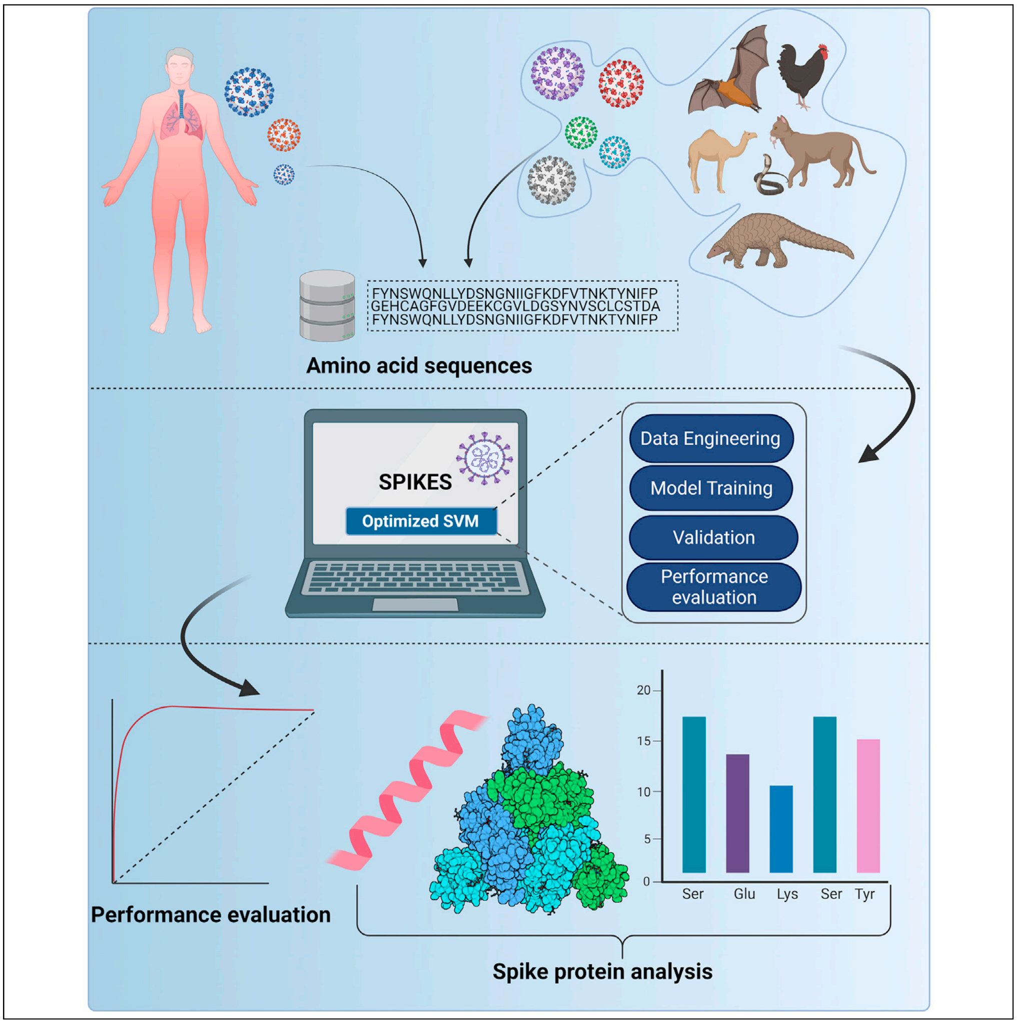
iScience

CellPress
OPEN ACCESS

Article

Tracking the amino acid changes of spike proteins across diverse host species of severe acute respiratory syndrome coronavirus 2



Srinivasulu Yerukala Sathipati, Sanjay K. Shukla, Shinn-Ying Ho

sathipathi.srinivasulu@marshfieldclinic.org

Highlights

Differences exist in the amino acids within the S protein of diverse host species CoVs

We developed SPIKES to identify informative properties of S protein

SARS-CoV-2 variants have amino acid changes that alter infection and transmission

The SPIKES identified changes in S protein properties from animal to human host CoVs

Yerukala Sathipati et al., iScience 25, 103560
January 21, 2022 © 2021 The Authors.
https://doi.org/10.1016/j.isci.2021.103560

CellPress
OPEN ACCESS

Article

Tracking the amino acid changes of spike proteins across diverse host species of severe acute respiratory syndrome coronavirus 2

Srinivasulu Yerukala Sathipati,[1,5,*] Sanjay K. Shukla,[1] and Shinn-Ying Ho[2,3,4]

SUMMARY

Knowledge of the host-specific properties of the spike protein is of crucial importance to understand the adaptability of severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) to infect multiple species and alter transmissibility, particularly in humans. Here, we propose a spike protein predictor SPIKES incorporating with an inheritable bi-objective combinatorial genetic algorithm to identify the biochemical properties of spike proteins and determine their specificity to human hosts. SPIKES identified 20 informative physicochemical properties of the spike protein, including information measures for alpha helix and relative mutability, and amino acid and dipeptide compositions, which have shown compositional difference at the amino acid sequence level between human and diverse animal coronaviruses. We suggest that alterations of these amino acids between human and animal coronaviruses may provide insights into the development and transmission of SARS-CoV-2 in human and other species and support the discovery of targeted antiviral therapies.

INTRODUCTION

The outbreak of SARS-CoV-2 in the city of Wuhan, China, in 2019 led to the coronavirus disease-2019 (COVID-19) pandemic causing millions of deaths worldwide. The World Health Organization has reported nearly 178 million confirmed cases and 3,864,180 deaths globally as of June 21, 2021 (World Health Organization, 2021). Despite the implementation of effective vaccines to control COVID-19, mutations in the SARS-CoV-2 genome, particularly in the spike (S) protein, led to emergence of variants against which the current vaccines may be partially effective. COVID-19 is the third outbreak of coronavirus (CoV) associated disease in the past 100 years; the first two epidemics were severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome (MERS), respectively. These CoVs are single-stranded positive-sense RNA viruses belonging to the *Coronaviridae* family (Gorbalenya et al., 2020) and infect both animals (bats, snakes, pangolin, etc.) and humans (Jones et al., 2008; Karesh et al., 2012; Cleaveland et al., 2001). SARS-CoV was transmitted to humans from exotic animals (Guan et al., 2003), whereas MERS was transmitted from dromedary camels (Sabir et al., 2016) with bats being the primary reservoir for both viruses (Cui et al., 2019; Perlman, 2020). The continued evolution and genetic diversity in variants of SARS-CoV have increased the possibility of crossing the species barrier and transmission of disease to humans (Li et al., 2005b). However, the steps involved in the natural selection and adaptability of SARS-CoV-2 to its human host from animals remains unclear. Thus, a comprehensive analysis of SARS-CoV 2 across diverse host species may help determine their diversity and transmissibility of COVID-19.

SARS-CoV-2 shares ~79% nucleotide sequence identity to SARS-CoV (Zhou et al., 2020) and is even more similar to several bat CoVs (Chen et al., 2020; Zhou et al., 2020). In contrast to the earlier SARS and MERS outbreaks, the SARS-CoV-2 outbreak has proven to be highly infectious (World Health Organization, 2021; Lu et al., 2015) and global in nature. Several molecular factors have contributed to increase the infectivity of SARS-CoV-2, particularly the recombination events in the S gene, which encodes the S glycoprotein (Hu et al., 2017). The S proteins of SARS-CoVs are critical for host cell recognition, entry, and infections (Shang et al., 2020). Increasing evidence has shown the role of S protein in binding with the host receptor and subsequent viral entry causing COVID-19 (Belouzard et al., 2012; Li, 2016). The S protein consists of two subunits S1, which is responsible for membrane binding (Millet and Whittaker, 2018), and S2 for membrane fusion (Li, 2012; Heald-Sargent and Gallagher, 2012; Wu et al., 2004). The S protein promotes entry into

[1]Center for Precision Medicine Research, Marshfield Clinic Research Institute, Marshfield, WI 54449, USA

[2]Institute of Bioinformatics and Systems Biology, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[3]Department of Biological Science and Technology, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[4]Center for intelligent Drug Systems and Smart Bio-Devices (IDS^{2}B), National Yang Ming Chiao Tung University, Hsinchu, Taiwan

[5]Lead contact

*Correspondence: sathipathi.srinivasulu@marshfieldclinic.org

https://doi.org/10.1016/j.isci.2021.103560

the host cell by binding to the receptor angiotensin-converting enzyme 2 (ACE2) (Walls et al., 2020), a cell surface enzyme with transcripts present in the lungs, heart, kidney, alveolar epithelial type 2 cells, and intestine (Donoghue et al., 2000; Zhang et al., 2020). The S proteins of SARS-CoV and SARS-CoV-2 have a high degree of homology and share ~76% of amino acid sequence similarity (Li et al., 2005a; Xu et al., 2020). Although the receptor-binding motif that binds to the ACE2 is similar to that in SARS-CoVs, SARS-CoV-2 recognizes the human ACE2 more efficiently than SARS-CoV, which helps to enhance its transmissibility (Wan et al., 2020). Various therapeutic strategies have been developed against the SARS-CoV-2. Choudhury et al. demonstrated a significant binding between the S protein and Toll-like receptors (TLRs), TL3, TL7, and TL9 TLR, that could serve as potential targets for SARS-CoV-2 (Choudhury et al., 2021; Choudhury and Mukherjee, 2020). These characteristics make the S protein a potential target for vaccine design and CoV disease treatments (Ge et al., 2021).

Analysis of the physicochemical properties (PCPs) and amino acid changes of the S protein across diverse species could reveal the evolutionary changes that contributed to increase the infectivity of SARS-CoV-2. Previously, potential roles of the PCPs in SARS-CoV-2 proteins, including molecular weight, thermal stability, and pH, have been examined to develop quality control measures for vaccines (Scheller et al., 2020). Hasan et al. compared the PCPs of protein ORF8 from SARS-CoV-2, pangolin-CoV, and bat-RaTG13-CoV and reported that some PCPs of SARS-CoV-2 ORF8 showed higher correlation with bat-RaTG13-CoV and less correlation with pangolin-CoV (Hassan et al., 2021). We have previously reported some informative PCPs, including normalized van der Waals volume, delta G values, normalized frequencies of turn in α/β class, normalized positional residues frequency at helix termini N″, and relative mutability, which were potential predictors for differentiating human and non-human host CoVs (Yerukala Sathipati and Ho, 2021). These changes in the properties of S protein is of particular value for the development and refinement of artificial intelligence (AI) and bioinformatics-based approaches to analyze and predict the emergence of SARS-CoV-2 variants, identify alterations in treatment response, and highlight areas of concern for increased transmissibility (Cave et al., 2021; Arora et al., 2020; Auwul et al., 2021; Brierley and Fowler, 2021).

Here, we propose an S protein predictor SPIKES to determine the host-species specificity of SARS-CoV-2 S proteins across human and diverse animal species, with the goal of providing comprehensive knowledge about the biochemical properties of the S protein at the amino acid sequence level to understand its functions in host cell infection. SPIKES was developed based on support vector machine to distinguish S proteins of CoVs across human and diverse animal species and uses an optimal feature selection algorithm, inheritable bi-objective combinatorial genetic algorithm (IBCGA) (Ho et al., 2004), to select informative properties from four major protein features including amino acid composition (AAC), dipeptide composition (DPC), pseudo amino acid composition (PseAAC), and PCPs. SPIKES was developed using information retrieved from the large-scale S protein sequences of human and animal host CoVs from the Global Initiative on Sharing All Influenza Data (GISAID) and the National Center for Biotechnology Information (NCBI) databases. The predictive ability of SPIKES was compared with those of seven machine learning algorithms, and the results showed its excellent discriminative ability. Identified PCPs and amino acid and dipeptide compositions were further analyzed to identify the critical changes that occurred in the S protein. Information obtained from SPIKES may be applicable for guiding the refinement of vaccine and therapeutic targets against emerging SARS-CoV-2 variants.

RESULTS

SPIKES prediction performance

SPIKES identifies the distinguishing features of the S proteins of human and animal-host coronaviruses. The Spike-training, Spike-test, and All-Spike datasets were used for the training and independent test evaluation. SPIKES produced four different SPIKES models, SPIKES-PCP, SPIKES-AAC, SPIKES-DPC, and SPIKES-PseAAC for the four feature descriptors of PCP, AAC, DPC, and PseAAC, respectively. The flowchart of SPIKES development and inputs are shown in Figure 1.

The performance of the four individual SPIKES models using the Spike-training and Spike-test datasets is given as follows. SPIKES-PCP had a 10-fold cross-validation (10-CV) accuracy, sensitivity, specificity, Matthews correlation coefficient (MCC), and area under the receiver operating curve (AUC) of 99.28%, 0.99, 0.99, 0.98, and 0.99, respectively, and obtained a test accuracy, sensitivity, specificity, MCC and AUCs of 94.44%, 0.95, 0.93, 0.88, and 0.98, respectively. SPIKES-AAC obtained a 10-CV accuracy, sensitivity, specificity, MCC, and AUCs of 97.63%, 0.97,0.97, 0.95, and 0.98, respectively; SPIKES-DPC obtained a10-CV

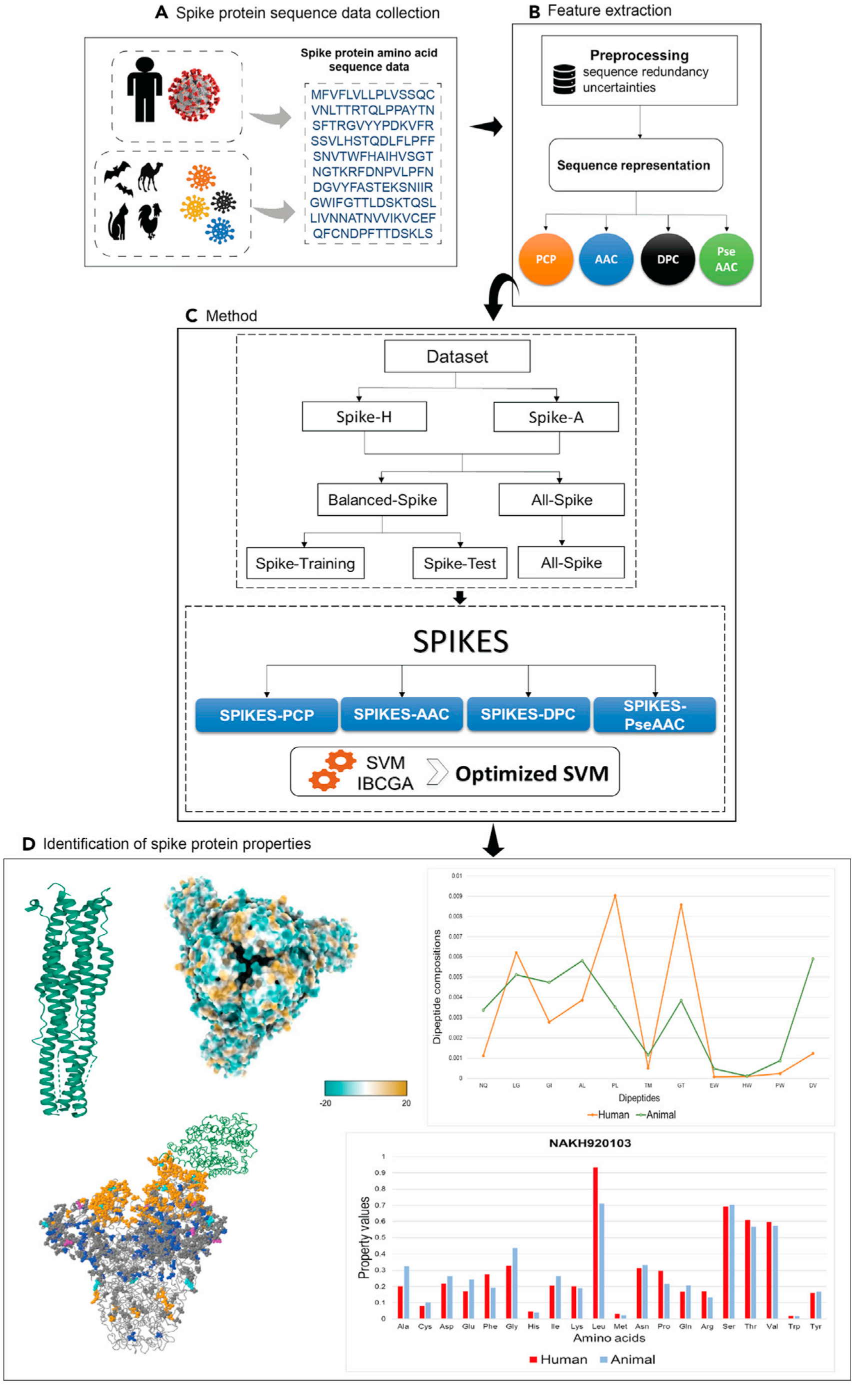


Figure 1. System overview of SPIKES

(A–D) (A) Collection of spike protein sequences, (B) preprocessing and feature extraction, (C) method in brief, and (D) the analysis of identified spike proteins between human and animal host coronaviruses.

accuracy, sensitivity, specificity, MCC, and AUCs of 99.05%, 0.99, 0.99, 0.98, and 0.99, respectively; and SPIKES-PseAAC obtained a 10-CV accuracy, sensitivity, specificity, MCC, and AUCs 98.81%, 0.98, 0.98, 0.97, and 0.99, respectively, as shown in Table 1.

We compared the four SPIKES models using the Spike-Balanced dataset with other machine learning methods, such as Naive Bayes (NB), Multilayer perceptron (MLP), Logistic Regression (LR), Sequential

Table 1. The prediction performance of SPIKES with some standard machine learning classifiers

Feature descriptor	Method	10-CV accuracy	Sensitivity	Specificity	MCC	AUC
PCP	Naive Bayes	88.62	0.98	0.82	0.78	0.95
	MLP	95.73	0.94	0.96	0.91	0.97
	Logistic-Regression	95.97	0.95	0.96	0.91	0.98
	SMO	92.18	0.96	0.88	0.84	0.92
	Simple Logistic	95.73	0.96	0.94	0.91	0.98
	J48	92.41	0.91	0.93	0.84	0.93
	Random Forest	96.2	0.97	0.96	0.93	0.99
	SPIKES-PCP	99.28	0.99	0.99	0.98	0.99
AAC	Naive Bayes	83.64	0.98	0.75	0.7	0.87
	MLP	93.6	0.93	0.93	0.87	0.97
	Logistic Regression	94.13	0.94	0.94	0.88	0.97
	SMO	88.86	0.98	0.82	0.79	0.88
	Simple Logistic	94.54	0.95	0.93	0.89	0.97
	J48	94.07	0.94	0.93	0.88	0.95
	Random Forest	96.2	0.96	0.95	0.92	0.99
	SPIKES-AAC	97.63	0.97	0.97	0.95	0.98
DPC	Naive Bayes	94.78	0.93	0.96	0.89	0.98
	MLP	95.73	0.94	0.96	0.91	0.96
	Logistic Regression	94.54	0.94	0.94	0.89	0.95
	SMO	92.89	0.93	0.92	0.85	0.92
	Simple Logistic	95.26	0.96	0.94	0.9	0.96
	J48	95.26	0.94	0.95	0.9	0.94
	Random Forest	95.73	0.94	0.97	0.91	0.98
	SPIKES-DPC	99.05	0.99	0.99	0.98	0.99
PseAAC	Naive Bayes	88.86	0.99	0.82	0.79	0.95
	MLP	95.49	0.96	0.94	0.91	0.98
	Logistic Regression	91.7	0.92	0.91	0.83	0.98
	SMO	93.12	0.98	0.88	0.86	0.93
	Simple Logistic	92.18	0.96	0.88	0.84	0.97
	J48	91.7	0.9	0.93	0.83	0.91
	Random Forest	96.20	0.97	0.95	0.92	0.98
	SPIKES-PseAAC	99.05	0.98	0.99	0.98	0.99

PCP, physicochemical property; AAC, amino acid composition; DPC, dipeptide composition; PseAAC, pseudo amino acid composition; AUC, area under the receiver operating curve; CV, cross-validation; MCC, Matthews correlation coefficient; MLP, multilayer perceptron; SMO, sequential minimal optimization.

Minimal Optimization (SMO), Simple Logistic (SL), J48 decision tree, and Random Forest. SPIKES-PCP performed better when compared with these machine learning methods with respect to 10-CV, sensitivity, specificity, MCC, and AUC. Although the ensemble classifier Random Forest using a large number of decision trees obtained an AUC of 0.99 comparable with that of SPIKES-PCP, the SPIKES outperformed it in terms of 10-CV, sensitivity, specificity, and MCC (Table 1). The other prediction models, SPIKES-AAC, SPIKES-DPC, and SPIKES-PseAAC, performed better when compared with these machine learning methods.

In addition, the prediction performance of SPIKES-PCP was compared with these machine learning methods using Spikes-all dataset. SPIKES-PCP obtained a 10-CV accuracy, sensitivity, specificity, MCC, and AUCs of 98.46%, 0.97, 0.98, 0.96, and 0.99, respectively, whereas well-known machine learning

Table 2. The prioritization of physicochemical properties

Rank	AAindex ID	Feature description	MED score
1	RACS820104	Average relative fractional occurrence in EL(i) (Rackovsky and Scheraga, 1982)	13.17
2	ROBB760101	Information measure for alpha-helix (Robson and Suzuki, 1976)	12.50
3	RACS820109	Average relative fractional occurrence in AL(i-1) (Rackovsky and Scheraga, 1982)	11.14
4	GEIM800105	Beta-strand indices (Geisow and Roberts, 1980)	9.79
5	QIAN880137	Weights for coil at the window position of 4 (Quian and Sejnowski, 1988)	7.09
6	PRAM820103	Correlation coefficient in regression analysis (Prabhakaran and Ponnuswamy, 1982)	6.41
7	JOND920102	Relative mutability (Jones et al., 1992)	5.74
8	NAKH920103	Amino acid composition of EXT of single-spanning proteins (Nakashima and Nishikawa, 1992)	4.39
9	OOBM850101	Optimized beta-structure-coil equilibrium constant (Oobatake et al., 1985)	3.04
10	CHAM830104	The number of atoms in the side chain labeled 2 + 1 (Charton and Charton, 1983)	2.36
11	ROBB760103	Information measure for middle helix (Robson and Suzuki, 1976)	0.33

AAindex ID, amino acid index identification; MED, main effect difference.

methods obtained 10-CV prediction accuracies, sensitivities, specificities, MCCs, and AUCs in the ranges of 92.30%–96.66%, 0.90–0.98, 0.90–0.97, 0.80–0.91, and 0.87–0.98, respectively. The predictive performance of SPIKES-PCP is better than these machine learning methods in terms of 10-CV, specificity, sensitivity, MCC, and AUC, whereas the sensitivity (0.98) of SMO is higher than the sensitivity (0.97) of SPIKES-PCP.

Furthermore, we used S protein of five variants of SARS-CoV-2, including Alpha (VOC 20212/01 GRY [B.1.1.7]), Beta (B.1.351, 20H/501Y.V2), Gamma (20J/501Y.V3[P.1, B.1.1.28.1]), Delta (B.1.617.2, 20A/452R), and Wuhan strain (hCoV/wuhan/WIV05/2019), as an independent validation set for SPIKES and seven machine learning methods previously mentioned. SPIKES achieved 100% (5/5) accuracy on distinguishing the variants as human CoVs. Among the seven machine learning methods, five of these, NB, MLP, LR, SMO, and Random Forest, obtained a test accuracy of 100% while distinguishing the variants as human CoVs, whereas two methods, J48 and SL, failed to distinguish the variants as human CoVs.

Prioritization of informative properties

SPIKES identified informative PCPs, AAC, DPC, and PseAAC that could distinguish the S proteins of humans and animal CoVs. SPIKES-PCP selected 11 PCPs that were potential descriptors of S proteins. We further prioritized the PCPs based on their predictive performance capability using main effect difference (MED) analysis. The MED scores and their ranks for the 11 PCPs, RACS820104, ROBB760101, RACS820109, GEIM800105, QIAN880137, PRAM820103, JOND920102, NAKH920103, OOBM850101, CHAM830104, and ROBB760103 are listed in Table 2.

PCPs of spike protein

SPIKES-PCP was employed to identify the biochemical and biophysical properties of S proteins across diverse species, as listed in Tables 2 and 3. The properties RACS820104 and RACS820109 were described as the "Average relative fractional occurrence in EL(i)" and "Average relative fractional occurrence in AL(i-1)," respectively. Rackovsky and Scherage examined the various structural features on the C^{α} length scale associated with some specific amino acids (Rackovsky and Scheraga, 1982). RACS820104 highlighted a group of amino acids consisting of Pro, Gly, His, Tyr, Cys, Asn, and Trp, which are responsible for

Table 3. The physicochemical properties based on feature frequency score

AAindex ID	Feature description	Frequency score
QIAN880129	Weights for coil at the window position of 4 (Quian and Sejnowski, 1988)	0.88
AURR980120	Normalized positional residue frequency at helix termini C4′ (Aurora and Rose, 1998)	0.86
NAKH920106	Amino acid composition of CYT of multi-spanning proteins (Nakashima and Nishikawa, 1992)	0.60
GEIM800105	Beta-strand indices (Geisow and Roberts, 1980)	0.56
YUTK870104	Activation Gibbs energy of unfolding, pH9.0 (Yutani et al., 1987)	0.42
LEVM780105	Normalized frequency of beta-sheet, unweighted (Levitt, 1978)	0.32
ROBB760103	Information measure for middle helix (Robson and Suzuki, 1976)	0.14
FAUJ880113	pK-a(RCOOH) (Fauchère et al., 1988)	0.12
JOND920102	Relative mutability (Jones et al., 1992)	0.12

AAindex ID, amino acid index identification.

nucleation of extended structures; this feature showed significant conformational behavior on C^{α} unit (Rackovsky and Scheraga, 1982). In contrast, RACS820109 described the amino acid conformational preferences at bends on the C^{α} unit. To examine the amino acid conformational preferences on the C^{α} unit of S proteins, we measured the properties for the S proteins of human and animal CoVs of RACS820104 and RACS820109. For RACS820104, slight differences were observed for amino acids Pro, Leu, Phe, Cys, Gly, Ala, and Ile in a range of 1%–2% between the S proteins of human and animal CoVs, whereas the largest differences were noticed for the amino acids Gly, Phe, Pro, and Asp with 11%, 5%, 3%, and 2%, respectively, between human and animal CoVs for RACS820109, as shown in Figure 2.

The property ROBB760101 was derived by Robson and Suzuki and is described as an ''information measure for alpha-helix'' (Robson and Suzuki, 1976). This measurement was based on an analysis of the conformational properties of amino acids in the alpha helix of 25 proteins using the information theory approach. The information measures for the alpha helical conformations shows that Glu and Ala are the strongest helix formers when compared with other amino acids. We calculated the conformational measurement for the alpha helices of S proteins using the property values of ROBB760101. Larger differences in the amino acids, Ala, Gly, Pro, Glu, and Leu, were observed for the alpha helix conformational preferences of S proteins, whereas no difference was observed for Val, His, Trp, and Asp between human and animal host CoVs. Previous research has shown that alpha helices are critical for SARS-CoV infection (Millet and Whittaker, 2018). In SARS-CoV, fusion domains are enriched in alpha helices and heptad repeats containing some hydrophobic residues involved in the membrane fusion process (Millet and Whittaker, 2018). The domain structure of S proteins is similar between SARS-CoV and SARS-CoV-2; however, the contact amino acid sites between SARS-CoV and human ACE-2 are different from those between SARS-CoV-2 and human ACE-2 (Walls et al., 2016, 2020). We performed protein sequence alignment of two helix fusion cores of S proteins, PDB: 6LXT (SARS-CoV-2) and PDB: 1WYY (SARS-CoV), using the SIM alignment tool (Huang and Miller, 1991), Expasy. The alignment analysis showed that both sequences shared 85.6% sequence similarity with an SIM alignment score of 536. The smaller changes in the amino acid core could affect the binding affinity of SARS-CoV-2 to its host receptor. The helix cores of SARS-CoV-2 and SARS-CoV are shown in Figure 3.

The property of GEIM800105 was also associated to the beta sheet protein secondary structure, described as ''Beta-strand indices'' by Geisow and Roberts (Geisow and Roberts, 1980). Evolution of the degree of individual preference of amino acids for a polypeptide were calculated using the Chou and Fasman method (Chou and Fasman, 1978) and revealed that conformational indices were not constant in α-helical, β, and α/β protein classes. We measured the conformational preferences for beta-strand indices using the property of GEIM800105 to

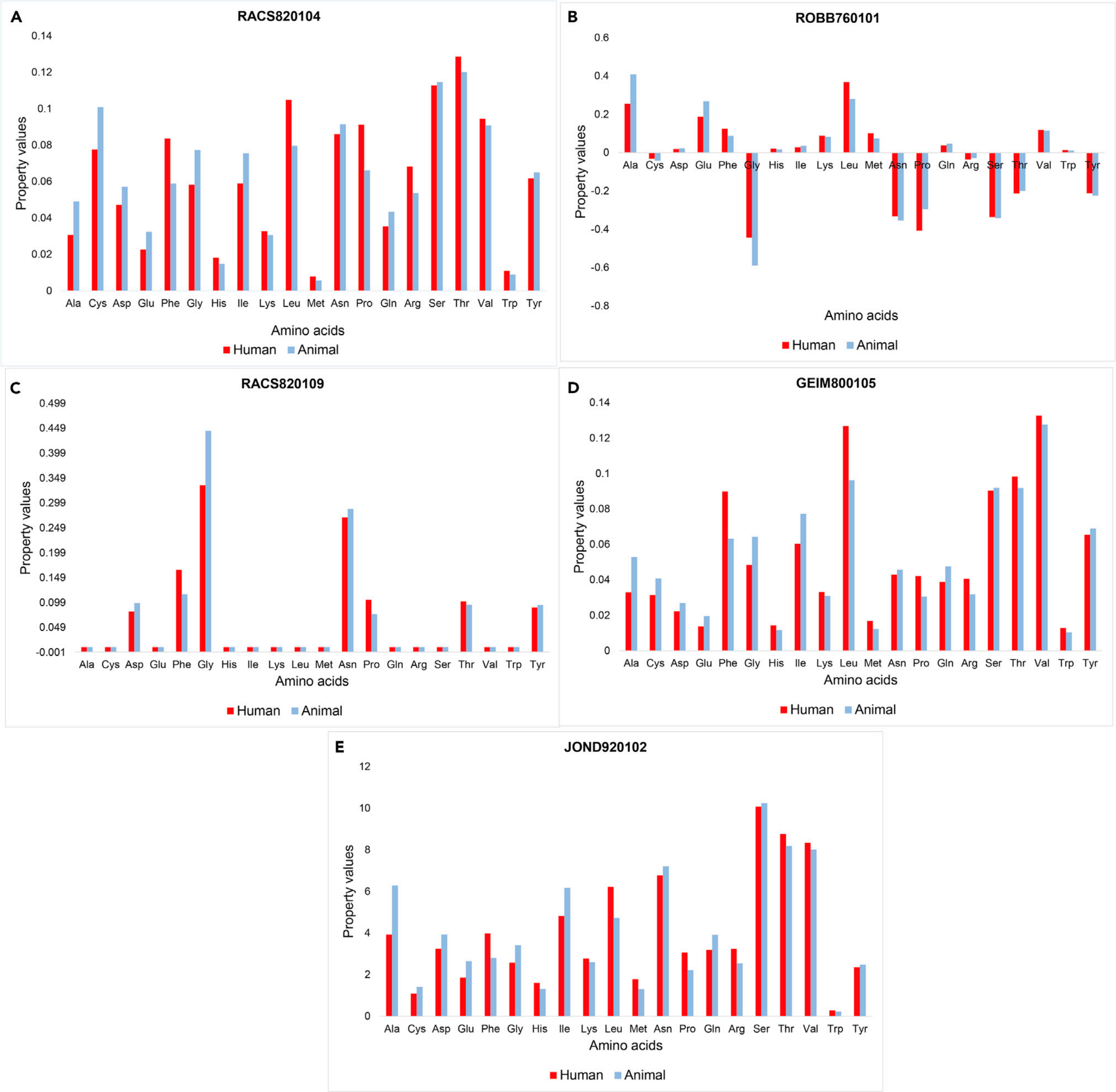


Figure 2. The comparison of physicochemical properties

The property comparison of spike protein between human and animal host coronaviruses (A) RACS820104, (B) ROBB760101, (C) RACS820109, (D) GEIM800105, and (E) JOND920102.

observe the difference between human and animal CoVs. A larger difference was observed for the amino acids, Leu, Phe, Ala, Ile, and Gly, between S proteins of human and animal CoVs.

The extracellular domain of S protein consists of S1 and S2 subunits, which mediate membrane binding and membrane fusion, respectively. We measured the differences in the identified 11 PCPs of S1 and S2 subunits of S protein. For each PCP, some amino acid compositions are significantly different between S1 and S2 subunits. For instance, in PCP1, amino acids Cys, Ile, Gln, Ala, and Thr showed a larger difference between S1 and S2 subunits, as shown in Figure S1. The significant amino acid compositional differences for the 11 PCPs between S1 and S2 domains are listed in Table S3.

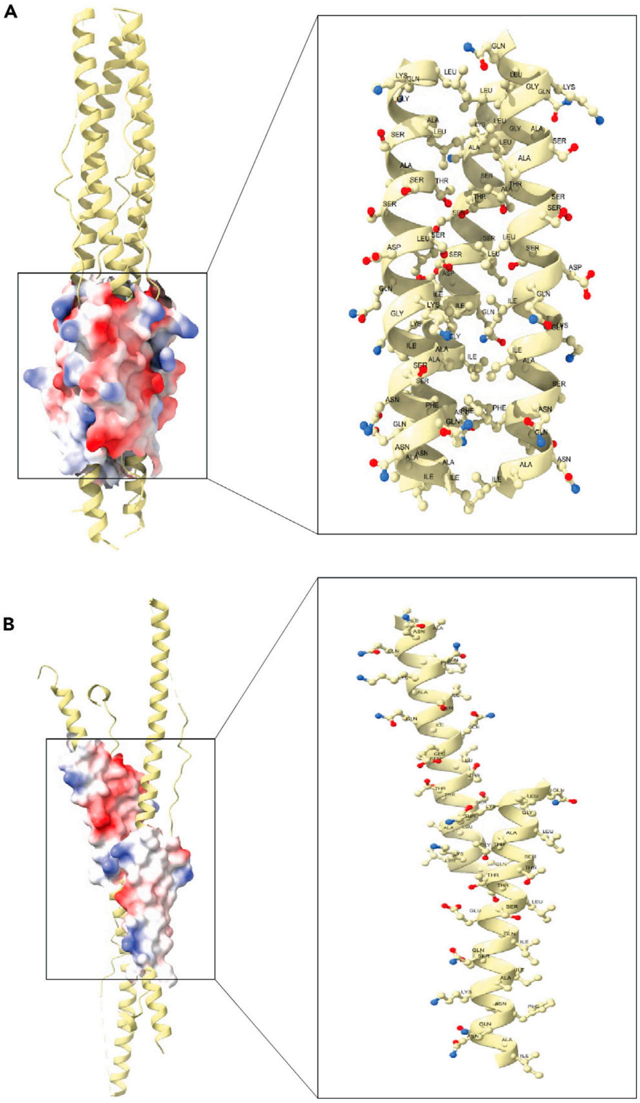


Figure 3. The helix core of spike protein

(A and B) (A) Structures of post-fusion core of 2019-nCoV S2 subunit (PDB: 6LXT) and (B) post-fusion hairpin conformation of the SARS-CoV spike glycoprotein (PDB: 1WYY). Close-up view of helix core (HR1 domain) from the helix bundle and arrangement of amino acids shown as ball-and-stick model.

Mutation analysis

An important property of JOND920102 is its degree of sequence differences among species, which is described as "Relative mutability" by Jones et al. (1992). Jones and colleagues generated mutation data matrices from a large number of protein sequences based on a mutation frequency matrix proposed by Dayhoff et al. (1978). The relative mutability of amino acids as calculated by Dayhoff et al. highlights the number of amino acid changes that occur in a given evolutionary interval. Genomic diversity and recurrent mutations might be the underlying mechanism for ongoing adaptation of SARS-CoV-2 to the human host (van Dorp et al., 2020). Several amino acid changes were noted between the receptor-binding domains of SARS-CoV and SARS-CoV-2 (Ortega et al., 2020). We calculated the relative mutability of S proteins between human and animal CoVs based on the information contained in the JOND920102 PCP. A larger difference was observed for the amino acids, Ala, Leu, Ile, Phe, Gly, Pro, Glu, Gln, Arg, and Asp, between the S proteins of human and animal host CoVs.

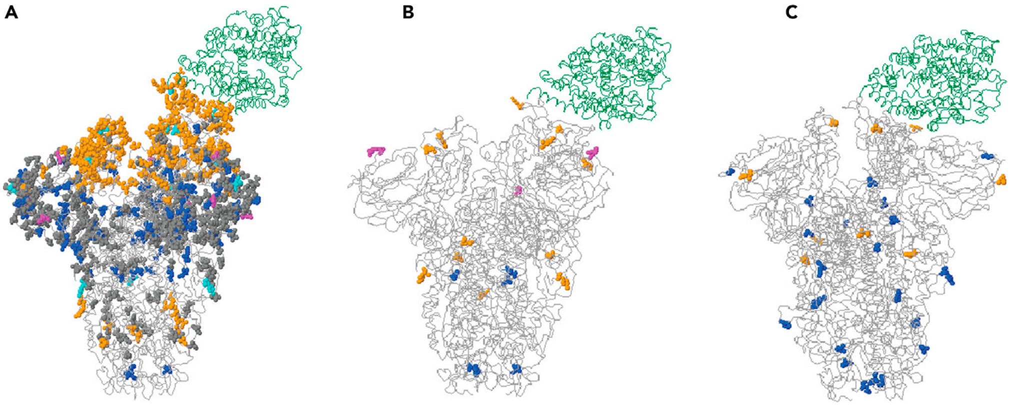


Figure 4. The comparison of mutations in spike protein across different variants of coronaviruses

Spike glycoprotein (PDB: 6ACJ) in complex with ACE2 (green ribbon) and amino acid changes occurred between (A) hCoV-19/bat/Yunnan/Prc3/2018 and hCoV/Wuhan/WIV05/2019, (B) Delta strain VOC G/452R.V3 (B.1.617+) and hCoV-19/Wuhan/WIV04/2019, and (C) Alpha strain VOC 20212/01 GRY (B.1.17) and hCoV-19/Wuhan/WIV04/2019. The mutations in different strains are shown in color balls.

Mutations in the S protein of different strains

Mutations in the S protein have led to the emergence of new strains with increased infectivity (Harvey et al., 2021). To determine the significance of compositional changes in amino acids in S protein from different strains of CoVs, we compared changes between the bat CoV strain, hCoV-19/bat/Yunnan/Prc3/2018 strain, and hCoV/wuhan/WIV05/2019 strain. The S protein of the hCoV-19/bat/Yunnan/Prc3/2018 was 82.7% identical to the S protein of the hCoV/wuhan/WIV05/2019 with 244 amino acid differences (Table S2). Next, the S protein of the hCoV/wuhan/WIV05/2019 was compared with the Delta variant, VOC G/452R.V3 [B.1.6.1.7+]. Seven amino acid differences were noted between these two variants. Furthermore, 12 amino acid differences were noted when hCoV-19/Wuhan/WIV04/2019 was compared with the S protein of the Alpha strain, VOC 20212/01 GRY (B.1.1.7) (Table S2). The comparison of mutations within the S proteins of different strains is shown in Figure 4.

Next, we examined the infectivity of new variants with amino acid changes in the S protein. The most commonly reported mutation in S protein of SARS-CoV-2 is an amino acid substitution from aspartic acid to glycine at the 614th position (D614G mutation) (Korber et al., 2020). As of now, 124 amino acid changes have been reported to occur in the S protein from SARS-CoV variants from at least 10 geographical locations (https://www.gisaid.org/). The specific amino acid changes in S protein have been implicated to increase the infectivity and virulence in new variants. We used GISAID data statistics to examine the amino acid changes in S protein that increased the infectivity in emerging new variants. The amino acid changes in Spike_T19R, Spike_E156G, and Spike_D950N increased the infectivity of the variant 452R-572K-681R, shown in Figure 5. The number and location of the amino acid changes of different variants within the S protein of SARS-CoV-2 are highlighted in Figures 5A and 5B.

Other important PCPs emerged from the SPIKES, specifically PRAM820103, NAKH920103, OOBM850101, CHAM830104, and ROBB760103, which appear to be better predictors for species-specific S proteins according to the MED analysis, as shown in Figure S2. The optimized SPIKES-PCP was generated after 50 independent runs of SPIKES-PCP. To explore the additional important PCPs beyond the eleven major properties, we measured the feature frequency scores across the 50 independent runs of SPIKES-PCP. There were nine PCPs that appeared frequently during the SPIKES-PCP optimization process. These nine features, including QIAN880129, AURR980120, NAKH920106, GEIM800105, YUTK870104, LEVM780105, and NAKH920103, have frequency scores in a range of 0.8–0.24. A detailed list of PCPs and feature frequency score are listed in Table 3.

Amino acid and dipeptide compositions

Although the overall structures of human and animal CoVs are similar, there are some key differences in the compositions of particular amino acids and dipeptides. To examine the compositional differences, amino acid compositions were compared between the S proteins of human and animal CoVs. A larger difference

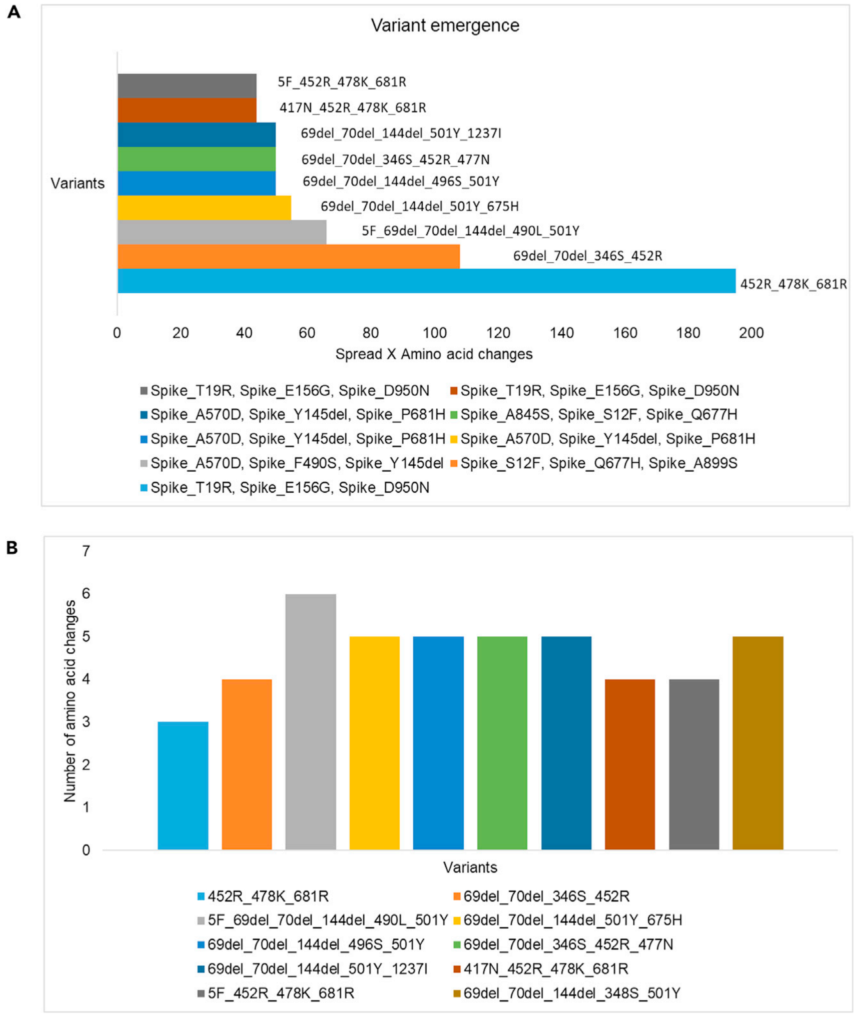


Figure 5. Recent variants emerging in SARS-CoV-2

(A) Top variants that are emerging with mutations in the spike protein.

(B) Number of amino acid changes that occurred in the variants over the 3 months (March, April, and May 2021).

in the amino acid compositions was observed for amino acids, including Leu, Ala, Gly, Phe, Arg, Asp, Cys, Glu, Gln, Ile, Met, Pro, and Thr, with a difference range of 1%–3%, whereas there was no difference noticed for the remaining amino acids, Asn, His, Lys, Ser, Trp, Tyr, and Val, between the S proteins of human and animal host CoVs, as shown in Figure 6A. The statistical significance of the amino acid compositions was analyzed using multiple testing corrections. We measured the false discovery rate (FDR) for the amino acid compositional differences between S proteins of human and animal host CoVs using the Benjamini, Krieger, and Yekutieli method (Benjamini et al., 2006). These amino acid compositional differences were significant between S proteins of human and animal host CoVs after FDR adjustment (FDR q-value <0.005). The FDR-adjusted q-values for the amino acid compositional differences are shown in Table S4.

A small change in these amino acid compositions may result in a notable change in structure and function (Bogatyreva et al., 2006; Tekaia and Yeramian, 2006). Previously, dipeptide compositions have been used to predict the infection risk of CoVs (Qiang et al., 2020). SPIKE-DPC identified some important dipeptides, NQ, LG, GI, AL, PL, TM, GT, EW, HW, PW, and DV, derived from dipeptide compositions that could accurately predict the S proteins of human and animal CoVs. We compared the dipeptides between the S proteins of human and animal CoVs and observed a larger difference for five dipeptides, PL, DV, GT, NQ, and GI, which showed more than a 20% difference between the S proteins of human and animal host CoVs, as shown in Figure 6B. Next, we measured the FDR for the dipeptide

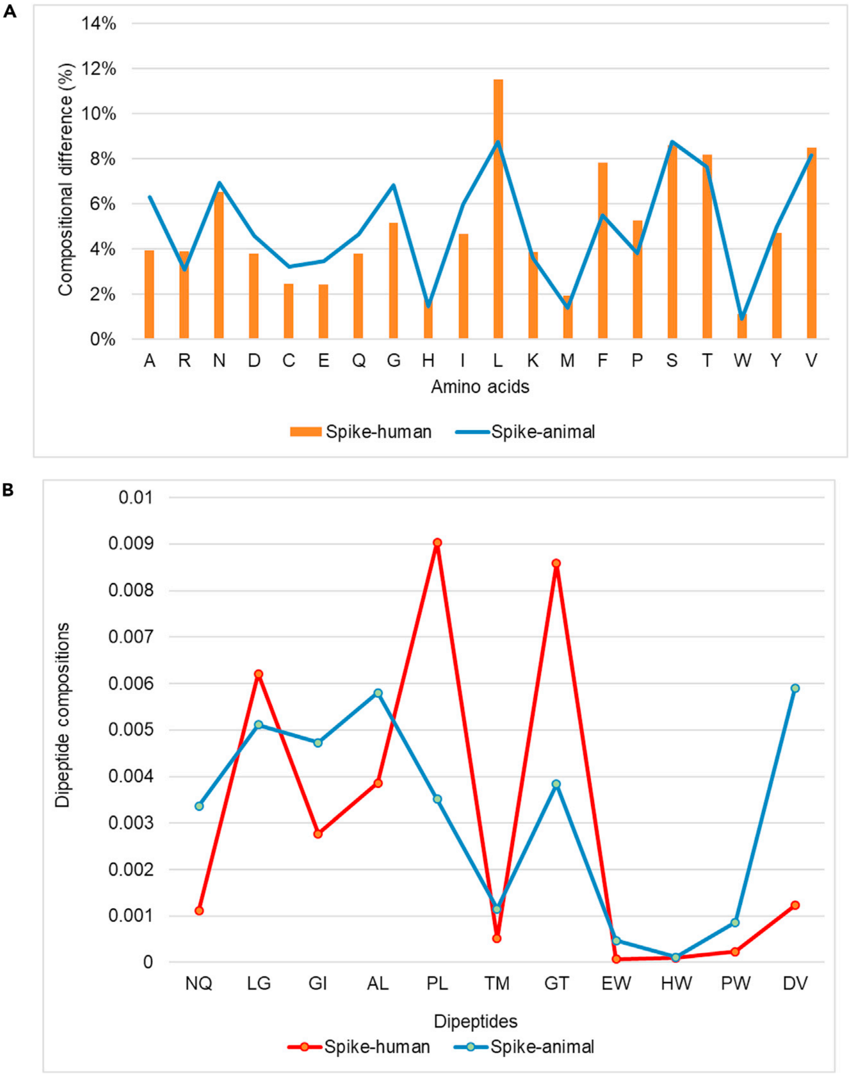


Figure 6. Compositional difference analysis

(A and B) (A) Amino acid compositional difference between spike proteins of human and animal corona viruses and (B) dipeptide compositional differences between spike proteins of human and animal coronaviruses.

compositions between S proteins of human and animal host CoVs. Among the 11 DPCs, except HW, the remaining showed significant differences between S proteins of human and animal host CoVs after FDR adjustment (FDR q-value <0.005). The FDR-adjusted q-values for the DPC compositions are shown in Table S5. These significant changes in the specific DPCs between human and animal host CoVs may have effect on S protein properties.

To further identify deviations in physical properties between human and animal CoVs, we compared some of the properties of S proteins, including molecular weight, number of charged residues, estimated half-life, stability, and aliphatic index using the structure of human (PDB: 6VXX_1) and animal host (GenBank: YP_009380521.1) S proteins. The S protein of human-host CoV has a molecular weight of 141,410.94, which is slightly larger than that of the animal host CoV at 126,219.51. The numbers of positively charged (Arg + Lys) and negatively charged (Asp + Glu) residues in the human S protein were 111 and 99, respectively, which were larger than those in the S protein of animal CoV with 101 and 66, respectively. In contrast, the aliphatic index in the S protein of human CoV was 83.32, which is slightly lower than that of animal CoV at 87.21. Despite these variations, the estimated half-life (30 h) of the two S proteins were similar and stable.

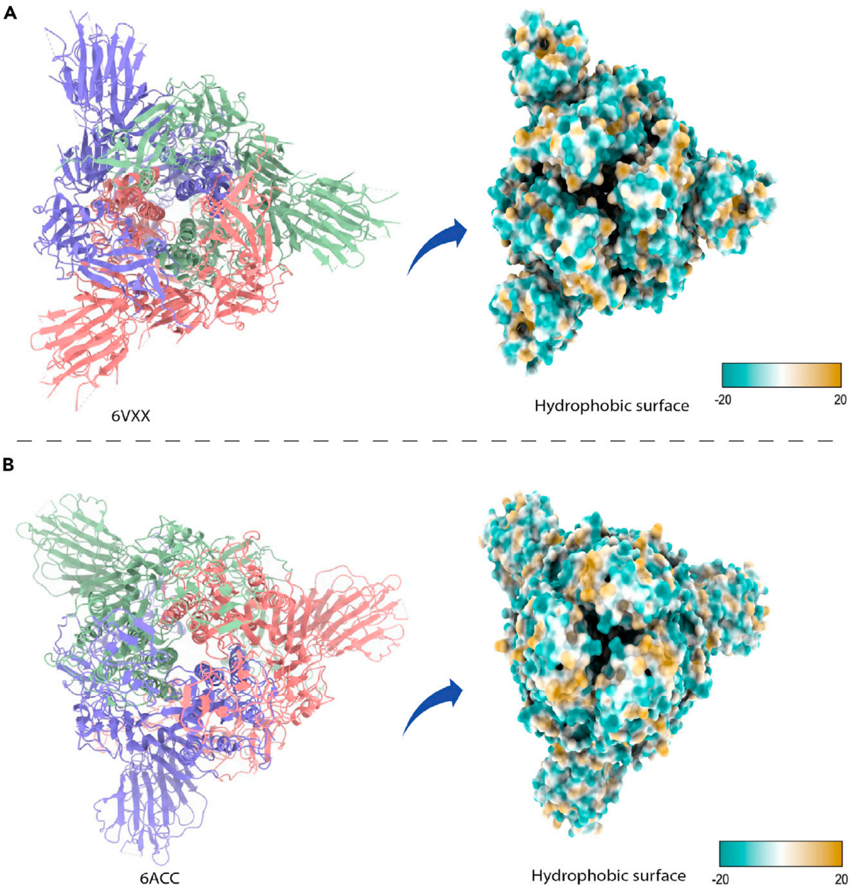


Figure 7. Surface hydrophobicity difference between PDB: 6VXX and 6ACC
Secondary structure and surface hydrophobicity of (A) PDB: 6VXX and (B) PDB: 6ACC, respectively.

Hydrophobicity of spike proteins

Previous research has highlighted the importance of hydrophobic regions in virus entry and cell fusion in CoVs (Chambers et al., 1990). The hydrophobic contact at the interfaces of the RBD-ACE2 complex contributes to the receptor binding affinity of the SARS-CoV-2 S protein (Wang et al., 2020b). We compared the hydrophobicity between the S proteins of human and animal-host CoVs using the hydrophobicity index proposed by Kyte and Doolittle (Kyte and Doolittle, 1982). We observed that the average hydrophobicity index for the hydrophobic amino acids of the S proteins in human host CoVs (0.18 ± 0.18) was slightly larger than those in animal CoVs (0.17 ± 0.14). Among the hydrophobic amino acids, a larger hydrophobic index difference was observed for the amino acids, Leu, Ile, Phe, Ala, Cys, and Pro, within the range of 11 to ±2%, between the S proteins of human and animal host CoVs, as shown in Figure 7.

DISCUSSION

Because of its fundamental role in infection, the S protein in SARS-CoV-2 is an important target for vaccine development and anti-viral therapies. To track the amino acid changes and properties of S proteins across diverse animal hosts, we proposed a prediction method called SPIKES. The advantage of SPIKES is 2-fold: one is to identify the properties of S protein of SARS-CoV-2 and the other is to determine its specificity across diverse species. SPIKES accurately distinguished S proteins of diverse species and identified some informative PCPs, AAC, DPC, and PseAAC, which possess diverse roles in SARS-CoVs. A comparison of the predictive performance of our method highlighted enhanced predictive ability of SPIKES to some well-known machine learning classifiers.

Analysis of informative properties associated with proteins revealed that the secondary structure properties found in PCPs ROBB760101, RACS820109, GEIM800105, and RACS820104 and relative mutability were important features of S proteins that showed differences between the S proteins of human and animal CoVs. Amino acid compositional changes were also observed for the remaining top-ranked PCPs, including QIAN880137, PRAM820103, NAKH920103, OOBM850101, CHAM830104, and ROBB760103, between the S proteins of human and animal CoVs. We also reported nine additional PCPs based on feature frequency scores. In our previous work (Yerukala Sathipati and Ho, 2021), we attempted to discover the properties of CoV proteins and identified some PCPs of interest based on differences in van der Waals volume (FAUJ880103), helix and beta turns properties (ONEK900101, PALJ810116, AURR980102, and MONM990101), and relative mutability (DAYM780201) that distinguish the human host from animal host species CoVs. In this study, we exclusively focused on S proteins. Our analyses showed that secondary structure properties and relative mutability were important PCPs that have the potential to determine and alter the species specificity of CoVs. The helix bundle in SARS-CoVs appears to play an essential role in protein fusion and virus entry. In the S proteins of SARS-CoV and MERS-CoV, heptad repeat 1 (HR1) and heptad repeat 2 (HR2) form a helix bundle that is essential for S protein fusion and entry into the host cell. A sequence alignment study by Xia and colleagues showed that the S2 subunits of SARS-CoV-2 and SARS-CoV S2 are highly conserved with more than 90% identity, whereas the HR1 core region showed nearly 38% difference due to mutations (Xia et al., 2020). The ROBB760101 property analysis revealed a larger difference in the compositions of amino acids, Ala, Gly, Pro, Glu, and Leu, at the alpha helix of S protein between human and animal CoVs. These amino acid differences may increase the affinity of S proteins to the ACE 2 receptor in SARS-CoV-2. In addition, hydrophobicity index analysis revealed that the larger difference in hydrophobic amino acids, including Leu, Phe, and Ile, was noticed in a range of 6%–11% between human and animal host CoVs. These amino acid changes in the S protein might influence the binding affinity to the S protein in the viral fusion process. Results from Li and colleagues indicate that hydrophobic interaction between SARS-CoV-2 and ACE2 is higher than those in SARS-CoV (Li et al., 2020), which supports our findings. We also observed significant differences in amino acids, Ala, Leu, Phe, Gly, Pro, Glu, Ile, and Gln, for the relative mutability of S proteins between human and animal host CoVs. Based on our analysis and previous evidences, these identified PCPs and amino acid compositional differences might affect the viral fusion in human host CoVs when compared with the animal host CoVs. However, further study is needed to validate the association between amino acid substitutions in the S protein and increased infectivity of SARS-CoV-2.

The unique properties of S protein of SARS-CoV-2, including higher binding affinity to its receptors (Wrapp et al., 2020; Yan et al., 2020); more amino acids at the interaction sites, which forms hydrogen bonds and van-der-Waal contacts; and antigenicity (Wang et al., 2020a; Lan et al., 2020), differ from those of other CoVs and facilitate receptor binding and membrane fusion between the virus and host to enhance disease transmission. Applications of AI and machine learning techniques may help predict amino acid changes that could lead to the emergence of more infective and transmissible SARS-CoV-2 variants. We developed a machine learning method SPIKES that identified changes at the amino acid and secondary structure levels that could determine the species specificity of S protein. Knowledge on PCPs will aid in developing appropriate quality control measures for vaccine designs (Scheller et al., 2020). The amino acid substitutions in S protein that interacts with ACE-2 receptor may play a key role in viral fusion and transmissibility of SARS-CoV-2 in humans. Owing to its crucial role in viral infection, S protein has therefore been selected as a potential target for S protein-based target therapies and vaccine development (Harvey et al., 2021). Hence, exploration of PCPs and amino acid changes in S proteins could be beneficial to understand the immune response and neutralize the antigenicity in SARS-CoV-2. Accordingly, this study aims to track the amino acid changes in S proteins from animal to humans that could facilitate the PCPs and amino acid compositional differences that may have affected the viral fusion process. We anticipate that the identified properties of these viruses will help in comprehensive understanding of S proteins and guide the implementation of S protein-based control measures.

Limitations of the study

This study has some limitations. First, this work was solely focused on amino acid sequence-based analysis of S protein of SRAS-CoV-2. However, the function of a protein could be influenced by other physiological factors in a host including stoichiometry of its complex with ACE2 receptor. Second, this study used data from GISAID until a certain period (May 2021) and therefore inclusion of data after the cutoff date could have further added to the significance of PCPs. Third, *in vivo*-based experimental validations could further strengthen the findings.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
    - ○ Lead contact
    - ○ Materials availability
    - ○ Data and code availability
- METHOD DETAILS
    - ○ Spike protein dataset
    - ○ Proposed SPIKES method
    - ○ Machine learning classifiers

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2021.103560.

ACKNOWLEDGMENT

This work was supported in part by the Marshfield Clinic Research Institute, Marshfield, WI. The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript. The authors acknowledge Emily Andreae, PhD, from the Marshfield Clinic Research Institute for manuscript editing assistance.

AUTHOR CONTRIBUTIONS

S.Y.S. and S.Y.H. designed the system and carried out the detailed study. S.Y.S. participated in the experimental design and implemented the experiments. S.K.S. and S.Y.S. participated in data analysis and interpretation. All authors participated in the manuscript preparation and approved the final manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no competing interests.

Received: June 28, 2021
Revised: November 2, 2021
Accepted: November 30, 2021
Published: January 21, 2022

REFERENCES

Arora, N., Banerjee, A.K., and Narasu, M.L. (2020). The role of artificial intelligence in tackling COVID-19. Future Virol. https://doi.org/10.2217/fvl-2020-0130.

Aurora, R., and Rose, G.D. (1998). Helix capping. Protein Sci *7*, 21–38. https://doi.org/10.1002/pro.5560070103.

Auwul, M.R., Rahman, M.R., Gov, E., Shahjaman, M., and Moni, M.A. (2021). Bioinformatics and machine learning approach identifies potential drug targets and pathways in COVID-19. Brief. Bioinform *22*, bbab120, https://doi.org/10.1093/bib/bbab120.

Belouzard, S., Millet, J.K., Licitra, B.N., and Whittaker, G.R. (2012). Mechanisms of coronavirus cell entry mediated by the viral spike protein. Viruses *4*, 1011–1033.

Benjamini, Y., Krieger, A.M., and Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. Biometrika *93*, 491–507.

Bogatyreva, N.S., Finkelstein, A.V., and Galzitskaya, O.V. (2006). Trend of amino acid composition of proteins of different taxa. J. Bioinform. Comput. Biol. *4*, 597–608.

Brierley, L., and Fowler, A. (2021). Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. PLoS Pathog. *17*, e1009149.

Cave, S., Whittlestone, J., Nyrup, R., O Heigeartaigh, S., and Calvo, R.A. (2021). Using AI ethically to tackle covid-19. BMJ *372*, n364.

Chambers, P., Pringle, C.R., and Easton, A.J. (1990). Heptad repeat sequences are located adjacent to hydrophobic regions in several types of virus fusion glycoproteins. J. Gen. Virol. *71*, 3075–3080.

Chang, C.-C., and Lin, C.-J. (2011). LIBSVM: a library for support vector machines. ACM Trans. Intell. Syst. Technol. *2*, 1–27.

Chen, Y.H., Yang, C.D., Tseng, C.P., Huang, H.D., and Ho, S.Y. (2015). GeNOSA: inferring and experimentally supporting quantitative gene regulatory networks in prokaryotes. Bioinformatics *31*, 2151–2158.

Charton, M., and Charton, B.I. (1983). The dependence of the Chou-Fasman parameters on amino acid side chain structure. J Theor. Biol *102*, 121–134.

Chen, Y., Liu, Q., and Guo, D. (2020). Emerging coronaviruses: genome structure, replication, and pathogenesis. J. Med. Virol. *92*, 418–423.

Chou, P.Y., and Fasman, G.D. (1978). Empirical predictions of protein conformation. Annu. Rev. Biochem. *47*, 251–276.

Choudhury, A., and Mukherjee, S. (2020). In silico studies on the comparative characterization of the interactions of SARS-CoV-2 spike glycoprotein with ACE-2 receptor homologs and human TLRs. J. Med. Virol. *92*, 2105–2113.

Choudhury, A., Das, N.C., Patra, R., and Mukherjee, S. (2021). In silico analyses on the comparative sensing of SARS-CoV-2 mRNA by the intracellular TLRs of humans. J. Med. Virol. *93*, 2476–2486.

Cleaveland, S., Laurenson, M.K., and Taylor, L.H. (2001). Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. Philos. Trans. R. Soc. Lond. B Biol. Sci. *356*, 991–999.

Cui, J., Li, F., and Shi, Z.-L. (2019). Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. *17*, 181–192.

Dayhoff, M., Schwartz, R., and Orcutt, B. (1978). 22 a model of evolutionary change in proteins. Atlas Protein Seq. Struct. *5*, 345–352.

Donoghue, M., Hsieh, F., Baronas, E., Godbout, K., Gosselin, M., Stagliano, N., Donovan, M., Woolf, B., Robison, K., Jeyaseelan, R., et al. (2000). A novel angiotensin-converting enzyme-related carboxypeptidase (ACE2) converts angiotensin I to angiotensin 1-9. Circ. Res. *87*, E1–E9.

Fauchère, J.L., Charton, M., Kier, L.B., Verloop, A., and Pliska, V. (1988). Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res *32*, 269–278. https://doi.org/10.1111/j.1399-3011.1988.tb01261.x.

Ge, J., Wang, R., Ju, B., Zhang, Q., Sun, J., Chen, P., Zhang, S., Tian, Y., Shan, S., Cheng, L., et al. (2021). Antibody neutralization of SARS-CoV-2 through ACE2 receptor mimicry. Nat. Commun. *12*, 250.

Geisow, M.J., and Roberts, R.D.B. (1980). Amino acid preferences for secondary structure vary with protein class. Int. J. Biol. Macromol. *2*, 387–389.

Gorbalenya, A.E., Baker, S.C., Baric, R.S., De Groot, R.J., Drosten, C., Gulyaeva, A.A., Haagmans, B.L., Lauber, C., Leontovich, A.M., et al. (2020). The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat. Microbiol. *5*, 536–544.

Guan, Y., Zheng, B.J., He, Y.Q., Liu, X.L., Zhuang, Z.X., Cheung, C.L., Luo, S.W., Li, P.H., Zhang, L.J., Guan, Y.J., et al. (2003). Isolation and characterization of viruses related to the SARS coronavirus from animals in southern China. Science *302*, 276–278.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I.H. (2009). The WEKA data mining software: an update. ACM SIGKDD Explor. Newsl. *11*, 10–18.

Harvey, W.T., Carabelli, A.M., Jackson, B., Gupta, R.K., Thomson, E.C., Harrison, E.M., Ludden, C., Reeve, R., Rambaut, A., Peacock, S.J., et al. (2021). SARS-CoV-2 variants, spike mutations and immune escape. Nat. Rev. Microbiol. *19*, 409–424.

Hassan, S.S., Aljabali, A.A.A., Panda, P.K., Ghosh, S., Attrish, D., Choudhury, P.P., Seyran, M., Pizzol, D., Adadi, P., Abd El-Aziz, T.M., et al. (2021). A unique view of SARS-CoV-2 through the lens of ORF8 protein. Comput. Biol. Med. *133*, 104380.

Heald-Sargent, T., and Gallagher, T. (2012). Ready, set, fuse! the coronavirus spike protein and acquisition of fusion competence. Viruses *4*, 557–580.

Ho, S.Y., Chen, J.H., and Huang, M.H. (2004). Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications. IEEE Trans. Syst. Man Cybern. B Cybern. *34*, 609–620.

Hu, B., Zeng, L.-P., Yang, X.-L., Ge, X.-Y., Zhang, W., Li, B., Xie, J.-Z., Shen, X.-R., Zhang, Y.-Z., Wang, N., et al. (2017). Discovery of a rich gene pool of bat SARS-related coronaviruses provides new insights into the origin of SARS coronavirus. PLoS Pathog. *13*, e1006698.

Huang, X., and Miller, W. (1991). A time-efficient, linear-space local similarity algorithm. Adv. Appl. Math. *12*, 337–357.

Jones, D.T., Taylor, W.R., and Thornton, J.M. (1992). The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. *8*, 275–282.

Jones, K.E., Patel, N.G., Levy, M.A., Storeygard, A., Balk, D., Gittleman, J.L., and Daszak, P. (2008). Global trends in emerging infectious diseases. Nature *451*, 990–993.

Karesh, W.B., Dobson, A., Lloyd-Smith, J.O., Lubroth, J., Dixon, M.A., Bennett, M., Aldrich, S., Harrington, T., Formenty, P., Loh, E.H., et al. (2012). Ecology of zoonoses: natural and unnatural histories. Lancet *380*, 1936–1945.

Kawashima, S., Pokarowski, P., Pokarowska, M., Kolinski, A., Katayama, T., and Kanehisa, M. (2008). AAindex: amino acid index database, progress report 2008. Nucleic Acids Res. *36*, D202–D205.

Korber, B., Fischer, W., Gnanakaran, S., Yoon, H., Theiler, J., Abfalterer, W., Hengartner, N., Giorgi, E., Bhattacharya, T., and Foley, B. (2020). Tracking changes in SARS-CoV-2 Spike: evidence that D614G increases infectivity of the COVID-19 virus. Cell *182*, 812–827.e19. https://doi.org/10.1016/j.cell.2020.06.043.

Kyte, J., and Doolittle, R.F. (1982). A simple method for displaying the hydropathic character of a protein. J. Mol. Biol. *157*, 105–132.

Lan, J., Ge, J., Yu, J., Shan, S., Zhou, H., Fan, S., Zhang, Q., Shi, X., Wang, Q., Zhang, L., and Wang, X. (2020). Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature *581*, 215–220.

Levitt, M. (1978). Conformational preferences of amino acids in globular proteins. Biochemistry *17*, 4277–4285. https://doi.org/10.1021/bi00613a026.

Li, F. (2012). Evidence for a common evolutionary origin of coronavirus spike protein receptor-binding subunits. J. Virol. *86*, 2856.

Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. Annu. Rev. Virol. *3*, 237–261.

Li, F., Li, W., Farzan, M., and Harrison, S.C. (2005a). Structure of SARS coronavirus spike receptor-binding domain complexed with receptor. Science *309*, 1864.

Li, W., Shi, Z., Yu, M., Ren, W., Smith, C., Epstein, J.H., Wang, H., Crameri, G., Hu, Z., Zhang, H., et al. (2005b). Bats are natural reservoirs of SARS-like coronaviruses. Science *310*, 676.

Li, J., Ma, X., Guo, S., Hou, C., Shi, L., Zhang, H., Zheng, B., Liao, C., Yang, L., Ye, L., and He, X. (2020). A hydrophobic-interaction-based mechanism triggers docking between the SARS-CoV-2 spike and angiotensin-converting enzyme 2. Glob. Challenges *4*, 2000067.

Lu, G., Wang, Q., and Gao, G.F. (2015). Bat-to-human: spike features determining 'host jump' of coronaviruses SARS-CoV, MERS-CoV, and beyond. Trends Microbiol. *23*, 468–478.

Millet, J.K., and Whittaker, G.R. (2018). Physiological and molecular triggers for SARS-CoV membrane fusion and entry into host cells. Virology *517*, 3–8.

Nakashima, H., and Nishikawa, K. (1992). The amino acid composition is different between the cytoplasmic and extracellular sides in membrane proteins. FEBS Lett. *303*. https://doi.org/10.1016/0014-5793(92)80506-c.

Oobatake, M., Kubota, Y., and Ooi, T. (1985). Optimization of amino acid parameters for correspondence of sequence to tertiary structures of proteins (commemoration issue dedicated to Professor Eiichi Fujita on the occasion of his retirement). Bull. Inst. Chem. Res. Kyoto Univ. *63*, 82–94.

Ortega, J.T., Serrano, M.L., Pujol, F.H., and Rangel, H.R. (2020). Role of changes in SARS-CoV-2 spike protein in the interaction with the human ACE2 receptor: an in silico analysis. EXCLI J. *19*, 410–417.

Perlman, S. (2020). Another decade, another coronavirus. N. Engl. J. Med. *382*, 760–762.

Prabhakaran, M., and Ponnuswamy, P.K. (1982). Shape and surface features of globular proteins. Macromolecules *15*, 314–320. https://doi.org/10.1016/0022-2836(88)90564-5.

Qiang, X.-L., Xu, P., Fang, G., Liu, W.-B., and Kou, Z. (2020). Using the spike protein feature to predict infection risk and monitor the evolutionary dynamic of coronavirus. Infect. Dis. Poverty *9*, 33.

Quian, N., and Sejnowski, T.J. (1988). Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. *202*, 865–884. ISSN 0022-2836. https://doi.org/10.1016/0022-2836(88)90564-5.

Rackovsky, S., and Scheraga, H.A. (1982). Differential geometry and polymer conformation. 4. Conformational and nucleation properties of individual amino acids. Macromolecules *15*, 1340–1346.

Robson, B., and Suzuki, E. (1976). Conformational properties of amino acid residues in globular proteins. J. Mol. Biol. *107*, 327–356.

Sabir, J.S., Lam, T.T., Ahmed, M.M., Li, L., Shen, Y., Abo-Aba, S.E., Qureshi, M.I., Abu-Zeid, M., Zhang, Y., Khiyami, M.A., et al. (2016). Co-circulation of three camel coronavirus species and recombination of MERS-CoVs in Saudi Arabia. Science *351*, 81–84.

Sathipati, S.Y., and Ho, S.-Y. (2021). Identification of the miRNA signature associated with survival in patients with ovarian cancer. Aging *13*, 12660–12690.

Scheller, C., Krebs, F., Minkner, R., Astner, I., Gil-Moles, M., and Wätzig, H. (2020). Physicochemical properties of SARS-CoV-2 for drug targeting, virus inactivation and attenuation, vaccine formulation and quality control. Electrophoresis *41*, 1137–1151.

Shang, J., Wan, Y., Luo, C., Ye, G., Geng, Q., Auerbach, A., and Li, F. (2020). Cell entry mechanisms of SARS-CoV-2. Proc. Natl. Acad. Sci. U S A *117*, 11727.

Srinivasulu, Y.S., Wang, J.-R., Hsu, K.-T., Tsai, M.-J., Charoenkwan, P., Huang, W.-L., Huang, H.-L., and Ho, S.-Y. (2015). Characterizing informative sequence descriptors and predicting binding affinities of heterodimeric protein complexes. BMC Bioinformatics *16*, S14.

Tekaia, F., and Yeramian, E. (2006). Evolution of proteomes: fundamental signatures and global trends in amino acid compositions. BMC Genomics *7*, 307.

Tsai, M.-J., Wang, J.-R., Ho, S.-J., Shu, L.-S., Huang, W.-L., and Ho, S.-Y. (2020). GREMA: modelling of emulated gene regulatory networks with confidence levels based on evolutionary intelligence to cope with the underdetermined problem. Bioinformatics *36*, 3833–3840.

van Dorp, L., Acman, M., Richard, D., Shaw, L.P., Ford, C.E., Ormond, L., Owen, C.J., Pang, J., Tan, C.C.S., Boshier, F.A.T., et al. (2020). Emergence of genomic diversity and recurrent mutations in SARS-CoV-2. Infect. Genet. Evol. *83*, 104351.

Walls, A.C., Tortorici, M.A., Bosch, B.-J., Frenz, B., Rottier, P.J.M., Dimaio, F., Rey, F.A., and Veesler, D. (2016). Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. Nature *531*, 114–117.

Walls, A.C., Park, Y.J., Tortorici, M.A., Wall, A., Mcguire, A.T., and Veesler, D. (2020). Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell *181*, 281–292.e6.

Wan, Y., Shang, J., Graham, R., Baric, R.S., and Li, F. (2020). Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of SARS coronavirus. J. Virol. *94*, e00127–20.

Wang, Q., Zhang, Y., Wu, L., Niu, S., Song, C., Zhang, Z., Lu, G., Qiao, C., Hu, Y., Yuen, K.-Y., et al. (2020a). Structural and functional basis of SARS-CoV-2 entry by using human ACE2. Cell *181*, 894–904.e9.

Wang, Y., Liu, M., and Gao, J. (2020b). Enhanced receptor binding of SARS-CoV-2 through networks of hydrogen-bonding and hydrophobic interactions. Proc. Natl. Acad. Sci. U S A *117*, 13967.

World Health Organization. https://covid19.who.int/.

Wrapp, D., Wang, N., Corbett Kizzmekia, S., Goldsmith Jory, A., Hsieh, C.-L., Abiona, O., Graham Barney, S., and Mclellan Jason, S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science *367*, 1260–1263.

Wu, X.D., Shang, B., Yang, R.F., Hao, Y., Hai, Z., Xu, S., Ji, Y.Y., Ying, L., Di Wu, Y., and Lin, G.M. (2004). The spike protein of severe acute respiratory syndrome (SARS) is cleaved in virus infected Vero-E6 cells. Cell Res. *14*, 400–406.

Xia, S., Zhu, Y., Liu, M., Lan, Q., Xu, W., Wu, Y., Ying, T., Liu, S., Shi, Z., and Jiang, S. (2020). Fusion mechanism of 2019-nCoV and fusion inhibitors targeting HR1 domain in spike protein. Cell Mol. Immunol. *17*, 1–3.

Xu, X., Chen, P., Wang, J., Feng, J., Zhou, H., Li, X., Zhong, W., and Hao, P. (2020). Evolution of the novel coronavirus from the ongoing Wuhan outbreak and modeling of its spike protein for risk of human transmission. Sci. China Life Sci. *63*, 457–460.

Yan, R., Zhang, Y., Li, Y., Xia, L., Guo, Y., and Zhou, Q. (2020). Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science *367*, 1444–1448.

Yerukala Sathipati, S., and Ho, S.-Y. (2017). Identifying the miRNA signature associated with survival time in patients with lung adenocarcinoma using miRNA expression profiles. Sci. Rep. *7*, 7507.

Yerukala Sathipati, S., and Ho, S.-Y. (2018). Identifying a miRNA signature for predicting the stage of breast cancer. Sci. Rep. *8*, 16138.

Yerukala Sathipati, S., and Ho, S.-Y. (2020). Novel miRNA signature for predicting the stage of hepatocellular carcinoma. Sci. Rep. *10*, 14452.

Yerukala Sathipati, S., and Ho, S.-Y. (2021). Identification and characterization of species-specific severe acute respiratory syndrome coronavirus 2 physicochemical properties. J. Proteome Res. *20*, 2942–2952.

Yerukala Sathipati, S., Sahu, D., Huang, H.-C., Lin, Y., and Ho, S.-Y. (2019). Identification and characterization of the lncRNA signature associated with overall survival in patients with neuroblastoma. Sci. Rep. *9*, 5125.

Yutani, K., Ogasahara, K., Tsujita, T., and Sugino, Y. (1987). Dependence of conformational stability on hydrophobicity of the amino acid residue in a series of variant proteins substituted at a unique position of tryptophan synthase alpha subunit. Proc Natl Acad Sci USA *84*, 4441–4444. https://doi.org/10.1073/pnas.84.13.4441.

Zhang, H., Penninger, J.M., Li, Y., Zhong, N., and Slutsky, A.S. (2020). Angiotensin-converting enzyme 2 (ACE2) as a SARS-CoV-2 receptor: molecular mechanisms and potential therapeutic target. Intensive Care Med. *46*, 586–590.

Zhou, P., Yang, X.-L., Wang, X.-G., Hu, B., Zhang, L., Zhang, W., Si, H.-R., Zhu, Y., Li, B., Huang, C.-L., et al. (2020). A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature *579*, 270–273.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
Spike protein sequences	Global Initiative on Sharing Avian Influenza Data	GISAID - Initiative https://www.gisaid.org/
Spike protein sequences	National Center for Biotechnology Information	National Center for Biotechnology Information (nih.gov)
Software and algorithms		
Support vector machine	Chang and Lin, 2011	LIBSVM – A Library for Support Vector Machines (ntu.edu.tw)
Protein structure visualization	UCSF Chimera	https://www.rbvi.ucsf.edu/chimera/
Inheritable bi-objective combinatorial genetic algorithm	Ho et al., 2004	Inheritable genetic algorithm for biobjective 0/1 combinatorial optimization problems and its applications - PubMed (nih.gov)
Machine learning methods	Waikato Environment for Knowledge Analysis	https://www.cs.waikato.ac.nz/ml/index.html

RESOURCE AVAILABILITY

Lead contact

Further information and request for resources and reagents should be directed to and will be fulfilled by the lead contact, Srinivasulu Yerukala Sathipati (sathipathi.srinivasulu@marshfiledclinic.org).

Materials availability

This study did not generate new unique reagents.

Data and code availability

The paper does not report original data.

Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

METHOD DETAILS

Spike protein dataset

We retrieved the S protein sequences of human-host CoVs (spike-H) and animal-host CoVs (Spike-A) for 96 diverse host species of the *Coronaviridae* family (as shown in Table S1) from GISAID and NCBI databases, respectively. Initial dataset consisted of 827,075 and 2,095 S protein sequences of spike-H and spike-A, respectively. Since the amino acid change is a crucial factor for disease transmission but to reduce the ambiguity, we considered 99% sequence identity in the redundancy reduction process. After sequence redundancy reduction and accounting for uncertainties, the final dataset, called All-Spike, consisted of 211 and 611 S protein sequences of spike-H and spike-A, respectively. A balanced dataset, called Balanced-Spike, consisted of the 211 sequences of spike-H and 211 sequences of spike-A, which were randomly chosen. To evaluate the predictive performance of SPIKES on an independent test, the Balanced-Spike was divided into a training (Spike-training) and test (Spike-test) in a 7:3 ratios.

Proposed SPIKES method

The SPIKES method was developed using SVM incorporating an optimal feature selection algorithm, IBCGA, to select *m* informative features from a large number *n* of candidate features. The optimized SVMs incorporated with IBCGA is well-suited for solving various biological modeling problems, such as cancer survival and stage predictions (Yerukala Sathipati and Ho, 2017, 2018, 2020; Yerukala Sathipati et al., 2019; Sathipati and Ho, 2021), protein function predictions (Srinivasulu et al., 2015; Yerukala Sathipati and Ho, 2021), and modelling gene regulatory networks (Tsai et al., 2020; Chen et al., 2015). IBCGA uses an intelligent evolutionary algorithm (IEA) to select a small set of informative features while optimizing predictive performance. In IBCGA, the genetic algorithm (GA) terms such as 'gene' and 'chromosome' were placed with 'GA-gene' and 'GA-chromosome' for distinction. The GA-chromosome consists of binary GA-genes for selecting *m* informative features of PCPs, dipeptides, AAC, and PseAAC, and two 4-bit GA-genes for encoding the parameters C and γ of SVM. IBCGA can simultaneously obtain a set of solutions, *Xr*, where $r=r_{end}$, $r_{end\,+}$ 1,…, r_{start} in a single run. In SPIKES, the radial basis function (RBF) kernel was used for the implementation of SVM (Chang and Lin, 2011). The scoring function of the RBF kernel is computed in the feature space between the two data points, x_i and y_j. The RBF kernel function is defined as follows:

$$K\left(x_i, y_j\right) = \exp\left(-\gamma\|x_i - y_j\|\right)^2 \qquad \text{(Equation 1)}$$

The feature process in the SPIKES method can be described in two parts, (1) feature representation and (2) feature selection, described as follows:

(1) Feature representation

SPIKES used four feature descriptors including physicochemical property (PCP), amino acid composition (AAC), and dipeptide composition (DPC), and pseudo amino acid composition (PseAAC).

(a) PCP representation

SPIKES adopted 531 PCPs retrieved from the AAindex database (Kawashima et al., 2008) as candidate features to distinguish S proteins of diverse species CoVs. The original CoVs' amino acid sequences were converted into AAindex numerical indices according to the 531 PCP values. The feature representation of the 531 PCPs is described as follows:

a) Collect the spike-H and spike-A protein sequences from the dataset.
b) Calculate the amino acid composition $f(aa_i)$ of a sequence for the i^{th} amino acid aa_i of 20 amino acids and encode the protein sequence of variable length into the feature vector with a length of 531 properties.
c) Calculate the feature value of the p^{th} physicochemical property, PCP(p), of a spike protein, where p=1, 2, …, 531.

$$PCP(p) = \sum_{i=1}^{20} f(aa_i).PCP_p(aa_i) \qquad \text{(Equation 2)}$$

where $PCP_p(aa_i)$ is the value of the aa_i amino acid of the p^{th} physicochemical property.

(b) AAC, DPC and PseAAC representation

The values of *f(aai)* were calculated for the spikes-H and spikes-A where i=1,…, 20. The feature set of DPC is represented as a feature vector of length 400 for the dipeptides, (i.e., AA, AC.… YY). The feature set of PseAAC is represented as a feature vector of length 80 for the AAC and PseAAC for hydrophilicity and hydrophobicity.

(2) Feature selection

Step 1: (Data Preparation) Compile the training sets from the spike-H and spike-A for developing and evaluating the SPIKES method, which is a combination of four predictive models, SPIKES-PCP, SPIKES-AAC, SPIKES-DPC, and SPIKES-PseAAC.

Step 2: (Initialization) Randomly generate an initial population of *Npop* individuals. *Npop* = 50, r_{start} = 50, r_{end} = 10, and $r = r_{start}$.

Step 3: (Evaluation) Evaluate the fitness value of all individuals using the fitness function that is the prediction accuracy in terms of 10-fold cross-validation (10-CV).

Step 4: (Selection) Use a conventional tournament selection method that selects the winner from two randomly selected individuals to generate a mating pool.

Step 5: (Crossover) Select two parents from the mating pool to perform an orthogonal array crossover operation of IEA.

Step 6: (Mutation) Apply a conventional bit mutation operator to GA-genes of SVM parameters and a swap mutation to the binary GA-genes for keeping *r* selected features. The best individual was not mutated for the elite strategy.

Step 7: (Termination test) If the stopping condition for obtaining the solution *Xr* is satisfied, output the best individual as the solution *Xr*. Otherwise, go to Step 2.

Step 8: (Inheritance) If $r > r_{end,}$ randomly change one bit from 1 to 10 in the binary genes for each individual. Decrease the number *r* by one and go to Step 2. Otherwise, stop the algorithm.

Step 9: (Output) Obtain a set of *m* features in total for PCPs, AAC, DPC, and PseAAC from the best solution *Xm* among the solutions *Xr*, where $r = r_{end}, r_{end} + 1, \ldots, r_{start}$.

Machine learning classifiers

The predictive performance of SPIKES was evaluated by comparison with seven popular machine learning methods such as Naïve Bayes (NB), Multilayer perceptron (MLP), Logistic regression (LR), Sequential minimal optimization (SMO), Simple logistic (SL), J48 decision tree, and Random forest. The Weka data mining software (Hall et al., 2009) was used for implementing the machine learning methods to distinguish Spike-H and Spike-A. The subset evaluator and the best first search feature selection was employed to design classifiers for feature descriptors PCP, AAC, DPC, and PseAAC.