S-PDB: Analysis and Classification of SARS-CoV-2 Spike Protein Structures

M. Saqib Nawaz
Shenzhen University, China
msaqibnawaz@szu.edu.cn

Philippe Fournier-Viger
Shenzhen University, China
philfv@szu.edu.cn

Yulin He
Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
Shenzhen University, China
yulinhe@gml.ac.cn

2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995562

***Abstract*—This paper proposes a novel and efficient method, called S-PDB, for the analysis and classification of Spike (S) protein structures of SARS-CoV-2 and other viruses/organisms in the Protein Data Bank (PDB). The method first finds and identifies protein structures in PDB that are similar to a protein structure of interest (SARS-CoV-2 S) via a protein structure comparison tool. The amino acid (AA) sequences of identified protein structures, downloaded from PDB, and their aligned amino acids (AAA) and secondary structure elements (ASSE), that are stored in three separate datasets, are then used for the reliable detection/classification of SARS-CoV-2 S protein structures. Three classifiers are used and their performance is compared by using six evaluation metrics. Obtained results show that two classifiers for text data (Multinomial Naive Bayes and Stochastic Gradient Descent) performed better and achieved high accuracy on the dataset that contains AAA of protein structures compared to the datasets for AA and ASSE, respectively.**

***Index Terms*—SARS-CoV-2, Spike, PDB, DALI, Classification.**

I. Introduction

The COVID-19 pandemic, caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) [1], still remains a health emergency of international concern. The World Health Organization (WHO)'s director-general said recently that the recent surge in COVID-19 cases shows that "this pandemic is nowhere near over"[1]. According to the latest WHO report[2], more than 630 million people have been infected by COVID-19, with approximately 6.5 million deaths worldwide. Regular emergence of SARS-CoV-2 variants [2] and their sub types are making it hard to find an effective therapeutic or vaccine that could offer long-term immunity. Many countries lifted COVID-19 restrictions and recommended citizens to take approved COVID-19 vaccines. While some countries such as China are still taking preventive, quarantine and isolation measures to reduce the transmission and reproduction rate.

SARS-CoV-2 can enter the host cell membrane when the Spike (S) protein interacts with the host angiotensin-converting enzyme 2 (ACE2) [3], [4]. Thus, S protein plays a fundamental role in the pathogenesis, transmission and virulence of the this virus and COVID-19 disease. Structural biology techniques such as Cryo-EM (electron microscopy) and Xray crystallography are generally used to find how S protein, through binding domains such as receptor-binding and N-terminal, interacts with the ACE2 receptor and binds to it. These techniques explains the three-dimensional (3D) structures of proteins and their conformational changes. Since the emergence of SARS-CoV-2 in December 2019, its proteins structures are deposited at a fast speed in online databases such as Protein Data Bank (PDB) [5] and Electron Microscopy Data Bank (EMDB) [6]. By using these databases, one can analyze viral structure of interest, their functions as well as the molecular basis. Moreover, researchers/scientists working on designing potential antibody therapies and antiviral drugs rely on structural models of the virus's proteins [7]. At the time of writing this paper, PDB contains 196,779 structures in total[3], in which more than 1,100 belong to the S protein of SARS-CoV-2.

We focus on the analysis and classification/detection of S protein structures of SARS-CoV-2 and other viruses/organisms considering their availability, in large number, in the PDB. In the literature, some studies focused on the classification and detection of SARS-CoV-2 genome sequences. For example, [8]–[10] take advantage of CpG (or CG)-based features for SARS-CoV-2 genomes classification. Representative genomic sequences of SARS-CoV-2 were discovered by Lopez-Rincon et al. [11] by coupling a deep learning method with explainable AI techniques. Naeem et al. [12] developed a classification system that utilized the discrete Cosine transform, discrete Fourier transform and seven moment invariants to extract features from 76 SARS-CoV-2 genome sequences. The classification method of Randhawa et al. [13] used an intrinsic SARS-CoV-2 genomic signature with a machine learning-based alignment-free (AF) method. Ahmed and Jeon [14] classified genome sequences of four viruses (SARS-CoV-1, SARS-CoV-2, MERS and Ebola) by using using machine learning algorithms. A convolutional neural network, inspired by a cockroach optimization algorithm was used [15] for multi-classification of genomes of two viruses (SARS-CoV-2 and Influenza). Singh et al. [16] used biomarkers, that were extracted from the genome sequences of coronaviruses on the basis of three-base periodicity, for the classification of SARS-

[1]livemint.com/science/health/virus-is-running-freely-who-chief-warns-against-covid-19-infections-surge-11657685970342.html

[2]COVID19.who.int

[3]www.rcsb.org/stats/growth/growth-released-structures

CoV-2 from other coronaviruses.

Most of the aforementioned studies focused on virus genome sequences and finding important features in them that are then used for classification. To the best of our knowledge, no study has been published on the analysis and classification of protein structures, particularly those of harmful viruses, in PDB. More specifically, a novel method called S-PDB is proposed in this paper to:

1) Find and analyze the structures in PDB that are similar to the S protein structures of SARS-CoV-2.
2) Detect the S protein structures of SARS-CoV-2 and other viruses/organisms.

Three types of classification is carried out that are based on (1) amino acids (AA) sequences, (2) aligned AA (AAA) sequences, and (3) aligned secondary structure elements (ASSE) sequences. AA sequences of protein structures are downloaded from PDB and DALI [17] is used to find similar protein structures in PDB and for the AA and SSE alignments. Multinomial Naive Bayes Text (MNBT), Stochastic Gradient Descent Text (SGDT) and ZeroR are used for classification and their efficacy is accessed with six evaluation metrics. We found that MNBT and SGDT performed better on AAA compared to AA and ASSE. This shows that information from sequence alignment can be used efficiently to classify protein structures instead of using their whole AA sequences.

The rest of the paper is organized into four sections: Section II discusses the SARS-CoV-2, the S protein and the tool used for the protein structures comparison. Section III presents the proposed S-PDB method along with the details for the datasets. Section IV presents and discusses the obtained results. Finally, Section V concludes the paper with some future research opportunities.

II. Background

This section provides a brief overview of SARS-CoV-2, the S protein and the DALI tool for protein structures comparison.

A. SARS-CoV-2

SARS-CoV-2 is a positive-strand RNA virus, with spherical to pleomorphic shape and length between 80-160 nm [18]. SARS-CoV-2 contains four structural proteins (1) Spike (S), (2) Envelope (E), (3) Membrane (M) and (4) Nucleocapsid (N) (Fig. 1). The outer structure is made by S, M, and E proteins. The E protein also plays a role in the maturation and production of SARS-CoV-2. The S and M proteins are also involved in the process of virus attachment during replication. N protein form the nucleocapsid inside the envelope. The SARS-CoV-2 virus can enter the human host cell membrane by interacting with the host ACE2 receptor.

The S protein, that comprises two subunits (S1 and S2) binds itself to the ACE2 receptor in the host cells. S1 contains binding domains, receptor binding domain (RBD) and n-terminal domain (NTD). S2 contains a fusion peptide, HR1 and HR2 domains which are responsible for the virus fusion. The S protein of this virus binds to ACE2 with higher affinity than its predecessor, SARS-CoV [4]. After binding, the entry

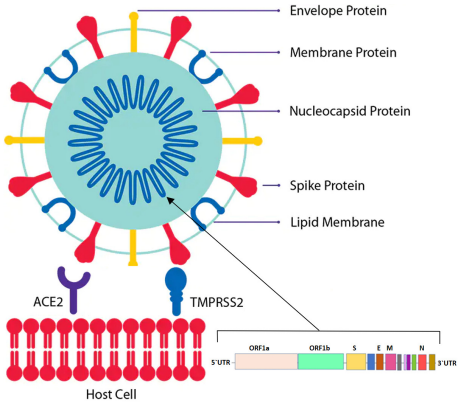

Fig. 1. SARS-CoV-2 structure and how it binds to host cell through the ACE2 receptor and TMPRSS2

depends on S protein priming process carried out by the type 2 serine protease (TMPRSS2) [19] that is present on the surface of the host cell (Fig. 1). Blocking or preventing the binding of S proteins with ACE2 receptors is considered the first and most important approach to block cell entry and stopping the COVID-19 disease. In its genome range, SARS-CoV-2 contains six to twelve open reading frames (ORFs). The main reading frame, ORF1ab that occupies two-thirds of the genome, is present at the 5'UTR (terminal region). Whereas at the 3'UTR, one third consists of genes that encode structural proteins (S, E, M and N).

The primary structure of a protein contains ordered sequence of AA residues. The secondary structure of a protein contains regions of AA chains stabilized by the hydrogen bonds from the polypeptide backbone. These bonds generate α-helix and β-sheets that contain β-strands. From the secondary structures, a protein can be folded into a stable 3D structure (the tertiary structure) [20].

B. DALI

DALI (Distance matrix alignment) software [17], [21] is used to find structures in PDB that are similar to the SARS-CoV-2 S protein structures and for their comparison and analysis. DALI optimizes a set of one-to one correspondences between two protein (sub)structures (A and B) that maximizes the DALI score:

$$DALI_{AB} = \sum_{i=1}^{LALI} \sum_{j=1}^{LALI} \left(\theta - \frac{2|d_{ij}^A - d_{ij}^B|}{d_{ij}^A + d_{ij}^B} \right) e^{-\left(\frac{d_{ij}^A + d_{ij}^B}{2D}\right)^2}$$

where $LALI$ represents the number of aligned residue pairs, θ = 0.2, D = 20 A and d_{ij}^A, d_{ij}^B are intra-molecular Cα–Cα distances in structures A and B respectively. For random pairwise comparison, the expected DALI$_{AB}$ score increases with the number of residues in the compared proteins. DALI Z-score is used to describe the statistical significance of a DALI$_{AB}$:

$$Z_{AB} = \frac{DALI_{AB} - m(L)}{\alpha(L)}$$

where $L = \sqrt{L_A L_B}$ is the geometric mean length of structures A and B. The relation between m (mean score), σ (standard deviation) and L was derived empirically from a large set of random pairs of structures. Fitting a polynomial give the following approximation:

$$m(L) = \begin{cases} 7.95 + .71L - 2.59E^{-4}L^3 - 1.92E^{-6}L^3 & \text{if } L \leq 400 \\ m(400) + L - 400, & \text{if } L > 400 \end{cases}$$

The empirical estimate for the standard deviation was $\sigma(L)$ = 0.5 × m(L). For every possible pair of domains (determined by the Puu algorithm [22]), the Z-score is computed and the highest value is reported as the Z-score of the protein pair. Thus DALI's Z-score is an optimized similarity score defined as the sum of equivalent residue-wise Cα-Cα distances among two proteins. For two proteins, the large Z-score indicates more similarity that corresponds to the optimal set of residue equivalence obtained by permuting the equivalent structural patterns by Monte Carlo optimization. A Z-score < 2 is considered as a spurious similarity and can be ignored [23].

DALI supports three types of database searches (PDB search, PDB25 and AF-DB) and two types of structure comparisons (pairwise and all against all). Proteins in secondary structure are traditionally characterized with three states: (1) helix (H), strand (E) and Coil (C). The Dictionary of Secondary Structure of Proteins (DSSP) [24] offers a finer classification of the secondary structures by extending the three general states into eight states. DALI uses the secondary structure assignments by DSSP.

III. S-PDB METHOD

The proposed S-PDB method (Fig. 2) for the analysis of protein structures and classification of S protein structures of viruses in PDB consists of two main steps:

1) *Similar protein structures identification and datasets creation*: This step consists of two main activities: (1) Identifying the protein structures that are similar to the SARS-CoV-2 S protein structures in PDB. This is done by using DALI. (2) The AA sequences of obtained similar S protein structures, downloaded from PDB, and the pairwise alignment of AA and SSE of protein structures, obtained through DALI, are stored in three datasets.
2) *S protein structures classification*: Sequence of AA, AAA and ASSE identified in Step (1)2 are used for the classification of S proteins that belong to SARS-CoV-2 and to other viruses and organisms. The classification task is composed of two main parts: (1) The training phase contains two phases, AA, AAA and ASSE representation and classifier training, that are performed sequentially. (2) The testing phase contains three phases, AA, AAA and ASSE representation, hypothesis prediction and evaluation.

The next two subsections provide more details for the two parts.

A. Similar proteins identification in PDB through DALI

The SARS-CoV-2 protein structure with PDB ID (PID) 6VSB [4] (deposited to PDB on 10 February 2020) is used as the query structure. The main reason to select 6VSB as a query structure is that it is one of the earliest S protein structures deposited in PDB. Thus, for 6VSB, DALI returned 397 structures in PDB90 search. Note that PDB90 is a non-redundant subset of PDB structures. In PDB90 subset, those structures in PDB are found that are less than 90% identical in sequence. After removing the same structures with different chains, the total structures reduced to 388.

The similar structures obtained can be divided into three types (families):

1) S protein structures of SARS-CoV-2,
2) S protein structures of other viruses and organisms, and
3) Protein (enzyme) structures for others.

In the similar protein structures obtained with PDB90 search in DALI, approximately 13.65% (53 out of 388) belong to the first type (S protein structures of SARS-CoV-2), approximately 12.37% (48 out of 388) belong to the second type (S protein structures of other viruses and organism). Remaining belonged to the third type (structures of others). The AA sequences of all 388 structures are then downloaded from the PDB. Some sequences have multiple AA sequences due to multiple chains. Thus the downloaded sequences are refined to only include the sequences for the chain which is similar to the query structure. Table I shows the number of structures that belong to each of the three families.

TABLE I
STRUCTURES DISTRIBUTION ACCORDING TO THEIR FAMILIES

Structures	Samples	AA	FAA	MinL, MaxL, ASL
SARS-CoV-2 S	53	20	L,S,T,V, G	127, 1380, 1074
Other S	48	20	S,L,V,T,G	135, 1469, 847
Others	287	22	L,G,S,V,A	69, 4646, 375
Total	388	22	L,S,G,V,T	69, 4646, 526

FAA: Frequent AA, MinL: Minimum Length, MaxL: Maximum Length, ASL: Average Sequence Length

The five most frequent AA in SARS-CoV-2 S are Leucine (L) (8.30%), Serine (S) (8.05%), Threonine (T) (7.43%), Valine (V) (7.42%) and Glycine (G) (7.11%). In Other S, the five most frequent AA are: S (8.52%), L (8.50%), Valine (V) (7.78%), T (7.39%) and G (6.76%). The five most frequent AA in Others are: L (8.05%), G (7.62%), S (7.42%), V (6.73%) and Alanine (A) (6.48%). The third family (Others) has 22 distinct AA because the amino acid B that can be either Asparagine (N) amino acid or Aspartic (D) amino acid was present in one structure and the amino acid X that can be any of the 20 AA was present once in multiple structures. As mentioned earlier, AA sequences of structures that belong to two families (SARS-CoV-2 S and Other S) are stored in a dataset. Similarly, the AAA and ASSE of structures that belong to SARS-CoV-2 S and Other S are also stored in their respective datasets. Thus, we have three datasets for: (1) AA sequences, (2) AAA and (3) ASSE. More details for the datasets of AAA and ASSE are provided in the results sections.

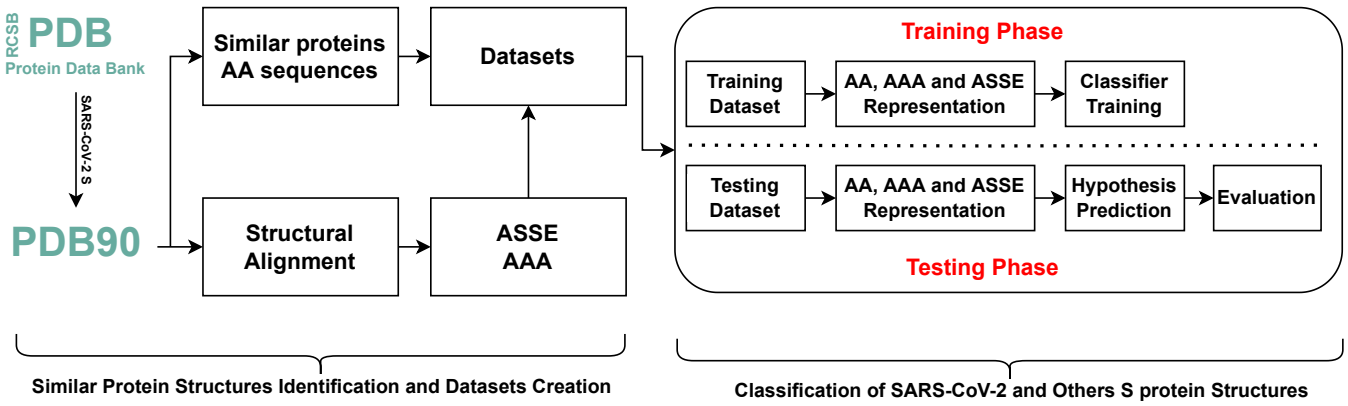

Fig. 2. Schematic of the S-PDB method for the analysis of proteins structures in PDB and for the classification of SARS-CoV-2 S vs Other S

B. Classification

The second step performed by the proposed S-PDB is to classify protein structures according to the first two types using the AA, AAA and ASSE sequences. Binary classification is carried out on three datasets to train a model to classify two structure types separately. For a selected structure type, binary classification assigns "class name" to each AA, AAA and ASSE sequences corresponding to that type.

Evaluation metrics: We use six metrics to evaluate the performance of classifiers, which are: (1) accuracy, (2) false positive rate (FPR), (3) recall, (4) precision, (5) F1 score and (6) Matthews correlation coefficient (MCC). In this work, the accuracy (ACC) is defined as the proportion of correctly classified S proteins structures of SARS-CoV-2 divided by the total S protein structures. The formal definition of ACC is:

$$ACC = \frac{TP + TN}{TP + TN + FP + FN}$$

whereas in the context of this paper,
TP stands for true positive, i.e. number of protein structures correctly identified as belonging to a given protein structures type,
FP stands for false positive, i.e. number of of protein structures incorrectly identified as belonging to a given protein structures type,
FN stands for false negative i.e. number of of protein structures incorrectly identified as not belonging to a given of protein structures type, and
TN stands for true negative i.e. number of of protein structures correctly identified as not belonging to a given protein structures type.

The other five measures, FPR, precision, recall, f-measure and MCC are calculated as follows:

$$FPR = \frac{FP}{FP + TN}$$

$$Precision(P) = \frac{TP}{TP + FP}$$

$$Recall(R) = \frac{TP}{TP + FN}$$

$$F - measure = 2 \times \frac{P \times R}{P + R}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Classifiers: Three machine learning algorithms are used, which are: (1) MNBT (Multinomial Naive Bayes Text), (2) SGDT (Stochastic Gradient Descent Text) and ZeroR [25]–[27]. MNBT is a probabilistic learning algorithm based on the Bayes theorem. SGDT is a generic optimization algorithm that uses stochastic gradient descent for learning a linear binary class SVM or binary class logistic regression on text data. ZeroR is a simple classification algorithm that relies on the target and all predictors are ignored. ZeroR is selected as a benchmark for MNBT and SGDT. Standard 10-fold cross validation is used to evaluate the performance of the classifiers.

IV. Results

The experiments are performed on a workstation with a fifth-generation Core i7 processor and 32 GB of RAM. The open-source WEKA software [27], developed in Java, is used to train the classifiers. WEKA is selected because it can run on various platforms and offers not only classifiers for machine learning but also tools for data preparation and meta learners. Moreover, it also provides a GUI, along with its CLI, that is very easy to use.

First, the results for the pairwise sequence alignment in DALI for the 12 similar structures against the query structure is presented. DALI aligned more than 970 AA in 12 structures and the first 200 AA alignment is shown in the upper block of Fig. 3. The most frequent AA in each column (structure) are colored. The uppercases letters represents those positions that are structurally equivalent with the query structure. The second part (lower block) in Fig. 3 shows the secondary structure states (assigned by the DSSP). From the lower part, the two most frequent SSE are: Coil or turn (L), followed by β-sheet (E) and α-helix (H). Note that two structure (6NB7 and 6CS2) belong to SARS-CoV, 6JX7 to Feline infectious peritonitis

Fig. 3. AAA and ASSE in 12 structures obtained by using DALI. 6VSB is used as the query structure

(FIP) virus, 5I08 to Human Coronavirus-HKU1, 6M5Y to sugar binding protein and 50CQ to hydrolase enzyme.

DALI also reports the root-mean square deviation (RMSD) of aligned Cα-atoms, LALI (number of aligned Cα-atoms), NRES (number of AA residues in the target structure) and IDEN (% identity of AAA) for similar protein structures (Table II). The goal in DALI is to maximize a geometrical similarity score, which is defined in terms of similarities of intramolecular distances. Therefore, DALI does not generate alignments with low RMSD. An alignment is considered "better" if it has both smaller RMSD and larger LALI. If both RMSD and LALI are smaller or larger, it is not possible to establish an order between the alignments. Along with the AA sequences for protein structures of SARS-CoV-2 and other viruses/organisms, their AAA and ASSE are also stored that are used for the classification.

TABLE II
DALI RESULTS FOR 12 STRUCTURES

Structures PID	Z-score	RMSD	LALI	NRES	IDEN
7AD1B	48.8	3.6	792	935	97
7Q6QA	48.2	1.6	965	1013	99
7RA8A	48.1	2.7	762	917	97
7N9CB	48.1	1.7	778	812	99
7SN3A	47.8	2.9	803	925	97
6NB7B	40.7	5.0	822	1032	77
6CS2C	35.7	2.0	797	893	78
6ZOZC	32.9	3.3	964	1070	99
6JX7A	28.0	8.7	604	1245	28
5I08A	27.4	3.9	801	958	31
6M5YA	8.4	8.1	132	270	8
5OCQB	5.7	3.8	136	279	6

Two classifiers (MNBT and SGDT) are used with three tokenization strategies: (1) WordTokenizor (WT), (2) NGram-Tokenzior (NGT) and (3) CharacterNGramTokenzior (CNGT). The first one is a simple technique to tokenize the strings. The second tokenzior splits a string into an n-gram with user specified minimum and maximum grams. Whereas, the third tokenizor splits a string into all character n-grams it contains on the basis of user specified maximum and minimum for n. In both NGT and CNGT, the maximum and minimum grams were set to 3 and 1 respectively. Obtained results are provided in Table III. Two strategies WT and NGT generated the same results for both classifiers. Whereas CNGT strategy performed better than WT and NGT on both classifiers. Interestingly, the results for the ZeroR on AA, AAA and ASSE for various parameters were the same as MNBT's results with WT and NGT strategies.

On AA dataset, SGDT with CNGT strategy performed better than MNBT with the same strategy. On AAA dataset SGDT with CNGT strategy performed similar to MNBT with the same strategy. On ASSE datasets, the MNBT with CNGT strategy performed better than SGDT with the same strategy. On three datasets, SGDT with different strategies was slow compared to MNBT. The confusion matrix in Fig. 4 is for the MNBT and SGDT with CNGT strategy. The format $AA\frac{AAA}{ASSE}$ is used. The entries outside bracket is for the MNBT and inside the bracket is for the SGDT.

TABLE III
CLASSIFIERS PERFORMANCE ON THREE DATASETS WITH DIFFERENT TOKENIZATION STRATEGIES

Type	P	MNBT		SGDT		ZeroR
		WT (NGT)	CNGT	WT(NGT)	CNGT	
AA	ACC	52.47	87.12	52.47	**91**	52.47
	FPR	0.525.	0.126	0.440	**0.091**	0.525
	P	?	0.873	0.587	**0.911**	?
	R	0.525	0.871	0.525	**0.911**	0.525
	F1	?	0.871	0.459	**0.911**	?
	MCC	?	0.744	0.119	**0.821**	?
AAA	ACC	52.47	**92**	52.47	**92**	52.47
	FPR	0.525.	**0.082**	0.525	**0.082**	0.525
	P	?	**0.921**	?	**0.921**	?
	R	0.525	**0.921**	0.525	**0.921**	0.525
	F1	?	**0.921**	?	**0.921**	?
	MCC	?	**0.842**	?	**0.842**	?
ASSE	ACC	52.47	**84.14**	52.47	71.28	52.47
	FPR	0.525.	**0.159**	0.525	0.291	0.525
	P	?	**0.842**	?	0.713	?
	R	0.525	**0.842**	0.525	0.713	0.525
	F1	?	**0.842**	?	0.712	?
	MCC	?	**0.682**	?	0.423	?

We performed paired t-test (corrected) in Weka to check which of three classifiers are significantly better than others. We selected ZeroR as the baseline. Both MNBT and SGDT with CNGT strategy performed significantly better than ZeroR. For MNBT and SGDT, the later performed better than the

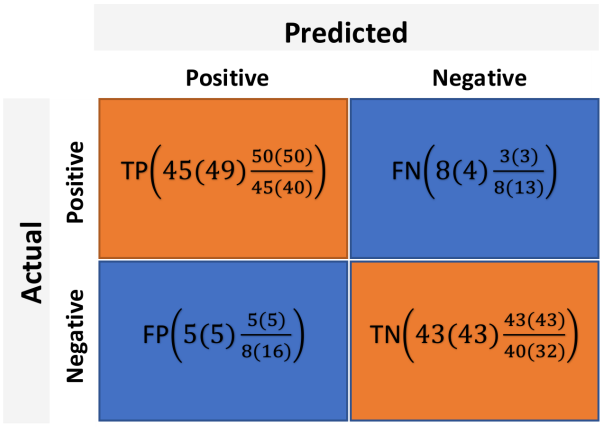

Fig. 4. Confusion matrix for two classifiers on three datasets

former on AA dataset while the opposite is true for the ASSE dataset. On AAA dataset, both classifiers' performance was the same. The test results confirmed that the difference in the performance of MNBT and SGDT on three datasets is not that significant.

TABLE IV
CLASSIFIERS PERFORMANCE ON DATASETS THAT CONTAIN VARYING NUMBER OF AAA AND ASSE

Type	P	MNBT	SGDT
AAA$\left(100\left(\frac{200}{300}\right)\right)$	ACC	89.2$\left(\frac{90.1}{92}\right)$	83.3$\left(\frac{91}{93}\right)$
	FPR	0.103$\left(\frac{0.098}{0.082}\right)$	0.169$\left(\frac{0.085}{0.073}\right)$
	P	0.897$\left(\frac{0.902}{0.921}\right)$	0.833$\left(\frac{0.915}{0.932}\right)$
	R	0.892$\left(\frac{0.901}{0.921}\right)$	0.833$\left(\frac{0.911}{0.931}\right)$
	F1	0.892$\left(\frac{0.901}{0.921}\right)$	0.833$\left(\frac{0.911}{0.931}\right)$
	MCC	0.789$\left(\frac{0.803}{0.842}\right)$	0.665$\left(\frac{0.826}{0.862}\right)$
ASSE$\left(100\left(\frac{200}{300}\right)\right)$	ACC	81.18$\left(\frac{78.21}{79.2}\right)$	79.2$\left(\frac{75.2}{73.2}\right)$
	FPR	0.194$\left(\frac{0.233}{0.210}\right)$	0.212$\left(\frac{0.246}{0.272}\right)$
	P	0.814$\left(\frac{0.803}{0.792}\right)$	0.793$\left(\frac{0.754}{0.733}\right)$
	R	0.812$\left(\frac{0.782}{0.792}\right)$	0.792$\left(\frac{0.752}{0.733}\right)$
	F1	0.811$\left(\frac{0.776}{0.792}\right)$	0.792$\left(\frac{0.753}{0.732}\right)$
	MCC	0.624$\left(\frac{0.581}{0.583}\right)$	0.583$\left(\frac{0.506}{0.463}\right)$

AAA and ASSE sequences are further analyzed by only considering some of their parts. For example, the first 100, 200 and 300 AAA and ASSE are used for the classification instead of whole AAA and ASSE sequences. The reason to consider three different numbers of patterns (100, 200 and 300) is to see whether different patterns count has any effect on the performance of classifiers. The results are listed in Table IV. The results for classifier metrics are shown with the following format: 100 $\left(\frac{200}{300}\right)$. For example, consider the first entry of 89.2$\left(\frac{90.1}{92}\right)$. It indicates that MNBT achieved ACC of 89.2% on the dataset that contains first 100 AAA of structures, 90.1% ACC on first 200 AAA of structures and 92% ACC on first 300 AAA of structures respectively. Note that the results in Table IV for the classifiers are with the CNGT strategy. We found some interesting results. For AAA, the performance of MNBT and SGDT increased with increase in the length of AAA. The opposite is true for ASSE, where MNBT and SGDT performance decreased with the increase in the ASSE length. Infect, for AAA and ASSE, SGDT performance was better at first 300 AAA and first 100, 200 and 300 ASSE compared to the whole AAA and ASSE sequences respectively. The datasets used for the classification experiments are available at *github.com/saqibdola/S-PDB*.

V. CONCLUSION

In this study, a novel method (named S-PDB) is developed that first finds the protein structures in PDB that are similar to the SARS-CoV-2 Spike proteins. The similar S protein structures of SARSCoV-2 and other viruses and organisms are then classified by using (1) AA sequences, (2) AAA and ASSE, that are obtained by using a protein structures comparison tool. Three classifiers were used to reliably predict/classify and their performance was checked against six metrics. We found that the two classifiers (MNBT and SGDT) performance on AAA was high, followed by AA and ASSE. This shows that information obtain from sequence alignment can be used efficiently to classify protein structures instead of using their whole AA sequences. The developed method is not limited to the S protein structures but can be used for other protein structures too. For future, some research directions are:

- Extending the method to (1) classify non-S protein structures in PDB and (2) analyze and classify the S protein structures of SARS-CoV-2 that belongs to various variant families such as Alpha, Delta Omicron, etc.
- Using alignment-free methods [28], [29] for comparison of AA sequences of protein structures.
- Using pattern mining techniques such as sequential pattern mining [30] and emerging or contrast pattern mining [31] to find similar (contrasting) frequent patterns in AA, AAA and ASSE for the analysis and classification of protein structures in PDB.

REFERENCES

[1] F. Wu et al. A new coronavirus associated with human respiratory disease in China. *Nature*, 579(7798):265–529, 2020.

[2] F. Konings et al. SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nature Microbiology*, 6:821–823, 2021.

[3] J. Lan et al. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. *Nature*, 581(7807):215-220, 2020.

[4] D. Wrapp Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. *Science*, 367(6483): 1260-1263, 2020.

[5] S. K. Burley et al. Protein Data Bank (PDB): The single global macromolecular structure archive. In: *Protein Crystallography. Methods in Molecular Biology*, vol 1607. Humana Press, NY, USA, 2017.

[6] C. L. Lawson et al. EMDataBank.org: Unified data resource for Cryo-EM. *Nucleic Acids Research*, 39 (S1, 1): D456–D464, 2011.

[7] S. K. Burley. Impact of structural biologists and the Protein Data Bank on small-molecule drug discovery and development. *Journal of Biological Chemistry*, 296:100559, 2021.

[8] H. Arslan Machine learning methods for COVID-19 prediction using human genomic data. *Proceedings*, 74(1): 20, 2021.

[9] H. Arslan and H, Arslan. A new COVID-19 detection method from human genome sequences using CpG island features and KNN classifier. *Engineering Science and Technology, an International Journal*, 24(4): 839-847, 2021.

[10] H. Arslan. COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus. Comput Ind Eng 161: 107666, 2021.

[11] Lopez-Rincon et al. Classification and specific primer design for accurate detection of SARS-CoV-2 using deep learning. *Scientific Reports*, 11, 947, 2021.
[12] S. M. Naeem et al. A diagnostic genomic signal processing (GSP)-based system for automatic feature analysis and detection of COVID-19. *Briefings in Bioinformatics*, 22(2):1197-1205, 2021.
[13] G. S. Randhawa et al. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *PLoS One*, 15(4): e0232391, 2020.
[14] I. Ahmed and G. Jeon. Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses. *Interdisciplinary Sciences: Computational Life Sciences*, 14(2): 504–519, 2022.
[15] M. A. El-dosuky, M. Soliman, and A. E. Hassanien. COVID-19 vs Influenza viruses: A cockroach optimized deep neural network classification approach. *International Journal of Imaging Systems and Technology*, 31: 472– 482, 2021.
[16] O. P. Singh et al. Classification of SARS-CoV-2 and non-SARS-CoV-2 using machine learning algorithms. *Computers in Biology and Medicine*, 136:104650, 2021.
[17] L. Holm. Dali server: Structural unification of protein families. *Nucleic Acids Research*, 50(W1): W210-W215, 2022.
[18] M. S. Nawaz et al. Using artificial intelligence techniques for COVID-19 genome analysis. *Applied Intelligence*, 51(5):3086-3103, 2021.
[19] M. Hoffmann et al. SARS-CoV-2 cell entry depends on ACE2 and TMPRSS2 and is blocked by a clinically proven protease inhibitor, *Cell*, 181(2):271-280.e8, 2020.
[20] A. Sekmen, K. A. Nasr and C. Jones. Subspace modeling for classification of protein secondary structure elements from Cα trace. In: *Proceedings of BIBM*, pp. 72-97, 2021 .
[21] L. Holm. DALI and the persistence of protein shape. *Protein Science*, 29: 128– 140, 2020.
[22] G. Csaba, F. Birzele and R. Zimmer. Protein structure alignment considering phenotypic plasticity. *Bioinformatics*, 24:98–104, 2008.
[23] L. Holm. Using DALI for protein structure comparison. In: *Structural Bioinformatics. Methods in Molecular Biology*, vol 2112. Humana, NY, USA, 2020.
[24] W. Kabsch and C. Sander. Dictionary of protein secondary structure: Pattern recognition of hydrogen bonded and geometrical features. *Biopolymers*, 22: 2577-2637, 1983.
[25] R. J. Urbanowicz and W. N. Browne. *Introduction to Learning Classifier Systems*. 1st Edition, Springer, 2017.
[26] Yang X-S. *Introduction to Algorithms for Data Mining and Machine Learning*. Elsevier; 2019
[27] E. Frank, M. A. Hall and I. H. Witten. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Fourth Edition. Morgan Kaufmann, 2016
[28] A. Zielezinski et al. Benchmarking of alignment-free sequence comparison methods. *Genome Biology*, 20: 144, 2019.
[29] M. S. Nawaz et al. COVID-19 genome analysis using alignment-free methods. In: *Proceedings of IEA/AIE*, pp. 316–328, 2021.
[30] P. Fournier-Viger et al. A survey of sequential pattern mining. *Data Science and Pattern Recognition*, 1(1):54–77, 2017.
[31] S. Ventura and J. M. Luna. *Supervised Descriptive Pattern Mining*, Springer, 2018.