Spike Protein Undeformable Motif shared by SARS-CoV-2 and SARS-CoV: Flexible Conformations Predicted by using Deep Neural Network–based Programs of Supersecondary Structure Codes

Hiroshi Izumi*

National Institute of Advanced Industrial Science and Technology (AIST), Tsukuba, Ibaraki 305-8569, Japan

A short running title: Undeformable Motif shared by SARS-CoV-2 and SARS-CoV

*Correspondence to: Hirosi Izumi; National Institute of Advanced Industrial Science and Technology (AIST), AIST Tsukuba West, 16-1 Onogawa, Tsukuba, Ibaraki 305-8569, Japan. E-mail: izumi.h@aist.go.jp

Acknowledgments: This work was supported partly by JSPS KAKENHI Grant Number JP19K05431. The author also thanks native-English-speaking professional editors from ELSS, Inc. for English proofreading.

ABSTRACT

A deep neural network-based program for sequence-based prediction of supersecondary structure codes (SSSCs), called SSSCPrediction (SSSCPred) was constructed. Furthermore, to predict the flexibility and conformational change of proteins, a comparison program of three deep-neural-network-based prediction systems (SSSCPred200, SSSCPred100, and SSSCPred) was developed. I compared the predicted and observed flexible conformations of SARS-CoV-2 and SARS-CoV spike proteins by using SSSCs and the comparison program. The SARS-CoV SSSC sequences of the receptor-binding motif predicted by the three deep-neural-network-based systems well reproduced those of the Protein Data Bank (PDB) data, including the structured loops. In contrast, the receptor-binding motif SSSCs of SARS-CoV-2 differs greatly from those of SARS-CoV, with that of SARS-CoV-2 being more flexible. Only one common identical motif (SSSC: SSSHSSHHHH) among all of the compared SSSC sequences, including predicted and observed ones, was found at the S2 subunit. This motif has an extremely rare and relatively undeformable conformation. The comparison program may be helpful to explore undeformable drug discovery targets of many unsolved protein structures.

Key words: Conformation; Deep neural network; SARS-CoV-2; Sequence-based prediction; Supersecondary structure code.

INTRODUCTION

There are around 110.3 million non-redundant protein sequences in the RefSeq database,[1,2] and many methods for sequence-based prediction of secondary and supersecondary structures have been developed in the past several years.[3-11] Further, many secondary structure prediction methods based on deep learning have also been reported.[12-16] However, the classification and prediction of fine structured loops other than α-helixes, β-strands, coiled coils,[17-19] and disordered regions[20,21] remains elusive.

Intrinsically disordered regions are protein regions that undergo conformation changes and therefore lack a stable three-dimensional structure. It has been shown that intrinsically disordered regions are important for protein–protein interactions and the binding of proteins to RNA and DNA; therefore, there is a need for an accurate way of predicting the conformations of these regions.[22,23]

In the past decade, a means of identifying and codifying supersecondary structures (supersecondary structure code; SSSC) has been developed that uses the concept of Ramachandran plot data[24-26] with ω angles, and the specification of positions of torsion angles in a protein derived from a fuzzy search of structural code homology using template patterns, represented as conformational codes 3a5c4a (α-helix-type conformation) and 6c4a4a (β-sheet-type conformation), to describe supersecondary structural motifs and their conformation.[27,28] SSSC is transcribed using the letters H, S, T, and D to refer to an α-helix-type conformation, a β-sheet-type conformation, other-type conformations, and disordered residues or the C-terminus, respectively.

The DSSP (Dictionary of Secondary Structure in Proteins) program has been used to standardize secondary structure assignments for all of the protein entries in the Protein Data Bank (PDB).[29,30] This program is able to identify hydrogen bonds between main chain carbonyl groups and amide

groups and then use that information to assign a secondary structure; however, it does not handle the fine characterization of loops or irregular regions very well because no hydrogen bonds exist in such regions. In contrast, supersecondary structure code is very sensitive to differences among similar protein supersecondary structures. For example, this code can distinguish the difference of characteristic loop structures between IgG immunoglobulin (SSSC: SHHSHSS) and IgM rheumatoid factor (SSSC: TTTSSSS).[27,28] Interferon α, β, and γ, GroEL, and ubiquitin-associated domains also have a unique common motif (SSSC: HHHTTSHHH).[27]

Recently, a deep-neural-network-based program for sequence-based prediction of SSSCs, called SSSCPrediction (SSSCPred) was constructed. For files containing the keywords ALKALINE, GLUCOSIDASE, OUTER MEMBRANE, ENVELOPE, PORIN, REPLICATION, INTERLEUKIN, and RIBOSOMAL PROTEIN, the concordance rate was <0.60; however, these keywords were sometimes found in files with concordance rate ≥ 0.90 and were associated with flexible conformations. Furthermore, to predict the flexibility and conformational change of proteins, a comparison program of three deep-neural-network-based prediction systems (SSSCPred200, SSSCPred100, and SSSCPred) was developed.

To develop a vaccine against the coronavirus disease 2019 (COVID-19), which is currently prevalent all over the world, structural information on the virus is required.[31] The sequence of severe acute respiratory syndrome coronavirus (SARS-CoV) moderately resembles that of SARS-CoV-2 (about 79% identity).[31] Several observed structures in SARS-CoV[32-37] and SARS-CoV-2,[38-44] including cryo-electron microscopy (Cryo-EM) structures, have been registered in the PDB database[30] and are thus available for use in comparing the predicted SSSCs of SARS-CoV-2.

Here, I show the construction of the deep-neural-network-based program for sequence-based prediction of SSSCs (SSSCPrediction) and the comparison program of three deep-neural-network-

based prediction systems (SSSCPred200, SSSCPred100, and SSSCPred) and that the receptor-binding motif (binding to human angiotensin-converting enzyme 2, ACE2) SSSCs of SARS-CoV-2 differs greatly from those of SARS-CoV, with that of SARS-CoV-2 being more flexible. I also describe the only shared, relatively undeformable motif (SSSC: SSSHSSHHHH), which in SARS-CoV-2 S2 subunit is associated with cell adhesion and cell division.

MATERIALS AND METHODS

Dataset

A total of 582,813 FASTA-format files containing the amino acid sequences and SSSCs of protein subunits were extracted from 139,932 PDB files[30] by using the SSSCview program (available online at https://researchmap.jp/multidatabases/multidatabase_contents/detail/256924/6216cafbe7d56e9a65c649886edcb0a3?frame_id=708960).[27] Of these FASTA files, 379,334 files containing subunits with more than or equal to 100 continuous amino acid residues were extracted, and from those files 150,000 files as training data for the deep neural network, 10,000 files as test data for the deep neural network, and three sets of 10,000 files as test data for the inference system were randomly selected.

From each FASTA file, a set of 100 continuous amino acid residues and the corresponding SSSC were randomly extracted. SSSC terms “H”, “S”, “T”, and “D” were converted to [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1], respectively, and a set of matrices (100, 4) was constructed. The amino acid sequence was also similarly converted.[45] The dataset for the deep neural network was prepared by using python.[46]

SSSCPrediction

Deep learning for the prediction of SSSCs from amino acid sequences was performed by using Neural Network Console (https://dl.sony.com/app/). The revised template of network "12_residual_learning.sdcproj" for the standard MNIST dataset was used to provide the initial structure of the deep neural network, which was then trained with our prepared training dataset. The obtained network is shown in Figure 1 (activation function: ReLU; cost function: HuberLoss; max epoch: 20; batch size: 64; precision: float; structure search: Network Feature + Gaussian Process; updater: Adam; update interval: 1 iteration; alpha: 0.001; beta1: 0.9; beta2: 0.999; epsilon: 1E-8). The obtained network and parameters were introduced to the SSSCPrediction inference system, and the system was set to examine amino acid sequences containing at least 100 amino acid residues. For each amino acid sequence, SSSC terms were predicted for every 50 continuous amino acid residues and for the initial and final 100 amino acid residues in the sequence. Then, the first 70 SSSC terms in the sequence were selected, followed by every 50 SSSC terms; any remaining SSSC terms at the end of the sequence were also selected. The other prepared three sets of 10,000 test data files for the SSSCPrediction inference system were then used to evaluate concordance rate.

Comparison of SSSCPrediction with Quick2D[6] was carried out by using an amino acid sequence of a PDB file (1a00_A). The method was benchmarked by using 612 and 17,169 protein subunits containing at least 100 amino acid residues in the CB513 and CullPDB datasets.[13] The CASP10, CASP11 and CASP12 datasets could not be used for the benchmark because all of the correct answer data could not be obtained for SSSCPrediction. The 150,000 training data files and the 10,000 test data files for the prediction of SSSCs from amino acid sequences using the deep neural network were also tested to evaluate concordance rate.

SSSCPrediction is available as a standalone program at https://researchmap.jp/multidatabases/multidatabase_contents/detail/256924/8fe07c64e364d8218108144f0d33c142?frame_id=708960.

Comparison program of three deep-neural-network-based prediction systems

I constructed two additional deep-neural-network-based prediction systems by using procedures similar to that used to construct SSSCPrediction (SSSCPred). A total of 582,666 FASTA-format files containing the amino acid sequences and SSSCs of protein subunits were extracted from 139,932 PDB files[30] by using the SSSCview program.[27] Of these FASTA files, 207,738 files containing subunits with more than or equal to 200 continuous amino acid residues were extracted, and from those files 150,000 files as training data for the deep neural network, 10,000 files as test data for the deep neural network, and 10,000 files as test data for the inference system were randomly selected for SSSCPred200. From each FASTA file, a set of 200 continuous amino acid residues and the corresponding SSSC were randomly extracted. SSSC terms “H”, “S”, “T”, and “D” were converted to [1,0,0,0], [0,1,0,0], [0,0,1,0], and [0,0,0,1], respectively, and a set of matrices (200, 4) was constructed. The amino acid sequence was also similarly converted. Deep learning for the prediction of SSSCs from amino acid sequences was performed by using Neural Network Console (https://dl.sony.com/app/). The revised template of network “12_residual_learning.sdcproj” for the standard MNIST dataset was used to provide the initial structure of the deep neural network, which was then trained with the prepared training dataset. The obtained network and parameters were introduced to the SSSCPred200 inference system, and the system was set to examine amino acid sequences containing at least 200 amino acid residues. For each amino acid sequence, SSSC terms were predicted for every 50 continuous amino acid

residues and for the initial and final 200 amino acid residues in the sequence. Then, the first 125 SSSC terms in the sequence were selected, followed by every 50 SSSC terms; any remaining SSSC terms at the end of the sequence were also selected.

Training data of 200 continuous amino acid residues and 150,000 subunits were used to construct SSSCPred200; those of 100 continuous amino acid residues and 350,000 subunits were used for SSSCPred100; and those of 100 continuous amino acid residues and 150,000 subunits were used for SSSCPred. The systems well reproduced many SSSCs of the PDB subunit data; the benchmarks (average concordance rates) of the three systems were as follows: for SSSCPred200, CullPDB,[13] 0.905 (9851 subunits) and CB513,[13] 0.911 (361 subunits); for SSSCPred100, CullPDB, 0.896 (17,169 subunits) and CB513, 0.907 (612 subunits); and for SSSCPred, CullPDB, 0.861 (17,169 subunits) and CB513, 0.882 (612 subunits).

RESULT AND DISCUSSION

Translation of amino acid sequences to SSSCs

The comparison of SSSCPrediction with Quick2D[6] was carried out by using the PDB file (1a00_A). As shown in Figure 2, the main difference between SSSCPrediction and Quick2D was found in the structured loop regions. Only SSSCPrediction could predict the fine loop conformations. Although a direct comparison could not be made because of the difference of correct data between SSSCPrediction and other prediction methods, the concordance rates for the translation of amino acid sequences to SSSCs using 612 and 17,169 protein subunits containing at least 100 amino acid residues in the CB513 and CullPDB datasets[13] for the benchmark of SSSCPrediction were 0.88 and 0.86, respectively.

The average concordance rate for the translation of amino acid sequences to SSSCs using the three test datasets comprising 10,000 FASTA files each was 0.90. A total of 3450 files in the test dataset had a concordance rate ≥ 0.95, and 6000 files had a concordance rate ≥ 0.90 (Figure 3). In the past three decades, much progress has been made in the development of accurate predictors of protein secondary structure. Recently, prediction accuracy has increased from about 82% to 84%, which is approaching the estimated upper accuracy limit of around 88%.[3,12-14] Although a direct comparison of accuracy is impossible due to the differences between secondary structures and supersecondary structures, these prediction accuracies are comparable.

The correlation between keywords in the training files and concordance rate was examined to understand more about the target subunits for SSSCPrediction. For files containing the keywords PROTEASOME, FAB, LYSOZYME, HEMOGLOBIN, MICROGLOBULIN, HLA, and MYOGLOBIN, the ratio of files with that keyword and concordance rate ≥ 0.90 to total no. of files with that keyword was extremely high (≥0.92; Table 1). In contrast, for files containing the keywords ALKALINE (ratio of files with that keyword and concordance rate ≥ 0.90 to total no. of files with that keyword: 4/97), GLUCOSIDASE (81/245), OUTER MEMBRANE (158/362), ENVELOPE (122/271), PORIN (126/262), REPLICATION (134/271), INTERLEUKIN (265/472), and RIBOSOMAL PROTEIN (1259/1908), the concordance rate was much lower (<0.60); however, these keywords were sometimes found in files with concordance rate ≥ 0.90 and were associated with flexible conformations, and there were no keywords found only in files with a low concordance rate. In the 379,334 files in the overall dataset, the keywords KINASE (4219/6080), TRANSFERASE (3812/6010), SYNTHASE (2868/4159), REDUCTASE (3050/4302), DEHYDROGENASE (2545/3815), HYDROGENASE (2732/4120), POLYMERASE (1863/2888), HYDROLASE (1199/2041), PROTEASE (1344/1765),

PHOSPHATASE (990/1,690), ISOMERASE (1279/1912), and OXIDASE (1086/1682) frequently appeared, and the ratios of files with that keyword and concordance rate ≥ 0.90 to total no. of files with that keyword ranged from 0.59 to 0.76. Thus, there were no keywords associated only with a low concordance rate, and many files with the same keywords were found to have different SSSCs.

To confirm whether the flexibility and conformational change of proteins can be predicted or not, I constructed two additional deep-neural-network-based prediction systems by using procedures similar to that used to construct SSSCPrediction (SSSCPred). The benchmarks (average concordance rates) of the three systems were as follows: for SSSCPred200, CullPDB,[13] 0.905 (9851 subunits) and CB513,[13] 0.911 (361 subunits); for SSSCPred100, CullPDB, 0.896 (17,169 subunits) and CB513, 0.907 (612 subunits); and for SSSCPred, CullPDB, 0.861 (17,169 subunits) and CB513, 0.882 (612 subunits). For CullPDB files, total no. of files with that concordance rate < 0.65 between SSSCPred200 and PDB data was 66. Of these CullPDB files, the ratio of files with that concordance rate < 0.70 between SSSCPred200 and SSSCPred100 data to total no. of files was 0.83 (see Table S1). For CB513 files, total no. of files with that concordance rate < 0.75 between SSSCPred200 and PDB data was 17. Of these CB513 files, the ratio of files with that concordance rate < 0.80 between SSSCPred200 and SSSCPred100 data to total no. of files was 0.59 (see Table S2). Exceptionally, in the CB513 files, the subunit with the keyword PHOSPHOGLYCERATE MUTASE 1 (3pgm_A) showed the high concordance rate (0.91) between SSSCPred200 and SSSCPred100 data in contrast with the low concordance rate (0.62) between SSSCPred200 and PDB data. In that case, the PDB file (1qhf_A) of the same keyword with the high concordance rate (0.96) between SSSCPred200 and PDB data was found. This means that the PDB files 3pgm_A and 1qhf_A have the identical amino acid sequence, but the SSSC

sequences, which reflect the subunit flexibility, are largely different. The value size of concordance rates among the three systems provides a good indication of the flexibility of the protein subunits.

Predicted and observed SSSC sequences of SARS-CoV-2 spike protein

I then compared the predicted and observed SSSC sequences of spike proteins of SARS-CoV-2 and SARS-CoV at the receptor-binding domain (Figure 4; see Figure S1 for complete sequences). The SSSC sequences of SARS-CoV predicted by the three deep-neural-network-based systems well reproduced those of the PDB data (6acc_A, 5xlr_A, 5x58_A, and 5wrg_A), including the structured loops. The observed SSSC sequence of SARS-CoV-2 main protease (6lu7_A) corresponded well to the predicted ones (av. 0.919, see Figure S2). In contrast with the relatively undeformable receptor-binding motif (binding to human ACE2) of SARS-CoV, the corresponding motif of SARS-CoV-2 indicated the possibility of conformational change between the α-helix and β-strand. This possibility was also supported by a Quick2D analysis, including a series of secondary structure predictions (Figure 4).[6] Actually, the receptor-binding motif SSSCs of SARS-CoV-2 with blanks for the Cryo-EM structure data of the entire SARS-CoV-2 spike protein (6vsb, 6vxx, and 6vyb) differs greatly from those of SARS-CoV, with that of SARS-CoV-2 being more flexible (Figure 4). On the other hand, the receptor-binding motif SSSCs of SARS-CoV-2 connected with human ACE2 for the Cryo-EM or X-ray structure data of the partial receptor-binding domain (6m17, 6vw1_E, 6lzg_B, 6m0j_E, and 6w41_C) are very similar to those of SARS-CoV. Wrapp and coworkers reported that although spike protein S1 of SARS-CoV-2 binds human ACE2 with higher affinity than that of SARS-CoV, several published SARS-CoV receptor-binding-domain-specific monoclonal antibodies do not have appreciable binding to that of SARS-CoV-2.[38] Yuan and coworkers described that a neutralizing antibody previously isolated from a

convalescent SARS patient, in complex with the receptor-binding domain of the SARS-CoV-2 spike protein, targets a highly conserved epitope, distal from the receptor-binding site, that enables cross-reactive binding between SARS-CoV-2 and SARS-CoV.[44] The observed SSSCs of the highly conserved epitope (6w41_C) resembles those of SSSCPred100, and the gap of the epitope SSSCs among the three systems is smaller than that of the receptor-binding-motif SSSCs (see Figure S1). It is suggested that although the binding of receptor-binding motif to human ACE2 stabilizes the connected conformation, the flexibility of receptor-binding motif in SARS-CoV-2 disturbs the appreciable binding of the SARS-CoV receptor-binding-motif-specific monoclonal antibodies.

The sequence identity of spike protein S2 between SARS-CoV-2 and SARS-CoV (aa 668 to 1255, about 90% identity) was greater than that of S1 (see Figure S1). Only one identical motif (SSSC: SSSHSSHHHH) among all of the compared SSSC sequences, including predicted and observed ones, was found at the S2 subunit (Figure 5). This motif is extremely rare: only 200 subunit files containing the SSSC sequence of the motif exist among all of the 582,666 PDB subunit files (see Figure S3). Usually, the number of subunits for a commonplace motif (SSSC: SSSHHTSSS) is about 140,000. Even for an already reported common motif (SSSC: SSSHHSHSSS) in antibodies and in major histocompatibility complex class I and II molecules, 34,039 subunits exist.[28] Apart from virus proteins, integrin αL (leukocyte function associated antigen 1),[47,48] and cell division protein kinase 2 (CDK2),[49] which are involved in cell adhesion and cell division, are the main proteins that have such a relatively undeformable motif (Figure 6). For CDK2 with cyclin A, an adenosine-5'-triphosphate (ATP) molecule interacts with this motif.[50] The SSSC of this motif in the free-form of CDK2 (1buh_A)[51] is identical to that in the ATP-binding form (1fin_A).[50] The relatively undeformable motif protrudes on the molecular surface,

and the amino acid sequence of the motif for SARS-CoV differs from the other proteins (6acc_A: LPPLLTDDMI; 3f74_A: YKTEFDFSDY; 3ig7_A: EFLHQDLKKF). Walls and coworkers found that the SARS-CoV-2 S glycoprotein harbors a furin cleavage site (**NSPRR**AR ↓ S) at the boundary between the S1/S2 subunits, which is processed during biogenesis and sets this virus apart from SARS-CoV and SARS-related CoVs.[41] Therefore, the relatively undeformable motif at the S2 subunit may be available for the drug discovery targets.

CONCLUSIONS

SSSCPrediction (SSSCPred) is a program for the prediction of SSSCs from amino acid sequences. SSSCPred was tested using three datasets each comprising 10,000 FASTA files containing the amino acid sequences and SSSCs of protein subunits and two datasets of subunits from CB513 and CullPDB. To confirm whether the flexibility and conformational change of proteins can be predicted or not, two additional deep-neural-network-based prediction systems (SSSCPred200 and SSSCPred100) were constructed. The value size of concordance rates among the three systems provides a good indication of the flexibility of the protein subunits. The comparison program may be helpful to explore undeformable drug discovery targets of many unsolved protein structures.

REFERENCES

1. Haft DH, DiCuccio M, Badretdin A, Brover V, Chetvernin V, O'Neill K, Li WJ, Chitsaz F, Derbyshire MK, Gonzales NR et al. Derbyshire MK, Gonzales NR, Gwadz M, Lu F, Marchler GH, Song JS, Thanki N, Yamashita RA, Zheng C, Thibaud-Nissen F, Geer LY, Marchler-Bauer A, Pruitt KD. RefSeq: an update on prokaryotic genome annotation and curation. Nucleic Acids Res 2018;46:D851–D860.

2. Pruitt KD, Tatusova T, Klimke W, Maglott DR. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res 2009;37:D32–D36.

3. Oldfield CJ, Chen K, Kurgan L. Computational prediction of secondary and supersecondary structures from protein sequences. Methods Mol Biol 2019;1958:73–100.

4. Montgomerie S, Sundararaj S, Gallin WJ, Wishart DS. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics 2006;7:301.

5. Zheng C, Kurgan L. Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinformatics 2008;9:430.

6. Zimmermann L, Stephens A, Nam SZ, Rau D, Kübler J, Lozajic M, Gabler F, Söding J, Lupas AN, Alva V. A completely reimplemented MPI bioinformatics toolkit with a new HHpred Server at its core. J Mol Biol 2018;430:2237–2243.

7. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:195–202.

8. Yan R, Xu D, Yang J, Walker S, Zhang Y. A comparative assessment and analysis of 20 representative sequence alignment methods for protein structure prediction. Sci Rep 2013;3:2619.

9. Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 2001;305:567–580.

10. Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. J Mol Biol 2004;338:1027–1036.

11. Käll L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. Bioinformatics 2005;21:i251–i257.

12. Heffernan R, Yang Y, Paliwal K, Zhou Y. Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Bioinformatics 2017;33:2842–2849.

13. Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. Sci Rep 2016;6:18962.

14. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sønderby CK, Sommer MOA, Winther O, Nielsen M, Petersen B, Marcatili P. NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. Proteins 2019;87:520–527.

15. Hanson J, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. Bioinformatics 2017;33:685–692.

16. Almagro Armenteros JJ, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, von Heijne G, Nielsen H. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 2019;37:420–423.

17. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. Science 1991;252:1162–1164.

18. Gruber M, Söding J, Lupas AN. Comparative analysis of coiled-coil prediction methods. J Struct Biol 2006;155:140–145.

19. Delorenzi M, Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. Bioinformatics 2002;18:617–625.

20. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. Bioinformatics 2015;31:857–863.

21. Dosztányi Z, Csizmók V, Tompa P, Simon I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J Mol Biol 2005;347:827–839.

22. Peng ZL, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic Acids Res 2015;43:e121.

23. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA- and RNA-binding residues. Nucleic Acids Res 2017;45:e84.

24. Ho BK, Brasseur R. The Ramachandran plots of glycine and pre-proline. BMC Struct Biol 2005;5:14.

25. Kleywegt GJ, Jones TA. Phi/psi-chology: Ramachandran revisited. Structure. 1996;4:1395–1400.

26. Ramachandran GN, Ramakrishnan C, Sasisekharan V. Stereochemistry of polypeptide chain configurations. J Mol Biol 1963;7:95–99.

27. Izumi H. Homology searches using supersecondary structure code. Methods Mol Biol 2019;1958:329–340.

28. Izumi H, Wakisaka A, Nafie LA, Dukor RK. Data mining of supersecondary structure homology between light chains of immunogloblins and MHC molecules: absence of the common conformational fragment in the human IgM rheumatoid factor. J Chem Inf Model 2013;53:584–591.

29. Kabsch W, Sander C. Dictionary of protein secondary structure—pattern-recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.

30. Touw WG, Baakman C, Black J, te Beek TAH, Krieger E, Joosten RP, Vriend G. A series of PDB-related databanks for everyday needs. Nucleic Acids Res 2015;43:D364–D368.

31. Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, Ma X, Zhan F, Wang L, Hu T, Zhou H, Hu Z, Zhou W, Zhao L, Chen J, Meng Y, Wang J, Lin Y, Yuan J, Xie Z, Ma J, Liu WJ, Wang D, Xu W, Holmes EC, Gao GF, Wu G, Chen W, Shi W, Tan W. Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. Lancet 2020;395:22–28.

32. Song WF, Gui M, Wang WQ, Xiang Y. Cryo-EM structure of the SARS coronavirus spike glycoprotein in complex with its host cell receptor ACE2. PLoS Pathog 2018;14:e1007236.

33. Gui M, Song WF, Zhou HX, Xu JW, Chen SL, Xiang Y, Wang X. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. Cell Res 2017;27:119–129.

34. Yuan Y, Cao DF, Zhang YF, Ma J, Qi JX, Wang QH, Lu G, Wu Y, Yan J, Shi Y, Zhang X, Gao GF. Cryo-EM structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. Nature Commun 2017;8:15092.

35. Kirchdoerfer RN, Wang N, Pallesen J, Wrapp D, Turner HL, Cottrell CA, Corbett KS, Graham BS, McLellan JS, Ward AB. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. Sci Rep 2018;8:15701.

36. Xu YH, Lou ZY, Liu YW, Pang H, Tien P, Gao GF, Rao Z. Crystal structure of severe acute respiratory syndrome coronavirus spike protein fusion core. J Biol Chem 2004;279:49414–49419.

37. Duquerroy S, Vigouroux A, Rottier PJM, Rey FA, Bosch BJ. Central ions and lateral asparagine/glutamine zippers stabilize the post-fusion hairpin conformation of the SARS coronavirus spike glycoprotein. Virology 2005;335:276–285.

38. Wrapp D, Wang N, Corbett KS, Goldsmith JA, Hsieh CL, Abiona O, Graham BS, McLellan JS. Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science 2020;367:1260–1263.

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

40. Shang J, Ye G, Shi K, Wan Y, Luo C, Aihara H, Geng Q, Auerbach A, Li F. Structural basis of receptor recognition by SARS-CoV-2. Nature 2020. https://doi.org/10.1038/s41586-020-2179-y.

41. Walls AC, Park YJ, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. Cell 2020. https://doi.org/10.1016/j.cell.2020.02.058.

42. Wang Q, Zhang Y, Wu L, Niu S, Song C, Zhang Z, Lu G, Qiao C, Hu Y, Yuen KY, Wang Q, Zhou H, Yan J, Qi J. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. Cell 2020. https://doi.org/10.1016/j.cell.2020.03.045.

43. Lan J, Ge J, Yu J, Shan S, Zhou H, Fan S, Zhang Q, Shi X, Wang Q, Zhang L, Wang X. Structure of the SARS-CoV-2 spike receptor-binding domain bound to the ACE2 receptor. Nature 2020. https://doi.org/10.1038/s41586-020-2180-5.

44. Yuan M, Wu NC, Zhu X, Lee CD, So RTY, Lv H, Mok CKP, Wilson IA. A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV. Science 2020. https://doi.org/10.1126/science.abb7269.

45. Jurtz VI, Johansen AR, Nielsen M, Armenteros JJA, Nielsen H, Sonderby CK, Winther O, Sonderby SK. An introduction to deep learning on biological sequence data: examples and solutions. Bioinformatics 2017;33:3685–3690.

46. Hopf TA, Green AG, Schubert B, Mersmann S, Scharfe CPI, Ingraham JB, Toth-Petroczy A, Brock K, Riesselman AJ, Palmedo P, Kang C, Sheridan R, Draizen EJ, Dallago C, Sander C,

Marks DS. The EVcouplings Python framework for coevolutionary sequence analysis. Bioinformatics 2019;35:1582–1584.

47. Zhang H, Astrof NS, Liu JH, Wang JH, Shimaoka M. Crystal structure of isoflurane bound to integrin LFA-1 supports a unified mechanism of volatile anesthetic action in the immune and central nervous systems. FASEB J 2009;8:2735–2740.

48. Qu A, Leahy DJ. The role of the divalent cation in the structure of the I domain from the CD11a/CD18 integrin. Structure 1996;4:931–942.

49. Helal CJ, Kang Z, Lucas JC, Gant T, Ahlijanian MK, Schachter JB, Richter KEG, Cook JM, Menniti FS, Kelly K, Mente S, Pandit J, Hosea N. Potent and cellularly active 4-aminoimidazole inhibitors of cyclin-dependent kinase 5/p25 for the treatment of Alzheimer's disease. Bioorg Med Chem Lett 2009;19:5703–5707.

50. Jeffrey PD, Russo AA, Polyak K, Gibbs E, Hurwitz J, Massagué J, Pavletich NP. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. Nature 1995;376:313–320.

51. Bourne Y, Watson MH, Hickey MJ, Holmes W, Rocque W, Reed SI, Tainer JA. Crystal Structure and Mutational Analysis of the Human CDK2 Kinase Complex with Cell Cycle–Regulatory Protein CksHs1. Cell 1996;84:863–874.

Table 1. Keywords included in the training dataset files that afforded high concordance rates

Keyword	Files with concordance rate ≥ 0.90 (A)	Total number of files (B)	A/B ratio
PROTEASOME	3283	3551	0.92
FAB	1989	2071	0.96
LYSOZYME	786	830	0.95
HEMOGLOBIN	760	825	0.92
MICROGLOBULIN	501	534	0.94
HLA	402	424	0.95
MYOGLOBIN	174	178	0.98

FIGURE LEGEND

Figure 1 Network architecture of SSSCPrediction.

Figure 2 Comparison of SSSCPrediction with Quick2D.[6] The PDB file (1a00_A) was used for the comparison (SSSCPrediction: H, α-helix-type conformation; S, β-sheet-type conformation; T, other-type conformation; and D, disordered residue or C-terminus. Quick2D: H, α-helix; E, β-strand; and D, disorder).

Figure 3 Distribution map of concordance rate. The average concordance rate of translation of amino acid sequences to SSSCs was 0.90.

Figure 4

Comparison of predicted and observed SSSC sequences of spike proteins in SARS-CoV-2 (first 19 lines) and SARS-CoV (next 12 lines) at the receptor-binding domain (red), including the receptor-binding motif (purple, binding to human ACE2). A comparison of the SSSCPred200 (SARS-CoV-2 and SARS-CoV) results with those of Quick2D (from 32 to 59 lines)[6] is also shown. The receptor-binding motif of SARS-CoV is more undeformable than that of SARS-CoV-2.

Figure 5

Comparison of predicted and observed SSSC sequences of spike proteins in SARS-CoV-2 (first 14 lines) and SARS-CoV (next 11 lines) at aa 851 to 935 (green, heptad repeat 1). A comparison of SSSCPred200 (SARS-CoV-2) results with those of Quick2D (from 26 to 39 lines)[6] is also shown. Only one undeformable motif, the structured loop (red and blue, SSSC: SSSHSSHHHH), was common to the compared SSSC sequences.

Figure 6

Common undeformable motif of structured loop (blue, SSSC: SSSHSSHHHH). (A) SARS-CoV (6acc, monomer), (B) SARS-CoV (6acc, trimer), (C) integrin αL (3f74), (D) leukocyte function-associated antigen 1 (1zop), and (E) cell division protein kinase 2 (3ig7). The relatively undeformable motif protrudes on the molecular surface.

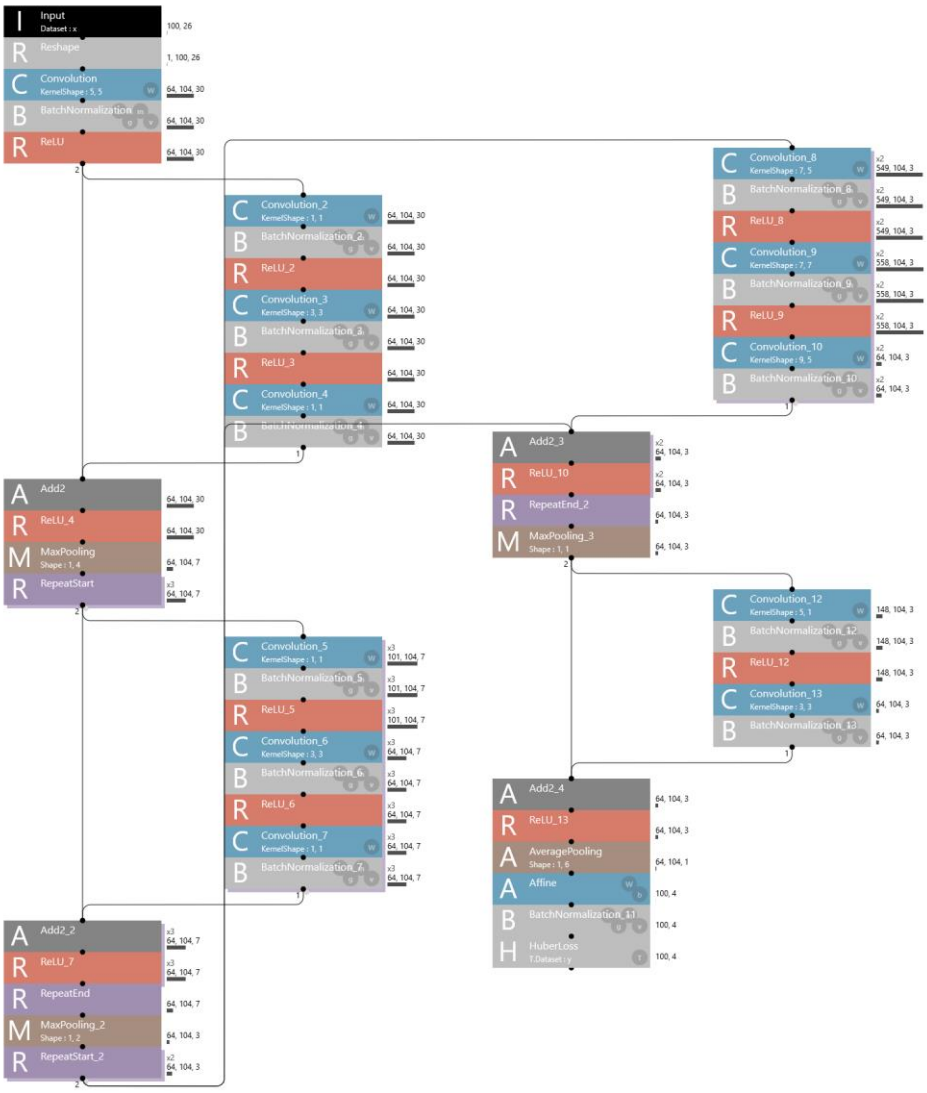
Input
Dataset : x
100, 26
Reshape
1, 100, 26
Convolution
KernelShape : 5, 5
64, 104, 30
BatchNormalization
64, 104, 30
ReLU
64, 104, 30
Convolution_2
KernelShape : 1, 1
64, 104, 30
BatchNormalization_2
64, 104, 30
ReLU_2
64, 104, 30
Convolution_3
KernelShape : 3, 3
64, 104, 30
BatchNormalization_3
64, 104, 30
ReLU_3
64, 104, 30
Convolution_4
KernelShape : 1, 1
64, 104, 30
BatchNormalization_4
64, 104, 30
Add2
64, 104, 30
ReLU_4
64, 104, 30
MaxPooling
Shape : 1, 4
64, 104, 7
RepeatStart
x3
64, 104, 7
Convolution_5
KernelShape : 1, 1
x3
101, 104, 7
BatchNormalization_5
x3
101, 104, 7
ReLU_5
x3
101, 104, 7
Convolution_6
KernelShape : 3, 3
x3
64, 104, 7
BatchNormalization_6
x3
64, 104, 7
ReLU_6
x3
64, 104, 7
Convolution_7
KernelShape : 1, 1
x3
64, 104, 7
BatchNormalization_7
x3
64, 104, 7
Add2_2
x3
64, 104, 7
ReLU_7
x3
64, 104, 7
RepeatEnd
64, 104, 7
MaxPooling_2
Shape : 1, 2
64, 104, 3
RepeatStart_2
x2
64, 104, 3
Convolution_8
KernelShape : 7, 5
x2
549, 104, 3
BatchNormalization_8
x2
549, 104, 3
ReLU_8
x2
549, 104, 3
Convolution_9
KernelShape : 7, 7
x2
558, 104, 3
BatchNormalization_9
x2
558, 104, 3
ReLU_9
x2
558, 104, 3
Convolution_10
KernelShape : 9, 5
x2
64, 104, 3
BatchNormalization_10
x2
64, 104, 3
Add2_3
x2
64, 104, 3
ReLU_10
x2
64, 104, 3
RepeatEnd_2
64, 104, 3
MaxPooling_3
Shape : 1, 1
64, 104, 3
Convolution_12
KernelShape : 5, 1
148, 104, 3
BatchNormalization_12
148, 104, 3
ReLU_12
148, 104, 3
Convolution_13
KernelShape : 3, 3
64, 104, 3
BatchNormalization_13
64, 104, 3
Add2_4
64, 104, 3
ReLU_13
64, 104, 3
AveragePooling
Shape : 1, 6
64, 104, 1
Affine
100, 4
BatchNormalization_11
100, 4
HuberLoss
T.Dataset : y
100, 4

Figure 1

Protein ID: Q_2483096

```
AA_QUERY   1          VLSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSD   85
SSSCPrediction        TSSHHHHHHHHHHHHHHTHHHHHHHHHHHHHHHHHTHHHHHHTHHSSHSSTSHHHHHHHHHHHHHHHHHHHTHHTHHHHHHHHHH
SS_PSIPRED               HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH HHHHHH          HHHHHHHHHHHHHHHHHH    HHHHHHHHHH
SS_SPIDER3               HHHHHHHHHHHHHHH  HHHHHHHHHHHHHHH   HHHH          HHHHHHHHHHHHHHHHHH    HHHHHHHHHH
SS_PSSPRED4                    HHHHHHH    HHHHHHHHHHHHHH           E     EEEE   HHHHHHHHHHHHHH   HHHHHHHHH
SS_DEEPCNF               HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH                    HHHHHHHHHHHHHH    HHHHHHHHHH
SS_NETSURFP2             HHHHHHHHHHHHHHHH HHHHHHHHHHHHHHH HHHHHH          HHHHHHHHHHHHHHHHHH    HHHHHHHHHH
DO_DISOPRED           DDDDD
AA_QUERY   86         LHAHKLRVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTVLTSKYR   141
SSSCPrediction        HHHHHHTSSHHHHHHHHHHHHHHHHHHTHHHTSHHHHHHHHHHHHHHHHHHHHHHTD
SS_PSIPRED            HHHHH    HHHHHHHHHHHHHHHHHH  HH  HHHHHHHHHHHHHHHHHHHHH
SS_SPIDER3            HHHHH    HHHHHHHHHHHHHHHHHH      HHHHHHHHHHHHHHHHHHHHH
SS_PSSPRED4           HHH         HHHHHHHHHHHHHHH       HHHHHHHHHHHHHHHHH
SS_DEEPCNF            HHH         HHHHHHHHHHHHHHH      HHHHHHHHHHHHHHHHHHH
SS_NETSURFP2          HHHHH    HHHHHHHHHHHHHHHHHH  HH  HHHHHHHHHHHHHHHHHHHHH
DO_DISOPRED                                                      D
```

Figure 2

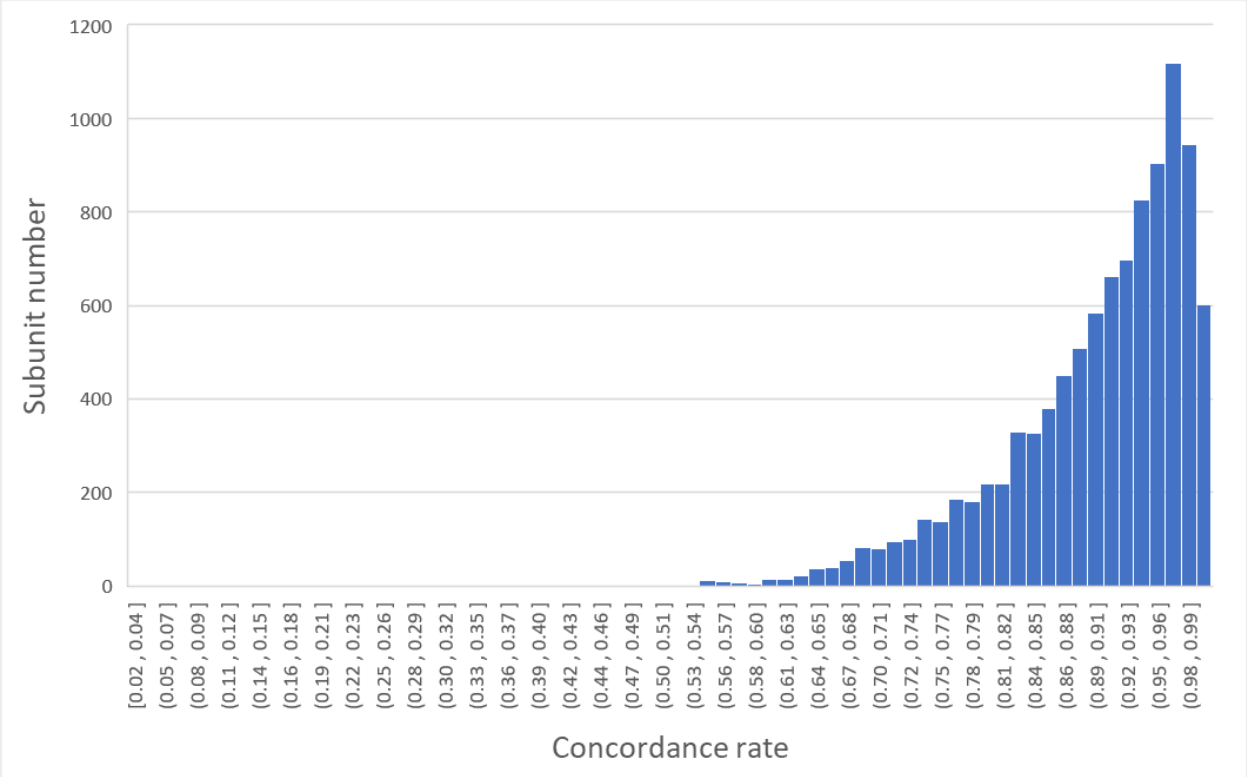
Subunit number
1200
1000
800
600
400
200
0
[0.02 , 0.04]
(0.05 , 0.07]
(0.08 , 0.09]
(0.11 , 0.12]
(0.14 , 0.15]
(0.16 , 0.18]
(0.19 , 0.21]
(0.22 , 0.23]
(0.25 , 0.26]
(0.28 , 0.29]
(0.30 , 0.32]
(0.33 , 0.35]
(0.36 , 0.37]
(0.39 , 0.40]
(0.42 , 0.43]
(0.44 , 0.46]
(0.47 , 0.49]
(0.50 , 0.51]
(0.53 , 0.54]
(0.56 , 0.57]
(0.58 , 0.60]
(0.61 , 0.63]
(0.64 , 0.65]
(0.67 , 0.68]
(0.70 , 0.71]
(0.72 , 0.74]
(0.75 , 0.77]
(0.78 , 0.79]
(0.81 , 0.82]
(0.84 , 0.85]
(0.86 , 0.88]
(0.89 , 0.91]
(0.92 , 0.93]
(0.95 , 0.96]
(0.98 , 0.99]
Concordance rate

Figure 3


AA_QUERY 426 PDDFTGCVIAWNSNNLDSKVGGNYNYLYRLFRKSNLKPFERDISTEIYQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYQPYRV 510
SSSCPred200 SHHSHTSSSSSSHHHHHHHHTTHHHSSSSSSSSSTSSSSSSSSHHHHSSSTSSHSHTHTTSSSSSSSSSSSSSSSHTHHSSSSSS
SSSCPred100 SHHTSSSSSSSSTSSHHHHSSHHHHHHHHHHHHHHSSHHHHHHHHHHHSSTSSSSTTSSTSSSSSSSSTHSSSHHSSSSHSSSSS
SSSCPred SHHHHTSSSSSSHHSHHSHHHHSHHHHHHHHHHHSHSSHHHHSHHSSSSSHHSSSSSTSSTSSSSSTSSTSTSSHHSTSTSSSSSS
6vsb_A STSSSHSSSSSSHHHHHTD THSSSTD TSSSSTSSHTD THHSSSSS
6vsb_B STSSSHSSSSSSHHHHHTD THSSSTD TSSSTSSTHSSSSTD TSSSHSSSSSSSHTD THHSSSSS
6vsb_C STSSSHSSSSSSHHHHHTD THSSSTD TSSSTSSSHSSTD TSSSHSSSSSSTD THHSSSSS
6m17_E STSHSHSSSSSSTHHHSTSHHTSTTSSSSHSSSTSSSSTSSSHSSSSHTSTSTSTSTTTHSTSSSHSSSHTSSHHSSHHHSSSSS
6m17_F STSHSHSSSSSSTHHHSTSHHTSTTSSSSHSSSTSSSSTSSSHSSSSHTSTSTSTSTTTHSTSSSHSSSHTSSHHSSHHHSSSSS
6vxx_A SHHSSTSSSSSSTHHHHSTD TSHHSSSTD TSTSSTTD TSHSSSSSSSHHTDTHHSSSSS
6vxx_B SHHSSTSSSSSSTHHHHSTD TSHHSSSTD TSTSSTTD TSHSSSSSSSHHTDTHHSSSSS
6vxx_C SHHSSTSSSSSSTHHHHSTD TSHHSSSTD TSTSSTTD TSHSSSSSSSHHTDTHHSSSSS
6vyb_A SHHSSTSSSSSSHHHHTD THHSSSSHSSHSSSSSTSSSTSTD THSSSSTSSHHTDTHHSSSSS
6vyb_B SHHSSTSSSSSSTHHHHSSHHTSHHSSSTD TSTSTD TSSSSTSSHHSSHHHSSSSS
6vyb_C SHHSSTSSSSSSTHHHHSTDTHHSSSTD TSTSSSHSTD TSSSSTSSHHSSHHHSSSSS
6vw1_E SHTSSSSSSSSSTHHHHSSHHTSHHSSSSHSSSTSSSSTSSSHSTSSSSHSHSSHTTSSSSTSSSHSSSSSSSHHSSHHHSSSSS
6lzg_B SHTSSTSSSSSSTHHHHSSSTTSHHSSSSHSSHSSSSSTSSSHSHSSSSHSHSSHTTSSSSTSSSHSSSSTSSHHSSHHHSSSSS
6m0j_E SHTSHSSSSSSSHHHHHSSHHTSHHSSSSHSSHSSSSSTSSSHSTSSSSHSHSSHTTSSSSTSSSHSSSSSSSHHSSHHHSSSSS
6w41_C SHTSTSSSSSSSTHHHTSSHHTSHHSSSSHSSHSSSSSTSSSHSHSSSSHSHSSSTTSSTSTSSSHSSSSSTSHHSSHHTSSSSS
AA_QUERY 426 PDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDISNVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRV 510
SSSCPred200 SHHSSHSSSSSSHHHHHSSHHTSHHHSSSHSSSTSSSSTSSSHSTSSSSHTTSSSS-TSSSTSSSHSSSHTSSHHSSHHHSSSSS
SSSCPred100 SHHSSHSSSSSSHHHHHSSHHTSHHSSSSHSSSTSSSSTSSSHSTSSSSHHTSSSS-TTSSTSSSHSSSHTSSHHSSHHHSSSSS
SSSCPred SHHHSHSSSSSSHHHHHSSHHTSHHSSSSSSSSSTSSSHTSSSHHTSSSSHHTSSSS-TSSSTSSSHSSSSSSSHHSSHHHSSSSS
6acc_A SHHTSHSSSSSSTHHHHSSHHTTHHSSSSHSSTTSSSSTSSSHSTSSSSHTTTSSS-TTSSTSSSHSSSHTSSHHSSHHHSSSSS
6acg_C SHHSSHSSSSSSTHHHHSSHHTSHHSSSSHSSTTSSSSTSSSHSTSSSSHTTTSSS-TTSSTTSSHSSSHTSSHHSHHHTSSSSS
6acj_C SHHTSHSSSSSSTHHHHSSHHTSTHSSSSHSSTTSSSSTSSSHSSSSSSHTTTSST-TTSSTTSSHSSSHTSSHHSSHHTSSSSS
6ack_C SHHTSHSSSSSSTHHHTSSHHTSHHSSSSHSSSTSSSSTSSSHSSSSSSHTTTSSS-TTSSTTSSHSSSHTSSHHSSHHHSSSSS
5xlr_A SHHSSHSSSSSSTHHHHSSHHTSTHSSSSHSSSTSSSSTSSSHSSSSSSHTTTSSS-TTSSTSSSHSSSHTSSHHSSHHTSSSSS
5x58_A SHHSSSSSSSSSTHHHHTSHHTSTHSSSSHSSSTSSSSTSSSHSTSSSSHHTSSSS-TTSSTTSSHSSTTSSSHHSSHHHSSSSS
6crv_A
5wrg_A SHTSSSSSSSSSTHHHHSSHHTSTHSSSSHSSTTSSSSTSSTHSSSSSSHHTSSSS-TSSSTSSSHSSSSSTSHHSSHHHSSSSS
SS_PSIPRED EEE EEEEEEEE EEEEEE EEEEEE
SS_SPIDER3 EEEEEE EEEEEEEE E EEEEEEE EEEEE
SS_PSSPRED4 EEEEE HHHHHHHHHH HHH HHHHHH EEEE EEE
SS_DEEPCNF EEEE HHHHHHH HH EEE
SS_NETSURFP2 EEEEEEE EEEEEEEEE EEE EEEEEEEEE EEEEEE
CC_MARCOIL
CC_COILS_W28
CC_PCOILS_W28
TM_TMHMM
TM_PHOBIUS
TM_POLYPHOBIUS
DO_NETSURFPD2
DO_DISOPRED
DO_SPOTD
SS_PSIPRED EEE EEEEE - EEEEE EE EE EEE
SS_SPIDER3 EEEEE EEEEE - EEEEE EEEEEE
SS_PSSPRED4 HHHHEEEEEE EEEEEEE - E EEEE
SS_DEEPCNF EEEE EEEEEE - EEEE
SS_NETSURFP2 EEEEEE EEEEEEE E E - EEE EEEEEEE
CC_MARCOIL
CC_COILS_W28
CC_PCOILS_W28
TM_TMHMM
TM_PHOBIUS
TM_POLYPHOBIUS
DO_NETSURFPD2
DO_DISOPRED
DO_SPOTD

Figure 4

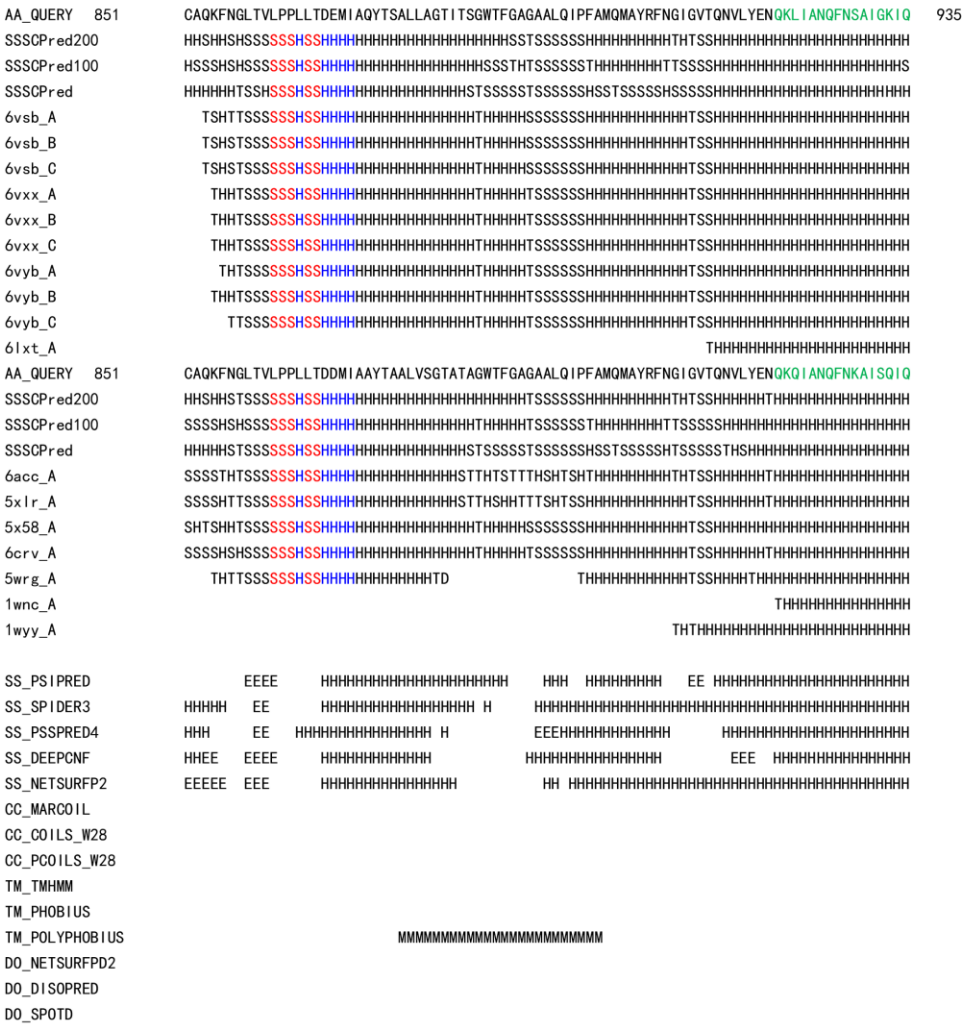
AA_QUERY 851 CAQKFNGLTVLPPLLTDEMIAQYTSALLAGTITSGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKLIANQFNSAIGKIQ 935
SSSCPred200 HHSHHSHSSSSSSHSSHHHHHHHHHHHHHHHHHHHHHSSTSSSSSSHHHHHHHHHHTHTSSHHHHHHHHHHHHHHHHHHHHH
SSSCPred100 HSSSHSHSSSSSSHSSHHHHHHHHHHHHHHHHHHHHSSSTHTSSSSSSTHHHHHHHHTTSSSSHHHHHHHHHHHHHHHHHHHS
SSSCPred HHHHHHTSSHSSSHSSHHHHHHHHHHHHHHHHHHSTSSSSSTSSSSSSHSSTSSSSSHSSSSSHHHHHHHHHHHHHHHHHHHHH
6vsb_A TSHTTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHSSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6vsb_B TSHSTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHSSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6vsb_C TSHSTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHSSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6vxx_A THHTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHTSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6vxx_B THHTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHTSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6vxx_C THHTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHTSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6vyb_A THTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHTSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6vyb_B THHTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHTSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6vyb_C TTSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHTSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6lxt_A THHHHHHHHHHHHHHHHHHHHH
AA_QUERY 851 CAQKFNGLTVLPPLLTDDMIAAYTAALVSGTATAGWTFGAGAALQIPFAMQMAYRFNGIGVTQNVLYENQKQIANQFNKAISQIQ
SSSCPred200 HHSHHSTSSSSSSHSSHHHHHHHHHHHHHHHHHHHHHHTSSSSSSHHHHHHHHHHTHTSSHHHHHHTHHHHHHHHHHHHHH
SSSCPred100 SSSSHSHSSSSSSHSSHHHHHHHHHHHHHHHHHHTHHHHHTSSSSSSTHHHHHHHHTTSSSSSHHHHHHHHHHHHHHHHHHHH
SSSCPred HHHHHSTSSSSSSHSSHHHHHHHHHHHHHHHHHSTSSSSSTSSSSSSHSSTSSSSSHTSSSSSTHSHHHHHHHHHHHHHHHHHH
6acc_A SSSSTHTSSSSSSHSSHHHHHHHHHHHHHHHHSTTHTSTTTHSHTSHTHHHHHHHHHTHTSSHHHHHHTHHHHHHHHHHHHHH
5xlr_A SSSSHTTSSSSSSHSSHHHHHHHHHHHHHHHHSTTHSHHTTTSHTSSHHHHHHHHHHHHTSSHHHHHHTHHHHHHHHHHHHHH
5x58_A SHTSHHTSSSSSSHSSHHHHHHHHHHHHHHHHHTHHHHHSSSSSSSHHHHHHHHHHHHTSSHHHHHHHHHHHHHHHHHHHHH
6crv_A SSSSHSHSSSSSSHSSHHHHHHHHHHHHHHHHHTHHHHHTSSSSSSHHHHHHHHHHHHTSSHHHHHHTHHHHHHHHHHHHHH
5wrg_A THTTSSSSSSHSSHHHHHHHHHHHHTD THHHHHHHHHHHHTSSHHHHTHHHHHHHHHHHHHHHHH
1wnc_A THHHHHHHHHHHHHHH
1wyy_A THTHHHHHHHHHHHHHHHHHHHHHHHH
SS_PSIPRED EEEE HHHHHHHHHHHHHHHHHHHHH HHH HHHHHHHHH EE HHHHHHHHHHHHHHHHHHHHH
SS_SPIDER3 HHHHH EE HHHHHHHHHHHHHHHHHH H HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
SS_PSSPRED4 HHH EE HHHHHHHHHHHHHHHH H EEEHHHHHHHHHHHH HHHHHHHHHHHHHHHHHHHH
SS_DEEPCNF HHEE EEEE HHHHHHHHHHHH HHHHHHHHHHHHHHH EEE HHHHHHHHHHHHHHH
SS_NETSURFP2 EEEEE EEE HHHHHHHHHHHHHHH HH HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
CC_MARCOIL
CC_COILS_W28
CC_PCOILS_W28
TM_TMHMM
TM_PHOBIUS
TM_POLYPHOBIUS MMMMMMMMMMMMMMMMMMMMM
DO_NETSURFPD2
DO_DISOPRED
DO_SPOTD

Figure 5

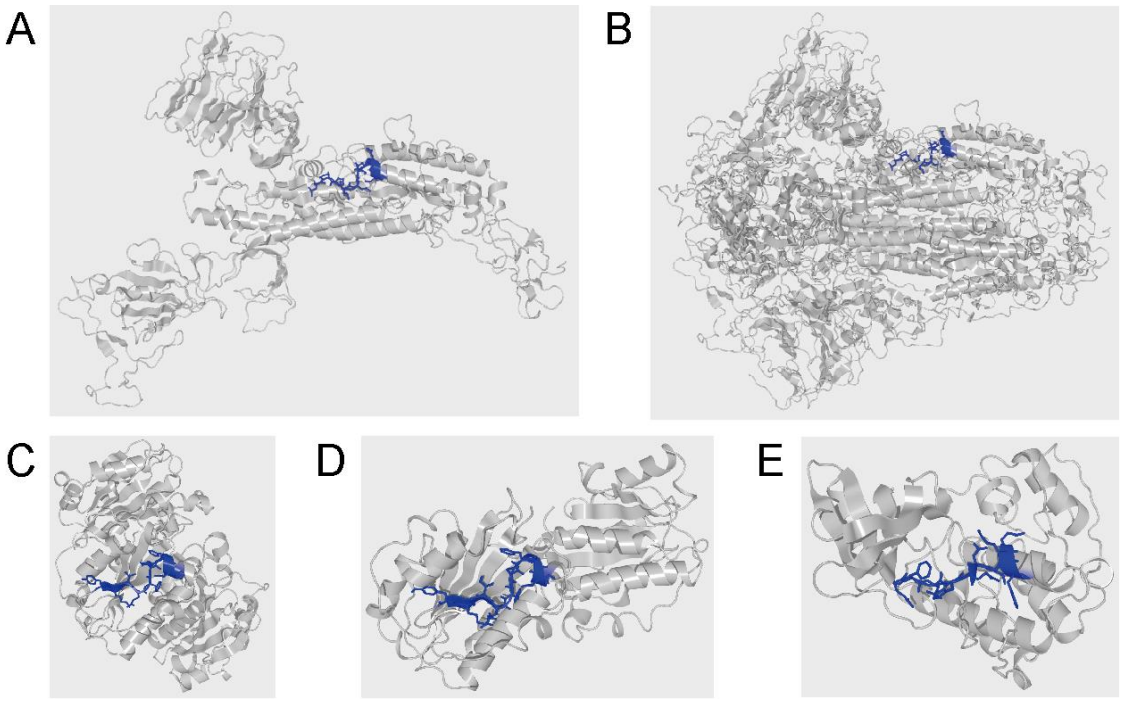

Figure 6