Received: 19 August 2021 | Revised: 19 October 2021 | Accepted: 25 October 2021

DOI: 10.1002/prot.26274

RESEARCH ARTICLE

PROTEINS WILEY

In silico prediction of Severe Acute Respiratory Syndrome Coronavirus 2 main protease cleavage sites

Zheng Rong Yang

School of Biosciences, University of Exeter, Exeter, UK

Correspondence
Zheng Rong Yang, School of Biosciences, University of Exeter, Exeter, UK.
Email: ron.zheng.rong.yang@gmail.com; z.r.yang@exeter.ac.uk

Abstract

One of the emerging subjects to combat the SARS-CoV-2 virus is to design accurate and efficient drug such as inhibitors against the viral protease to stop the viral spread. In addition to laboratory investigation of the viral protease, which is fundamental, the in silico research of viral protease such as the protease cleavage site prediction is critically important and urgent. However, this problem has yet to be addressed. This article has, for the first time, investigated this problem using the pattern recognition approaches. The article has shown that the pattern recognition approaches incorporating a specially tailored kernel function for dealing with amino acids has the outstanding performance in the accuracy of cleavage site prediction and the discovery of the prototype cleavage peptides.

KEYWORDS
kernel function, machine learning, pattern recognition, SARS-CoV-2 main protease, viral cleavage site

1 | INTRODUCTION

SARS-CoV-2 is a single-stranded RNA genome and belongs to the coronavirus family and is composed of 23 ORFs.[1] Among them, ORF1a and ORF1b are translated to two polyproteins,[2] which can be cleaved by the viral proteases to generate 16 nonstructural proteolytic proteins.[3] The cleavage in these ORFs is mainly carried out by the chymotrypsin like 3CL cysteine protease (main protease). The papain-like protease (PLpro) carries out three cleavages.[2] A protease works when it interacts with a specific site in the amino acid sequence of a polyprotein. The site at which a polyprotein is cleaved by a protease is called a protease cleavage site. A collection of the consecutive residues surrounding a protease cleavage site is called a cleaved peptide expressed as $P_m \cdots P_2 P_1 \downarrow P'_1 P'_2 \cdots P'_n$. In this expression, $\downarrow$ stands for the cleavage, $P_{1 \le i \le m}$ stands for a N-terminal residue and $P'_{1 \le j \le n}$ stands for a C-terminal residue. The coronavirus main protease cleavage always happens at the amino acid glutamine in a polyprotein, that is, $P_1 = Q$. However, the cleavage of a protease on subsites allows certain variation of the amino acid distribution, that is, the tolerances.[4,5] Only when the fitness between a protease and a substrate is satisfied, the protease will bind to a polyprotein at the substrate to cleave the polyprotein. This is called the lock-and-key mechanism.[6,7] The glutamines whose substrates do not fit the structure requirement for a protease will not be targeted by the protease for the cleavage. Laboratory identification of the protease cleavage sites within a polyprotein is the best technology, but it is more expensive and time-consuming.[8,9] It is better to use an efficient in silico approach to screen out a subset of the glutamine residues which are the most probable protease cleavage sites. The best in silico approach is to use a pattern recognition approach model. The main data used for the in silico protease cleavage site prediction are peptides as aforementioned. In this study, a peptide of a main protease cleaved glutamine residue is called a cleaved main protease peptide, or a cleaved peptide for short in the rest discussion. Such a glutamine residue, which interacts with the main protease, is called a main protease cleavage site, or a cleavage site for short in the rest discussion. Other glutamine residues are called the uncleavable sites and the peptides at these sites are called the uncleavable peptides.

Various pattern recognition approaches have been employed to construct in silico predictive models for the protease cleavage pattern discovery based on the available known cleaved peptides for different viral proteases, but yet SARS-CoV-2 viral protease so far. For

instance, a logistic linear regression model and a linear discriminant analysis model were used to predict the HIV-1 protease cleavage sites,[10,11] a decision tree model was constructed for the tryptic cleavage site prediction,[12] a support vector machine model was constructed for the caspase cleavage site prediction,[13] a multi-layer perception (or artificial neural network) model was constructed for predicting various protease cleavage sites[14] and a random forest model was constructed for predicting various protease cleavage sites as well,[15] to name a few.

The collection of the cleaved peptides is straightforward. All the experimentally verified cleavage sites for a protease can be used to generate the cleaved peptides. To collect uncleavable peptides, a specific rule must be followed, that is, they should contain sufficient background information for them to be compared with the cleaved peptides.[16] For instance, the coronavirus main protease only cleaves at a sequence where the P_1 residue is Q.[2,17,18] Therefore, a uncleavable peptide for the coronavirus main protease cleavage site prediction must target the uncleavable glutamine residues only.

Having known that the prediction of the SARS-CoV-2 main protease cleavage sites requires the glutamine peptides as inputs, the next issue is how to present glutamine peptides to a pattern recognition model. Most pattern recognition approaches only accept numerical data as the inputs. Therefore, the amino acids of the peptides must be transformed to some numerical data at first. This is called an amino acid encoding process. There have been many different methods for transforming the amino acids to numerical values. The mostly well-employed methods in the literature include the binary encoding approach,[19] the descriptor encoding approach[20–22] and the profile encoding approach,[6,23] to name a few.

In addition to these approaches used to transform amino acids to numerical data, a question is whether there is an alternative to handle the non-numeric amino acids in a pattern recognition model, which can be more biologically sound. It has been found that the structure of a protease will not be varying very fast during an evolution[24] (Yen et al.,[25]). Most importantly, a protease will have some degree of the tolerance to target a cleavage site in a polyprotein for the interaction even if genetic evolution may have occurred.[26,27] Therefore, the correlation between the cleaved peptides should be statistically significant compared with the correlation between uncleavable peptides or the correlation between uncleavable and cleaved peptides. Based on this understanding, the kernel function[28,29] can be used to map the original non-numerical peptide space to a numerical kernel space based on the correlations between the available peptides and the cleaved peptides. This has led to the development of the bio-kernel function as an alternative approach to deal with non-numerical amino acids. The Supporting Information Document S2 provides the details of the bio-kernel function.

As aforementioned, in silico protease cleavage site prediction is a pattern recognition problem. Though the SARS-CoV-2 main protease has been researched in the laboratory,[30,31] the in silico prediction of the cleavage sites of this protease has yet to be addressed. This study, for the first time, examines the in silico prediction of the SARS-CoV-2 main protease cleavage sites using the pattern recognition approaches, especially incorporating the kernel function. This article will show two important findings. First, the SARS-CoV-2 main protease cleavage sites are predictable with high accuracy by an in silico model because the cleaved glutamine peptides have reserved an excellent cleavage pattern for separating the cleaved peptides from the uncleavable peptides. Second, the pattern recognition approaches incorporating with the kernel function works the best.

2 | MATERIALS AND METHODS

In total, all available 64 SARS-CoV-2 protein sequences which contain the main protease cleavage sites were downloaded from NCBI using the keywords, {([coronavirus] AND main protease) AND cleavage}, on the April 4, 2021. Table S1 lists these 64 sequences. There were 273 main protease cleavage sites within part of these 64 sequences. A cleaved peptide was generated for each cleavage site, which is expressed by $P_5P_4P_3P_2P'_1P'_2P'_3P'_4P'_5$. In this notation, the residue P_1 was removed because the coronavirus main protease always targets the amino acid glutamine (Q).[2,17,18] After removing the duplicated peptides, 116 non-redundant cleaved peptides were maintained for the study.

Correspondingly, non-redundant uncleavable peptides were also randomly selected from these 64 sequences. The following rule was used to select uncleavable peptides. Suppose one sequence had K cleavage sites and $M > K$ non-cleavage glutamine residues for the main protease. K of M non-cleavage glutamines were randomly selected to generate K uncleavable peptides. This generated 273 uncleavable peptides. The duplicated 9-mer uncleavable peptides were also removed resulting in 259 non-redundant uncleavable peptides. Therefore the data set was composed of 375 9-mer peptides for the in silico prediction of the SARS-CoV-2 main protease cleavage sites in this study.

In addition to 273 uncleavable glutamines (Q), the rest 5071 glutamines (hence, 5071 9-mer uncleavable peptides) were not abandoned. After redundancy clearance, these 5071 uncleavable peptides were reduced to 2360 nonredundant uncleavable peptides. These 2360 non-redundant uncleavable peptides were saved for the blind test of the constructed models. In theory, all of these 2360 blind uncleavable peptides were expected to be classified as the uncleavable ones using a pattern recognition model if it was well constructed.

Table 1 shows these 116 cleaved peptides and the proteins as well as the cleavage sites.

Figure 1 shows the sequence logo for the 116 9-mer cleaved peptides. Figure S1 shows the sequence logo for the 259 9-mer uncleavable peptides. Comparing these two sequence logos, it can be seen that the uncleavable peptides had no trend of the amino acid composition. However, the cleaved peptides displayed a distinct trend of the amino acid composition. For instance, the residue P_2 (labeled by 4 in Figure 1) was mainly occupied by the amino acid leucine (L) in addition to valine (V), methionine (M), and isoleucine (I). The reside P'_1 (labeled by 5 in Figure 1) was mainly occupied by the amino acid

TABLE 1 The cleaved peptides and the proteins and the cleavage sites

Peptides	Protein and sites
SAVLQSGFRK	R1AB_SARS2#3263,R1A_SARS2#3263,R1AB_SARS#3240,R1A_SARS#3240
GVTFQSAVKR	R1AB_SARS2#3569,R1A_SARS2#3569
VATVQSKMSD	R1AB_SARS2#3859,R1A_SARS2#3859,R1AB_SARS#3836,R1A_SARS#3836,R1AB_BC279#3842, 1AB_BCRP3#3834, R1A_BC279#3842
RATLQAIASE	R1AB_SARS2#3942,R1A_SARS2#3942,R1AB_SARS#3919,R1A_SARS#3919,R1AB_BC279#3925, R1AB_BCRP3#3917, R1A_BC279#3925
AVKLQNNELS	R1AB_SARS2#4140,R1A_SARS2#4140,R1AB_SARS#4117,R1A_SARS#4117,R1AB_BC279#4123, R1AB_BCRP3#4115, R1A_BC279#4123
TVRLQAGNAT	R1AB_SARS2#4253,R1A_SARS2#4253,R1AB_SARS#4230,R1A_SARS#4230,R1AB_BC279#4236, R1AB_BCRP3#4228, R1A_BC279#4236
EPMLQSADAQ	R1AB_SARS2#4392,R1A_SARS2#4392
HTVLQAVGAC	R1AB_SARS2#5324,R1AB_SARS#5301,R1AB_BC279#5307,R1AB_BCRP3#5299
VATLQAENVT	R1AB_SARS2#5925,R1AB_SARS#5902,R1AB_BC279#5908,R1AB_BCRP3#5900
FTRLQSLENV	R1AB_SARS2#6452,R1AB_SARS#6429,R1AB_CVMA5#6503,R1AB_CVMJH#6507, R1AB_CVM2#6451,R1AB_BC279#6435, R1AB_BCRP3#6427
YPKLQSSQAW	R1AB_SARS2#6798
GVTFQGKFKK	R1AB_SARS#3546,R1A_SARS#3546,R1AB_BC279#3552,R1A_BC279#3552
EPLMQSADAS	R1AB_SARS#4369,R1A_SARS#4369
YPKLQASQAW	R1AB_SARS#6775,R1AB_BC279#6781,R1AB_BCRP3#6773
TSFLQSGIVK	R1AB_CVMA5#3333,R1AB_CVBQ#3246,R1AB_CVBLU#3246,R1AB_CVMJH#3336, R1AB_CVM2#3279,R1A_CVMA5#3333, R1A_CVMJH#3336,R1A_CVHOC#3246, R1A_CVHN5#3284,R1A_CVHN2#3304,R1A_CVHN1#3334,R1A_CVBM#3246
GVKLQSKRTR	R1AB_CVMA5#3635,R1AB_CVMJH#3639,R1AB_CVM2#3582,R1A_CVMA5#3635, R1A_CVMJH#3639
VSQIQSRLTD	R1AB_CVMA5#3921,R1AB_CVMJH#3927,R1AB_CVM2#3869,R1A_CVMA5#3921, R1A_CVMJH#3927
LQALQSEFVN	R1AB_CVMA5#4013,R1AB_CVMJH#4019,R1AB_CVM2#3961,R1A_CVMA5#4013, R1A_CVMJH#4019,R1A_CVHN5#3966, R1A_CVHN2#3986,R1A_CVHN1#4016
TVVLQNNELM	R1AB_CVMA5#4207,R1AB_CVMJH#4213,R1A_CVMA5#4207,R1A_CVMJH#4213
TVRLQAGTAT	R1AB_CVMA5#4317,R1AB_CVBQ#4232,R1AB_CVBLU#4232,R1AB_CVMJH#4323, R1AB_CVM2#4265,R1A_CVMA5#4317, R1A_CVMJH#4323,R1A_CVHOC#4232, R1A_CVBM#4232
GSQFQSKDTN	R1AB_CVMA5#4454,R1AB_CVMJH#4460,R1AB_CVM2#4402,R1A_CVMA5#4454, R1A_CVMJH#4460
SAVLQSVGAC	R1AB_CVMA5#5382
NPRLQCTTNL	R1AB_CVMA5#5982,R1AB_CVMJH#5988,R1AB_CVM2#5930
YPRLQAAADW	R1AB_CVMA5#6877,R1AB_CVMJH#6881,R1AB_CVM2#6825
NSTLQSGLRK	R1AB_CVPPU#2878,R1A_CVPPU#2878
GVNLQAGKVK	R1AB_CVPPU#3180,R1A_CVPPU#3180
ISTVQSKLTE	R1AB_CVPPU#3474,R1A_CVPPU#3474
TTILQSVASA	R1AB_CVPPU#3557,R1A_CVPPU#3557
TTKLQNNEIM	R1AB_CVPPU#3752,R1A_CVPPU#3752
TVRLQAGKPT	R1AB_CVPPU#3863,R1A_CVPPU#3863
RTSMQSFTVD	R1AB_CVPPU#3998,R1A_CVPPU#3998
STVLQAAGMC	R1AB_CVPPU#4927
KIGLQAKPET	R1AB_CVPPU#5526
SKALQSLENV	R1AB_CVPPU#6045
YPQLQSAEWN	R1AB_CVPPU#6384
GSTLQAGLRK	R1AB_CVH22#2965,R1A_CVH22#2965
GVNLQSGKTT	R1AB_CVH22#3267,R1A_CVH22#3267
VSTVQSKLTD	R1AB_CVH22#3546,R1A_CVH22#3546
DSILQSVASS	R1AB_CVH22#3629,R1A_CVH22#3629
VVKLQNNEIM	R1AB_CVH22#3824,R1A_CVH22#3824,R1A_CVHNL#3799

(Continues)

TABLE 1 (Continued)

Peptides	Protein and sites
TVRLQAGKQT	R1AB_CVH22#3933,R1A_CVH22#3933,R1A_BC512#3976,R1A_PEDV7#3965
RTAIQSFDNS	R1AB_CVH22#4068,R1A_CVH22#4068
STVLQAAGLC	R1AB_CVH22#4995
MTDLQSESSC	R1AB_CVH22#5592
EVNLQGLENI	R1AB_CVH22#6110
YPQLQSAEWK	R1AB_CVH22#6458
GIKLQSKRTR	R1AB_CVBQ#3549,R1AB_CVBLU#3549,R1A_CVHOC#3549,R1A_CVBM#3549
VSQFQSKLTD	R1AB_CVBQ#3836,R1AB_CVBLU#3836,R1A_CVHOC#3836,R1A_CVBM#3836
NTVLQALQSE	R1AB_CVBQ#3925,R1AB_CVBLU#3925,R1A_CVHOC#3925,R1A_CVBM#3925
ATVLQNNELM	R1AB_CVBQ#4122,R1AB_CVBLU#4122,R1A_CVHOC#4122
DTTVQSKDTN	R1AB_CVBQ#4369,R1AB_CVBLU#4369,R1A_CVHOC#4369,R1A_CVBM#4369
SAVMQSVGAC	R1AB_CVBQ#5297,R1AB_CVBLU#5297,R1AB_CVMJH#5388,R1AB_CVM2#5330
ETRVQCSTNL	R1AB_CVBQ#5900,R1AB_CVBLU#5900
FTKLQSLENV	R1AB_CVBQ#6421,R1AB_CVBLU#6421
YPRLQAASDW	R1AB_CVBQ#6795,R1AB_CVBLU#6795
SVVLQNNELM	R1AB_CVM2#4155
EPMMQSADAS	R1AB_BC279#4375,R1AB_BCRP3#4367,R1A_BC279#4375
GVTFQGKFKR	R1AB_BCRP3#3544
VSRLQAGFKK	R1AB_IBVM#2781
GVRLQSSFVR	R1AB_IBVM#3088,R1AB_IBVBC#3086
IATVQSKLSD	R1AB_IBVM#3381
STVLQSVTQE	R1AB_IBVM#3464,R1AB_IBVBC#3462
DVALQNNELM	R1AB_IBVM#3674
VVVLQSKGHE	R1AB_IBVM#3785,R1AB_IBVBC#3783
KPSVQSVAVA	R1AB_IBVM#3930
PTTLQSCGVC	R1AB_IBVM#4870,R1AB_IBVBC#4868
VASLQGTGLF	R1AB_IBVM#5470
FSALQSIDNI	R1AB_IBVM#5991,R1AB_IBVBC#5989
YPQLQSAWTC	R1AB_IBVM#6329,R1AB_IBVBC#6327
VSRLQSGFKK	R1AB_IBVBC#2779
IATVQAKLSD	R1AB_IBVBC#3379
DVVLQNNELM	R1AB_IBVBC#3672
KSSVQSVAGA	R1AB_IBVBC#3928
ETSLQGTGLF	R1AB_IBVBC#5468
GVKLQSKTKR	R1A_CVHN5#3587,R1A_CVHN2#3607,R1A_CVHN1#3637
VSQIQSKLTD	R1A_CVHN5#3874,R1A_CVHN2#3894,R1A_CVHN1#3924
NAVMQNNELM	R1A_CVHN5#4160,R1A_CVHN2#4180,R1A_CVHN1#4210
TIRLQAGVAT	R1A_CVHN5#4270,R1A_CVHN2#4290,R1A_CVHN1#4320
SVAVQSKDLN	R1A_CVHN5#4407,R1A_CVHN1#4457
GVAVQSKDLN	R1A_CVHN2#4427
SAALQAGLTR	R1A_BCHK9#3103
GVKLQGKFQS	R1A_BCHK9#3409
VSTIQSNMTD	R1A_BCHK9#3699
NSVLQAVASE	R1A_BCHK9#3782
PVKLQNNELM	R1A_BCHK9#3982
TVRLHAGSAT	R1A_BCHK9#4094

TABLE 1 (Continued)

Peptides	Protein and sites
EINLQARDEC	R1A_BCHK9#4233
NSTLQSGLKK	R1A_CVHNL#2939
GVNLQSGKVI	R1A_CVHNL#3242
ISTVQSKLTD	R1A_CVHNL#3521
SSTLQSVASS	R1A_CVHNL#3604
TIRLQAGKQT	R1A_CVHNL#3908
RTTIQSVDIS	R1A_CVHNL#4043
SSVLQSGLVK	R1A_BCHK4#3291,R1A_BC133#3298,R1A_BCHK5#3338
GVVMQSGVKR	R1A_BCHK4#3597,R1A_BC133#3604,R1A_BCHK5#3644
VATVQSKLTD	R1A_BCHK4#3889,R1A_BC133#3896
SSVLQATLTE	R1A_BCHK4#3972,R1A_BC133#3979
AVKLQNNEIH	R1A_BCHK4#4171,R1A_BC133#4178
TVRLQAGANT	R1A_BCHK4#4281,R1A_BC133#4288
NTVPQSKDTN	R1A_BCHK4#4420,R1A_BC133#4427
ATALQNNELM	R1A_CVBM#4122
IASVQSKLTD	R1A_BCHK5#3936
PSVLQATLSE	R1A_BCHK5#4019
AVTLQNNEIR	R1A_BCHK5#4218
TVRLQAGSNT	R1A_BCHK5#4328
TTIPQSKDSN	R1A_BCHK5#4467
NSTLQAGLRK	R1A_BC512#3012,R1A_PEDV7#2997
GVTLQSGKVS	R1A_BC512#3314
ISSVQSKLTD	R1A_BC512#3590,R1A_PEDV7#3579
SSVLQSVAAT	R1A_BC512#3673
IIKLQNNEII	R1A_BC512#3868
RAVIQSVDSG	R1A_BC512#4111
GVNLQGGYVS	R1A_PEDV7#3299
NSMLQSVAST	R1A_PEDV7#3662
IVKLQNNEII	R1A_PEDV7#3857
RSIMQSTDMA	R1A_PEDV7#4100

Note: The # key is used to separate between a protein and a cleavage site. Multiple protein sequences may contain an identical peptide. For instance, the peptide SAVLQSGFRK was found in four protein sequences (R1AB_SARS2, R1A_SARS2, R1AB_SARS, R1A_SARS).

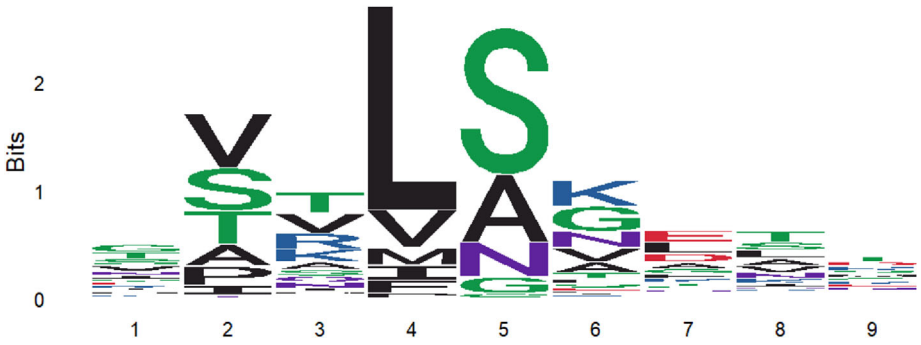


FIGURE 1 The sequence logo of 116 cleaved peptides, where the integers from 1 to 9 represent the residues, $P_5P_4P_3P_2P'_1P'_2P'_3P'_4P'_5$ in order. Note that the glutamine (Q) has been omitted

serine (S) in addition to alanine (A), asparagine (N), and glycine (G). Therefore, it is expected that two types of peptides (cleaved vs. uncleavable) should not be very difficult to separate in this data set.

Based on the comparison between Figure 1 and Figure S1, it can be seen that residues P_5, P'_3, P'_4, and P'_5 (labeled by 1, 7, 8, and 9 in Figure 1) may not have a significant contribution to the discrimination between the cleaved and uncleavable peptides. Therefore,

another data set used for the SARS-CoV-2 main protease cleavage site prediction in this study was based on the peptide structure of five residues, that is, $P_4P_3P_2P'_1P'_2$. After reducing the peptide size from 9 to 5, the redundancy among peptides was checked again. This led to 87 non-redundant cleaved 5-mer peptides, 256 non-redundant uncleavable 5-mer peptides and 2061 non-redundant blind uncleavable 5-mer peptides. Table 2 summarizes the number of peptides.

A pattern recognition model, which is a classifier in this case, can thus be constructed to examine the discriminative power for either data set, that is, the 9-mer peptide set and the 5-mer set for the purpose of the in silico prediction of the SARS-CoV-2 main protease cleavage sites. To construct a classifier, three issues were considered. The first issue was how to present (encode) the amino acids into a model. This is because most pattern recognition algorithms only accept numerical data. Different methods of dealing non-numerical amino acids have different efficiency and reliability. The binary-encoding, descriptors, profiling and the bio-kernel function approaches were considered in this study. Although there are many others, these have been the most popularly used in the literature. The second issue was the selection of the pattern recognition approaches. There is normally no rule-of-thumb for determining which is superior in advance and there is a need for careful examination of each. Seven most popularly used and representative as well as matured pattern recognition approaches have been employed in this study. The third issue was how to evaluate such a model when it has been constructed. The cross-validation method as well as the ROC analysis approach was employed in this study. Refer to the extended methods for details of these three methods.

The kernel function has been well exercised in the machine-learning field.[28,29] A naïve description of the kernel function is briefly described here. One of the most promising advantages of the kernel function is that it can transform a nonlinearly separable space to a linearly separable space. For instance, two classes of data points in the original space (A, B, α, and β) in the left panel of Figure 2 are nonlinearly separable. When using two data points (α and β) as the kernels, the distances between four data points and these two kernels can be calculated. Based on the distances, a new space is formulated shown in the right panel of Figure 2. It can be seen that these four data points in this new kernel space coordinated by α and β become linearly separable.

TABLE 2 The peptide data used for this study

	9-mer		5-mer
	Raw	Reduced	Reduced
Cleaved	273	116	87
Non-cleaved	273	259	256
Blind	5071	2360	2061

Note: "Raw" stands for the number of all the peptides and "Reduced" stands for the number of non-redundant peptides.

3 | RESULTS AND DISCUSSION

The first question for any pattern recognition model is whether the variables themselves possess a good discriminative power. Therefore, Figure 3 shows the bio-SOM map constructed for the 9-mer peptides data without employing the classification labels of cleaved or uncleavable. Among 225 neurons (cells), 178 were mapped by at least one peptide. The occupancy rate of 225 cells was about 79%. Among these 178 cells, 170 cells were mapped by only one class of peptides, either cleaved peptides or uncleavable peptides. In other words, these 170 cells were pure for one class of peptides. In total, 353 were mapped to these 170 cells. The total accuracy of separating the cleaved peptides from the uncleavable peptides was 94.13% (353/375). This thus demonstrated that the discriminative power between the cleaved peptides and the uncleavable peptides in this data set should be greater than 94.13% in a well-constructed supervised pattern recognition model. In addition to the discriminative power demonstrated by the bio-SOM map, the distribution of two classes of peptides was also consistent regarding the biological knowledge of the peptides. The cleaved peptides occupied a smaller number of cells while the uncleavable peptides occupied a greater number of cells. This is because each peptide was aligned with the cleaved peptides. If the cleaved peptides held a good amino acid composition trend, the correlation between the cleaved peptides should be very high, but the correlation between the uncleavable peptides and the cleaved peptides should be very low. In other words, the cleaved peptides maintained a more conserved amino acid composition trend to occupy a smaller area in the bio-SOM map while the uncleavable peptides had a random distribution of amino acids that occupied a greater area in the bio-SOM map.

Figure S3 shows the SOM model constructed for the binary-encoded data. It has the same model structure as the bio-SOM map shown in Figure 3. The number of cells occupied by at least one peptide was 154, that is, the occupancy rate was 68.89%. The purity rate of the cells was 91.47%, which was lower than 94.13%. Figure S4A shows the SOM model constructed based on the descriptor-encoded data. The number of cells occupied by at least one peptide was 148, that is, the occupancy rate was 65.78%. The purity rate of the cells was 75.47%, which was much lower than 94.13%. Figure S4B shows the SOM model constructed based on the profile-encoded data. The number of cells occupied by at least one peptide was 141, that is, the occupancy rate was 62.67%. The purity rate of the cells was 78.67%, which was also much lower than 94.13%.

It can be seen that the bio-SOM model maximally explored the discriminative power from this SARS-CoV-2 protease peptide data, being 94.13%. Therefore, 94.13% should be considered as the benchmark when evaluating the supervised pattern recognition models constructed for the in silico prediction of the SARS-COV-2 main protease cleavage sites.

Table 2 shows the performance of two sets of 23 supervised pattern recognition models. For the 9-mer peptides data, eight models had the total prediction accuracy over 94.13%. Therefore, the discriminative power was well explored in these supervised models. All the

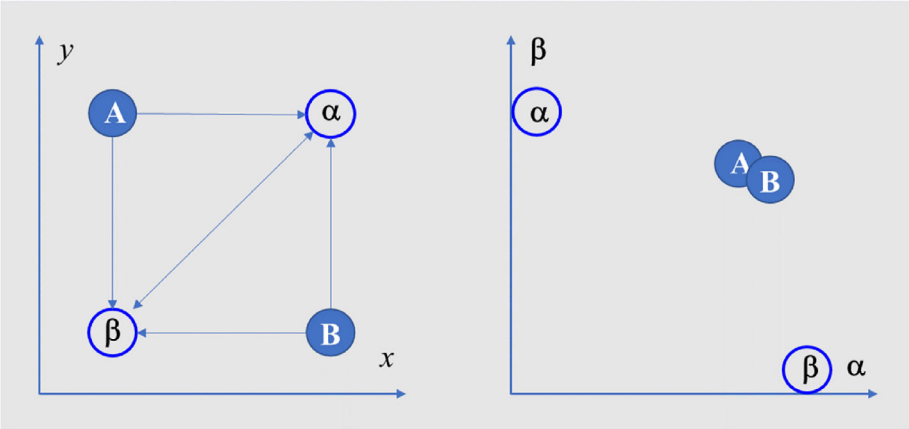
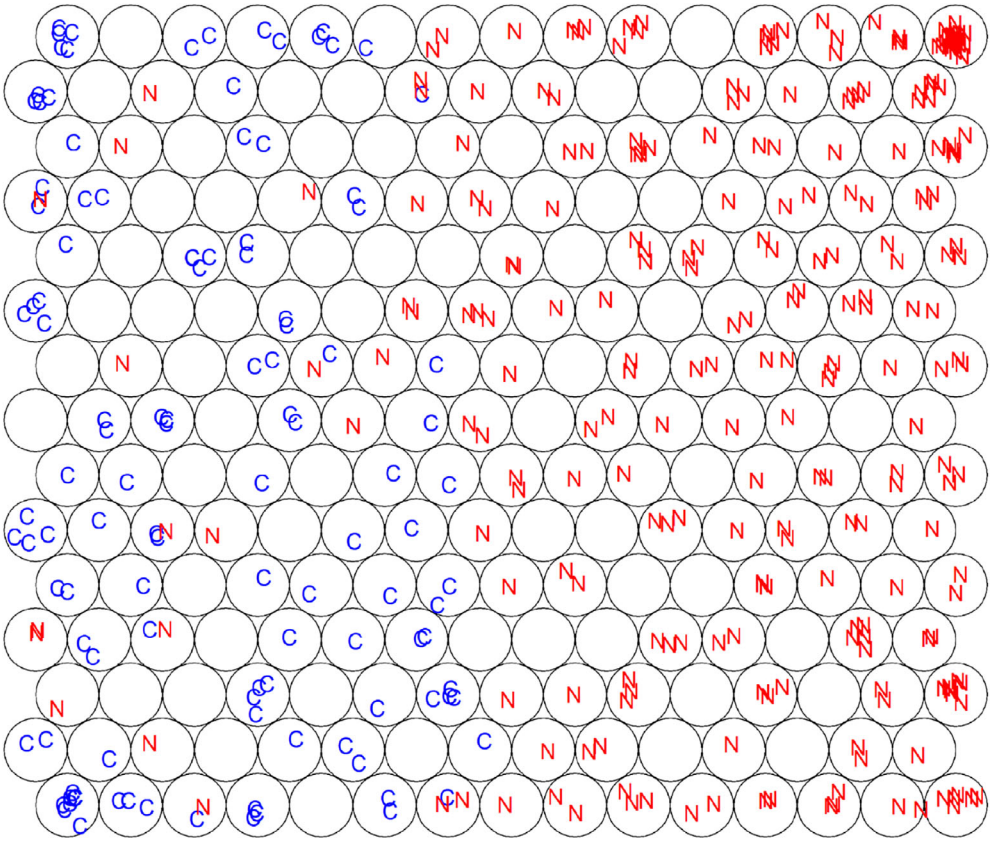


FIGURE 2 A naïve description of the kernel function approach. The left panel shows the original data space coordinated by *x* and *y*, in which four data points are labeled by A, B, α, and β. A and B belong to one class while α and β belong to the other class. They are nonlinearly separable because it is impossible to separate these two classes using one straight line. Suppose α and β are selected as the kernels. The distances between four data points and two kernels are calculated. The right panel shows the distribution of four data points based on four sets of distances using the kernel function. In this new space, two coordinates are no longer *x* and *y*, but α and β. It can be seen that this new space of four data points becomes linearly separable

FIGURE 3 The bio-SOM map of 225 neurons constructed for the 9-mer peptides data. "N" stands for the uncleavable peptides and "C" stands for the cleaved peptides. One circle stands for one neuron or one cell. The printed letter in a cell, which is either N or C, stands for a peptide, which has been mapped to the cell. For instance, two cleaved peptides were mapped to the first cell at the bottom row while one uncleavable peptide was mapped to the third cell at the bottom row. These two cells were pure for one class. However, the second cell at the top row contained two cleaved peptides and one uncleavable peptide. Thus, this cell was not pure for one class



profile-encoded models were not able to achieve the total prediction accuracy greater than 94.13%.

Five models had no Type I error on the blind data. All were the bio-kernel models. Four best models had the total prediction accuracy greater than 94.71% and had no Type I error. They were bio-FOREST, bio-MLP, bio-SVM, and bio-RVM. Among them, the bio-SVM model was the best. Its AUC was 1, its MCC was 0.9938, and its total prediction accuracy was 99.73%. The total prediction accuracy of bio-SVM was about 5% greater than that of bio-SOM, which was a significant increase.

For the 5-mer peptides data, eight models had the total prediction accuracy greater than 94.13% and five models had no Type I error on the blind data. The best 5-mer model was the bio-SVM model. Its AUC

was 0.9999, its MCC was 0.9770, and its total prediction accuracy was 99.42%, which was 5% greater than the benchmark accuracy 94.13%.

Comparing all the models, it can be seen that the bio-kernel models (bio-FOREST, bio-MLP, bio-SVM, and bio-RVM) performed the best for the SARS-CoV-2 main protease cleavage site prediction. Other models either failed to have the total prediction accuracy greater than 94.13% or failed to have 0% Type I error rate in the blind data set testing.

Figure S5 shows the ROC curves of the bio-SVM models. They were consistent with the figures included in Table 3. Figure S6 shows the densities estimated for the predictions on the blind data using the bio-SVM models. It can be seen that the prediction values were all smaller than the threshold 0.5, which was the default threshold when prediction values were between 0 and 1 for the discrimination between two classes of peptides. This means all of uncleavable glutamine residues in the blind test data set were accurately predicted as uncleavable ones.

Figure 4 shows the prediction spectra of four bio-kernel models for the protein R1AB_SARS2. All demonstrated the excellent discriminative power between the cleaved glutamines and uncleavable glutamines. The bio-RVM model was outstanding because the prediction values of all the uncleavable glutamines were almost zero.

As indicated in the earlier studies[30,31] that the main protease cleavage pattern, if recognized, can help the design of the SARS-CoV-2 drug (inhibitors). Below, I show how the pattern recognition approaches can be used for this task. The inductive pattern recognition models such as a decision tree model or a random forest model provide excellent interpretation capabilities. Such a model can help discover the residues or the peptides which play an important role in the in silico cleavage site prediction. For instance, a decision tree model was constructed to extract the cleavage knowledge for the hepatitis C virus[32] and to discover the tryptic cleavage pattern.[12] However, in these applications, the data used in a decision tree model were the peptide residues, that is, one residue was one variable. For instance, a resulting decision tree model based on the 5-mer SARS-CoV-2 main protease cleavage peptides data in this study can thus explain which of the 5 residues (either P_4 or P_3 or P_2 or P'_1 or P'_2) play an important discriminatory role that separates the cleaved glutamine residues from the uncleavable glutamine residues for the coronavirus main protease.

TABLE 3 The model performance. "Type I" stands for the Type I error rate

	9-mer				5-mer			
Models	**AUC**	**MCC**	**Type I**	**Total**	**AUC**	**MCC**	**Type I**	**Total**
NO + C5.0	0.9636	0.8311	6%	92.27%	0.9458	0.6525	5%	93.00%
NO + FOREST	0.9940	0.9385	5%	**95.73%**	0.9890	0.9070	6%	**95.75%**
BIN + MLP	0.9696	0.8775	4%	**94.93%**	0.9804	0.8654	4%	**94.46%**
BIN + SVM	0.9691	0.8370	8%	89.87%	0.9651	0.7743	12%	88.92%
BIN + RVM	0.9963	0.9501	1%	**96.27%**	0.9809	0.8829	3%	**95.92%**
DES + Linear	0.9798	0.8436	9%	93.33%	0.9626	0.7623	6%	90.38%
DES + C5.0	0.9527	0.8462	5%	93.07%	0.9605	0.7960	4%	93.00%
DES + FOREST	0.9897	0.9137	3%	**96.27%**	0.9863	0.8942	3%	**96.21%**
DES + MLP	0.9658	0.8202	5%	90.93%	0.9639	0.7965	6%	90.96%
DES + SVM	0.9827	0.8561	6%	93.60%	0.9700	0.8055	4%	92.13%
DES + RVM	0.9663	0.8215	6%	92.27%	0.9559	0.7861	6%	91.84%
PSE + Linear	0.9428	0.7382	5%	88.53%	0.9487	0.7349	5%	88.05%
PSE + C5.0	0.8672	0.6333	5%	83.20%	0.9070	0.7821	7%	90.96%
PSE + FOREST	0.9773	0.8319	3%	92.27%	0.9735	0.8286	4%	92.13%
PSE + MLP	0.9467	0.7608	6%	89.60%	0.9520	0.7302	5%	89.50%
PSE + SVM	0.9727	0.8339	3%	91.20%	0.9519	0.7760	3%	87.46%
PSE + RVM	0.9472	0.7608	6%	88.27%	0.9475	0.7395	6%	88.63%
bio-Bayesian	0.9745	0.8419	3%	93.33%	0.9624	0.7474	3%	90.67%
bio-C5.0	0.9357	0.7805	**0%**	90.93%	0.9394	0.7708	**0%**	84.62%
bio-FOREST	0.9889	0.9380	**0%**	**95.73%**	0.9835	0.8970	**0%**	**95.63%**
bio-MLP	0.9843	0.9041	**0%**	**96.80%**	0.9648	0.8605	**0%**	**94.46%**
bio-SVM	1.0000	0.9938	**0%**	**99.73%**	0.9999	0.9770	**0%**	**99.42%**
bio-RVM	0.9834	0.8932	**0%**	**94.40%**	0.9792	0.8790	**0%**	**95.34%**

Note: "NO" means no encoding process was used. "BIN" stands for the binary-encoded data. "DES" stands for the descriptor-encoded data. "PSE" stands for the profile-encoded data. The percentages in bold were greater than 94.13% of the bio-SOM model.

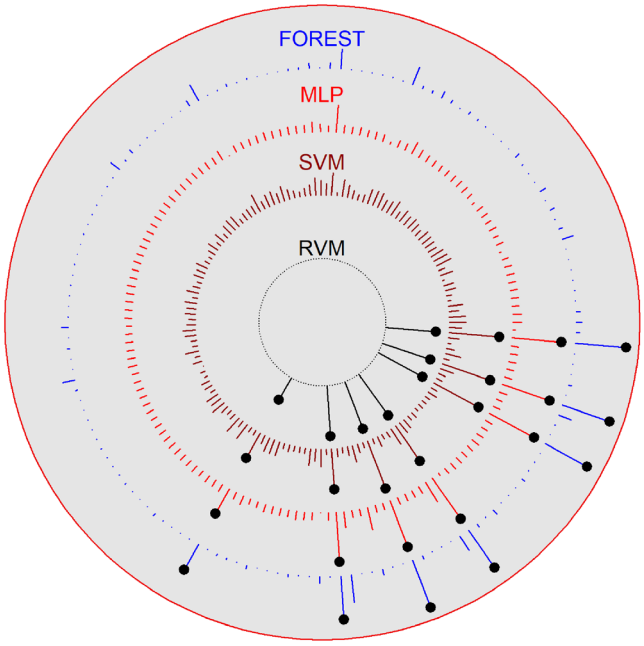


FIGURE 4 The prediction spectra of four bio-kernel models (bio-FOREST, bio-MLP, bio-SVM, and bio-RVM) for the protein R1AB_SARS2. The protein had seven main protease cleavage sites. The heights of the bars stand for the predicted values which have been normalized between 0 and 1. The bars with the dots on the top stand for the true cleavage sites

Rather than using the raw residues as the variables, a bio-kernel inductive pattern recognition model employs the cleaved peptides as the variables. Thus, a bio-kernel inductive pattern recognition model (bio-C5.0 and bio-FOREST) was able to discover which cleaved peptides were the most significant ones for the discrimination between the cleaved peptides and the uncleavable peptides. These most discriminating cleaved peptides were then the most probable prototypes as the targets for the drug (inhibitors) design.[6,33]

The bio-C5.0 model and the bio-FOREST model constructed for the 9-mer peptides are shown in Figures S7 and S8, respectively. The bio-C5.0 model employed nine cleaved peptides and the most important cleaved peptide was GVNLGSGKTT. The bio-FOREST tree employed 11 cleaved peptides and the most important cleaved peptide was DTTVGSKDTN. Among them, eight were significant because their p values were less than .01. The decision tree algorithm partitions a space using the orthogonal partitioning rules while the random forest algorithm partitions a space using the non-orthogonal partitioning rules. Therefore, the resulting bio-C5.0 and bio-FOREST models were different. Figures S9 and S10 show the sequence logos of the cleaved peptides employed by these two trees. It can be been that most cleaved peptides selected in the trees had leucine (L) in the residue P_2 and S (serine) in the residue P'_1. Compared with Figure 1, these sequence logos show clear amino acid composition trends. The cleaved peptides selected by these tree models represent their importance (measured by the p values) to discriminate between the cleaved peptides and the uncleavable peptides. In terms of the use of the bio-kernel technique, such a selected cleaved peptide has the following property. The majority of the cleaved peptides have high alignment scores with this selected cleaved peptide while the majority of the uncleavable peptides have low alignment scores with this selected cleaved peptide. Therefore, such a selected cleaved peptide is very different from uncleavable peptides and these two selected cleaved peptides are most different from uncleavable peptides compared with other cleaved peptides.

The consensus peptide of the bio-C5.0 model (Figure S7) was −V−LS−−−− and the consensus peptide of the bio-FOREST model (Figure S8) was −−−LS−−−−. The latter was identical with the consensus sequence derived from all cleaved peptides. This means that the main protease cleavage rule can be simplified to either −V−LSG↓Q−−−− or −−−LSG↓Q−−−−, where ↓ stands for the cleavage site. If only checking whether the consensus peptide was present in the peptides, these two consensus peptides can be used to scan all the peptides to examine whether they matched. Table S2 shows the confusion matrices. It can be seen that although the consensus peptide −−−LS−−−− was far less than perfect, it outperformed the consensus peptide −V−LS−−−−. This is not a surprise because the residue P_4 was not mainly occupied by the amino acid valine (V). Figure S11 shows how the 20 amino acids were distributed at the residue P_4 among the SARS-CoV-2 main protease cleaved peptides. The amino acid valine (V) had the largest frequency (29%) at this residue, the frequency of the serine (S) was 23%, and the frequency of the threonine (T) was 18%. The difference between three frequency values was actually not insignificant.

To further validate the discriminative power of the consensus peptides, the mutation matrix BLOSUM62 was used to calculate the homology alignment scores between the consensus peptides and all the peptides. The calculation method is included in the Supporting Information Document S7. Figure S12 shows the estimated densities of the homology alignment scores. It can be seen that the density of the uncleavable and the density of the cleaved peptides had a higher degree of overlap using the consensus peptide −V−LS−−−−. The overlap degree was smaller when using the consensus peptide −−−LS−−−−. Therefore, the bio-FOREST model may be able to deliver more robust decision-making rules for this data set.

In addition to identifying the consensus peptides, another question was whether the cleaved peptides can be ranked in terms of their discriminative power. A random forest model can rank the variables. In the bio-kernel space, the variables were the cleaved peptides. Therefore, the bio-FOREST model can rank the cleaved peptides in terms of their discriminative power between the cleaved and uncleavable peptides. The increase in node purity measurement of the bio-FOREST model was used to rank the cleaved peptides. Figures S13 and S14 show the amino acid distribution trend of the top 10 cleaved peptides selected by the bio-FOREST model. Again, the residue P_2 was always occupied by the amino acid leucine (L) and the residue P'_1 was almost occupied by the amino acid serine (S). This was again consistent with what has been discussed above.

The next issue was whether the top-ranked cleaved peptides can reserve a good discriminative power after the cleaved peptides were ranked. If so, later drug (inhibitor) design can have a parsimonious

structure to investigate. Based on the top 10 cleaved peptides selected by the bio-FOREST model, the parsimonious models were constructed. In a parsimonious model, only top 10 cleaved peptides were used as the kernel peptides. These 10 cleaved peptides were used for the following analysis. Four models which demonstrated the best performance shown in Table 3 were re-constructed to examine whether these parsimonious models had the performance significantly decreased or not. These four models were bio-FOREST, bio-MLP, bio-SVM, and bio-RVM. Table 4 shows the result. It can be seen that the performance was indeed worse, but the difference between the full models and the parsimonious models was insignificant. The decreased accuracy was not a surprise because the rest of the unused cleaved peptides may carry some extra discriminative power though minor. For instance, from the full bio-FOREST model to the parsimonious bio-FOREST model, the AUC value decreased from 0.9889 to 0.9744 and the MCC value decreased from 0.9380 to 0.8581 and the Type I error rate for the blind data was still 0%. From the full bio-SVM model to the parsimonious bio-SVM model, the AUC decreased from 1 to 0.99, the MCC decreased from 0.9938 to 0.8775, the total prediction accuracy decreased from 99.73% to 94.67% and the Type I error rate increased from 0% to 2%. The parsimonious bio-MLP and bio-RVM models also had a decrease in accuracy. Therefore, a parsimonious model sacrificed the accuracy slightly. However, using less than 10% variables, the decrease in the prediction accuracy was insignificant.

This study has shown that the SARS-CoV-2 main protease cleavage pattern has been well-reserved in peptides. For the 9-mer peptide data, only one model had the AUC value below 0.9 and only five models had the total prediction accuracy below 90%. For the 5-mer peptide data, no model had the AUC value below 0.9 and six models had the total prediction accuracy below 90%. Importantly, the bio-SOM model demonstrated 94.13% total prediction accuracy meaning that the amino acid composition trend or pattern inherent in the cleaved peptides was significant in terms of the discriminative power between the cleaved peptides and the uncleavable peptides. The use of a supervised model further explored the discriminative power when the model was able to capture the complexity within the peptide data.

Based on the above analysis of the in silico analysis results, it can be seen that the pattern recognition models incorporating the bio-kernel function outperformed other models which employed various amino acid encoding approaches. The reason may be due to the use of the amino acid mutation matrix which can make the mutual relationship between peptides more biologically sound. Mapping a difficult-to-model space to a kernel space for efficient data modeling including regression analysis and classification analysis has been well exercised in the pattern recognition area. Mapping a non-numerical peptide space to a numerical bio-kernel space has two benefits. First, the difficulty of handling non-numerical peptides is eased. Second, which is more important, discovering the most probable prototypes for SARS-CoV-2 drug (inhibitor) design can benefit. This feature may not be possible using any approach other than the bio-kernel models.

4 | CONCLUSION

Predicting protease cleavage sites in silico aims to generate a predictive model in a computer based on the known cleaved and uncleavable glutamine residues (peptides). Therefore, it is a typical pattern recognition problem. The basic assumption of modeling peptides for a protease cleavage problem using an in silico approach is that there should be sufficient known cleaved peptides verified in laboratory and the most importantly the cleaved peptides should well cover the amino acid composition trend for a specific protease to recognize. If there are insufficient known cleaved peptides or the available known cleaved peptides have not yet well covered the amino acid composition trend for a specific protease, efficiently predicting protease cleavage sites in silico would be impossible. The pattern recognition approaches have been well used for predicting protease cleavage sites in silico where the peptide data can well-satisfy the above two conditions. The benefits of using a pattern recognition approach for this kind of problem are obvious. *First*, some complex or nonlinear pattern can be well explored using a nonlinear pattern recognition model. For instance, in the 9-mer models shown in Table 3 of this article, the MCC of the DES + Linear model was 0.8436, but it was 0.9137 in the DES + FOREST model. The former was a linear model while the latter was a nonlinear model. The MCC was 0.7382 in the PSE + Linear model, but was 0.8339 in the PSE + SVM model. Again, the former was a linear model and the latter was a nonlinear model. *Second*, a pattern recognition model can be used to deliver useful information for the drug or inhibitor design if the cleavage knowledge can be well discovered. It has been well recognized that the bio-kernel models can generate a predictive model with a better generalization capability in addition to biological sound content. The bio-kernel models used in this study have shown their powerfulness for predicting the SARS-CoV-2 main protease cleavage sites in silico.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/prot.26274.

TABLE 4 The performance comparison between the full models and the parsimonious models

	Parsimonious models				Full models			
Algorithm	**AUC**	**MCC**	**Type I**	**Total**	**AUC**	**MCC**	**Type I**	**Total**
bio-FOREST	0.9744	0.8581	**0%**	93.60%	0.9889	0.9380	**0%**	**95.73%**
bio-MLP	0.9487	0.8046	2%	89.33%	0.9843	0.9041	**0%**	**96.80%**
bio-SVM	0.9900	0.8775	2%	**94.67%**	1.0000	0.9938	**0%**	**99.73%**
bio-RVM	0.9652	0.8204	2%	92.27%	0.9834	0.8932	**0%**	**94.40%**

Note: The values in bold stand for the best models.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available in NCBI at https://www.ncbi.nlm.nih.gov/guide/proteins/. These data were derived from the following resources available in the public domain: - https://www.ncbi.nlm.nih.gov/guide/proteins/.

REFERENCES

1. Finkel Y, Mizrahi O, Nachshon A, et al. The coding capacity of SARS-CoV-2. *Nature*. 2021;589:125-130.
2. Koudelka T, Boger J, Henkel A, et al. N-terminomics for the identification of *in vitro* substrates and cleavage site specificity of the SARS-CoV-2 main protease. *Proteomics*. 2021;21:e2000246.
3. Lei J, Hilgenfeld R. RNA-virus proteases counteracting host innate immunity. *FEBS Lett*. 2017;591:3190-3210.
4. Bohorquez H, Suarez CF, Patarroyo MF. Mass & secondary structure propensity of amino acids explain their mutability and evolutionary replacements. *Sci Rep*. 2017;7:7717.
5. Koehl P, Levitt M. Structure-based conformational preferences of amino acids. *Proc Natl Acad Sci U S A*. 1999;96:12524-12529.
6. Chou KC. Prediction of human immunodeficiency virus protease cleavage sites in proteins. *Anal Biochem*. 1996;233:1-14.
7. Speicher DW, Weglarz L, De Silva TM. Properties of human red cell spectrin heterodimer (side-to-side) assembly and identification of an essential nucleation site. *J Biol Chem*. 1992;267:14775-14782.
8. Vidmar R, Vizovisek M, Turk D, Turk B, Fonovic M. Protease cleavage site fingerprinting by label-free in-gel degradomics reveals pH-dependent specificity switch of legumain. *EMBO J*. 2017;36:2455-2465.
9. Vizovisek M, Vidmar R, Van Quickelberghe E, et al. Fast profiling of protease specificity reveals similar substrate specificities for cathepsins K, L and S. *Proteomics*. 2015;15:2479-2490.
10. Nanni L, Lumini A. A reliable method for HIV-1 protease cleavage site prediction. *Neurocomputing*. 2006;69:838-841.
11. Singh O, Su EC. Prediction of HIV-1 protease cleavage site using a combination of sequence, structural, and physicochemical features. *BMC Bioinf*. 2016;17:s478.
12. Fannes T, Vandermarliere E, Schietgat L, Degroeve S, Martens L, Ramon J. Predicting tryptic cleavage from proteomics data using decision tree ensembles. *J Proteome Res*. 2013;12:2253-2259.
13. Wee LJ, Tan TW, Ranganathan S. SVM-based prediction of caspase substrate cleavage sites. *BMC Bioinf*. 2006;5:S14.
14. Blom N, Hansen J, Blaas D, Brunak S. Cleavage site analysis in picornaviral polyproteins: discovering cellular targets by neural networks. *Prot Sci*. 1996;5:2203-2216.
15. Li BQ, Cai YD, Feng KY, Zhao GJ. Prediction of protein cleavage site with feature selection by random forest. *PLoS One*. 2012;7:e45854.
16. Schechter I, Berger A. On the size of active sites in proteases. *Biochem Biophys Res Commun*. 1967;27:157-162.
17. Rut W, Groborz K, Zhang L, et al. SARS-CoV-2 M^{pro} inhibitors and activity-based probes for patient-sample imaging. *Nat Chem Biol*. 2021;17:222-228.
18. Zhang L, Lin D, Kusov Y, et al. Alpha-Ketoamides as broad-spectrum inhibitors of coronavirus and enterovirus replication: structure-based design, synthesis, and activity assessment. *J Med Chem*. 2020;63:4562-4578.
19. Wu C, Berry M, Shivakumar S, McLarty J. Neural networks for full-scale protein sequence classification: sequence encoding with singular value decomposion. *Mach Learn*. 1995;21:177-193.
20. Fontaine NT, Cadet XF, Vetrivel I. Novel descriptors and digital signal processing- based method for protein sequence activity relationship study. *Int J Mol Sci*. 2019;20:e5640.
21. Kidera A, Konishi Y, Oka M, Ooi T, Scheraga HA. Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *J Protein Chem*. 1985;4:23-55.
22. Lin Z, Long H, Bo Z, Wang Y, Wu Y. New descriptors of amino acids and their application to peptide QSAR study. *Peptides*. 2008;29:1798-1805.
23. Nanni L, Lumini A. Coding of amino acids by texture descriptors. *Artif Intell Med*. 2010;48:43-50.
24. Buller AR, Townsend CA. Intrinsic evolutionary constraints on protease structure, enzyme acylation, and the identity of the catalytic triad. *Proc Natl Acad Sci U S A*. 2013;110:E653-E661.
25. Yen YT, Kostakioti M, Henderson IR, Stathopoulos C. Common themes and variations in serine protease autotransporters. *Trends Microbiol*. 2008;16(8):370–379. http://dx.doi.org/10.1016/j.tim.2008.05.003
26. Adams MJ, Antoniw JF, Beaudoin F. Overview and analysis of the polyprotein cleavage sites in the family *Potyviridae*. *Mol Plant Pathol*. 2005;6:471-487.
27. da Silva PG, Mesquita JR, de São José Nascimento M, Ferreira VAM. Viral, host and environmental factors that favor anthropozoonotic spillover of coronaviruses: an opinionated review, focusing on SARS-CoV, MERS-CoV and SARS-CoV-2. *Sci Total Environ*. 2021;750:141483.
28. Hofmann T, Scholkopf B, Smola AJ. Kernel methods in machine learning. *Annals Stat*. 2008;36:1171-1220.
29. Muller KR, Mika S, Ratsch G, Tsuda K, Scholkopf B. An introduction to kernel-based learning algorithms. *IEEE Trans Neural Netw*. 2001;12:181-201.
30. Sacco MD, Ma C, Lagarias P, et al. Structure and inhibition of the SARS-CoV-2 main protease reveal strategy for developing dual inhibitors against M^{pro} and cathepsin L. *Sci Adv*. 2020;6:eabe0751.
31. Ullrich S, Nitsche C. The SARS-CoV-2 main protease as drug target. *Bioorg Med Chem Lett*. 2020;30:127377.
32. Narayanan A, Wu X, Yang ZR. Mining viral protease data to extract cleavage knowledge. *Bioinformatics*. 2002;18:S5-S13.
33. Mothay D, Ramesh KV. Binding site analysis of potential protease inhibitors of COVID-19 using AutoDock. *Virus*. 2020;31:1-6.

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of the article at the publisher's website.

How to cite this article: Yang ZR. In silico prediction of Severe Acute Respiratory Syndrome Coronavirus 2 main protease cleavage sites. *Proteins*. 2022;90(3):791-801. doi:10.1002/prot.26274