5. Exploration of pocket amino acid composition

The amino acid composition of a protein pocket determines the interactions available for ligand binding, and understanding the composition of the potential binding site is of great importance for structure-based drug development. In order to explore the property of the potential ligand binding pockets of 20 coronavirus protein, as well as finding the conserved protein pockets with the consistent pocket amino acid composition among 7 human coronaviruses, we counted the frequency of each amino acid around the pocket.

2.6. Sequence similarity between highly conserved coronavirus proteins and human proteins

To assess the potential off-target problem, we performed sequence similarity searches between the 6 highly conserved proteins of coronaviruses and all the human proteins (using the protein sequences of SARS-CoV-2 as the representative of the coronaviruses protein family) using the search tool BLASTp provided by NCBI (https://www.ncbi.nlm.nih.gov/BLAST/). The amino acids that are identical to human proteins are shown in the 3D protein structure, and those in binding pockets are highlighted to facilitate the design of highly selective anti-coronavirus drugs in the future.

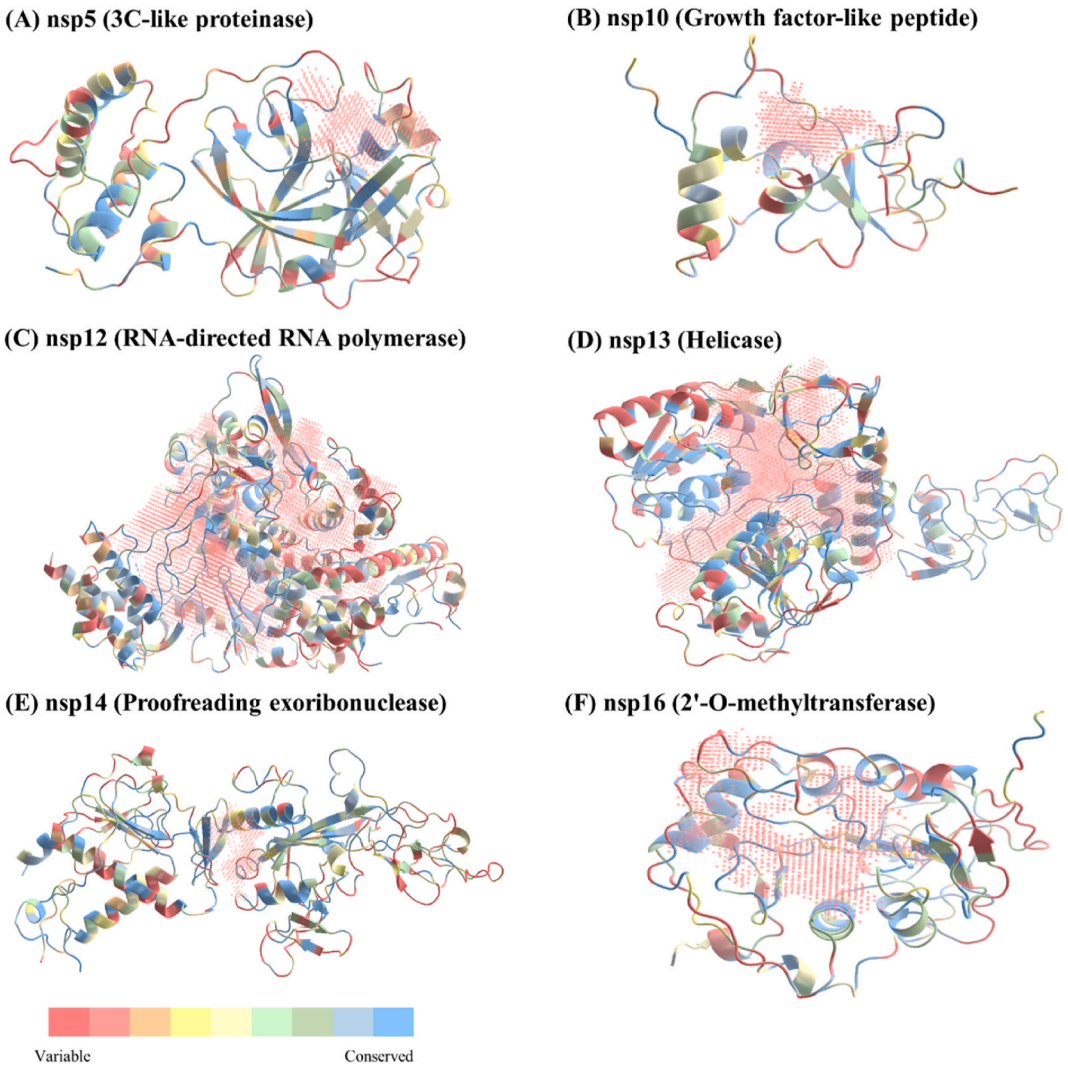

Fig. 2. Conservative analysis of coronavirus proteins. The 3D structures of (A) nsp5, (B) nsp10, (C) nsp12, (D) nsp13, (E) nsp14, (F) nsp16 presented by cartoon models (pink indicates the variable while blue indicates the conserved residues). The potential ligand binding pockets are displayed in dots. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

3. Results and discussion

3.1. Sequence alignment and conservative assessment

The protein information of the 7 coronaviruses is shown in Table 1. The sequence information of the 140 proteins was obtained from UniProt, but only 42 proteins were found in PDB. The 3D structures of the rest 98 proteins were modeled by AlphaFold v2.0. As examples, the 20 proteins of SARS-CoV-2 were depicted in Fig. 1B and C (for 16 non-structural proteins) and 1D (for 4 structural proteins).

The MSA results and conservation scores were used to rank the conservation degree of 20 types of coronavirus proteins. The ranking results by both the average of sequence similarity percentages and the average value of each amino acid's conservation scores are summarized in Table 2. There are 10 proteins with average identity percentage scores higher than 50% between SARS-CoV-2 and other coronaviruses, which all ranked in the top 10 by amino acid conservation scores (Table 2). Among these 10 proteins, the nsp11 monomer has no potential binding site because of its short sequence. It suggesting that the remaining 9 proteins deserve further assessment as potential targets for pan-coronavirus drug development, which are nsp13 (Helicase), nsp12 (RNA-directed RNA polymerase), nsp16 (2′-O-methyltransferase), nsp14 (Proofreading exoribonuclease), nsp10 (Growth factor-like

Table 3
Pocket conservation ranking results of the 18 proteins among the 7 human coronaviruses.

Order	Protein Name	Average of Protein Pocket Amino Acid Conservation Scores	Variance of Protein Pocket Amino Acid Conservation Scores
1	nsp16	6.638	8.679
2	nsp14	6.333	7.944
3	M protein	6.294	4.590
4	nsp13	6.110	10.149
5	nsp1	6.000	2.625
6	nsp5	5.975	8.124
7	nsp12	5.958	8.895
8	nsp10	5.875	7.026
9	S protein	5.815	4.062
10	N protein	5.584	5.126
11	nsp3	5.571	3.959
12	nsp4	5.562	4.837
13	nsp9	5.500	7.625
14	nsp15	5.338	7.764
15	nsp8	5.227	4.903
16	nsp2	5.218	2.507
17	nsp7	4.905	6.086
18	nsp6	4.676	3.705

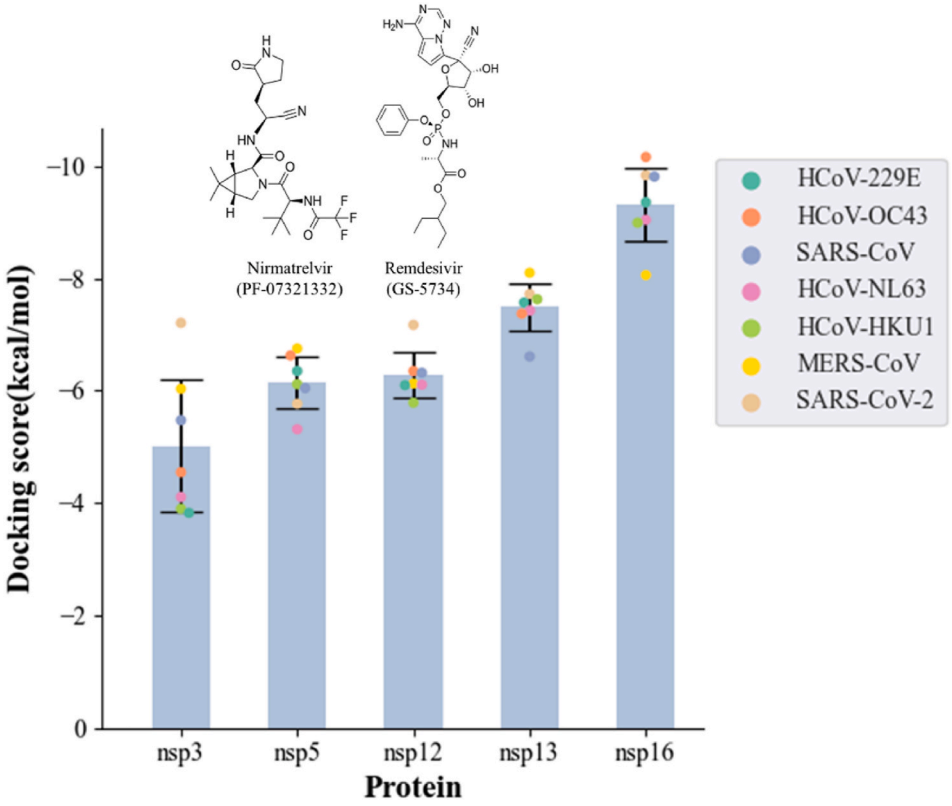

Fig. 3. The docking scores of SARS-CoV-2 ligands to the homology proteins of different coronaviruses (kcal/mol). The 2D structures represent the FDA approved SARS-CoV-2 nsp5 inhibitor Nirmatrelvir and nsp12 inhibitor Remdesivir.

peptide), nsp5 (3C-like proteinase), nsp9, nsp15 (Uridylate-specific endoribonuclease) and nsp7. We projected the discrete grades on the protein sequence (Fig. S1), the cartoon structure and the surface structure of the 20 proteins of the SARS-CoV-2 (Fig. S2). As examples, we show the cartoon structures of the 6 relatively conservative proteins in Fig. 2.

3.2. Pockets conservative assessment and verification

With D3Pockets, we found that nsp11 of HCoV-HKU1 and SARS-CoV-2, E protein of MERS-CoV and SARS-CoV-2 have no possible ligand binding pockets. Therefore, the monomer structures of these two types of proteins are not potential targets. Accordingly, we only scored and ranked the pocket conservative property of the other 18 types of protein. Table 3 shows that 6 of the above 9 conservative proteins are ranked within top 10 by pocket amino acid conservation scores, which are nsp16 (2′-O-methyltransferase), nsp14 (Proofreading exoribonuclease), nsp13 (Helicase), nsp5 (3C-like proteinase), nsp12 (RNA-directed RNA polymerase) and nsp10 (Growth factor-like peptide), indicating that the 6 proteins should be potential drug targets for pan-coronavirus drug development. We also used Fpocket to calculate the pockets of these six proteins and compared them with those predicted by D3pockets (Fig. S3, Table S1) [31,32].

There are 5 ligand-protein structures of SARS-CoV-2 available from PDB, which are nsp3, nsp5, nsp12, nsp13, nsp16. To validate whether a protein ligand of a coronavirus, e.g. SARS-CoV-2, is also good pan-ligand to the homology protein of other coronaviruses, we performed molecular docking study for the 5 ligands to other homology proteins. The docking results are shown in Fig. 3 and Table S2. The variance of the docking scores of nsp12, nsp5, nsp13, nsp16 are all within 0.5 kcal/mol, in this perspective we considered that these four proteins have higher similarities in the same ligand binding pockets of different coronaviruses. In addition, as shown in Fig. 3, these four types of proteins have good binding ability to ligand molecules, especially nsp14 and nsp16. The successful marketing of SARS-CoV-2 nsp5 inhibitors (such as Nirmatrelvir, an antiviral medication developed by Pfizer which is part of the nirmatrelvir/ritonavir combination sold under the brand name Paxlovid) and nsp12 inhibitors (such as Remdesivir, an antiviral nucleotide analogue developed by Gilead Sciences) proves the druggability of nsp5 and nsp12. Therefore, we have reason to believe that nsp14 and nsp16 are also target proteins with high research value. In contrast, the docking scores for different nsp3 (PL-PRO) are quite different with variance of 1.4 kcal/mol, suggesting that the ligand has different binding affinity to the nsp3 protein from different coronaviruses. All the results revealed that the strategy for identifying potential pan-coronaviruses targets in this study is reasonable.

3.3. Pocket amino acids analysis

We counted the frequency of each amino acid around the pocket, for investigating the pocket property of the 20 coronavirus proteins (Fig. 4, Fig. S4). As shown in Fig. S4I and Fig. S4L, the sequence length of nsp11 and E protein is short, and the pockets calculated in the form of monomer protein are of little reference value as we mentioned above. From the statistical results of the pocket amino acids, we can summarize some characteristics: A) the pocket amino acid residues between SARS-CoV and SARS-CoV-2 are more consistent than others (Fig. 4), B) the frequency of CYS in the pocket of nsp10 is significantly higher than that of all other protein pockets (Fig. 4B), indicating that covalent inhibitors target nsp10 can be designed, C) the acidic amino acid ASP has a higher content in the nsp12 pockets of various coronaviruses (Fig. 4C), thus molecules that target nsp12 are preferably positively charged, D) the amino acid frequency of the nsp13 pocket is basically the same among the seven coronaviruses (Fig. 4D), with the smallest difference and the best conservation, followed by nsp5, nsp14 and nsp16 (Fig. 4A, E, 4F).

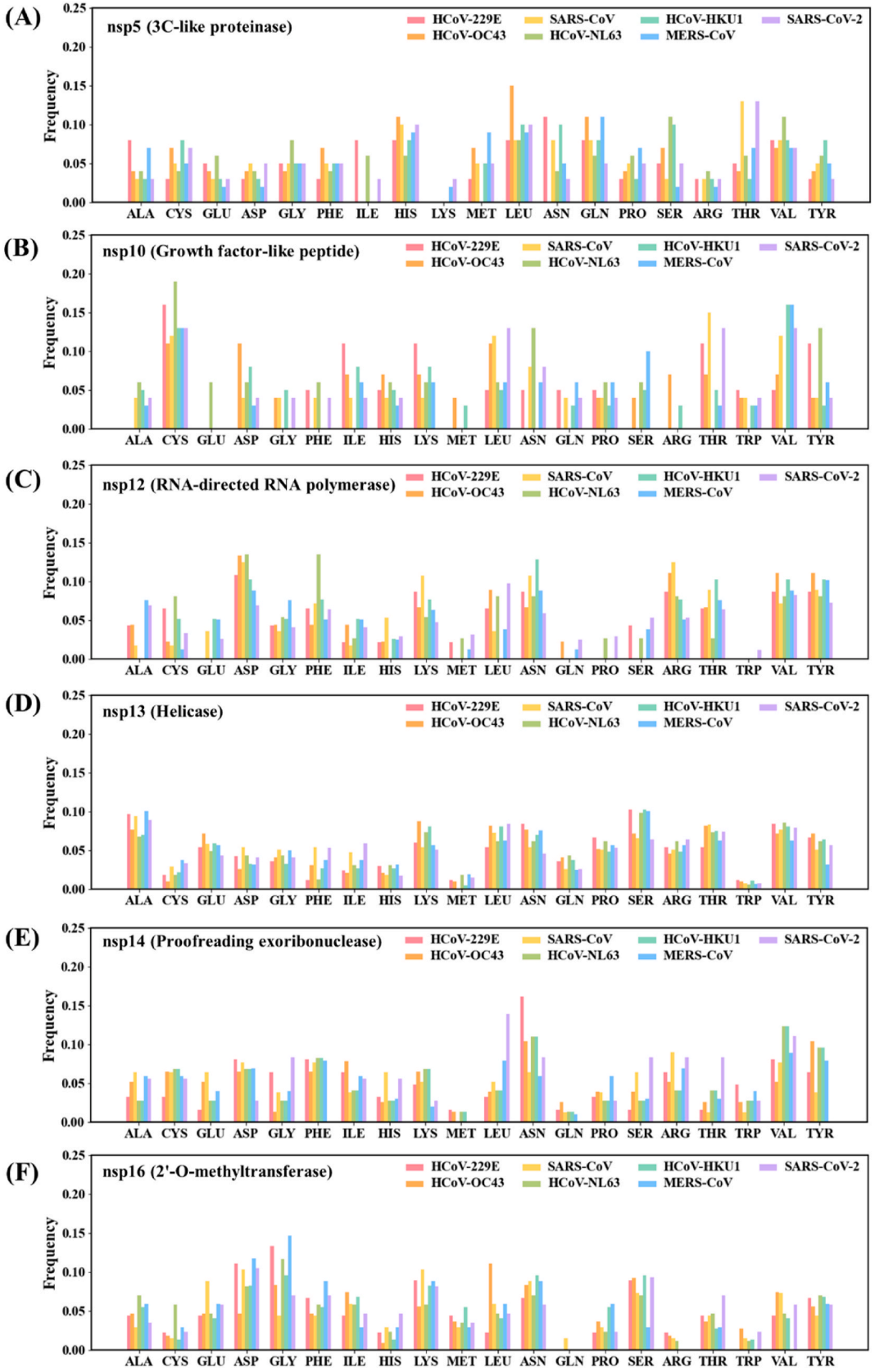

Fig. 4. Ratio of 20 amino acids that form the pockets of coronavirus (A) nsp5, (B) nsp10, (C) nsp12, (D) nsp13, (E) nsp14, (F) nsp16. Amino acid ratios of HCoV-229E, HCoV-OC43, SARS-CoV, HCoV-NL63, HCoV-HKU1, MERS-CoV, SARS-CoV-2 are shown in rose pink, pale orange, light mustard, olive green, light teal, sky blue, pale violet respectively. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

Table 4
Information of the top three pieces proteins with the smallest except value in the BLAST results of SARS-CoV-2 nsp5 (3C-like proteinase), nsp10 (Growth factor-like peptide), nsp12 (RNA-directed RNA polymerase), nsp13 (Helicase), nsp14 (Proofreading exoribonuclease) and nsp16 (2′-O-methyltransferase).

Protein	Human-derived protein description	Total Score	Query Cover	E Value	Per. Similarity[a]
nsp5	Cadherin-12 isoform 3 [Homo sapiens]	26.6	29%	125	20.64%
	Cadherin-12 isoform 2 precursor [Homo sapiens]	26.2	29%	135	21.84%
	Cadherin-12 isoform 1 preproprotein [Homo sapiens]	26.2	29%	138	21.58%
nsp10	E3 ubiquitin-protein ligase MYCBP2 isoform X4 [Homo sapiens]	28.5	10%	8.6	22.48%
	E3 ubiquitin-protein ligase MYCBP2 isoform X17 [Homo sapiens]	28.5	10%	8.7	22.48%
	E3 ubiquitin-protein ligase MYCBP2 isoform X11 [Homo sapiens]	28.5	10%	8.7	22.48%
nsp12	WASH complex subunit 2C isoform 25 [Homo sapiens]	29.6	9%	45	20.86%
	WASH complex subunit 2A isoform X14 [Homo sapiens]	29.6	4%	46	20.86%
	WASH complex subunit 2A isoform X16 [Homo sapiens]	29.6.	4%	46	20.86%
nsp13	Protein ZGRF1 isoform X10 [Homo sapiens]	38.1	34%	0.073	22.55%
	Protein ZGRF1 isoform X9 [Homo sapiens]	37.7	34%	0.11	22.55%
	Protein ZGRF1 isoform X8 [Homo sapiens]	37.4	34%	0.12	22.55%
nsp14	Alstrom syndrome protein 1 isoform 2 [Homo sapiens]	28.9	5%	55	19.79%
	Alstrom syndrome protein 1 isoform 1 [Homo sapiens]	28.9	5%	55	19.79%
	Betaine—homocysteine S-methyltransferase 1 [Homo sapiens]	26.6	3%	206	12.87%
nsp16	Potassium voltage-gated channel subfamily C member 2 isoform X4 [Homo sapiens]	24.3	6%	514	25.00%
	Sphingomyelin phosphodiesterase 3 [Homo sapiens]	24.3	27%	525	24.01%
	Sphingomyelin phosphodiesterase 3 isoform X2 [Homo sapiens]	23.9	27%	701	25.73%

[a] Per. Similarity was calculated by Clustal W via the webserver Clustal Omega [24].

3.4. The similarity between coronavirus nsp5, nsp10, nsp12, nsp13, nsp14, nsp16 and human-derived proteins

As mentioned above, nsp5 (3C-like proteinase), nsp10 (Growth factor-like peptide), nsp12 (RNA-directed RNA polymerase), nsp13 (Helicase), nsp14 (Proofreading exoribonuclease) and nsp16 (2′-O-methyltransferase) are relatively conserved proteins. Therefore, we preliminarily believed that these types of proteins can be the preferred targets when designing broad-spectrum drugs for coronaviruses. In order to further explore the rationality of this hypothesis, we performed sequence similarity searches for these six types of proteins respectively by BLASTp. We listed the top three pieces of human-derived protein information with the smallest E value in the search results for each protein in Table 4. In this table it can be seen that nsp5, nsp10, nsp12, nsp14, and nsp16 have very low similarity with human proteins, while nsp13 has a lower expect value with protein ZGRF1 isoform family.

Therefore, we ran MSA between nsp13 (represented by the nsp13 sequence of SRAS-CoV-2) and ten proteins in ZGRF1 isoform protein family. It showed that the sequence similarity percentage of SARS-CoV-2 nsp13 and protein ZGRF1 isoform X10 is 22.55%. With the MSA result, 111 amino acids (accounting for about 18.5% of the total length of nsp13) that completely matched with human protein were marked on the structure of SARS-CoV-2 nsp13 (Fig. S5A). Among them, 82 amino acid residues are located around the protein ligand binding pocket of SARS-CoV-2 nsp13 (Fig. S5B). However, even the nsp13, which has the highest sequence similarity with human-derived proteins, the similarity has not reached 30%. In fact, as shown in the last column of Table 4, these six conserved proteins have similarity less than 25% with human-derived proteins. Therefore, we still believed that these six proteins are potential drug targets for pan-coronavirus drug development.

4. Conclusion

Although the first human coronavirus was isolated as early as the 1960s, because HCoV-229E and HCoV-OC43 have low pathogenicity and low infectivity, the large-scale serious infections that have been caused by coronaviruses until the 21st century, especially after the global pandemic of COVID-19 that people have truly realized how harmful the coronavirus will cause to human. In order to deal with the current pathogenic human coronaviruses and even the coronaviruses that may be harmful to human in the future, researches related to pan-coronavirus drug discovery will be key issues for researchers.

In this study, we performed MSA for the major proteins of currently known human coronaviruses, and evaluated the sequence conservation of these proteins among seven human coronaviruses. We found that nsp13, nsp12, nsp16, nsp14, nsp10, nsp5, nsp9, nsp15 and nsp7 are highly conserved throughout the 7 coronaviruses. Among them, 6 proteins, viz., nsp16, nsp14, nsp13, nsp5, nsp12 and nsp10, have more conserved ligand binding pockets than other proteins. We also found that the 6 highly conserved proteins are significantly different from all the human-derived proteins in both protein sequence similarity and ligand binding pocket structure. Overall, we believe all the work we have done can help people understand the structural and non-structural proteins of human coronaviruses, and more importantly, can provide information for the future development of pan-coronavirus drugs.

Declaration of competing interest

The authors declare no competing financial interest.

Acknowledgment

Funding: This work has been supported by the National Key R&D Program of China (2016YFA0502301 & 2017YFB0202601) and Natural Science Foundation of Shanghai (21ZR1475600).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105455.

References

[1] S. Krishnamoorthy, B. Swain, S.S. Gunthe, et al., SARS-CoV, MERS-CoV, and 2019-nCoV viruses: an overview of origin, evolution, and genetic variations, VirusDis 31 (4) (2020) 411–423, https://doi.org/10.1007/s13337-020-00632-9.
[2] L.H.W. Lum, P.A. Tambyah, Outbreak of COVID-19 – an urgent need for good science to silence our fears? Singap. Med. J. 61 (2) (2020) 55–57, https://doi.org/10.11622/smedj.2020018.
[3] C.A. Sánchez, H. Li, K.L. Phelps, K.J. Olival, P. Daszak, et al., A strategy to assess spillover risk of bat SARS-related coronaviruses in Southeast Asia, medRxiv (2021), https://doi.org/10.1101/2021.09.09.21263359.
[4] L. Wu, L. Zhou, M. Mo, et al., SARS-CoV-2 Omicron RBD shows weaker binding affinity than the currently dominant Delta variant to human ACE2, Signal

Transduct, Targeted Ther 7 (2022) 8, https://doi.org/10.1038/s41392-021-00863-2.
[5] H. Yang, C. Qin, Y. Li, L. Tao, J. Zhou, C. Yu, F. Xu, Z. Chen, F. Zhu, Y. Chen, Therapeutic target database update 2016: enriched resource for bench to clinical drug target and targeted pathway information, Nucleic Acids Res. 44 (2016) D1069–D1074, https://doi.org/10.1093/nar/gkv1230.
[6] P. V'kovski, A. Kratzel, S. Steiner, H. Stalder, V. Thiel, Coronavirus biology and replication: implications for SARS-CoV-2, Nat. Rev. Microbiol. 19 (2021) 155–170, https://doi.org/10.1038/s41579-020-00468-6.
[7] M. Pal, G. Berhanu, C. Desalegn, V. Kandi, Severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2): an update, Cureus 12 (3) (2020), e7423, https://doi.org/10.7759/cureus.7423.
[8] D. Schoeman, B.C. Fielding, Coronavirus envelop protein: current knowledge, Virol. J. 16 (2019) 69, https://doi.org/10.1186/s12985-019-1182-0.
[9] A.R. Fehr, S. Perlamn, Coronaviruses, An overview of their replication and pathogenesis, Coronaviruses 1282 (2015) 1–23, https://doi.org/10.1007/978-1-4939-2438-7_1.
[10] B. Malone, N. Urakova, E. Snijder, E.A. Campbell, Structures and functions of coronavirus replication–transcription complexes and their relevance for SARS-CoV-2 drug design, Nat. Rev. Mol. Cell Biol. (2021) 1–19, https://doi.org/10.1038/s41580-021-00432-z.
[11] E.J. Snijder, P.J. Bredenbeek, et al., Unique and conserved features of genome and proteome of SARS-coronavirus, an early split-off from the coronavirus group 2 Lineage, J. Mol. Biol. 331 (5) (2003) 991–1004, https://doi.org/10.1016/s0022-2836(03)00865-9.
[12] J. Mohan, T. Wollert, Membrane remodeling by SARS-CoV-2 - double-enveloped viral replication, Fac Rev 10 (2021) 17, https://doi.org/10.12703/r/10-17.
[13] C. Bai, Q. Zhong, G.F. Gao, Overview of SARS-CoV-2 genome-encoded proteins, Sci. China Life Sci. 10 (2021) 1–15, https://doi.org/10.1007/s11427-021-1964-4.
[14] H. Yang, Z. Rao, Structural biology of SARS-CoV-2 and implications for therapeutic development, Microbiology 19 (2021) 685–700, https://doi.org/10.1038/s41579-021-00630-8.
[15] Y.L. Siu, K.T. Teoh, F. Kien, J.S.M. Peiris, Nal B el at., the M, E, and N structural proteins of the severe acute respiratory syndrome coronavirus are required for efficient assembly, trafficking, and release of virus-like particles, J. Virol. 82 (22) (2008) 11318–11330, https://doi.org/10.1128/JVI.01052-08.
[16] S. Satarker, M. Nampoothiri, Structural proteins in sever acute respiratory syndrome coronavirus-2, Arch. Med. Res. 51 (2020), https://doi.org/10.1016/j.arcmed.2020.05.012, 5, 482-49.
[17] J.S. Kahn, K. Mclntosh, History and recent advances in coronavirus discovery, Pediatr. Infect. Dis. J. 24 (2005) S223–S227, https://doi.org/10.1097/01.inf.0000188166.17324.60.
[18] E. Mahase, Covid-19: first coronavirus was described in the BMJ in 1965, BMJ 369 (2020) m1547, https://doi.org/10.1136/bmj.m1547.
[19] L. Hoek, Human coronaviruses, what do they cause? Antivir. Ther. 12 (4 Pt B) (2007) 651–658.
[20] E.R. Gaunt, A. Hardie, E.C.J. Claas, P. Simmonds, K.E. Templeton, Epidemiology and clinical presentations of the four human coronaviruses 229E, HKU1, NL63, and OC43 detected over 3 years using a novel multiplex real-time PCR method, J. Clin. Microbiol. 48 (8) (2020) 2940–2947, https://doi.org/10.1128/JCM.00636-10.
[21] D.X. Liu, J.Q. Liang, T.S. Fung, Human coronavirus-229E, -OC43, -NL63, and -HKU1 (Coronaviridae), Encyclop. Virol. 2 (2021) 428–440, https://doi.org/10.1016/B978-0-12-809633-8.21501-X.
[22] J. Jumper, R. Evans, et al., Highly accurate protein structure prediction with AlphaFold, Nature 596 (2021) 583–589, https://doi.org/10.1038/s41586-021-03819-2.
[23] J. Thompson, D. Higgins, T. Gibson, W. Clustal, Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, Nucleic Acids Res. 22 (22) (1994) 4673–4690, https://doi.org/10.1093/nar/22.22.4673.
[24] F. Madeira, Y. Park, et al., The EMBL-EBI search and sequence analysis tools APIs in 2019, Nucleic Acids Res. 47 (2019) W636–W641, https://doi.org/10.1093/nar/gkz268.
[25] H. Ashkenazy, O. Chay, T. Pupko, et al., ConSurf 2016: an improved methodology to estimate and visualize evolutionary conservation in macromolecules, Nucleic Acids Res. 44 (2016) W344–W350, https://doi.org/10.1093/nar/gkw408.
[26] I. Mayrose, D. Graur, N. Ben-Tal, T. Pupko, Comparison of site-specific rate-inference methods for protein sequences: empirical Bayesian methods are superior, Mol. Biol. Evol. 21 (9) (2004) 1781–1791, https://doi.org/10.1093/molbev/msh194.
[27] Z. Chen, X. Zhang, C. Peng, et al., D3Pockets: a method and web server for systematic analysis of protein pocket dynamics, J. Chem. Inf. Model. 59 (2019) 3353–3358, https://doi.org/10.1021/acs.jcim.9b00332.
[28] R.A. Friesner, R.B. Murphy, M.P. Repasky, L.L. Frye, J.R. Greenwood, T.A. Halgren, P.C. Sanschagrin, D.T. Mainz, Extra precision Glide: docking and scoring incorporating a model of hydrophobic Enclosure for protein-ligand complexes, J. Med. Chem. 49 (21) (2006) 6177–6196, https://doi.org/10.1021/jm051256o.
[29] R.A. Friesner, J.L. Banks, R.B. Murphy, T.A. Halgren, J.J. Klicic, D.T. Mainz, M. P. Repasky, E.H. Knoll, D.E. Shaw, M. Shelley, J.K. Perry, P. Francis, P.S. Shenkin, Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy, J. Med. Chem. 47 (7) (2004) 1739–1749, https://doi.org/10.1021/jm0306430.
[30] T.A. Halgren, R.B. Murphy, R.A. Friesner, H.S. Beard, L.L. Frye, W.T. Pollard, J. L. Banks, Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening, J. Med. Chem. 47 (7) (2004) 1750–1759, https://doi.org/10.1021/jm030644s.
[31] Vincent Le Guilloux, Schmidtke Peter, Pierre Tuffery, Fpocket, An open source platform for ligand pocket detection, BMC Bioinf. 10 (168) (2009). https://10.1186/1471-2105-10-168.
[32] P. Schmidtke, V. Le Guilloux, J. Maupetit, P. Tuffery, Fpocket: online tools for protein ensemble pocket detection and tracking, Nucleic Acids Res. 38 (2010) W582–W589. https://10.1093/nar/gkq383.

Minfei Ma is a postgraduate at Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences. Her research interests are computer-aided drug design and bioinformatics.

Yanqing Yang is a postgraduate at Shanghai Institute of Materia Medica. His research interests are molecular docking, virtual screening and molecular dynamics. His affiliation is with CAS Key Laboratory of Receptor Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China; School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China.

Leyun Wu is a postgraduate at Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences. Her research interests are computer-aided drug design and molecular dynamics.

Liping Zhou is a Ph.D. student at Shanghai Institute of Materia Medica. Her research interests are molecular dynamics simulation. Her affiliation is Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China; School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China.

Yulong Shi is a postgraduate at Shanghai Institute of Materia Medica. His research interest is molecular docking method development. His affiliation is with CAS Key Laboratory of Receptor Research; Drug Discovery and Design Center, Shanghai Institute of Materia Medica, Chinese Academy of Sciences, Shanghai, 201203, China; School of Pharmacy, University of Chinese Academy of Sciences, No.19A Yuquan Road, Beijing, 100049, China.

Jiaxin Han is a postgraduate at Nanjing University of Chinese Medicine. His research interest is molecular docking and virtual screening. His affiliation is with School of Chinese Materia Medica, Nanjing University of Chinese Medicine, Nanjing, 210046, China

Professor Zhijian Xu got his Ph.D. degree at Shanghai Institute of Materia Medica in 2012. His research interests include computer-aided drug design, computational chemistry, computational biology and artificial intelligence. More information could be found at the website: https://www.researchgate.net/profile/Zhijian_Xu

Professor Weiliang Zhu received his Ph.D. degree from Shanghai Institute of Materia Medica in 1998. His main research fields are computer-aided drug design, computational biology, computational chemistry and pharmaceutical chemistry, with a special focus on the theoretical research and method development of drug design.