

Spike Protein Rigid Motif shared by SARS-CoV and SARS-CoV-2 (2019-nCoV): Flexible Conformations Predicted by using Supersecondary Structure Codes

Hiroshi Izumi*

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

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

*Correspondence to: Hiroshi 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

I compared the predicted and observed flexible conformations of SARS-CoV and SARS-CoV-2 (2019-nCoV) spike proteins by using supersecondary structure codes (SSSCs) and a comparison program of three deep-neural-network-based prediction systems (SSSCPred200, SSSCPred100, and SSSCPred). The SARS-CoV SSSC sequences predicted by the three deep-neural-network-based systems well reproduced those of the Protein Data Bank (PDB) data, including the structured loops. Only one common identical motif (SSSC: SSSHSSHHHH) among all of the compared SSSC sequences, including predicted and observed ones, was found at the S2 position. This motif has an extremely rare rigid conformation. The antibody or ligand binding to the spike protein S2 of SARS-CoV near the rigid motif may also have a more accessible effect on SARS-CoV-2 than those binding to the receptor-binding motif of SARS-CoV have.

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

INTRODUCTION

To develop a vaccine against the coronavirus disease 2019 (COVID-19), which is currently prevalent mainly in Wuhan, China, structural information on the virus is required.¹ Recently, a deep-neural-network-based program for sequence-based prediction of supersecondary structure codes (SSSCs), called SSSCPrediction (https://researchmap.jp/multidatabases/multidatabase_contents/detail/256924/8fe07c64e364d8218108144f0d33c142?frame_id=708960) was constructed.²⁻⁴ An SSSC is transcribed by using the letters H, S, T, and D to refer to an α -helix-type conformation, a β -sheet-type conformation, other types of conformations, and disordered residues or the C-terminus, respectively. 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. The sequence of severe acute respiratory syndrome coronavirus (SARS-CoV) moderately resembles that of SARS-CoV-2 (about 79% identity).¹ Several observed structures in SARS-CoV, including cryo-electron microscopy structures,⁵⁻¹⁰ have been registered in the Protein Data Bank (PDB) database¹¹ and are thus available for use in comparing the predicted SSSCs of SARS-CoV-2.

Here, I show that the receptor-binding motif (binding to human angiotensin-converting enzyme 2, ACE2) SSSCs of SARS-CoV differs greatly from those of SARS-CoV-2, with that of SARS-CoV-2 being more flexible. I also describe the only shared, relatively rigid motif (SSSC: SSSHSSHHH), which in SARS-CoV is associated with cell adhesion and cell division.

MATERIALS AND METHODS

I constructed two additional deep-neural-network-based prediction systems by using procedures similar to that used to construct SSSCPrediction (SSSCPred).² A total of 582,666 FASTA-format files containing the amino acid sequences and SSSCs of protein subunits were extracted from 139,932 PDB files¹¹ by using the SSSCview program (available online at https://researchmap.jp/multidatabases/multidatabase_contents/detail/256924/6216cafbe7d56e9a65c649886edcb0a3?frame_id=708960).³ 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,¹² 0.905 (9851 subunits) and CB513,¹² 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). The differences in concordance rates among the three systems provides a good indication of the flexibility of the protein subunits.

RESULT AND DISCUSSION

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 1; 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 newest 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 rigid 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.¹³⁻²⁷

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 position (Figure 2). 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: SSSHSHSSS) in antibodies and in major histocompatibility complex class I and II molecules, 34,039 subunits exist.⁴ Apart from virus proteins, integrin α L (leukocyte function associated antigen 1),^{28,29} and cell division protein kinase 2 (CDK2),³⁰ which are involved in cell adhesion and cell division, are the main proteins that have such a rigid motif (Figure 3). For CDK2 with cyclin A, an adenosine-5'-triphosphate molecule interacts with this motif.³¹ The rigid motif protrudes on the molecular surface, and the amino acid sequence of the motif for SARS-CoV differs from the other proteins (6acc_A: LPLLTDDMI; 3f74_A: YKTEFDFSDY; 3ig7_A: EFLHQDLKKF). Therefore, the antibody or ligand binding to the spike protein S2 of SARS-CoV near the rigid motif may also have a more accessible effect on SARS-CoV-2 than those binding to the receptor-binding motif of SARS-CoV have.

During a format check of the Journal's style, the Cryo-EM structure data of the SARS-CoV-2 spike protein (6vsb) was registered.³² The rigid motif (SSSC: SSSHSSHHHH) was confirmed at the S2 position (see Figure S4). Actually, the receptor-binding motif SSSCs of SARS-CoV-2 with blanks differs greatly from those of SARS-CoV, with that of SARS-CoV-2 being more flexible (see Figure S4). 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. This would be associated with the flexible receptor-binding motif of SARS-CoV-2.

REFERENCES

1. 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.
2. Izumi H. SSSCPrediction: a deep neural network–based program for the prediction of supersecondary structure codes from amino acid sequences. *BMC Bioinformatics* to be submitted.
3. Izumi H. Homology searches using supersecondary structure code. *Methods Mol Biol* 2019;1958:329–340.
4. Izumi H, Wakisaka A, Nafie LA, Dukor RK. Data mining of supersecondary structure homology between light chains of immunoglobulins and MHC molecules: absence of the common conformational fragment in the human IgM rheumatoid factor. *J Chem Inf Model* 2013;53:584–591.
5. 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.

6. 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.
7. 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.
8. 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.
9. 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.
10. 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.
11. 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.
12. Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. *Sci Rep* 2016;6:18962.

13. 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.
14. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;292:195–202.
15. 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.
16. 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.
17. 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.
18. Lupas A, Van Dyke M, Stock J. Predicting coiled coils from protein sequences. *Science* 1991;252:1162–1164.
19. Gruber M, Söding J, Lupas AN. Comparative analysis of coiled-coil prediction methods. *J Struct Biol* 2006;155:140–145.
20. Delorenzi M, Speed T. An HMM model for coiled-coil domains and a comparison with PSSM-based predictions. *Bioinformatics* 2002;18:617–625.

21. 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.
22. Käll L, Krogh A, Sonnhammer EL. A combined transmembrane topology and signal peptide prediction method. *J Mol Biol* 2004;338:1027–1036.
23. Käll L, Krogh A, Sonnhammer EL. An HMM posterior decoder for sequence feature prediction that includes homology information. *Bioinformatics* 2005;21:i251–i257.
24. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* 2015;31:857–863.
25. 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.
26. 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.
27. 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.
28. 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.

29. 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.
30. 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.
31. 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.
32. 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;eabb2507.

FIGURE LEGEND

Figure 1

Comparison of predicted and observed SSSC sequences of spike proteins in SARS-CoV-2 (first four lines) and SARS-CoV (next nine 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) results with those of Quick2D (from 14 to 27 lines)¹³⁻²⁷ is also shown. The receptor-binding motif of SARS-CoV is more rigid than that of SARS-CoV-2.

Figure 2

Comparison of predicted and observed SSSC sequences of spike proteins in SARS-CoV-2 (first four lines) and SARS-CoV (next eleven lines) at aa 851 to 935 (green, heptad repeat 1). A comparison of SSSCPred200 (SARS-CoV-2) results with those of Quick2D (from 16 to 29 lines)¹³⁻²⁷ is also shown. Only one rigid motif, the structured loop (red and blue, SSSC: SSSHSSHHHH), was common to the compared SSSC sequences.

Figure 3

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

AA_QUERY 426	PDDFTGCVIAWNSNNLDSKVGGNYNLYRLFRKSNLKPFERDISTEIQAGSTPCNGVEGFNCYFPLQSYGFQPTNGVGYPYRV	510
SSSCPred200	SHHSHTSSSSSSHHHHHHHHHTTHHHSSSSSSSSSTSSSSSSSSHHHHSSSTSSHHTHTTSSSSSSSSSSSSSHTHSSSSSS	
SSSCPred100	SHHTSSSSSSSTSSHHSSSHHHHHHHHHHHHSSHHHHHHHHHSSTSSSTSSSTSSSSSTHSSSHSSSHSSSSSS	
SSSCPred	SHHHHTSSSSSSSHHSSSHHHHHHHHHHHHSSHHHHHHSSSSSHSSSSSTSSSTSSSSSTSTSTSSHHTSTSSSSSS	
AA_QUERY 426	PDDFMGCVLAWNTRNIDATSTGNYNYKYRYLRHGKLRPFERDINSVPFSPDGKPCT-PPALNCYWPLNDYGFYTTTGIGYQPYRV	510
SSSCPred200	SHHSSSSSSSSHHHHHSSHTSHHHSSSHSSSTSSSTSSSHSTSSSHHTTSSSS-TSSSTSSSHSSHTSSHHSHHHHSSSS	
SSSCPred100	SHHSSSSSSSSHHHHHSSHTSHHSSSSHSSSTSSSTSSSHSTSSSHHTTSSSS-TTSSTSSSHSSHTSSHHSHHHHSSSS	
SSSCPred	SHHHSSSSSSSHHHHHSSHTSHHSSSSSSSSTSSHTSSSHHTSSSHHTTSSSS-TSSSTSSSHSSSSSSHHSSHHHHSSSS	
6acc_A	SHHTSHSSSSSTHHHHSSHTTHSSSSSHSSTTSSSTSSSHSTSSSHHTTSSSS-TTSSTSSSHSSHTSSHHSHHHHSSSS	
5xlr_A	SHHSSSSSSSSTHHHHSSHTSTHSSSSSHSSSTSSSTSSSHSSSSSHHTTSSSS-TTSSTSSSHSSHTSSHHSHHHTSSSS	
5x58_A	SHHSSSSSSSSTHHHTSHHTSTHSSSSSHSSSTSSSTSSSHSTSSSHHTTSSSS-TTSSTSSSHSTTSSHHSHHHHSSSS	
6crv_A		
5wrg_A	SHTSSSSSSSSSTHHHHSSHTSTHSSSSSHSSTTSSSTSSSTHSSSSSHHTSSSS-TSSSTSSSHSSSTSSHHSHHHHSSSS	
SS_PSIPRED	EEE EEEEEEE EEEEE EEEEE EEEEE	
SS_SPIDER3	EEEEEE EEEEEEE E EEEEEEE EEEEE	
SS_PSSPRED4	EEEE HHHHHHHHH HHH HHHHH EEEE EEE	
SS_DEEPCNF	EEEE HHHHHH HH	
SS_NETSURFP2	EEEEEE EEEEEEE EEE EEEEEEE EEEEE	
CC_MARCOIL		
CC_COILS_W28		
CC_PC0ILS_W28		
TM_TMHHM		
TM_PHOBIUS		
TM_POLYPHOBIUS		
DO_NETSURFPD2		
DO_DISOPRED		
DO_SPOTD		

Figure 1

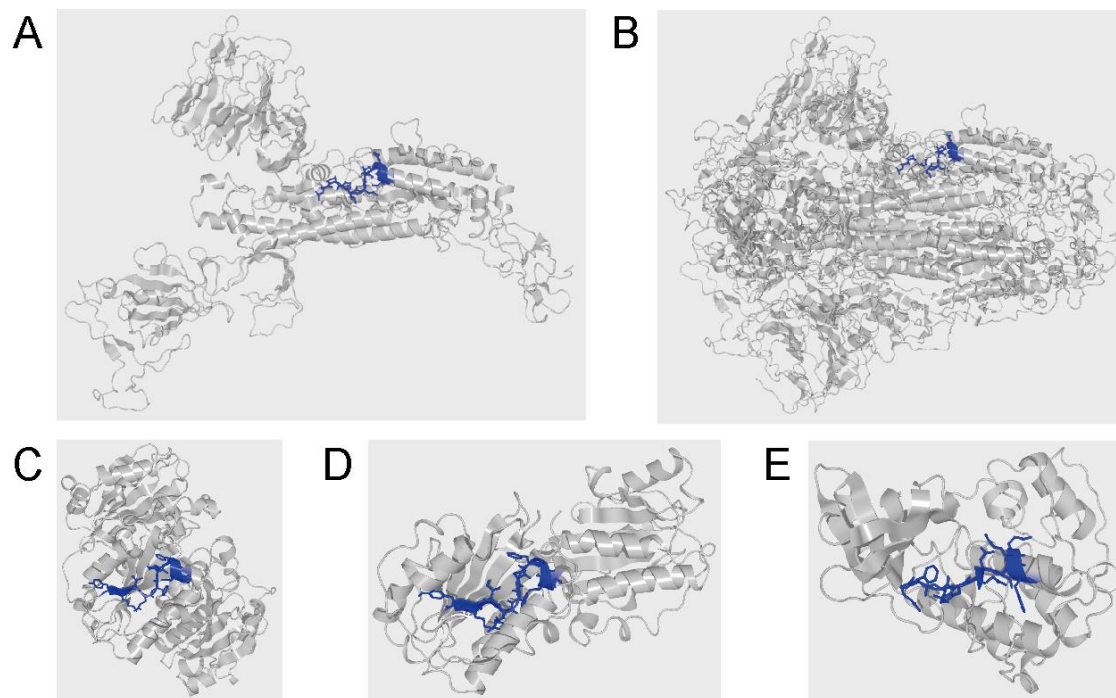


Figure 3