SSSCPreds: Deep Neural Network-Based Software for the Prediction of Conformational Variability and Application to SARS-CoV-2

Hiroshi Izumi,* Laurence A. Nafe, and Rina K. Dukor

ABSTRACT: Amino acid mutations that improve protein stability and rigidity can accompany increases in binding affinity. Therefore, conserved amino acids located on a protein surface may be successfully targeted by antibodies. The quantitative deep mutational scanning approach is an excellent technique to understand viral evolution, and the obtained data can be utilized to develop a vaccine. However, the application of the approach to all of the proteins in general is difficult in terms of cost. To address this need, we report the construction of a deep neural network-based program for sequence-based prediction of supersecondary structure codes (SSSCs), called SSSCPrediction (SSSCPred). Further, to predict conformational flexibility or rigidity in proteins, a comparison program called SSSCPred that consists of three deep neural network-based prediction systems (SSSCPred, SSSCPred100, and SSSCPred200) has also been developed. Using our algorithms we calculated here shows the degree of flexibility for the receptor-binding motif of SARS-CoV-2 spike protein and the rigidity of the unique motif (SSSC: SSSHSHHHHH) at the S2 subunit and has a value independent of the X-ray and Cryo-EM structures. The fact that the sequence flexibility/rigidity map of SARS-CoV-2 RBD resembles the sequence-to-phenotype maps of ACE2-binding affinity and expression, which were experimentally obtained by deep mutational scanning, suggests that the identical SSSC sequences among the ones predicted by three deep neural network-based systems correlate well with the sequences with both lower ACE2-binding affinity and lower expression. The combined analysis of predicted and observed SSSCs with keyword-tagged datasets would be helpful in understanding the structural correlation to the examined system.

INTRODUCTION

In general, the effects of amino acid mutation on functions such as binding between proteins and expression are correlated. The correlation between expression and binding suggests that mutations that improve stability and rigidity accompany increases in binding affinity. Therefore, conserved amino acids located on the protein surface can be more successfully targeted by antibodies. For this purpose, a quantitative deep mutational scanning approach is an excellent technique to understand viral evolution, and the obtained data can be utilized to develop a vaccine. However, there are approximately 110.3 million nonredundant protein sequences in the RefSeq database, and the application of the approach to all of the proteins in general is difficult in terms of cost. To address this need, we report the construction of a deep neural network-based program for sequence-based prediction of supersecondary structure codes (SSSCs), called SSSCPrediction (SSSCPred).

However, the classification and prediction of fine-structured loops other than α-helices, β-strands, coiled coils, and disordered regions remain elusive. There currently is no way to evaluate whether a particular protein sequence is flexible with the shape when cryo-electron microscopy (Cryo-EM) or X-ray structure of that sequence is not available as a guide. SSSCPreds, described in this work, is the first, and to date only, program that can simultaneously predict locations of protein flexibility or rigidity and the shapes of those regions with high accuracy. It does this by comparing different 3D conformation prediction programs that are based on protein sequences. The detail of conformations could not be discussed by using only the appearance of a molecular model, but rather a comparison of the observed SSSC sequences with the predicted ones obtained...
from the examined systems as embodied in SSSCPreds, as described here, would be necessary.

In the past decade, a means of identifying and codifying supersecondary structures (supersecondary structure code, SSSC) has been developed by us that uses the concept of Ramachandran plot data with \( \omega \) angles and the specification of positions of torsion angles in a protein. These data are derived from a fuzzy search of structural code homology using template patterns, represented as conformational codes, such as 3a5c4a (\( \alpha \)-helix-type conformation) and 6c4a4a (\( \beta \)-sheet-type conformation), to describe supersecondary structural motifs and their conformation.\(^{25-27} \) The SSSC is transcribed as a conformation propensity using the letters \( \alpha \) for \( \alpha \)-helix-type conformation (H), a \( \beta \)-sheet-type conformation (S), a variety of other-type conformations (T), and disordered residues or C-terminus (D). This code has been approved as a protocol for a molecular biology database\(^{28} \) and can be used to distinguish the flexibility and conformational flexibility of the SARS-CoV-2 protein obtained in a single crystal (X-ray) or in a frozen aqueous medium (Cryo-EM). Nevertheless, despite the lack of structural constraint, the result of our SSSCPreds prediction methodology is completely consistent with the reported X-ray and Cryo-EM structures.

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.\(^{30} \) The protein sequence of severe acute respiratory syndrome coronavirus (SARS-CoV) moderately resembles that of SARS-CoV-2 (about 79% identity).\(^{30} \) Several observed structures of spike proteins in SARS-CoV\(^{31-36} \) and SARS-CoV-2\(^{37-41} \) including Cryo-EM structures except the full postfusion state S2 proteins, have been registered in the PDB database\(^{44} \) and are thus available for use in comparing the predicted SSSCs of SARS-CoV-2. On the other hand, the sequence identity of ORF8 (Open Reading Frame 8) proteins between SARS-CoV\(^ {45,46} \) and SARS-CoV-2\(^ {37,48} \) (about 24% identity) is very low. Although the structure of SARS-CoV-2 has been reported recently,\(^ {48} \) the structure of SARS-CoV has been unsolved yet. ORF8 disrupts IFN-I signaling when exogenously overexpressed in cells.\(^ {47} \) The other report that ORF8 of SARS-CoV-2, but not ORF8 or ORF8a/b of SARS-CoV, downregulates MHC-I in cells has been published.\(^ {47} \) For SARS-CoV, the full 122 amino acid protein encoded by ORF8 induces ATF6-dependent transcription, which triggers the expression of chaperones.\(^ {45} \) A 29-nucleotide deletion (\( \Delta \)29), splitting of SARS-CoV ORF8 into ORF8a and ORF8b, is correlated with milder disease.\(^ {48} \) In this paper we present, for the first time, a sequence flexibility/rigidity map was obtained from a deep learning comparison program, SSSCPreds, that uses input from three structure prediction programs SSSCPred, SSSCPred100, and SSSCPred200. This sequence flexibility/rigidity map resembles the sequence-to-phenotype maps of the SARS-CoV-2 receptor-binding domain (RBD). As a particularly important and urgent demonstration of SSSCPreds, we assess the flexibility of the SARS-CoV-2 RBD and ORF8 and the rigidity of the nearby S2 region.

Figure 1. Comparison of SSSCPrediction with Quick2D.\(^ {8} \) The PDB file (1a00_A) was used for comparison (SSSCPrediction: H, \( \alpha \)-helix-type conformation; S, \( \beta \)-sheet-type conformation; T, other-type conformation; D, disordered residue or C-terminus. Quick2D: H, \( \alpha \)-helix; E, \( \beta \)-strand; D, disordered residue).
RESULTS AND DISCUSSION

Translation of Amino Acid Sequences to SSSCs. The comparison of SSSCPrediction with Quick2D was carried out by using the PDB file (1a00_A: HEMOGLOBIN ALPHA CHAIN). As shown in Figure 1, 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 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 3,450 files in the test dataset had a concordance rate of ≥0.95, and 6,000 files had a concordance rate of ≥0.90 (Figure 2). In the past three decades, much progress has been made in the development of accurate predictors of the protein secondary structure. Recently, prediction accuracy has increased from about 82 to 84%, which is approaching the estimated upper accuracy limit of around 88%.5,14−16 Again as stated above, although a direct comparison of accuracy is impossible due to the difference of correct data between SSSCPrediction and other prediction methods, these prediction accuracies, or concordance rates, are comparable.

The correlation between keywords of subunit names in the training files and concordance rates 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 those keywords and a concordance rate of ≥0.90 to the total number of files with those keywords was extremely high (≥0.92; Table 1). In contrast, for files containing the keywords ALKALINE (ratio of files with that keyword and a concordance rate of ≥0.90 to the total number 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 a concordance rate of ≥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 (3050/3815), DEHYDROGENASE (2545/3815), HYDROGENASE (2732/4120), POLYMERASE (1863/2888), HYDROLASE (1199/2041), PROTEASE (1344/1765), PHOSPHATASE (990/1690), ISOMERASE (1279/1912), and OXIDASE (1086/1682) frequently appeared, and the ratios of files with those keywords and a concordance rate of ≥0.90 to the total number of files with those keywords ranged from 0.59 to 0.76.

Figure 2. Distribution map of the number of subunits per concordance interval. The average concordance rate of translation of amino acid sequences to SSSCs was 0.90.

| keyword       | files with a concordance rate of ≥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      |
Thus, there were no keywords associated only with a low concordance rate, and an identical amino acid sequence of a flexible protein may possibly have different SSSC sequences.

To confirm whether the flexibility and conformational change of proteins can be predicted or not, two additional deep neural network-based prediction systems were constructed by using procedures similar to that used to construct SSSC-Pred

Figure 3. Sequence flexibility/rigidity map of SARS-CoV-2 RBD and SARS-CoV RBD. The identical SSSC sequences among the predicted ones by three deep neural network-based systems and the corresponding observed ones are colored (blue: \(\alpha\)-helix-type conformation; red: \(\beta\)-sheet-type conformation; green: a variety of other-type conformations). A comparison of the SSSC-Pred result (SARS-CoV-2) with that of Quick2D is also shown.
Figure 4. Sequence flexibility/rigidity map of the connecting regions of S2 subunits for SARS-CoV-2 and SARS-CoV. The identical SSSC sequences among the predicted ones by three deep neural network-based systems and the corresponding observed ones are colored (blue: α-helix-type...
conformation; red: β-sheet-type conformation; green: a variety of other-type conformations). A comparison of the SSSCPreds result (SARS-CoV-2) with that of Quick2D is also shown. UH: upstream helix; L: linker region; FP: fusion peptide; CR: connecting region; HR1: heptad repeat 1.

(SSSCPred). The benchmarks (average concordance rates) of the three systems were as follows: for SSSCPred200, 0.905 (CullPDB;15 9851 subunits) and 0.911 (CB513;15 361 subunits); for SSSCPred100, 0.896 (CullPDB; 17,169 subunits) and 0.907 (CB513; 612 subunits); for SSSCPred, 0.861 (CullPDB; 17,169 subunits) and 0.882 (CB513; 612 subunits). For CullPDB files, the total number of files with a concordance rate of <0.65 between SSSCPred200 and PDB data was 66. Of these CullPDB files, the ratio of files with a concordance rate of <0.70 between SSSCPred200 and SSSCPred100 data to the total number of files was 0.83 (see Table S1). For CB513 files, the total number of files with a concordance rate of <0.75 between SSSCPred200 and PDB data was 17. Of these C5B13 files, the ratio of files with a concordance rate of <0.80 between SSSCPred200 and SSSCPred100 data to the total number of files was 0.59 (see Table S2). Exceptionally, in the CB513 files, the subunit with the keyword PHOSPHOGLYCERATE MUTASE 1 (3pjm_A) showed the highest 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 files (1q1f_A, 1q3_A, 4pjm_A, and 5pjm_A) of the same keyword with high concordance rates (0.96, 0.97, 0.96, and 0.99) between SSSCPred200 and PDB data were found. This means that the PDB files 3pjm_A and 1q1f_A have the identical amino acid sequence, but the SSSC sequences, which reflect the subunit flexibility, are largely different. For CullPDB files, the total number of files with a concordance rate of <0.65 between SSSCPred200 and SSSCPred100 was 80. Of these CullPDB files, the ratio of files with a concordance rate of <0.75 between SSSCPred200 and PDB data to the total number of files was 0.80 (64/80). The value size of concordance rates among the three systems (SSSCPreds) provides a good indication of the flexibility of the protein subunits.

**Predicted and Observed SSSC Sequences of SARS-CoV-2 Proteins. Spike Protein RBD.** We then compared the predicted and observed SSSC sequences of spike proteins of SARS-CoV-2 and SARS-CoV at the receptor-binding domain (Figure 3; 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, Sdx_A, and 5x58_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 (aa 437 to 508) 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 3).8 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 (6w6b and 6wxx) differ greatly from those of SARS-CoV, with those of SARS-CoV-2 being more flexible (Figure 3). 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, 6wv1_E, 6lg_B, 6m0j_E, and 6w41_C) are very similar to those of SARS-CoV. Wrapp and co-workers 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-2.37 Yuan and co-workers 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.43 The observed SSSC of the highly conserved epitope (6w41_C) resembles that of SSSCPred100, and the gap of the epitope SSSCs among the three systems is smaller than that of the receptor-binding motif SSSCs (underlined SSSCs in Figure 3). 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 flexibility/rigidity map of SARS-CoV-2 RBD resembles the sequence-to-phenotype maps, which were experimentally obtained by deep mutational scanning (Figure 3).4 This suggests that the identical SSSC sequences among the predicted ones by three deep neural network-based systems (SSSCPred200, SSSCPred100, and SSSCPred) with the observed ones correlate well with the sequences of both lower ACE2-binding affinity and lower expression.1 The sequence flexibility/rigidity map obtained from SSSCPreds would be therefore available for no-cost mutation-site selection of proteins in general.

**Spike Protein S2.** 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: SSSSHSSHHHH) among all of the compared SSSC sequences, including predicted and observed ones, was found at the S2 subunit (Figure 4). 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: SSSHSSHSSS) in antibodies and in major histocompatibility complex class I and II molecules, 34,039 subunits exist.29 Apart from virus proteins, integrin αL (leukocyte function-associated antigen 1)30,31 and cell division protein kinase 2 (CDK2),51 which are involved in cell adhesion and cell division, are the main proteins that have such a relatively undeformable motif (Figure 5). For CDK2 with cyclin A, an adenosine-5′-triphosphate (ATP) molecule interacts with this motif.52 The SSSC of this motif in the free form of CDK2 (1buh_A)53 is identical to that in the ATP-binding form (1n_A).52 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: LPPLLTDMDI; 3p74_A: YKTEFDFSDY; 3ig7_A: EFLHQLDKKF). The motif (SSSC: SSSHSSHSSS) was also found in the subunits of...
reaction center protein H, glyceraldehyde-3-phosphate dehydrogenase, glutamine phosphoribosylpyrophosphatase, green fluorescent protein fpS12, and insulin-degrading enzyme. Furthermore, the other coronavirus, murine hepatitis virus, has the identical motif (3jcl_A). This motif may correlate with proteins recognizing phosphates such as phospholipids. Walls and co-workers found that the SARS-CoV-2 S glycoprotein harbors a furin cleavage site (NSPARRAR) at the boundary between the S1/S2 subunits, which is processed during biogenesis and sets this virus apart from SARS-CoV and SARS-related CoV. Therefore, the relatively undeformable motif at the S2 subunit may be available for the drug discovery targets. The unique motif belongs to the connecting region of S2 subunit, and the entire structure of the postfusion state S2 subunit of SARS-CoV-2 has not been available. The predicted conformations of heptad repeat 1 (HR1) and HR2 at the S2 subunit well correspond to the observed ones in the partial structure of postfusion state (see Figure S1). The SSSC sequence change from “STSSHHHSSSS” to “STTTSSSSSSS” contributes to the transformation to postfusion hairpin state (Figure 6). The SSSCPreds analysis also suggests that the sequence “AGFIKQYGDCLDGIAARDLI” in the connecting region of S2 subunit easily forms successive stable α-helix-type conformations (Figure 4) by the interaction of fusion peptide (FP) and unique motif with the host membrane (Figure 6). The prediction results near the connecting region of S2 subunit including the relatively undeformable motif would be useful to understand in terms of the conformational change from prefusion native state to postfusion hairpin state via prehairpin intermediate state.

ORF8 Protein. The sequence identity of ORF8 proteins between SARS-CoV and SARS-CoV-2 (about 24% identity) is very low, and the function is in debate. The related data of ORF8 SSSC sequences are not contained in the training datasets at all. As shown in Figure 7, the concordance rate of SSSCs between SSSCPred100 and SSSCPred data for SARS-CoV-2 ORF8 (0.75) is larger than that for SARS-CoV ORF8a/b (0.64). The predicted SSSCs of SARS-CoV-2 ORF8 with many stable β-strands such as an immunoglobulin-like fold quite correspond to the observed ones (Figure 7). It is suggested that three sets of intramolecular disulfide bonds per monomer immobilize the thermodynamically unstable conformations (Figure 8), especially one of those in the motif of the amino acid sequence “YTVSCLPFT” (SSSC: HSHSSHSSS). Furthermore, the predicted SSSCs explain well that the flexibility of β3 motif for SARS-CoV ORF8a/b interferes with the crystallization (Figure 7).

As shown in Figure 8, a very rare motif (SSSC: SSSHSHHTHSS) with the amino acid sequence “RCSFYEDF-LEY” in the observed SSSCs of SARS-CoV-2 was found (7j1l_A). A total of 1372 subunit files containing the SSSC sequence exist among all of the 582,666 PDB subunit files (see Figure S4), and the main subunits are endothiapepsin, DNA damage-binding protein 1, isocitrate dehydrogenase [NADP] cytoplasmic, DNA-directed RNA polymerase subunit α, 70 kDa peptidylprolyl isomerase, purple acid phosphatase, α-1,6-mannanase, interferon regulatory factor 4, β-fructofuranosidase, ethanolamine utilization protein eutI, crispr-associated helicase, hemagglutinin-neuraminidase glycoprotein, plasminogen, sigma A, and catalase-peroxidase 2 (see Figure S4). A 2-(N-morpholino)-ethanesulfonic acid molecule interacts with this motif in a subunit of DNA damage-binding protein 1 (4a08_A). Although the predicted SSSCs suggest that the motif is flexible, the motif may correlate with the regulation of phosphorylated substances such as DNA, RNA, and phosphorylated proteins and may contribute to the difference that ORF8 of SARS-CoV-2, but not ORF8 or ORF8a/b of SARS-CoV, downregulates MHC-I in cells.

**CONCLUSIONS**

The deep neural network-based program for sequence-based prediction of SSSCs (SSSCPred) and the comparison program (SSSCPreds) of three deep neural network-based prediction systems (SSSCPred200, SSSCPred100, and SSSCPred) to predict the flexibility and conformational change of proteins were constructed. The degree of flexibility for the
receptor-binding motif of SARS-CoV-2 spike protein and the rigidity of the unique motif at the S2 subunit cannot be found only from the Cryo-EM and X-ray structures. This methodology provides a verified path to analyze other protein structures in a similar way when there may not be X-ray or Cryo-EM structures available such as SARS-CoV ORF8a/b. The sequence flexibility/rigidity map obtained from the combined analysis of

Figure 7. Sequence flexibility/rigidity map of SARS-CoV-2 ORF8, SARS-CoV ORF8a/b, and SARS-CoV GZ02 ORF8. The identical SSSC sequences between SSSCPred100 and SSSCPred and the corresponding observed ones are colored (blue: α-helix-type conformation; red: β-sheet-type conformation; green: a variety of other-type conformations). A comparison of the SSSCPreds result (SARS-CoV-2) with that of Quick2D is also shown.
predicted and observed SSSCs with keyword-tagged datasets would be useful to understand viral evolution.

## COMPUTATIONAL 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 by using the SSSCview program (available online at https://staff.aist.go.jp/izuemi/h/SSSCPreds/index-e.html). 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 as before 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. The dataset for the deep neural network was prepared by using Python.

**SSSCPrediction.** Deep learning for the prediction of SSSCs from amino acid sequences was performed by using Neural Network Console 1.40 (https://dl.sony.com/app/). Neural Network Console is convenient artificial intelligence (AI) software and can automatically optimize trained networks using a Gaussian process. The revised template of network “12_residual_learning.sdcproj” for the standard MNIST (Modified National Institute of Standards and Technology) dataset was used to provide the initial structure of the deep neural network, which was then trained with our prepared training dataset. The obtained trained network is shown in Figure 9 (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 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 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 the concordance rate using agreements of H, S, T, and D symbols.

Comparison of SSSCPrediction with Quick2D was carried out by using an amino acid sequence of a PDB file (1a00_A: HEMOGLOBIN ALPHA CHAIN). 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. The CB513 is a nonredundant dataset, suitable for development of algorithms for prediction of the secondary protein structure. CB513 has been made for learning of the neural network for prediction of the secondary protein structure. The CullPDB dataset is a large nonhomologous sequence set produced by using the PISCES server, which culls subsets of protein sequences from the Protein Data Bank based on sequence identity and structural quality criteria. 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 the concordance rate.

**Comparison Program (SSSCPreds) of Three Deep Neural Network-Based Prediction Systems.** Two additional deep neural network-based prediction systems were constructed by using procedures similar to that used to construct SSSCPrediction (SSSCPred). As before, 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 by using the SSSCview program. 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 as before 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 1.40 (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 (Figure 9) 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 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 50 SSSC terms; any remaining SSSC terms at the end of the sequence were also selected.

The three prediction programs for SSSCPreds (SSSCPred200, SSSCPred100, and SSSCPred) were obtained as follows. 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 sequence sampling range of SSSCPred200, SSSCPred100, and SSSCPred is considered to be optimal for the prediction of protein stability and rigidity from the sequences described above from the PDB. Consideration of longer sequences such SSSCPred300 would be limited to far fewer available sequences of such length, and consideration of shorter sequences, such as SSSCPred50 or SSSCPred20, would involve peptide sequences that are heavily influenced by solution environment factors that, if accounted for, would carry the correlations into a domain beyond the three programs SSSCPred200, SSSCPred100, and SSSCPred described here for stable protein sequences.

SSSCPreds is available as a standalone program at https://staff.aist.go.jp/izumi.h/SSSCPreds/index-e.html.

ASSOCIATED CONTENT

Supporting Information
The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acsomega.0c04472.

SSSCPred200 files with low concordance rates for CullPDB and CB513, comparison of predicted and observed SSSC sequences in SARS-CoV-2 proteins (spike protein and main protease), 200 subunit files containing SSSC sequence of common undeformable S2 motif, and 1372 subunit files containing SSSC sequence of common ORF8 motif (PDF)
References

1. Starr, T. N.; Greaney, A. J.; Hilton, S. K.; Ellis, D.; Crawford, K. H. D.; Dingens, A. S.; Navarro, M. J.; Bowen, J. E.; Tortorici, M. A.; Walls, A. C.; King, N. P.; Veesler, D.; Bloom, J. D. Deep mutational scanning of SARS-CoV-2 receptor binding domain reveals constraints on folding and ACE2 binding. Cell 2020, 182, 1295–1310.

2. Davenport, T. M.; Gorman, J.; Joyce, M. G.; Zhou, T.; Soto, C.; Guttman, M.; Moquin, S.; Yang, Y.; Zhang, B.; Donia-Rose, N. A.; Hu, S. L.; Mascia, J. R.; Kwong, P. D.; Lee, K. K. Somatic hypermutation-induced changes in the structure and dynamics of HIV-1 broadly neutralizing antibodies. Structure 2016, 24, 1346–1357.

3. Haft, D. H.; DiCuccio, M.; Badrettin, A.; Brover, V.; Chetverin, V.; O’Neill, K.; Li, W.; Chitsaz, F.; Derbyshire, M. K.; Gonzales, N. R.; Gwadz, M.; Lu, F.; Marchler, G. H.; Song, J. S.; Thakur, N.; Yamashita, R. A.; Zheng, C.; Thibaud-Nissl, F.; Geer, L. Y.; Marchler-Bauer, A.; Pruitt, K. D. RefSeq: an update on prokaryotic genome annotation and reference sequences: current status, policy and new initiatives. Nucleic Acids Res. 2018, 46, D851–D860.

4. Pruitt, K. D.; Tatusova, T.; Klimke, W.; Maglott, D. R. NCBI Reference Sequences: current status, policy and new initiatives. Nucleic Acids Res. 2009, 37, D32–D36.

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

6. Montgomerie, S.; Sundararaj, S.; Gallin, W. J.; Wishart, D. S. Improving the accuracy of protein secondary structure prediction using structural alignment. BMC Bioinformatics 2006, 7, 301.

7. 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.

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

9. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 1999, 292, 195–202.

10. 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.

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

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

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

14. 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.

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

16. Klausen, M. S.; Jespersen, M. C.; Nielsen, H.; Jensen, K. K.; Jurtz, V. I.; Sønderby, C. K.; Sommer, M. O. A.; 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.

17. Hansson, 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.

18. Almagro Armenteros, J. J.; Tsirigos, K. D.; Sønderby, C. K.; Petersen, T. N.; 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.

19. Zhang, C.; Zheng, W.; Huang, X.; Bell, E. W.; Zhou, X.; Zhang, Y. Protein Structure and Sequence Reanalysis of 2019-nCoV Genome at Its Spike Protein Insertions and HIV-1. Proteins: Structure, Function, and Bioinformatics 2020, 88, 1351–1360.

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

21. Gruber, M.; Söding, J.; Lupas, A. N. Comparative analysis of coiled-coil prediction methods. J. Struct. Biol. 2006, 153, 140–145.

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

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

24. Dostánová, Z.; Cisímová, V.; Tompa, P.; Simón, I. The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. J. Mol. Biol. 2005, 347, 827–839.

25. Ho, B. K.; Brasseur, R. The Ramachandran plots of glycine and pre-proline. BMC Struct. Biol. 2005, 5, 14.

26. Kleywegt, G. J.; Jones, T. A. Phi/psi-chology: Ramachandran plots. J. Mol. Biol. 2005, 13, 395–400.

27. Ramachandran, G. N.; Ramakrishnan, C.; Sasisekharan, V. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 1963, 7, 95–99.

28. Izumi, H. Homology searches using supersecondary structure code. Methods Mol. Biol. 2019, 1958, 329–340.

29. Izumi, H.; Wakisaka, A.; Nafie, L. A.; Dukor, R. K. Data mining of supersecondary structure homology between light chains of immuno-globulins and MHC molecules: absence of the common conformational fragment in the human IgM rheumatoid factor. J. Chem. Inf. Model. 2013, 53, 584–591.
Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res.* 2017, 27, 119–129.

(31) Yuan, Y.; Cao, D.; Zhang, Y.; Ma, J.; Qi, Q.; Wang, Q.; Lu, G.; Wu, Y.; Yan, J.; Shi, Y.; Zhang, X.; Gao, G. F. Cryo-electron microscopy structures of the SARS-CoV spike glycoprotein reveal a prerequisite conformational state for receptor binding. *Cell Res.* 2017, 27, 119–129.

(32) Guo, M.; Song, W.; Zhou, H.; Xu, J.; Chen, S.; 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.

(33) Yuan, Y.; Cao, D.; Zhang, Y.; Ma, J.; Qi, Q.; Wang, Q.; Lu, G.; Wu, Y.; Yan, J.; Shi, Y.; Zhang, X.; Gao, G. F. Cryo-electron microscopy structures of MERS-CoV and SARS-CoV spike glycoproteins reveal the dynamic receptor binding domains. *Nat. Commun.* 2017, 8, 15092.

(34) Kirchdoerfer, R. N.; Wang, N.; Pallesen, J.; Wrapp, D.; Turner, H. L.; Cottrell, C. A.; Corbett, K. S.; Graham, B. S.; McLellan, J. S.; Ward, A. B. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci. Rep.* 2018, 8, 15701.

(35) Xu, Y.; Lou, Z.; Liu, Y.; Pang, H.; Tien, P.; Gao, G. F.; Rao, Z. Crystal structure of severe acute respiratory syndrome coronavirus spike protein fusion core. *J. Biol. Chem.* 2004, 279, 49414–49419.

(36) Duquerroy, S.; Vigouroux, A.; Rottier, P. J. M.; Rey, F. A.; Bosch, B. J. Central ions and lateral asparagine/glutamine zippers stabilize the post-fusion hairpin conformation of the SARS coronavirus spike glycoprotein. *Virology* 2005, 335, 276–285.

(37) Wrapp, D.; Wang, N.; Corbett, K. S.; Goldsmith, J. A.; Hsieh, C. L.; Abiona, O.; Graham, B. S.; McLellan, J. S.; Ward, A. B. Stabilized coronavirus spikes are resistant to conformational changes induced by receptor recognition or proteolysis. *Sci. Rep.* 2018, 8, 15701.

(38) Yan, R.; Zhang, Y.; Li, Y.; Xia, L.; Guo, Y.; Zhou, Q. Structural basis for the recognition of SARS-CoV-2 spike by full-length human ACE2. *Science* 2020, 367, 1260–1263.

(39) Zhang, J.; Ge, Y.; Shi, K.; Wan, Y.; Luo, C.; Ashara, H.; Geng, Q.; Auerbach, A.; Li, F. Structural basis of receptor recognition by SARS-CoV-2. *Nature* 2020, 221.

(40) Walle, A. C.; Park, Y.; Tortorici, M. A.; All, A.; McGuire, A. T.; Veesler, D. Structure, Function, and Antigenicity of the SARS-CoV-2 Spike Glycoprotein. *Cell* 2020, 281.

(41) Wang, Q.; Zhang, Y.; Wu, L.; Niu, S.; Song, C.; Zhang, Z.; Lu, G.; Qiao, C.; Hu, Y.; Yuen, K. Y.; Wang, Q.; Zhou, H.; Yan, J.; Qi, J. Structural and Functional Basis of SARS-CoV-2 Entry by Using Human ACE2. *Cell* 2020, 894.

(42) 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, 215.

(43) Yuan, M.; Wu, N. C.; Zhu, X.; Lee, C.-C. D.; So, R. T. Y.; Lv, H.; Mok, C. K. P.; Wilson, I. A. A highly conserved cryptic epitope in the receptor-binding domains of SARS-CoV-2 and SARS-CoV. *Science* 2020, 630.

(44) Tove, W. G.; Baakman, C.; Black, J.; te Beek, T. A. H.; Krieger, E.; Joosten, R. P.; Vriend, G. A series of PDB-related databanks for everyday needs. *Nucleic Acids Res.* 2015, 43, D364–D368.

(45) Muth, D.; Corman, V. M.; Roth, H.; Binger, T.; Dijkman, R.; Gottula, L. T.; Glora-Rausch, F.; Ballonni, A.; Battilani, M.; Rüthart, D.; Toplak, I.; Ameneros, R. S.; Pfeifer, A.; Thié, V.; Drexler, J. F.; Müller, M. A.; Drosten, C. Attenuation of replication by a 29 nucleotide deletion in SARS coronavirus acquired during the early stages of human-to-human transmission. *Sci. Rep.* 2018, 8, 15177.

(46) Lau, S. K. P.; Feng, Y.; Chen, H.; Luk, H. K. H.; Yang, W. H.; Li, K. S. M.; Zhang, Y. Z.; Huang, Y.; Song, Z. Z.; Chow, W. N.; Fan, R. Y. Y.; Ahmed, S. S.; Yeung, H. C.; Lam, C. S. F.; Cai, J. P.; Wong, S. S. Y.; Chan, J. F. W.; Yuen, K. Y.; Zhang, H. L.; Woo, P. C. Y. Severe Acute Respiratory Syndrome (SARS) Coronavirus ORF8 Protein Is Acquired from SARS-Related Coronavirus from Greater Horseshoe Bats through Recombination. *J. Virol.* 2015, 89, 10532–10547.

(47) Li, J. Y.; Liao, C. H.; Wang, Q.; Tan, Y. J.; Luo, R.; Qi, Y.; Ye, X. Y. The ORF6, ORF8 and nucleocapsid proteins of SARS-CoV-2 inhibit type I interferon signaling pathway. *Virus Res.* 2020, 286, 195074.

(48) Flower, T. G.; Buflalo, C. Z.; Heo, R. M.; Allaire, M.; Ren, X.; Hurley, J. H. Structure of SARS-CoV-2 ORF8, a rapidly evolving coronavirus protein implicated in immune evasion. *BioRxiv* 2020, 1.

(49) Zhang, H.; Astrof, N. S.; Liu, J. H.; Wang, J. H.; 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. *EASEB J.* 2009, 23, 2735–2740.

(50) Qu, A.; Leahy, D. J. The role of the divalent cation in the structure of the I domain from the CD11a/CD18 integrin. *Structure* 1996, 4, 931–942.

(51) Helal, C. J.; Kang, Z.; Lucas, J. C.; Gant, T.; Ahlijianian, M. K.; Schachter, J. B.; Richter, K. E. G.; Cook, J. M.; Menniti, F. S.; 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.

(52) Jeffrey, P. D.; Russo, A. A.; Polvak, G.; Gibbs, E.; Hurwitz, J.; Massague, J.; Pavletich, N. P. Mechanism of CDK activation revealed by the structure of a cyclin-A-CDK2 complex. *Nature* 1995, 376, 313–320.

(53) Boume, T.; Watson, M. H.; Hickey, M. J.; Hickey, J.; Rocque, W.; Reed, S. I.; Tainer, J. A. Crystal Structure and Mutational Analysis of the Human CDK2 Kinase Complex with Cell Cycle–Regulatory Protein CksHsl. *Cell* 1996, 84, 863–874.

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

(55) Xia, S.; Liu, M.; Wang, C.; Xu, W.; Lan, Q.; Feng, S.; Qi, F.; Luo, L.; Du, L.; Liu, S.; Qin, C.; Sun, F.; Shi, Z.; Zhu, Y.; Jiang, S.; Lu, L. Inhibition of SARS-CoV-2 (previously 2019-nCoV) infection by a highly potent pan-coronavirus fusion inhibitor targeting its spike protein that harbors a high capacity to mediate membrane fusion. *Cell Res.* 2020, 30, 343–355.

(56) Jurtz, V. I.; Johansen, A. R.; Nielsen, M.; Armenteros, J. J.; Nielsen, H.; Sonderby, C. K.; Winther, O.; Sonderby, S. K. An introduction to deep learning on biological sequence data: examples and solutions. *Bioinformatics* 2017, 33, 3685–3690.

(57) Hofp, T. A.; Green, A. G.; Schubert, B.; Mersmann, S.; Schärfe, C. P. I.; Ingraham, J. B.; Toth-Petroczy, A.; Brock, K.; Riesselman, A. J.; Palmedo, P.; Kang, C.; Sheridan, R.; Draizen, E. J.; Dallago, C.; Sander, C.; Marks, D. S. The EVcouplings Python framework for coevolutionary sequence analysis. *Bioinformatics* 2019, 35, 1582–1584.