Protein secondary structure assignment using *pc*-polyline and convolutional neuron network

SSE assignment by P2PSSE

Lincong Wang[1]*, Chen Cao[2], Shuxue Zuo[1]

[1]The College of Computer Science and Technology, Jilin University, Changchun, Jilin, China, and [2]Department of Biochemistry & Molecular Biology, Alberta Children's Hospital Research Institute, University of Calgary, Calgary, Canada

Abstract

Motivation: The assignment of protein secondary structure elements (SSEs) underpins the structural analysis and prediction. The backbone of a protein could be adequately represented using a *pc*-polyline that passes through the centers of its peptide planes. One salient feature of *pc*-polyline representation is that the secondary structure of a protein becomes recognizable in a matrix whose elements are the pairwise distances between two peptide plane centers. Thus a *pc*-polyline could in turn be used to assign SSEs.

Results: Using convolutional neuron network (CNN) here we confirm that a *pc*-polyline indeed contains enough information for it to be used for the accurate assignments of six types of secondary structure elements: α-helix, β-sheet, β-bulge, 3_{10}-helix, turn and loop. The applications to three large data sets show that the assignments made by our CNN-based P2PSSE program agree very well with those by DSSP, STRIDE and quite well with those by five other programs. The analyses of the assignments by P2PSSE and those by other programs raise some general questions about the characterizations of protein secondary structure. In particular the analyses illustrate the difficulty with giving a quantitative and consistent definition for each of the six SSE types especially for 3_{10}-helix,

*Corresponding author: Lincong Wang, Email: wanglincong@jlu.edu.cn, wlincong@gmail.com.

β-bulge, turn or loop in terms of either backbone H-bond patterns, or backbone dihedral angles, or C_α-polylines or *pc*-polylines. The difficulty suggests that the SSE space though being dominated by the regions for the six SSE types is to a certain degree continuous.

Availability: The program is available at https://github.com/wlincong/p2pSSE.

Keywords: Protein secondary structure, peptide plane, secondary structure assignment, hydrogen bond, convolutional neuron network, machine learning.

1 Introduction

Given a structure the current methods for the assignment of protein secondary structure elements (SSEs) could be described in terms of data input and algorithm.

Historically the existence of protein secondary structure was first predicted by Pauling and Corey [1, 2] based on the common chemical structure of the twenty amino acids and the requirement for optimal hydrogen bonding (H-bonding) [1] interaction between backbone CO atoms and NH atoms. The prediction has been confirmed splendidly by the ever-increasing number of experimentally-determined structures [3]. The usefulness of secondary structure for protein structure classification, prediction, design and visualization has been well-documented, and is due in large part to the assumption that the number of SSE types in all the naturally-occurring proteins is small, that is, the space composed of all the SSEs in them is discontinuous. At present it is generally accepted that there are mainly six SSE types, α-helix, 3_{10}-helix, β-sheet, β-bulge, turn and loop, defined in terms of either the characteristic patterns of H-bonding interaction between backbone CO and NH atoms [1, 2, 4], or the lack of them. Specifically as used by H-bond based SSE assignment programs such as DSSP [4], STRIDE [5] and SECSTR [6], the current definitions for α-helix and β-sheet assume that their backbone H-bonding interactions[2] are locally optimal, and for 3_{10}-helix, β-bulge and turn at least sub-optimal [7], while

[1] Abbreviations: H-bond, hydrogen bond; H-bonding, hydrogen bonding; SSE, secondary structure element; *pc*-polyline, a polyline that links the centers of the peptide planes of a protein; C_α-polyline, a polyline that links the C_α atoms of a protein; PDB, Protein Data Bank; CNN, convolutional neuron network; CO atom, carbonyl oxygen; NH atom, amide proton; ϕ, ψ angles, backbone dihedral angles; SI, Supplementary Information.

[2] The backbone H-bonding interaction of a fragment or a sheet that is composed of two or more fragments (strands) is defined as the H-bonding interaction of all the pairs of backbone NH atom and CO atom that belong to the fragment or

a loop is defined as any backbone fragment that lacks the H-bond patterns characteristic of the other five SSE types. However, the structure of a protein even in crystalline state has a total free energy that is at least somewhat close to its global minimum in solution. Since the expression for global free energy is dominated by the terms other than backbone H-bonding interaction it is to be expected that an SSE type corresponds not to one but an ensemble of backbone conformations. In other words, the backbone H-bonding energies likely differ from each other for the different fragments of the *same* length and having been assigned to the *same* SSE type by an H-bond based method. Furthermore at present it is very challenging to compute the precise bonding energy for an H-bond in a protein due in part to the difficulty of obtaining accurate values for dielectric constants and in part to the lack of the coordinates for the protons in the vast majority of the currently-available crystal structures. Consequently the SSE assignments by an H-bond based method will likely be ambiguous for the residues in those conformations whose backbone H-bonding interaction energy could not be computed accurately.

In order to reduce the possible inconsistencies in assignment originated from the uncertainty in H-bond energy calculation some H-bond based assignment programs [5, 8] have added to the backbone H-bond patterns geometrical constraints such as backbone ϕ, ψ angles. Take a step further about a half dozen assignment programs rely only on geometrical constraints. The program DISICL [9] uses only backbone ϕ, ψ angles while KAKSI [10] uses both ϕ, ψ angles and C_α-polyline[3]. Most of the other ones rely on both C_α-polyline [11, 12, 13, 10, 14, 15, 16, 7] *and* either ϕ, ψ angles or other types of angles computed from the coordinates for consecutive CA atoms. However, neither C_α-polyline nor backbone angles are capable of providing precise and quantitative definitions for all the six SSE types. The lack of precise definitions makes at least in practice algorithmically challenging the problem of assigning an SSE type to each residue. The difficulty manifests itself by the internal inconsistencies in assignment by any one of these programs, and by the disagreements among the different ones. The discrepancies among these programs are particularly pronounced for the residues in the HGTU twilight zone that borders α-helix, 3_{10}-helix, turn, loop regions *and* in the EBU twilight zone that

the sheet. In this paper the three dimensional structure of a protein subsequence will be called a fragment. The backbone conformation of a fragment is specified by either a set of consecutive C_α atoms in C_α-polyline representation or a set of consecutive peptide plane centers in *pc*-polyline representation.

[3]For brevity in this paper using a C_α-polyline means using the geometric features such as lengths and angles extracted from the polyline.

borders β-sheet, β-bulge and loop regions [17, 18].

From an algorithmic perspective, the previous SSE programs are either deterministic [4, 5, 6, 11, 12, 13, 10, 14, 7] using a greedy or heuristic approach, or non-deterministic using either an information-theoretical approach [8] or a supervised-learning approach [16, 19]. The latter takes the advantage of the availability of a large number of experimentally-determined protein structures. For example PCASSO [16] had used 258 geometrical features extracted from a C_α-polyline to first train a random forest on 282 PDBs and then to assign the residues in 194 PDBs to three SSE types only: helix, sheet and loop. A convolutional neuron network (CNN) approach has also been applied to characterize different types of loops [19]. A key advantage of these machine-learning based approaches [20] is that they do not require a precise and quantitative definition for SSE type.

In a previous paper [21] we have shown that instead of using a C_α-polyline the backbone of a protein is better represented by a *pc*-polyline that passes through the centers of its peptide planes. Most interestingly it is found that the geometrical property of a *pc*-polyline is closely related to the protein's SSE composition and further to its backbone H-bonding interaction. Thus the geometrical features of a *pc*-polyline are expected to be useful for SSE assignment. To further evaluate the relationship between *pc*-polyline and protein secondary structure, and to develop a novel assignment algorithm using *pc*-polyline we have trained a dozen of CNN models using the geometrical features extracted from a large set $\mathscr{L}$ of *pc*-polylines[4], and six labels that correspond, respectively, to α-helix (H), β-sheet (E), β-bulge (B), 3_{10}-helix (G), turn (T) and loop (U). The labels are obtained from DSSP[5], the current de facto H-bond based standard for SSE assignment. The best trained CNN model is then used by our P2PSSE to assign the six SSE types for two different large sets of *pc*-polylines, set $\mathbf{P}$ of $4,172$ *pc*-polylines with a total of $\approx 1.1 \times 10^6$ residues, and set $\mathbf{Q}$ of $4,703$ *pc*-polylines with a total of $\approx 1.1 \times 10^6$ residues. The crystal structures of both $\mathbf{P}$ and $\mathscr{L}$ have higher qualities than those of $\mathbf{Q}$. The assignments by P2PSSE agree very well with those by DSSP with respective overall agreements of 96.6% for $\mathscr{L}$, 92.6% for $\mathbf{P}$ and 91.7% for $\mathbf{Q}$. At individual SSE type level and for all the three sets the agreements between P2PSSE and DSSP from the best to the worst are in the order of α-helix, β-sheet, loop, turn, 3_{10}-helix and β-bulge.

[4] Set $\mathscr{L}$ has $14,607$ *pc*-polylines with a total of $\approx 3.7 \times 10^6$ residues.

[5] DSSP assigns SSEs to eight types: H, G, I (π-helix), E, B, T, S and C. We label both S and C as U while ignore all the Is.

To better assess the performances of P2PSSE we have also compared its assignments with those by other six programs [5, 13, 22, 7, 16, 10] where STRIDE [5] is based mainly on H-bonds while the other five ones use only geometric data. Except for STRIDE and SEGNO the four programs, P-SEA, PROSS, PCASSO and KAKSI could at best identify three SSE types: *helix, sheet* and *loop*. The program SEGNO assigns a residue to one of the following five SSE types: α-helix, 3_{10}-helix, π-helix, poly-proline helix and β-sheet. For the purpose of comparison both our assignments and those by the other six programs are grouped into the above three SSE types. With such a grouping P2PSSE agrees best with DSSP with 97.2% for $\mathscr{L}$, 94.7% for $\mathbf{P}$ and 93.0% for $\mathbf{Q}$, and agrees very well with both STRIDE and PCASSO. Furthermore the agreements between P2PSSE and the four other programs are between 85.7% and 81.6%. These agreements are almost identical to their agreements with DSSP. Taken together the results show that the assignments by P2PSSE are highly accurate, and furthermore are not biased towards DSSP even though the assignments by the latter have been used as labels to train the CNN models. In terms of both input data and algorithm our method is most similar to PCASSO. However, compared with the latter P2PSSE uses far less geometrical features (data)[6], 20 features per residue on average, and is able to assign not three but all the six SEE types. In terms of data input P2PSSE is likely the most accurate SSE assignment program among all the currently-available ones that use geometric data only.

2 Materials and Methods

2.1 The data sets

The crystal structures are first downloaded from the PDB [3], then processed using our molecular analysis and visualization program, and finally the *pc*-polylines for individual chains are computed using P2P as described previously [21]. If a structure contains several identical chains only the first chain is selected. The chains with any gaps[7] are excluded. Protons are added using the program REDUCE [23] to any PDB that lacks their coordinates.

[6] In this paper we will use geometrical features and geometrical data interchangeably.

[7] A gap in a protein chain means that either one or several consecutive interior residues have no ATOM statement in the PDB file.

We separate the computed *pc*-polylines into two sets: a learn (train) set $\mathscr{L}$ and an assignment set $\mathscr{A}$ where $\mathscr{L}$ is used for learning and $\mathscr{A}$ for assignment. To assess model's data sensitivity, we have trained a series of CNN models using *four* sets of the geometrical features extracted using respectively $T_{pp} = 6.0, 7.0, 8.0, 9.0$Å where T_{pp} is a threshold for the pairwise distance between two peptide plane centers. For brevity such a distance will be called a *p2p distance* and denoted as d_{pp}. The four learning sets and the four assignment sets are denoted respectively as $\mathscr{L}_6, \mathscr{L}_7, \mathscr{L}_8, \mathscr{L}_9$, and as $\mathscr{A}_6, \mathscr{A}_7, \mathscr{A}_8, \mathscr{A}_9$. Each of the four assignment sets is further divided into two sets: set $\mathbf{P}$ computed on a set of high quality structures and set $\mathbf{Q}$ of low quality structures[8]. Each structure in $\mathscr{L}$ and $\mathbf{P}$ has a resolution ≤ 2.5Å and a R-free factor < 0.285 and $< 70\%$ sequence identity with any others. Each structure in $\mathbf{Q}$ has a resolution between $2.5-3.5$Å and a R-free factor ≥ 0.285 and $< 70\%$ sequence identity with any others. The sets $\mathscr{L}$, $\mathbf{P}$ and $\mathbf{Q}$ have respectively $3,686,465$, $1,135,055$ and $1,120,200$ residues.

2.2 The geometrical features extracted from a *pc*-polyline

For easy exposition let the vector between adjacent peptide plane centers $\mathbf{c}_i$ and $\mathbf{c}_{i+1}$ be $\mathbf{v}_{pp}(i, i+1)$, the p2p distance between peptide plane center[9] i and center j be $d_{pp}(i,j)$ and their sequence distance be $s_{pp}(i,j)$. For residue i its geometrical features are (a) all the p2p distances $d_{pp}(i,j)$s that satisfy $d_{pp}(i,j) < T_{pp}$ where T_{pp} is a predefined threshold as described above, (b) their corresponding $s_{pp}(i,j)$s, and (c) the five angles[10] $\theta(i,k)$ between $\mathbf{v}_{pp}(i,i+1)$ and five other $\mathbf{v}_{pp}(k,k+1)$s where $k-i = 1,2,3,4,5$. Both the *pc*-polylines and their geometrical features are computed using the program P2P (section S1 of Supplementary Information (SI)). The number of geometrical features increases with threshold T_{pp}. For each residue the number of $d_{pp}(i,j)$s is equal to the number of $s_{pp}(i,j)$s and both depend on the number of neighboring peptide planes that have $d_{pp} < T_{pp}$. The set of geometrical features extracted from all the individual *pc*-polylines in $\mathscr{L}$ are concatenated into a single file with no specific labels for both termini of a protein chain. This file is used for training CNN models. Likewise the geometrical features extracted from all the individual *pc*-polylines in either $\mathbf{P}$ or $\mathbf{Q}$ are concatenated into single files. The last two files are used for SSE assignments.

[8]By abuse of notation $\mathscr{L}$, $\mathbf{P}$ and $\mathbf{Q}$ without subscript may mean a set of protein structures, or a set of computed *pc*-polylines, or any one of the four sets of geometrical features extracted from a set of *pc*-polylines. Their meanings should be clear by the context.

[9]The peptide plane i for residue i is defined as the peptide plane between residue i and $i+1$.

[10]The residues at and near the C-terminal of a chain may have less than five angles, a default value of π is used for the missing angles.

2.3 The training of CNN models

The CNN models are trained and tested using TensorFlow 1.4 [24] with simple architectures and routine hyperparameters (section S1 of SI) [20]. The labels are the following six SSE types: α-helix (H), 3_{10}-helix (G), β-sheet (E), β-bulge (B), turn (T) and loop (U) [11] assigned by DSSP. To evaluate model's data sensitivity the same architecture and hyperparameters are used for training and testing the four CNN models on respective sets $\mathscr{L}_6, \mathscr{L}_7, \mathscr{L}_8$ and $\mathscr{L}_9$. Up to a dozen of training and testing are performed in order to find a model's best hyperparameters. Out of $\mathscr{L}_6, \mathscr{L}_7, \mathscr{L}_8$ and $\mathscr{L}_9$ the best agreement with DSSP is obtained with a model trained on $\mathscr{L}_8$. In the following unless stated otherwise the CNN model used by P2PSSE for SSE assignment will mean this particular one. The corresponding two sets of geometrical features extracted respectively from the *pc*-polylines in $\mathbf{P}$ and $\mathbf{Q}$ will be denoted by $\mathbf{P}_8$ and $\mathbf{Q}_8$.

2.4 The SSE assignments by P2PSSE and the comparisons with previous methods

For each residue P2PSSE assigns its SSE type to the label with the largest value among the six probabilities, $p(\mathrm{H}), p(\mathrm{E}), p(\mathrm{B}), p(\mathrm{G}), p(\mathrm{T})$ and $p(\mathrm{U})$. As detailed late in section 3 the SSEs assigned by the CNN model trained on $\mathscr{L}_8$ agree very well with those by DSSP. Consequently our analyses[12] of the SSE assignments by both DSSP and P2PSSE focus on their disagreements rather than their agreements. The interesting differences are further examined at individual fragment level using our molecular analysis and visualization program. To further evaluate the performances of P2PSSE we have also compared its assignments on all the three sets ($\mathscr{L}$, $\mathbf{P}$ and $\mathbf{Q}$) with the assignments by six other programs [5, 13, 22, 7, 16, 10].

3 Results and Discussion

In this section we first describe the geometrical data for $\mathscr{L}_6, \mathscr{L}_7, \mathscr{L}_8$ and $\mathscr{L}_9$ used to train CNN models. We then analyze the assignments by both DSSP and P2PSSE first at residue level and then at fragment level. The analyses focus on their differences in α-helix assignment and in β-sheet assignment. In

[11] The six SSE types are abbreviated respectively as H, E, B, G, T, and U.

[12] Simple python scripts and a C++ program have been written specifically for the analyses.

particular we examine in somewhat detail the α-helices assigned by P2PSSE that could not be matched to any α-helix by DSSP and vice versa. Finally we describe the key differences in assignments by P2PSSE and six other programs. For these comparisons all the residues are assigned only to three SSE types: helix, sheet and loop. Our discussion centers on the differences in assignment and how the differences are related to the HGTU twilight zone and to the EBU twilight zone in the SSE space.

3.1 The geometrical data extracted from *pc*-polylines

The data include three types of geometrical properties: p2p distance (d_{pp}), sequence distance (s_{pp}) and angle. The number of d_{pp}s for a residue depends on T_{pp} and on whether the residue is buried or on surface. As shown in Fig. 1 and Figure S1 of SI the average numbers of d_{pp}s per residue are 2.87 for $\mathscr{L}_6$, 4.90 for $\mathscr{L}_7$, 7.01 for $\mathscr{L}_8$ and 10.11 for $\mathscr{L}_9$. Their respective medians are 4, 6, 7 and 12. Since the best agreement with DSSP is achieved by $\mathscr{L}_8$ it seems that neither $\mathscr{L}_6$ nor $\mathscr{L}_7$ has enough features while $\mathscr{L}_9$ has too many noisy features. The average number of geometrical features per residue is 20 for $\mathscr{L}_8$ consisting of 8 d_{pp}s, 7 s_{pp}s and 5 angles per residue. By comparison PCASSO [16] uses 258 features per residue, a 12-fold more than our average.

3.2 The analyses of the SSE assignments at residue level by P2PSSE and DSSP

In this section we describe the assignments for all the six SSE types by both DSSP and P2PSSE with an emphasis on their differences. The main goal here is to assess the consistencies of the criteria used by DSSP, and by extension to assess the difficulty of the SSE assignment problem itself. The rational for such an analysis is that the level of difficulty to train a CNN model correlates with the degree of errors in the labels, in this case, the SSE assignments by DSSP. We focus on the assignments for $\mathscr{L}_8$ with references to the assignments for both $\mathbf{P}_8$ and $\mathbf{Q}_8$. In theory to best assess the DSSP's criteria we should use the assignments by P2PSSE for all the three sets, $\mathscr{L}_8$, $\mathbf{P}_8$ and $\mathbf{Q}_8$. However since the agreements in assignment with DSSP for both $\mathbf{P}_8$ and $\mathbf{Q}_8$ are only slightly worse than that for $\mathscr{L}_8$ the conclusions and discussions based on the performance of P2PSSE on $\mathscr{L}_8$ should be general. With the above consideration and because of space limitations, the main text only analyzes in somewhat detail the assignments for $\mathscr{L}_8$ while those for $\mathbf{P}_8$ and $\mathbf{Q}_8$ are only mentioned briefly. The assignments for

$\mathbf{P}_8$ and $\mathbf{Q}_8$ are detailed respectively in sections S3 and S4 of SI. In addition the analyses of the β-sheet assignments for $\mathbf{P}_8$, $\mathbf{Q}_8$ and $\mathscr{L}_8$ by both P2PSSE and DSSP are presented respectively in sections S3, S4 and S5 of SI.

3.2.1 The agreements at residue level between P2PSSE and DSSP

Overall P2PSSE performs very well on all the three sets, $\mathscr{L}_8$, $\mathbf{P}_8$ and $\mathbf{Q}_8$, though the overall agreements for all the six types decrease somewhat from 96.6% for $\mathscr{L}_8$ (Table 1) to 92.6% for $\mathbf{P}_8$ (Table S1 in SI) and to 91.2% for $\mathbf{Q}_8$ (Table S6 of SI). At individual type level the agreements from the best to the worst are H, E, U, T, G and B (Table 1, and Tables S1 and S6 of SI). Except for β-bulge, both the accuracy (precision) and the recalls for the other five types are ≥ 0.92 for $\mathscr{L}_8$, ≥ 0.83 for $\mathbf{P}_8$ and ≥ 0.75 for $\mathbf{Q}_8$. The recall value for β-bulge assignment is only 0.773 for $\mathscr{L}_8$, much smaller than those for the other five SSE types. Both the accuracy and recalls for β-bulge assignment in either $\mathbf{P}_8$ or $\mathbf{Q}_8$ are even lower with 0.594/0.410 for $\mathbf{P}_8$ (Table S1 of SI) and 0.587/0.384 for $\mathbf{Q}_8$ (Table S6 of SI). Among the possible reasons for the low agreements are: (a) a β-bulge is well-defined by neither backbone H-bond nor *pc*-polyline, and (b) the number of β-bulge residues in $\mathscr{L}$ is too small: only $41,385$, that is at least 3.6-fold smaller than those for the other five types (Table 1). Though it is well-known that the accuracy of a CNN model increases with the number of data in the train set the nondeterministic nature of neuron network makes it difficult to pinpoint either (a) or (b). On the other hand the high agreements with DSSP for the other five types by the same CNN model suggest that the low agreement in β-bulge assignment is likely due to the inconsistencies in the criteria used by DSSP for β-bulge assignment.

3.2.2 The distributions of the assignments at residue level by P2PSSE not agreed to by DSSP and vice versa

As described above P2PSSE first computes the six probabilities, $p(\mathrm{H}), p(\mathrm{E}), p(\mathrm{B}), p(\mathrm{G}), p(\mathrm{T})$ and $p(\mathrm{U})$, for a residue and then chooses the type with the largest probability as its assignment. However if we also count as an agreement when the type with the second largest probability agrees with DSSP, then the agreements between P2PSSE and DSSP increase respectively to 99.7% for $\mathscr{L}_8$, 98.7% for $\mathbf{P}_8$ and 98.3% for $\mathbf{Q}_8$. In other words, the percentages that neither the type with the largest probability nor the type with the second largest agrees with DSSP are $< 1.7\%$ for all the three sets. The increases

in agreement and the small differences among $\mathscr{L}_8$, $\mathbf{P}_8$ and $\mathbf{Q}_8$ point to the continuity in the HGTU zone and in the EBU zone in the SSE space. Their existence is further supported by the following analyses of the distributions of the assignments at both residue and fragment levels by P2PSSE not agreed to by DSSP and vice versa.

As shown in Table 2 P2PSSE assigns a DSSP-assigned H to $\mathcal{T}, G$ and $\mathcal{U}$ [13] with very small and decreasing percentages: 1.26%, 0.28% and 0.14%, and to $\mathcal{E}$ and $\mathcal{B}$ with rare possibilities. Likewise as shown in Table 3 DSSP assigns a P2PSSE-assigned $\mathcal{H}$ to T, G and U with very small and decreasing percentages: 0.65%, 0.22% and 0.03%, and to E and B with rare possibilities. As to β-sheet residues P2PSSE assigns a DSSP-assigned E to $\mathcal{U}, B$ and $\mathcal{T}$ with very small and decreasing percentages: 3.13%, 0.62% and 0.04%, and to $\mathcal{H}, G$ with rare possibilities. Likewise DSSP assigns a P2PSSE-assigned $\mathcal{E}$ to U, B and T with very small and decreasing percentages: 2.03%, 0.14% and 0.08%, and to G and H with rare possibilities (Table 3). Similar trends exist for both the α-helix and β-sheet assignments by P2PSSE and by DSSP for $\mathbf{P}_8$ (Tables S2 and S3 of SI) and for $\mathbf{Q}_8$ (Tables S7 and S8 of SI) though the percentages for the "wrong" assignments[14], that is the differences in assignment, do increase slightly.

The existence of the above "wrong" assignments shows that in terms of either H-bond pattern or the geometrical features extracted from a *pc*-polyline some Ts, Gs and Us are so similar to Hs, and some Us and Bs are so similar to Es that neither the criteria used by DSSP nor those by P2PSSE are capable of separating them from each other with certainty. In other words, as to be confirmed late at fragment level, there exist two twilight zones in the SSE space, the HGTU zone borders H, G, T and U regions and the EBU zone borders E, B and U regions. Furthermore though these two zones are well separated from each other, they are connected through a small numbers of Us and Ts. In other words the SSE space is continuous to some degree.

[13] For easy distinguish we use latex mathcal font to label the six SSE types assigned by P2PSSE.

[14] For brevity when we regard the assignment by DSSP (or by P2PSSE) as the standard then the assignment by P2PSSE (or by DSSP) that does not agree with the standard will be designated as "wrong".

3.3 The comparisons of α-helix assignments by DSSP and P2PSSE

The generally-accepted definition of an α-helix in terms of H-bond pattern requires the existence of at least one $i \to i+4$ H-bond and its length[15] be at least 4. Since no restraints on the length of any of the six types are imposed on P2PSSE, unlike the lengths of the α-helices by DSSP, the length of an α-helix by P2PSSE could be < 4. As a piece of evidence for the power of convolutional neuron network and for the close relationship between the geometrical properties of a *pc*-polyline and the protein's secondary structure, P2PSSE has assigned only very small percentages of the residues to the α-helices with < 4 residues: $1,051$ (0.08%) residues for $\mathscr{L}_8$, $2,872$ (0.73%) for $\mathbf{P}_8$ (Table S4 of SI) and $3,028$ (0.75%) for $\mathbf{Q}_8$ (Table S9 of SI). In the following we exclude any α-helix with < 4 residues.

As shown in Table 4 at helix level P2PSSE assigns $110,009$ α-helices for $\mathscr{L}_8$, that is $1,678$ less than the total number of α-helices $(111,687)$ assigned by DSSP. Out of the $110,009$ α-helices, $89,139$ (81.0%) agree exactly with DSSP. Except for 172 (0.16%) α-helices that could not be matched to any α-helix by DSSP, all the remaining ones agree with DSSP to some extents. Specifically each of the $7,574$ α-helices by P2PSSE is a part of an α-helix by DSSP, each of the 186 α-helices extends both termini of a DSSP-assigned α-helix, while $11,916$ and $1,110$ α-helices extend, respectively, the N-termini and the C-termini of the matched DSSP-assigned α-helices. If all these α-helices (plus those in the 9th and 10th columns in Table 4) are counted as agreement, then there are $109,837$ (99.84%) P2PSSE-assigned α-helices in total that agree with DSSP. One salient feature of P2PSSE is that it extends the N-termini of $11,916$ (10.8%) DSSP-assigned α-helices. In addition the average length of the α-helices by P2PSSE is 11.3, slightly longer than the average (11.1) by DSSP.

DSSP assigns $111,687$ α-helices for $\mathscr{L}_8$. Out of them, $1,754$ could not be matched with any α-helix by P2PSSE, and $13,212$ are embedded completely inside the α-helices by P2PSSE, and only $6,355$ extend the N-termini of the α-helices by P2PSSE (Table 4). Similar results are obtained for $\mathbf{P}_8$ (Table S4 of SI) and $\mathbf{Q}_8$ (Table S9 of SI). As is well-documented [10, 18], the current SSE assignment programs disagree largely on the assignments of both termini for an α-helix due apparently to the

[15]The length of a SSE type, that is, the length of an α-helix, a 3_{10}-helix, a β-strand, a β-bulge, a turn and a loop, is defined as the number of consecutive residues all been assigned to the same SSE type.

fact that many of the terminal residues are in the HGTU twilight zone. On the other hand since no geometrical features are used to define terminal residues and since our CNN models have been trained with more than 1.2×10^6 α-helix residues where the vast majority of them are the interior ones, the criteria used by P2PSSE for the assignments of α-helix termini should be the same as those for the interior residues. In other words, the assignments for both termini by P2PSSE are likely more reliable than those by DSSP. Due to space limitations we will only describe in somewhat detail the 172 α-helices by P2PSSE that could not be matched to any α-helix by DSSP and the $1,754$ α-helices by DSSP that could not be matched to any by P2PSSE.

3.3.1 The α-helices assigned by P2PSSE but not by DSSP

As shown in the last column of Table 4, there are 172 helices in $\mathscr{L}_8$ that could be assigned only by P2PSSE. We have examined all of them in somewhat detail. Out of the 172 helices, only five have lengths > 4. Overall it is unclear to us why DSSP does not assign them to α-helix since most of them have $i \to i+3, i \to i+4$ backbone H-bonds according to a formula for H-bonding energy computation used in DSSP. On the other hand, most of them have more than one type of H-bond, some of them even have $i \to i+2, i \to i+5$ H-bonds. One may intend to conclude that the existence of different H-bond types is the reason why DSSP does not assign them to α-helix if not for the cases where DSSP does not hesitate to assign α-helices even there exist different types of H-bonds. One possibility is that the amide protons if missing from the original PDB have been added differently: we use REDUCE while DSSP has it own method to protonate a structure. Less than a half dozen (see Fig. 2c for an example) have no detectable H-bonds. The vast majority of the residues in the 172 helices have relatively large $p(\mathrm{H})$ values (Figs. 2a, 2c and Table 5). However some do have relatively small $p(\mathrm{H})$s (Fig. 2e and Table 5). In general an α-helix with smaller $p(\mathrm{H})$s has more pronounced distortions (Fig. 2e) in helical geometry [25] because for such a helix its $p(T)$s are relatively large in general. On the other hand, as illustrated in Fig. 2, the vast majority of these helices do have helical geometries if not better than then at least as good as those for a typical short α-helix assigned by DSSP. It is likely that DSSP program has some inconsistencies in the criteria used for α-helix assignment because its α-helix definition in terms of H-bond pattern is qualitative rather than quantitative.

3.3.2 The α-helices assigned by DSSP but not by P2PSSE

Out of $111,687$ α-helices by DSSP, $1,754$ could not be matched to any α-helix by P2PSSE (Table 4). We have examined all of them in somewhat detail. Out of the $1,754$ α-helices only 9 include neither $\mathcal{G}$ nor $\mathcal{H}$ assigned by P2PSSE, only 72 do not include any $\mathcal{H}$ by P2PSSE. More than a third (653) include at least one residue assigned to $\mathcal{G}$ by P2PSSE. Most of the $1,754$ α-helices are on surface, and thus more flexible than the buried ones [26]. The flexibility may explain the distortions in their helical geometries. In the following we examine in detail the assignments by P2PSSE and by DSSP for four fragments to illustrate the backbone conformations formed by the residues in the HGTU twilight zone.

Helix Y124-R133 in 5ktm by DSSP. As shown in Figs. 3a and 3b, this helix in 5ktm includes both $i \rightarrow i+3$ and $i \rightarrow i+4$ H-bonds, and the latter may explain why DSSP has assigned the four residues to Hs. In addition they do assume a helical geometry though with some distortions (Fig. 3a). However, there apparently exists a turn from Y126 to D130 (Fig. 3a and 3b). The geometrical features of the four residues extracted from the *pc*-polyline appear to resemble those for a typical DSSP-assigned α-helix but also share some features for a typical DSSP-assigned turn. As shown in Table 6, all the four residues have $p(\mathrm{H})$s as their second largest probabilities. This may explain why P2PSSE assigns them to $\mathcal{T}$s but DSSP assigns them to Hs. This helix represents a backbone conformation in the HGTU zone.

Helix L296-D299 in 5mx9 by DSSP. As shown in Fig. 3c and 3d, one $i \rightarrow i+4$ H-bond between G297 and Y301 may be the reason why the 4-residue fragment, L296-D299, is assigned as an α-helix by DSSP. However geometrically this fragment looks more like a turn than a typical α-helix. Except for G297 whose $p(\mathrm{T}) = 0.8015$ none of the six probabilities for the other three residues are > 0.53 (Table 6). Only L296 and D299 have their $p(\mathrm{H})$s as the second largest. This helix provides an example of a backbone conformation in the HGTU zone.

Helix R500-T503 in 1cza and helix G14-Y17 in 2z2i by DSSP. The fragment R500-T503 in 1cza is assigned as an α-helix while their P2PSEE's assignments are $\mathcal{TTTG}$ with the $p(\mathrm{H})$s for R500, K501 and Q502 as the second largest (Table 6). A single $i \rightarrow i+4$ H-bond between R500 and H504 may explain why DSSP assigns them as an α-helix (Figs. 3e and 3f). The fragment G14-Y17 in 2z2i

has both helical and turn geometries (Figs. 3g and 3h). It has no $i \to i+4$ H-bond but two $i \to i+3$ H-bonds. Except for the last residue Y17, the largest probability for each of the other three residues is < 0.5. Only two of them have $p(\mathrm{H})$s as their second largest probabilities. Both fragments could serve as examples in the HGTU zone.

3.4 The comparisons of β-sheet assignments by DSSP and P2PSSE for $\mathscr{L}_8$

In this section we describe the distributions of the β-sheet assignments at residue level by P2PSSE not agreed to by DSSP and vice versa. Due to space limitation the comparisons at fragment level and three examples of β-strands in the the EBU twilight zone are presented in section S5 of SI.

A β-strand by P2PSSE could have a single residue while those by DSSP have at least two residues[16]. In the following we only analyze and compare with DSSP the β-strands with ≥ 2 residues. As shown in Table 7 P2PSSE assigns $149,909$ β-strands with a total of $796,747$ residues and an average length of 5.31 residues. Out of them $119,049$ (79.4%) agree exactly with DSSP while only $1,249$ (0.83%) could not be matched to any strand by DSSP, and $29,611$ of them could be partially matched to a strand by DSSP. Likewise DSSP assigns $149,655$ strands with a total of $798,642$ residues and an average length of 5.34 residues, both numbers are slightly larger than those by P2PSSE. Out of them $119,049$ (79.5%) agree exactly with P2PSSE while $2,398$ (1.60%) could not be matched to any strand by P2PSSE, and $28,208$ of them could only be partially matched to a strand by P2PSSE.

3.5 The comparisons with six other SSE assignment programs

The CNN models are trained using the six labels from DSSP and thus the assignments by P2PSSE are possibly biasd towards the latter. To evaluate the extent of bias, we have compared the assignments on all the three sets, $\mathscr{L}_8$, $\mathbf{P}_8$ and $\mathbf{Q}_8$, by P2PSSE and by six other programs [5, 13, 22, 7, 10, 16]. Among them STRIDE [5] is most close to DSSP but requires both H-bond and dihedral angle while P-SEA, SENGO, PROSS, PICCASO [13, 22, 7, 16] uses only geometrical data (distances and angles) computed from C_α-polylines. KAKSI [10] uses backbone dihedral angles. From algorithmic viewpoint these five

[16]Though rare there do exist some β-strands with a single residue by DSSP. We ignore them here.

programs are deterministic but greedy or heuristic in nature. Among the six programs our P2PSSE is most similar to PCASSO in that both are based on supervised machine-learning and the input data for both are purely geometrical data computed from either *pc*-polyline or C_α-polyline. As far as input is concerned the geometric features used by P2PSSE are most similar to those used by SABA [15] since the center of a peptide plane is close to the middle position of two consecutive C_α atoms. However we are not able to compare our assignments with those by SABA due to the inaccessibility of the latter. Likely due to the deficiencies in the purely geometric data used as inputs, the four programs, P-SEA, PROSS, KAKSI and PCASSO could only assign the residues to three types: helix, sheet and loop while SEGNO assigns a residue to one of the following five SSE types: α-helix, 3_{10}-helix, π-helix, poly-proline helix and β-sheet. For comparison only we label both P2PSSE-assigned $\mathcal{H}$ and $\mathcal{G}$ as helix, $\mathcal{E}$ and $\mathcal{B}$ as sheet and $\mathcal{T}$ and $\mathcal{U}$ as loop.

As shown in Table 8 with all the assignments by different programs being grouped into only the three types (Table S12 of SI), the assignments by P2PSSE for the train set $\mathscr{L}_8$ agree very well with both DSSP (97.2%) and STRIDE (93.9%) and quite well with PCASSO (91.7%). Furthermore, the agreements between P2PSSE and P-SEA, SENGO, PROSS, KAKSI are almost identical to their agreements with DSSP. Taken together the results show that the bias of the CNN model towards DSSP is very small. A more objective assessment of the performances of P2PSSE is to compare the assignments on sets $\mathbf{P}_8$ and $\mathbf{Q}_8$ that have not been used in learning. As shown in Table 9 for $\mathbf{P}_8$ the agreements between P2PSSE and DSSP, STRIDE, PCASSO decrease slightly to 94.7%, 92.3% and 91.4% while the agreements with other four programs remain almost the same. The small decreases in agreement from the train set $\mathscr{L}_8$ to the evaluation set $\mathbf{P}_8$ indicate that the training of the CNN model is quite adequate but not perfect. Further tuning of both the architecture and the hyperparameters for training CNN model models may increase somewhat assignment agreements with other programs and thus improve the performances of P2PSSE.

H-bond based programs such as DSSP, STRIDE and P2PSSE require the coordinates for backbone NHs but except for ultra-high resolution crystal structures the vast majority of the crystal structures currently available have no coordinates for them. Their coordinates must be computed theoretically. In P2PSSE these NHs are added using REDUCE [23] while both DSSP and STRIDE have their own ways

for protonation. The accuracy of protonation depends on the quality of the structure. To evaluate the impacts on assignment the accuracy of protonation in particular and structural quality in general we have selected a set $\mathbf{Q}_8$ of crystal structures with low qualities as judged by X-ray resolution and R-free factor. As shown in Table 10 compared with those for $\mathbf{P}_8$ the agreements for $\mathbf{Q}_8$ between P2PSSE and DSSP, STRIDE indeed decrease slightly to 93.0%, 91.0%. However somehow the agreement with PCASSO increases by 2.1% while the agreements between P2PSSE and the other four programs only change a bit. In P2PSSE NH is the only proton of and one of the six atoms used to compute the center of a peptide plane. Thus the impacts on its assignment accuracy of the possible errors in protonation should be minor. In the contrary a similar error in protonation will likely have larger impacts on assignment accuracy for both DSSP and STRIDE. That may explain why the agreement with PCASSO increases for set $\mathbf{Q}_8$ since PCASSO does not need protonation.

3.6 The HGTU and EBU zones and the continuity of the SSE space

Overall it is striking that even though backbone H-bonding energy is only a small part of the total free energy for a statistical system composed of all the protein molecules and the solvent molecules, it appears that it may account largely for the existence of protein secondary structure [1, 2, 7]. On the one hand the excellent agreements between P2PSSE and both DSSP and STRIDE as detailed above show that (a) the CNN models have been properly trained, (b) a *pc*-polyline contains enough geometrical features to uniquely determine the SSE types for the vast majority of residues, (c) the criteria used by DSSP for the assignments of different SSE types agree with each other to a great extent especially for α-helix and β-sheet, and (d) DSSP algorithm is deterministic. On the other hand the disagreements point to (a) the impropriety in model training, (b) the insufficiency of the geometrical features extracted from a *pc*-polyline for SSE assignment, (c) the inherent inconsistency in DSSP assignment criteria that makes it impossible to perfectly train a CNN model, and (d) the DSSP algorithm is greedy and P2PSSE is nondeterministic. Trained CNN models have been mainly used to label new samples, in this case SSE assignment. However, a well-trained CNN model could at least to some extent be used to assess the quality of the original labels, in this case, DSSP assignments. Since their agreements are much higher than their disagreements and the CNN models have been trained with a large number of samples (residues), it is likely that the disagreements are due at least in part to the inconsistencies in the criteria used by both DSSP and P2PSSE. The inconsistencies originate ultimately from the diffi-

culty with giving quantitative, consistent and practical definitions for the six SSE types, especially for 3_{10}-helix, β-bulge turn and loop, in terms of either backbone H-bond pattern or *pc*-polyline. And by extension the inconsistencies suggest that there exist not one but an ensemble of backbone conformations for any of the six types and the SSE space though being dominated by the six regions is somewhat continuous with different intermediate backbone conformations in two major twilight zones. The conformations in the first zone, the HGTU zone, share some criteria for α-helix, 3_{10}-helix, turn and loop while those in the second zone, the EBU zone, share some criteria for β-sheet, β-bulge and loop. The existence of these two zones suggests that the SSE space is somewhat continuous. The current uncertainties in SSE definition set the limitations for any SSE assignment programs and by extension set the limitations for programs in protein secondary structure prediction, structure classification [27], structure prediction [28] and protein design [29, 30]. Our analyses suggest that the future study of protein secondary structure should focus on the twilight zones of the SSE space rather than the well-characterized six regions.

4 Conclusion

A series of CNN models have been trained and evaluated using the geometrical features computed from a large set of *pc*-polylines and six labels that correspond to the assignments by DSSP for the six types of SSEs: α-helix, β-sheet, β-bulge, 3_{10}-helix, turn and loop. The applications of P2PSSE that uses the best-trained CNN model to assign the SSEs for two large sets of protein structures not used in the training show that the assignments by P2PSSE agree very well with those by DSSP for at least five SSE types. The comparisons with six other previous assignment programs confirm that the pure geometrical features computed from a *pc*-polyline could be used to accurately assign α-helix, β-sheet, 3_{10}-helix, turn and loop. The detailed analyses of the assignments by both DSSP and P2PSSE show that at least in practice the SSE space though dominated by six large regions is continuous with two twilight zones. The continuity of the SSE space sets the limitations for SSE assignment and for the quantification of protein secondary structure, and by extension sets the limitations for any program that relies on SSE assignment.

References

[1] L. Pauling, R. B. Corey, and H. R. Branson. The structure of proteins: Two hydrogen-bonded helical configurations of the polypeptide chain. *Proceedings of the National Academy of Sciences of the United States of America*, 37(4):205–211, 1951.

[2] L. Pauling and R. B. Corey. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proceedings of the National Academy of Sciences of the United States of America*, 37(11):729–7240, 1951.

[3] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28(1):235–242, 01 2000.

[4] W. Kabsch and C. Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22(12):2577–2637, 1983.

[5] D. Frishman and P. Argos. Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4):566–579, 1995.

[6] M. N. Fodje and S. Al-Karadaghi. Occurrence, conformational features and amino acid propensities for the π-helix. *Protein Engineering, Design and Selection*, 15(5):353–358, 05 2002.

[7] R. Srinivasan and G. D. Rose. A physical basis for protein secondary structure. *Proceedings of the National Academy of Sciences*, 96(25):14258–14263, 1999.

[8] A. S. Konagurthu, A. M. Lesk, and L. Allison. Minimum message length inference of secondary structure from protein coordinate data. *Bioinformatics*, 28(12):i97–i105, 2012.

[9] G. Nagy and C. Oostenbrink. Dihedral-based segment identification and classification of biopolymers ii: polynucleotides. *Journal of chemical information and modeling*, 54(1):278—288, January 2014.

[10] J. Martin, G. Letellier, A. Marin, J. Taly, A. G. De Brevern, and J. Gibrat. Protein secondary structure assignment revisited: a detailed analysis of different assignment methods. *BMC Structural Biology*, 5(1):17, 2005.

[11] F. M. Richards and C. E. Kundrot. Identification of structural motifs from protein coordinate data: Secondary structure and first-level supersecondary structure*. *Proteins: Structure, Function, and Bioinformatics*, 3(2):71–84, 1988.

[12] S. M. King and W. C. Johnson. Assigning secondary structure from protein coordinate data. *Proteins: Structure, Function, and Bioinformatics*, 35(3):313–320, 1999.

[13] G. Labesse, N. Colloc'h, J. Pothier, and J. P. Mornon. P-SEA: a new efficient assignment of secondary structure from c alpha trace of proteins. *Computer Applications in the Biosciences*, pages 291–295, 1997.

[14] I. Majumdar, S. S. Krishna, and N. V. Grishin. PALSSE: A program to delineate linear secondary structural elements from protein structures. *BMC bioinformatics*, 6(1):202, 2005.

[15] S. Y. Park, M. J. Yoo, J. Shin, and K. Cho. Saba (secondary structure assignment program based on only alpha carbons): a novel pseudo center geometrical criterion for accurate assignment of protein secondary structures. *BMB Rep.*, 44:118—122, 2011.

[16] M. L. Sean, A. T. Frank, and C. L. Brooks III. PCASSO: A fast and efficient c_α-based method for accurately assigning protein secondary structure elements. *Journal of Computational Chemistry*, 35(24):1757–1761, 2014.

[17] N. Colloc'h, C. Etchebest, E. Thoreau, B. Henrissat, and J. Mornon. Comparison of three algorithms for the assignment of secondary structure in proteins: the advantages of a consensus assignment. *Protein engineering*, 6(4):377–382, 1993.

[18] C. A. Andersen and B. Rost. Secondary structure assignment. *Structural Bioinformatics*, 44:459–484, 2009.

[19] C. Fang, Y. Shang, and D. Xu. Improving protein gamma-turn prediction using inception capsule networks. *Scientific Reports*, 8:15741–15751, 12 2018.

[20] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016.

[21] L. Wang, Y. Zhang, and S. Zou. The characterization of pc-polylines representing protein backbones. *Proteins: Structure, Function, and Bioinformatics*, 88(2):307–318, 2020.

[22] M. V. Cubellis, F. Cailliez, and S. C. Lovell. Secondary structure assignment that accurately reflects physical and evolutionary characteristics. *BMC Bioinformatics*, 6:S4—S8, 2005.

[23] J. M. Word, S. C. Lovell, J. S. Richardson, and D. C. Richardson. Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation1. *Journal of Molecular Biology*, 285(4):1735–1747, 1999.

[24] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[25] C. Cao, S. Xu, and L. Wang. An algorithm for protein helix assignment using helix geometry. *PLOS ONE*, 10(7):1–20, 07 2015.

[26] L. Wang, Y. Pang, T. Holder, J. R. Brender, A. V. Kurochkin, and E. R. P. Zuiderweg. Functional dynamics in the active site of the ribonuclease binase. *Proceedings of the National Academy of Sciences*, 98(14):7684–7689, 2001.

[27] R. Kolodny, D. Petrey, and B. Honig. Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. *Current Opinion in Structural Biology*, 16(3):393 – 398, 2006. Nucleic acids/Sequences and topology.

[28] J. Moult, F. Krzysztof, A. Kryshtafovych, T. Schwede, and A. Tramontano. Critical assessment of methods of protein structure prediction (casp)—round xii. *Proteins: Structure, Function, and Bioinformatics*, 86(S1):7–15, 2018.

[29] P. Gainza, H. M. Nisonoff, and B. R. Donald. Algorithms for protein design. *Current Opinion in Structural Biology*, 39:16 – 26, 2016. Engineering and design Membranes.

[30] P. S. Huang, S. E. Boyken, and D. Baker. The coming of age of de novo protein design. *Nature*, 537:320–327, 2016.

Supplementary Information

The SSE assignment using *pc*-polyline and convolutional neuron network (CNN) is achieved in two steps using respectively sub-program P2P for *pc*-polyline computation and sub-program P2PASSIGN for SSE assignment using a best-trained CNN model. In section S1 we present these two programs. In section S2 we describe two train sets, $\mathscr{L}_7$ and $\mathscr{L}_9$, of the geometrical features extracted from $\mathscr{L}$

using respectively $T_{pp} = 7.0$Å and $T_{pp} = 9.0$Å. In section S3 and S4 we analyze respectively the assignments by both DSSP and P2PSSE for $\mathbf{P}_8$ and $\mathbf{Q}_8$, both of which have not been used in model training. In section S5 we examine the β-sheet assignments for $\mathscr{L}_8$ by both DSSP and P2PSSE.

S1 The programs for the computation of *pc*-polylines, the training of CNN models and SSE assignment

In this section we briefly describe (1) the C++ program P2P for the computation of the geometrical data used to train CNN models *and* to assign SSEs, (2) a Python script for the training of CNN models, and (3) the Python script P2PASSIGN that uses a best-trained CNN model to assign SSEs.

P2P is written in C++ and is a module of an in-house molecular analysis and visualization program. Given a protonated protein structure P2P takes a fraction of a second to compute its *pc*-polyline and to extract the geometrical data from the *pc*-polyline. The CNN models were originally trained using TensorFlow1.4 [24]. The training of each model using TensorFlow1.4 took about 14 hours on a Dell T5810 workstation equipped with a Nvidia RTX2080 card. The same model trained using TensorFlow2.3 took about 9 hours on a Dell T5820 workstation with the same card. The SSE assignments that uses the model trained with the following hyperparameters and architecture agree best with those by DSSP. All the assignments described in the main text and in this SI are based on this particular model.

```
model = Sequential()
model.add(Conv1D(filters=64, kernel_size=3, padding=valid, input_shape=(noOfFeature,1), activation=relu))
model.add(Conv1D(filters=128, kernel_size=3, activation=relu))
model.add(AveragePooling1D(pool_size=2))
model.add(Conv1D(filters=128, kernel_size=3, activation=relu))
model.add(Conv1D(filters=160, kernel_size=2, activation=relu))
model.add(AveragePooling1D(pool_size=2))

model.add(Dropout(0.25))
model.add(Flatten())

model.add(Dense(1024, activation=relu))
model.add(Dropout(0.25))
model.add(Dense(512, activation=relu))
model.add(Dense(512, activation=relu))
model.add(Dense(256, activation=relu))
model.add(Dense(noOfSSEtype, activation=softmax))

Adam = adam(lr = 0.0001, beta_1=0.9, beta_2=0.999, epsilon=1e-08, amsgrad=True)
model.compile(loss=categorical_crossentropy, optimizer=Adam, metrics=accuracy])
```

The program P2PASSIGN that assigns the SSEs using the best-trained CNN model and the geometrical data computed by P2P is written in Python. It takes only a fraction of a second to assign the SSEs for a typical protein structure if the time for loading the model is excluded.

The two programs P2P and P2PASSIGN, the best-trained CNN model, two Charmm force field files, a protonated protein structure in PDB format and the output files for both P2P and P2PASSIGN as well as the instructions for their usages could all be downloaded from Github (https://github.com/wlincong/p2pSSE).

S2 The geometrical features of $\mathscr{L}_7$ and $\mathscr{L}_9$

Figure S1 depicts the distributions of the number of residues vs. the number of p2p distances for the residues in $\mathscr{L}_7$ and in $\mathscr{L}_9$.

S3 The analyses of the SSE assignments by P2PSSE and DSSP for $\mathbf{P}_8$

In this section we first describe the agreements at residue level between P2PSSE and DSSP. We then present the distributions of the assignments also at residue level by P2PSSE not agreed to by DSSP and vice versa. Finally we analyze their assignments for α-helix and β-sheet. Overall the agreements between DSSP and P2PSSE for $\mathbf{P}_8$ though lower than those for $\mathscr{L}_8$ remain to be high.

S3.1 The agreements at residue level between P2PSSE and DSSP

The agreements between P2PSSE and DSSP for all the six SSE types are listed in Table S1. The agreements from the best to the worse are H, E, U, T, G and B. The overall agreements remain to be high and except for β-bulge the agreements for all the other five types are better than 0.83% as judged by either accuracy (precision) or recall.

S3.2 The distributions of the assignments at residue level by P2PSSE not agreed to by DSSP and vice versa

As shown in Table S2 P2PSSE assigns a DSSP-assigned H to $\mathcal{T}$ (1.71%), $\mathcal{G}$ (0.49%) and $\mathcal{U}$ (0.25%) with very small and decreasing possibilities, and to $\mathcal{E}$ and $\mathcal{B}$ with rare possibilities. Compared with the assignments by P2PSSE for $\mathscr{L}_8$ (Table 2 of the main text) the percentages of "wrong" assignments to $\mathcal{T}, G, U$ by P2PSSE increase between 1.36-fold and 1.78-fold. Likewise as shown in Table S3 DSSP assigns a P2PSSE-assigned $\mathcal{H}$ to T (1.26%), G (0.43%) and U (0.10%) with very small and decreasing possibilities, and to E and B with rare possibilities.

As to β-sheet residues P2PSSE assigns a DSSP-assigned E to $\mathcal{U}$ (7.00%), $\mathcal{B}$ (1.35%) and $\mathcal{T}$ (0.10%) with small and decreasing possibilities, and to $\mathcal{H}$ and $\mathcal{G}$ with rare possibilities. Compared with the assignments by P2PSSE for $\mathscr{L}_8$ (Table 2 of the main text) the percentages of "wrong" assignments by P2PSSE to $\mathcal{U}$ and $\mathcal{B}$ increase respectively 2.24-fold to 2.18-fold. Likewise DSSP assigns a P2PSE-

assigned $\mathcal{E}$ to U (5.46%), B (0.56%) and T (0.18%) with small and decreasing possibilities, and to G and H with rare possibilities (Table S3).

S3.3 The α-helix assignments by both DSSP and P2PSSE

As shown in Table S4 at helix level P2PSSE assigns $34,558$ α-helices for $\mathbf{P}_8$, that is 297 more than the α-helices $(34,261)$ by DSSP. Out of the $34,558$ α-helices, $23,876$ (69.0%) agree exactly with DSSP. Except for 589 α-helices (1.70%) that could not be matched to any α-helix by DSSP, all the remaining ones agree with DSSP to some extents. Specifically each of the $5,013$ α-helices by P2PSSE is a part of an α-helix by DSSP, each of the 128 α-helices extends both termini of a DSSP-assigned α-helix, while $4,249$ and 595 α-helices extend, respectively, the N-termini and the C-termini of DSSP-assigned α-helices. If all these α-helices (plus those in the 9th and 10th columns in Table S4) are counted as agreement, then there are $34,113$ P2PSSE-α-helices (98.71%) in total that agree with DSSP.

S3.4 The β-sheet assignments by both DSSP and P2PSSE

As shown in Table S5 at strand level P2PSSE assigns $45,056$ β-strands for $\mathbf{P}_8$, that is 12 more than the β-strands $(45,044)$ by DSSP. Out of the $45,056$ β-strands, $25,053$ (55.6%) agree exactly with DSSP. Except for $1,277$ β-strands (2.83%) that could not be matched to any β-strand by DSSP, all the remaining ones agree with DSSP to some extents. Specifically each of the $8,503$ β-strands by P2PSSE is a part of a β-strand by DSSP, each of the 653 β-strands extends both termini of a DSSP-assigned β-strand, while $5,643$ and $2,528$ β-strands extend, respectively, the N-termini and the C-termini of the DSSP-assigned β-strands. If all these β-strands (plus those in the 9th and 10th columns in Table S5) are counted as agreement, then there are $43,779$ β-strands (97.27%) in total that agree with DSSP.

S4 The analyses of SSE assignments by P2PSSE and DSSP for $\mathbf{Q}_8$

This section presents the SSE assignments by both P2PSSE and DSSP for the set of low quality structures $\mathbf{Q}_8$. Overall the agreements between DSSP and P2PSSE for $\mathbf{Q}_8$ are slightly worse for H, E, U, T and modestly worse for G and B than those for $\mathbf{P}_8$. It shows the quality of a crystal structure has some but no large effects on the performance of P2PSSE. In the following we first describe the agreements at residue level between P2PSSE and DSSP. We then present the distributions of the assignments at

residue level by P2PSSE not agreed to by DSSP and vice versa. Finally we compare the assignments at fragment level for both α-helix and β-sheet.

S4.1 The agreements at residue level between P2PSSE and DSSP

The agreements between P2PSSE and DSSP for the six SSE types are listed in Table S6. The agreements from the best to the worse are H, E, U, T, G and B. Except for β-bulge the overall agreements for the other five types remain to be good. The largest decrease in agreement between P2PSSE and DSSP is for β-bulge.

S4.2 The distributions of the assignments at residue level by P2PSSE but not agreed to by DSSP and vice versa

As shown in Table S7 P2PSSE assigns a DSSP-assigned H to $\mathcal{T}$ (2.31%), $\mathcal{G}$ (0.49%) and $\mathcal{U}$ (0.46%) with very small and decreasing possibilities, and to $\mathcal{E}$ and $\mathcal{B}$ with rare possibilities. Compared with the assignments by P2PSSE for $\mathbf{P}_8$ (Table S2) the percentages of "wrong" assignments by P2PSSE to $\mathcal{T}, G, U$ increase slightly. Likewise as shown in Table S8 DSSP assigns a P2PSSE-assigned $\mathcal{H}$ to T (1.30%), G (0.46%) and U (0.13%) with very small and decreasing possibilities, and to E and B with rare possibilities. These percentages are almost the same as those for $\mathbf{P}_8$ (Table S3).

As to β-sheet residues P2PSSE assigns a DSSP-assigned E to $\mathcal{U}$ (7.54%), $\mathcal{B}$ (1.33%) and $\mathcal{T}$ (0.14%) with small and decreasing possibilities, and to $\mathcal{H}$ and $\mathcal{G}$ with rare possibilities. Compared with the assignments by P2PSSE for $\mathbf{P}_8$ (Table S2) the percentages of "wrong" assignments by P2PSSE to $\mathcal{T}, G, U$ also increase slightly. Likewise DSSP assigns a P2PSSE-assigned $\mathcal{E}$ to U (6.33%), B (0.51%) and T (0.13%) with small and decreasing possibilities, and to G and H with rare possibilities (Table S8). These percentages are very similar to those for $\mathbf{P}_8$ (Table S3).

S4.3 The α-helix assignments by both DSSP and P2PSSE

As shown in Table S9 at helix level P2PSSE assigns $33,450$ α-helices for $\mathbf{Q}_8$, that is 174 more than the $33,286$ α-helices by DSSP. Out of the $33,450$ α-helices, $21,139$ (63.2%) agree exactly with DSSP. Except for 627 α-helices (1.87%) that could not be matched to any α-helix by DSSP, all the remaining ones agree with DSSP to some extents. Specifically each of the $5,059$ α-helices by P2PSSE is a part of

an α-helix by DSSP, each of the 323 α-helices extends both termini of a DSSP-assigned α-helix, while $5,230$ and 941 α-helices extend, respectively, the N-termini and the C-termini of DSSP-assigned α-helices. If all these α-helices (plus those in the 9th and 10th columns in Table S9) are counted as agreement, then there are $32,823$ α-helices (98.13%) in total that agree with DSSP. Overall the agreements between P2PSSE and DSSP for the assignments of the α-helices in $\mathbf{Q}_8$ are slightly worse than those in $\mathbf{P}_8$.

S4.4 The β-sheet assignments by both DSSP and P2PSSE

As shown in Table S10 at strand level P2PSSE assigns $43,412$ β-strands for $\mathbf{Q}_8$, that is 73 more than the β-strands ($43,339$) by DSSP. Out of the $43,412$ β-strands, $22,666$ (52.2%) agree exactly with DSSP. Except for $1,408$ β-strands (3.24%) that could not be matched to any β-strand by DSSP, all the remaining ones agree with DSSP to some extents. Specifically each of the $9,213$ β-strands by P2PSSE is a part of a β-strand by DSSP, each of the 684 β-strands extends both termini of a DSSP-assigned β-strand, while $5,195$ and $2,730$ β-strands extend, respectively, the N-termini and the C-termini of DSSP-assigned β-strands. If all these β-strands (plus those in the 9th and 10th columns in Table S10) are counted as agreement, then there are $42,004$ (96.86%) β-strands in total that agree with DSSP.

S5 The β-sheet assignments at fragment level by P2PSSE and DSSP for $\mathscr{L}_8$ and the EBU zone

In this section we first analyze the β-strand assignments for $\mathscr{L}_8$ at fragment level by P2PSSE and DSSP. Then we illustrate the EBU twilight zone by describing in somewhat detail three examples from the two sets of the β-strands that are assigned by one program but could not be matched to any by the other.

S5.1 The β-strands assigned by DSSP but not agreed by P2PSSE

Out of the $2,398$ unmatched β-strands (Table 7 of the main text) by DSSP, $2,079$ have only two residues, 228 have three residues, 56 have four residues, 25 have five residues, 6 have six residues and 4 have seven residues. None of them have eight or more residues. Out of the $2,398$ β-strands,

only 165 include neither $\mathcal{E}$ nor $\mathcal{B}$ assigned by P2PSSE and only 332 have no $\mathcal{E}$s assigned by P2PSSE . There are $2,066$ strands that include only a single $\mathcal{E}$s assigned by P2PSSE. Almost all the strands that have $\mathcal{B}$s assigned by P2PSSE include only one $\mathcal{B}$s.

The majority of the $2,398$ unmatched β-strands belong to the sheets with only two 2-residue strands. The remaining ones are side strands of a few residues long. For any sheet with more than two strands, there are two different types of strands: the *internal* strands that form H-bonds with its two neighbors and the *side* strands that form H-bonds with only a single neighbor. As with the α-helices not assigned by P2PSSE , the majority of these unmatched short β-strands are on surface. Geometrically these unmatched β-strands have relatively small curvatures and are likely not to be parallel with their partners (Figures S2 and S3).

As shown in Figures S2a and S2b the residues Y73-K78 in 6a02 (pdbid) assigned to E by DSSP may be assigned to $\mathcal{B}$, $\mathcal{U}$, and $\mathcal{T}$ but only the two termini have their $p(\mathrm{E})$s as the second largest probabilities (Table S11). A shown in Figures S2c and S2d this 2-residue strand has distorted sheet geometry. P2PSSE assigns both residues to $\mathcal{U}$. However, both residues have $p(E)$s as their second largest probabilities. These two strands illustrate the backbone conformations in the EBU twilight zone.

S5.2 The β-strands assigned by P2PSSE but not agreed by DSSP

There are $1,249$ β-strands assigned by P2PSSE that could not be matched to any β-strand by DSSP. The vast majority of them have good sheet geometry though some of them may lack the typical inter-strand H-bonds required by DSSP. As illustrated in Figure S3 this 3-residue strand has good sheet geometry and there also exist DSSP required H-bonds. So it is unclear to us why DSSP does not assign this β-sheet. On the other hand, the sheet geometry is somewhat distorted and all the $p(E)$s are less than 0.805 (Table S11). In addition all the three residues have $p(U)$s as the second largest probabilities. This strand is an example of backbone conformation in the EBU twilight zone.

S6 The regrouping into three SSE types of the assignments by P2PSSE and seven previous programs

For the purpose of comparison the SSE assignments by P2PSSE and the previous four programs, DSSP, STRIDE, SENGO, PROSS, are regrouped into three types, *helix, sheet and loop* (Table S12) since P-SEA, KAKSI, PCASSO could only assign SSEs to three types.

Figure 1: **The distributions of the number of residues vs. the number of p2p distances (d_{pp}s) for the residues in $\mathscr{L}_6$ (a) and in $\mathscr{L}_8$ (b)**. The average numbers of d_{pp} per residue for $\mathscr{L}_6$ and $\mathscr{L}_8$ are respectively 2.87 and 7.01 while their medians are respectively 4 and 7. The x-axis is the number of d_{pp}s per residue while the y-axis is the number of residues. Please see Figure S1 of SI for the distributions for $\mathscr{L}_7$ and $\mathscr{L}_9$.

Figure 2: **A 5-residue α-helix in 1ah7, a 4-residue α-helix in 5zxm and a 4-residue α-helix in 5n13 by P2PSSE only.** The Y132-F132 fragment in 1ah7 is depicted respectively as a cartoon in (**a**) and as a helix (Q125-A129) in (**b**). The three H-bonds, $i \to i+3, i \to i+4$ and $i \to i+5$, are indicated by the dash lines in azure. All the five residues Q125-P126-M127-H128-A129 are assigned to Ts by DSSP but all have relatively large $p(\mathrm{H})$s while their $p(\mathrm{T})$s are rather small (Table 5). The T311-R316 fragment in 5zxm is depicted respectively as a cartoon in (**c**) and as a helix (T311-D314) in (**d**). No H-bond is detected. The four residues T311-G312-D313-D314 are assigned to UTTT by DSSP but all have relatively large $p(\mathrm{H})$s and the $p(\mathrm{T})$s for the three DSSP-assigned Ts are rather small (Table 5). The V415-I422 fragment in 5n13 is depicted respectively as a cartoon in (**e**) and as a helix (D416-A419) in (**f**). The three H-bonds, $i \to i+3, i \to i+4$ and $i \to i+5$, are indicated by the dash lines in azure. All the four residues D416-T417-V418-A419 are assigned to Ts by DSSP but all have $p(\mathrm{H}) > p(\mathrm{T})$ (Table 5). However the differences between their $p(\mathrm{H})$s and $p(\mathrm{T})$s are small. The H-bonds are computed as described in DSSP. Specifically if a pair of backbone NH atom and CO atom has an H-bond energy ≤ -555.55 (a threshold adopted from DSSP) then they form an H-bond. The cartoons and helices for visualization are computed using *pc*-polylines rather than C_α-polylines. All the molecular figures in both the main text and SI are prepared using our molecular visualization program.

Figure 3: **Four 4-residue α-helices in 5tkm, 5mx9, 1cza and 2z2i assigned by DSSP only.** The Y124-R133 fragment in 5tkm is depicted respectively as a cartoon in (**a**) and as a helix (S127-D130) in (**b**). The two H-bonds, $i \to i+3, i \to i+4$, are indicated by the dash lines in azure. Each of the four residues has $p(\mathrm{T})$ as the largest probability and $p(\mathrm{H})$ the second largest and the difference between them are small (Table 6). The F293-D302 fragment in 5xm9 is depicted respectively as a cartoon in (**c**) and as a helix (L296-D299) in (**d**). A single $i \to i+4$ H-bond is detected. P2PSSE assigns them to $\mathcal{UTTT}$. The $p(\mathrm{H})$s for L296 and D299 are the second largest. The L487-A507 fragment in 1cza is depicted respectively as a cartoon in (**e**) and as a helix (R500-T503) in (**f**). A single H-bond $i \to i+4$ between R500 and H504 may explain DSSP's assignment to an α-helix. P2PSSE assigns them to $\mathcal{TTTG}$. The G11-R19 fragment in 2z2i is depicted respectively as a cartoon in (**g**) and as a helix (G14-Y17) in (**h**). There exist two $i \to i+3$ H-bonds but no $i \to i+4$ in this fragment. P2PSSE assigns them to $\mathcal{GTGT}$.

Figure S1: **The distributions of the number of residues vs. the number of p2p distances the residues in $\mathscr{L}_7$ (a) and in $\mathscr{L}_9$ (b)**. The average numbers of p2p distance per residue for $\mathscr{L}_7$ and $\mathscr{L}_9$ are respectively 4.90 and 10.11 while their medians are respectively 6 and 12.

Figure S2. **A 6-residue β-strand in 6a02 and a 2-residue β-strand in 1iv8 assigned by DSSP only**. The six residues in 6a02 form two inter-strand H-bonds and two $i \rightarrow i+3$ H-bonds (Figures (**a**) and (**b**)). The sheet geometry is distorted with an almost 45 degree intersection angle between the two strands. The N-terminus of the strand assigned by DSSP follows immediately an α-helix. Figures (**c**) and (**d**) depict the strand R191-R192 in 1iv8. It has a single inter-strand H-bond and a distorted sheet geometry with an almost 45 degree intersection angle between the two strands. P2PSSE assigns both residues to $\mathcal{U}$s. Both strands serve as examples for the backbone conformations in the EBU twilight zone.

Figure S3: **A 3-residue β-strand in 4qpw by P2PSSE only**. All the three residues have $p(U)$s as the second largest probabilities (Table S11). This strand is an example of a backbone conformation in the EBU twilight zone.

SSE	DSSP	P2PSSE	Agreed	Accuracy	Recall
H	1,240,520	1,250,015	1,229,134	0.983	0.991
E	798,650	810,598	780,397	0.963	0.977
B	41,385	34,817	32,003	0.919	0.773
G	149,701	150,034	138,180	0.921	0.923
T	432,617	430,552	400,729	0.931	0.926
U	1,023,592	1,010,449	980,803	0.971	0.958

Table 1: **The agreements by SSE type between the assignments by DSSP and P2PSSE for $\mathscr{L}_8$**. The 2nd and 3rd columns list respectively the numbers of residues assigned to each of the six types by DSSP and by P2PSSE. The 4th column lists the numbers of residues assigned to the same SSE type by both. The 5th and 6th columns list the accuracy and the recalls for the assignments by P2PSSE with respect to those by DSSP.

SSE	$\mathcal{H}$	$\mathcal{E}$	$\mathcal{B}$	$\mathcal{G}$	$\mathcal{T}$	$\mathcal{U}$	Total
H	**1,229,134**	3	3	2,811	8,176	393	1,250,015
E		**780,397**	1,114	65	654	16,415	810,598
B			**32,003**	19	181	4,261	34,817
G				**138,180**	7,344	644	150,034
T					**400,729**	7,933	430,552
U						**980,803**	1,010,449

Table 2: **The distributions of the assignments by DSSP agreed to and not agreed to by P2PSSE for $\mathscr{L}_8$**. The last column lists the total number of residues assigned by DSSP for each of the six types. The 2nd-7th rows list, respectively, the agreed and the "wrong" assignments with P2PSSE as the standard. The numbers of assignments agreed to by both are in bold face.

SSE	H	E	B	G	T	U	Total
$\mathcal{H}$	**1,229,134**	5	3	3,499	15,679	1,695	1,240,520
$\mathcal{E}$		**780,397**	4,918	23	280	24,977	798,650
$\mathcal{B}$			**32,003**	11	93	1,593	41,385
$\mathcal{G}$				**138,180**	7,903	1,056	149,701
$\mathcal{T}$					**400,729**	13,468	432,617
$\mathcal{U}$						**980,803**	1,023,592

Table 3: **The distributions of the assignments by P2PSSE agreed to and not agreed to by DSSP for $\mathscr{L}_8$**. The last column lists the total number of residues assigned by P2PSSE for each of the six types. The 2nd–7th rows list, respectively, the agreed and the "wrong" assignments with DSSP as the standard. The numbers of assignments agreed to by both are in bold face.

Method	Residues	α-Helices	Exact	Embedded	Nter/Cter	Nter	Cter	n-Cter	Nter-c	No match
P2PSSE	1,239,469	110,009	89,139	7,574	186	11,916	1,110	165	82	172
DSSP	1,240,520	111,687	89,139	13,212	79	6,355	1,140	82	165	1,754

Table 4: **The α-helix assignments by P2PSSE and by DSSP for $\mathscr{L}_8$.** The α-helices listed in the table all have ≥ 4 residues. The 2nd and 3rd columns list respectively the total numbers of residues in the assigned helices and the total numbers of the assigned helices. The 4th column lists the numbers of exactly-agreed helices. The 5th column lists the numbers of the helices by one program each of which could be embedded completely inside a helix by the other. The 6th column lists the helices by one program each of which extends both termini of a helix by the other. The 7th–8th columns list, respectively, the helices by one program each of which extends the N-terminus only and the C-terminus only of a helix by the other. The 9th column lists the helices by one program each of which has its N-terminus inside a helix but its C-terminus extends beyond the C-terminus of the helix by the other. The 10th column lists the helices by one program each of which has its C-terminus inside a helix but its N-terminus extends beyond the N-terminus of the helix by the other. The last column lists the helices by one program that could not be matched to any helix by the other.

pdbid	Residue	p(H)	p(E)	p(B)	p(G)	p(T)	p(U)	DSSP	P2PSSE
1ah7	Q125	**0.9996**	0.0000	0.0000	0.0000	0.0000	0.0000	T	$\mathcal{H}$
	P126	**0.9349**	0.0000	0.0000	0.0036	*0.0614*	0.0000	T	$\mathcal{H}$
	M127	**0.8472**	0.0000	0.0000	0.0334	*0.1156*	0.0034	T	$\mathcal{H}$
	H128	**0.7744**	0.0000	0.0000	0.0000	*0.2241*	0.0000	T	$\mathcal{H}$
	A129	**0.6319**	0.0000	0.0000	0.0000	*0.3637*	0.0041	T	$\mathcal{H}$
5zxm	T311	**0.9961**	0.0000	0.0000	0.0000	0.0000	*0.0026*	U	$\mathcal{H}$
	G312	**0.8826**	0.0000	0.0000	0.0450	*0.0704*	0.0019	T	$\mathcal{H}$
	D313	**0.9138**	0.0000	0.0000	0.0407	*0.0442*	0.0012	T	$\mathcal{H}$
	D314	**0.9887**	0.0000	0.0000	0.0052	*0.0061*	0.0000	T	$\mathcal{H}$
5n13	D416	**0.4799**	0.0000	0.0000	0.0832	*0.4366*	0.0000	T	$\mathcal{H}$
	T417	**0.7560**	0.0000	0.0000	0.0412	*0.2023*	0.0000	T	$\mathcal{H}$
	V418	**0.5120**	0.0000	0.0000	0.0000	*0.4877*	0.0000	T	$\mathcal{H}$
	A419	**0.6918**	0.0000	0.0000	0.0076	*0.2988*	0.0017	T	$\mathcal{H}$

Table 5: **The probabilities for the six SSE types for the residues in α-helices Q125-A129 in 1ah7, T311-D314 in 5zxm and D416-A419 in 5n13 by P2PSSE.** The largest probability for each residue is in bold face while the second largest in *italics*. In all the three α-helices the labels with the second largest probabilities agree with their DSSP assignments. For easy distinguish we render the P2PSSE assignments in latex mathcal font. In this paper a protein structure is represented by its *pdbid*.

pdbid	Residue	p(H)	p(E)	p(B)	p(G)	p(T)	p(U)	DSSP	P2PSSE
5ktm	S127	*0.0239*	0.0000	0.0000	0.0041	**0.9620**	0.0101	H	$\mathcal{T}$
	E128	*0.3537*	0.0000	0.0000	0.0121	**0.6306**	0.0035	H	$\mathcal{T}$
	R129	*0.4826*	0.0000	0.0000	0.0209	**0.4958**	0.0000	H	$\mathcal{T}$
	D130	*0.4322*	0.0000	0.0000	0.0012	**0.5653**	0.0013	H	$\mathcal{T}$
5mx9	L296	*0.4443*	0.0000	0.0174	0.0115	0.0139	**0.5125**	H	$\mathcal{U}$
	G297	0.0793	0.0000	0.0000	*0.0913*	**0.8015**	0.0280	H	$\mathcal{T}$
	Q298	0.1653	0.0000	0.0000	*0.3084*	**0.5221**	0.0042	H	$\mathcal{T}$
	D299	*0.3754*	0.0000	0.0000	0.1495	**0.4620**	0.0126	H	$\mathcal{T}$
1cza	R500	*0.1079*	0.0000	0.0000	0.0064	**0.8844**	0.0013	H	$\mathcal{T}$
	K501	*0.2795*	0.0000	0.0000	0.0463	**0.6732**	0.0010	H	$\mathcal{T}$
	Q502	*0.4215*	0.0000	0.0000	0.0572	**0.5202**	0.0012	H	$\mathcal{T}$
	T503	0.1221	0.0000	0.0000	**0.4430**	*0.4201*	0.0148	H	$\mathcal{G}$
2z2i	G14	0.2176	0.0000	0.0000	**0.3829**	*0.3200*	0.0795	H	$\mathcal{G}$
	A15	0.2724	0.0000	0.0000	*0.3341*	**0.3927**	0.0000	H	$\mathcal{T}$
	N16	0.1448	0.0000	0.0000	**0.4777**	*0.3717*	0.0058	H	$\mathcal{G}$
	Y17	*0.2540*	0.0000	0.0000	0.0000	**0.7451**	0.0000	H	$\mathcal{T}$

Table 6: **The probabilities for the six SSE types for the residues in segments S127-D130 in 5tkm, L296-D299 in 5mx9, R500-T503 in 1cza and G14-Y17 in 2z2i**. DSSP assigns all of them to Hs while P2PSSE assigns them to $\mathcal{G}$s, $\mathcal{T}$s and $\mathcal{U}$. The largest probability for each residue is in bold face while the second largest in *italics*.

Method	Residues	β-strands	Exact	Embedded	Nter / Cter	Nter	Cter	n-Cter	Nter-c	No match
P2PSSE	796,747	149,909	119,049	13,666	488	9,805	5,019	793	366	1,249
DSSP	798,642	149,655	119,049	15,312	495	6,426	6,745	366	793	2,398

Table 7: **The β-sheet assignments by DSSP and P2PSSE for $\mathscr{L}_8$**. The items have the same meanings as those in Table 4.

	STRIDE	P-SEA	SENGO	PROSS	PCASSO	KAKSI	P2PSSE
DSSP	95.0	80.5	85.8	83.2	91.4	82.5	**97.2**
STRIDE		81.3	85.8	83.9	89.9	83.3	**93.9**
P-SEA			85.1	85.8	80.6	81.9	**81.0**
SENGO				87.8	84.5	81.6	**85.7**
PROSS					82.8	82.3	**83.1**
PCASSO						81.9	**91.7**
KAKSI							**82.6**

Table 8: **The comparisons of the assignments by P2PSSE and seven previous programs for $\mathscr{L}_8$**. The last column lists the agreements between P2PSSE and the seven programs. The agreement between two programs for SSE type X is computed as follows. Let n_c be the number of residues assigned to X by both programs, nc_1 the number of residues assigned to X by program 1 only, and nc_2 the number of residues assigned to X by program 2 only. Their agreement is defined as $\frac{n_c}{nc_1+n_1+nc_2}$.

	STRIDE	P-SEA	SENGO	PROSS	PCASSO	KAKSI	P2PSSE
DSSP	94.9	80.9	85.8	83.7	91.4	82.8	**94.7**
STRIDE		81.7	85.7	84.3	89.6	83.5	**92.3**
P-SEA			85.2	86.0	80.7	82.0	**81.6**
SENGO				88.0	84.2	81.6	**85.6**
PROSS					83.0	82.6	**83.6**
PCASSO						81.8	**91.4**
KAKSI							**82.7**

Table 9: **The comparisons of the assignments by P2PSSE and seven previous programs for $\mathbf{P}_8$.** The value in each cell has the same meaning as in Table 8.

	STRIDE	P-SEA	SENGO	PROSS	PCASSO	KAKSI	P2PSSE
DSSP	94.1	80.9	85.0	83.3	92.8	83.4	**93.0**
STRIDE		81.4	84.9	83.6	91.2	83.9	**91.0**
P-SEA			85.0	85.1	82.3	82.6	**81.6**
SENGO				87.9	85.4	82.3	**84.5**
PROSS					83.6	82.9	**82.7**
PCASSO						84.1	**93.5**
KAKSI							**83.2**

Table 10: **The comparisons of the assignments by P2PSSE and seven previous programs for $\mathbf{Q}_8$.** The value in each cell has the same meaning as in Table 8.

SSE	DSSP	P2PSSE	Agreed	Accuracy	Recall
H	392,768	395,456	385,714	0.975	0.982
E	242,510	248,345	227,425	0.916	0.938
B	12,312	8,505	5,049	0.594	0.410
G	44,881	46,531	38,492	0.827	0.858
T	132,238	132,434	114,332	0.863	0.865
U	310,346	303,784	280,508	0.923	0.904

Table S1: **The agreements by SSE type between the assignments by DSSP and P2PSSE for $\mathbf{P}_8$.** The 2nd and 3rd columns list respectively the numbers of residues assigned to each of the six types by DSSP and by P2PSSE. The 4th column lists the numbers of residues assigned to the same SSE type by both. The 5th and 6th columns list the accuracy and the recalls for the assignments by P2PSSE with respect to those by DSSP.

SSE	$\mathcal{H}$	$\mathcal{E}$	$\mathcal{B}$	$\mathcal{G}$	$\mathcal{T}$	$\mathcal{U}$	Total
H	**385,714**	6	7	1,699	4,936	406	392,768
E		**227,425**	1,351	66	426	13,235	242,510
B			**5,049**	28	232	3,632	12,312
G				**38,492**	3,843	542	44,881
T					**114,332**	5,461	132,238
U						**280,508**	310,346

Table S2: **The distributions of the assignments by DSSP agreed to and not agreed to by P2PSSE for $\mathbf{P}_8$.** The last column lists the total number of residues assigned by DSSP for each of the six types. The 2nd–7th rows list, respectively, the agreed and the "wrong" assignments with P2PSSE as the standard.The numbers of assignments agreed to by both are in bold face.

SSE	H	E	B	G	T	U	Total
$\mathcal{H}$	**385,714**	7	25	1,952	6,775	983	395,456
$\mathcal{E}$		**227,425**	3,346	32	252	17,284	248,345
$\mathcal{B}$			**5,049**	20	84	1,994	8,505
$\mathcal{G}$				**38,492**	5,334	912	46,531
$\mathcal{T}$					**114,332**	8,665	132,434
$\mathcal{U}$						**280,508**	303,784

Table S3: **The distributions of the assignments by P2PSSE agreed and not agreed to by DSSP for $\mathbf{P}_8$.** The last column lists the total number of residues assigned by P2PSSE for each of the six types. The 2nd-7th rows list, respectively, the agreed and the "wrong" assignments with DSSP as the standard. The numbers of assignments agreed to by both are in bold face.

Method	Residues	α-Helices	Exact	Embedded	Nter / Cter	Nter	Cter	n-Cter	Nter-c	No match
P2PSSE	392,583	34,558	23,876	5,013	128	4,249	595	139	113	589
DSSP	392,768	34,261	23,876	4,972	173	3,865	975	113	139	387

Table S4: **The α-helix assignments by P2PSSE and by DSSP for $\mathbf{P}_8$.** The items in the table have the same meanings as those in Table 4 of the main text.

Method	Residue	β-Strand	Exact	Embedded	Nter / Cter	Nter	Cter	n-Cter	Nter-c	No match
P2PSSE	240,005	45,056	25,053	8,503	653	5,643	2,528	1,142	568	1,277
DSSP	242,509	45,044	25,053	8,824	704	4,116	3,683	568	1,142	2,115

Table S5: **The β-sheet assignments by P2PSSE and by DSSP for $\mathbf{P}_8$.** The items in the table have the same meanings as those in Table 4 of the main text.

SSE	DSSP	P2PSSE	Agreed	Accuracy	Recall
H	396,270	401,860	388,738	0.967	0.981
E	233,314	238,229	216,734	0.910	0.929
B	11,589	7,578	4,447	0.587	0.384
G	39,724	44,139	33,270	0.754	0.838
T	125,141	125,846	102,073	0.811	0.816
U	314,162	302,548	276,853	0.915	0.881

Table S6: **The agreements by SSE type between the assignments by DSSP and P2PSSE for $\mathbf{Q}_8$.** The items have the same meanings as those in Table S1.

SSE	$\mathcal{H}$	$\mathcal{E}$	$\mathcal{B}$	$\mathcal{G}$	$\mathcal{T}$	$\mathcal{U}$	Total
H	**388,738**	3	4	1,818	5,185	522	396,270
E		**216,734**	1,188	85	537	1,4764	233,314
B			**4,447**	26	305	3,633	11,589
G				**33,270**	3,838	581	39,724
T					**10,2073**	6,195	125,141
U						**76,853**	314,162

Table S7: **The distribution of the assignments by DSSP agreed to and not agreed to by P2PSSE for $\mathbf{Q}_8$.** The items in this table have the same meanings as those in Table S2.

SSE	H	E	B	G	T	U	Total
$\mathcal{H}$	**388,738**	6	10	1,985	9,265	1,856	401,860
$\mathcal{E}$		**216,734**	3,168	33	326	17,965	238,229
$\mathcal{B}$			**4,447**	17	101	1,821	7,578
$\mathcal{G}$				**33,270**	7,181	1,759	44,139
$\mathcal{T}$					**10,2073**	13,908	125,846
$\mathcal{U}$						**276,853**	302,548

Table S8: **The distribution of the assignments by P2PSSE agreed to and not agreed to by DSSP for $\mathbf{Q}_8$.** The items in this table have the same meanings as those in Table S3.

Method	Residues	α-Helices	Exact	Embedded	Nter / Cter	Nter	Cter	n-Cter	Nter-c	No match
P2PSSE	398,832	33,450	21,139	5,059	323	5,230	941	217	206	627
DSSP	396,270	33,286	21,139	6,494	224	3,818	1,017	206	217	41

Table S9: **The α-helix assignments by P2PSSE and by DSSP for $\mathbf{Q}_8$.** The items in the table have the same meanings as those in Table 4 of the main text.

Method	Residues	β-strands	Exact	Embedded	Nter / Cter	Nter	Cter	n-Cter	Nter-c	No match
P2PSSE	229,669	43,412	22,666	9,213	684	5,195	2,730	1,232	679	1,408
DSSP	233,304	43,339	22,666	8,609	835	4,315	4,063	679	1,232	2,215

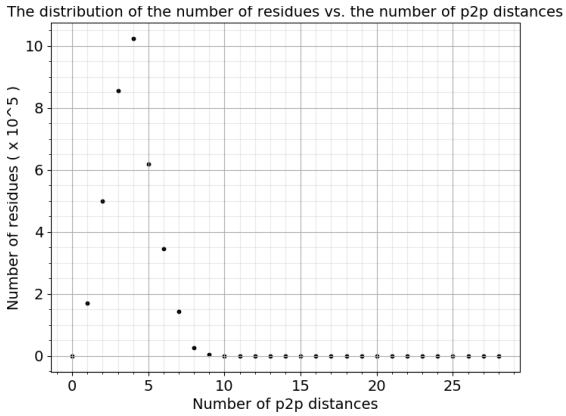
Table S10: **The β-sheet assignments by P2PSSE and by DSSP for $\mathbf{Q}_8$.** The items in the table have the same meanings as those in Table 4 of the main text.

pdbid	Residue	p(H)	p(E)	p(B)	p(G)	p(T)	p(U)	DSSP	P2PSSE
6a02	Y73	0.0000	*0.3556*	**0.5990**	0.0000	0.0000	0.0454	E	$\mathcal{B}$
	V74	0.0000	0.2011	**0.5312**	0.0000	0.0000	*0.2677*	E	$\mathcal{B}$
	K75	0.0000	0.1869	0.0000	0.0000	**0.4366**	*0.3765*	E	$\mathcal{T}$
	S76	0.0000	0.0050	0.0000	0.0000	**0.9614**	*0.0336*	E	$\mathcal{T}$
	D77	0.0000	0.0000	0.0000	0.0000	0.0000	**0.9997**	E	$\mathcal{U}$
	K78	0.0000	*0.4620*	0.0000	0.0000	0.0000	**0.5371**	E	$\mathcal{U}$
1iv8	R191	0.0000	*0.3836*	0.0574	0.0000	0.0000	**0.5590**	E	$\mathcal{U}$
	R192	0.0000	*0.2029*	0.0022	0.0000	0.0495	**0.7447**	E	$\mathcal{U}$
4qpw	A217	0.0000	**0.7076**	0.0027	0.0000	0.0000	*0.2896*	U	$\mathcal{E}$
	D218	0.0000	**0.8045**	0.0000	0.0000	0.0000	*0.1955*	U	$\mathcal{E}$
	N219	0.0000	**0.5377**	0.0108	0.0000	0.0000	*0.4515*	U	$\mathcal{E}$

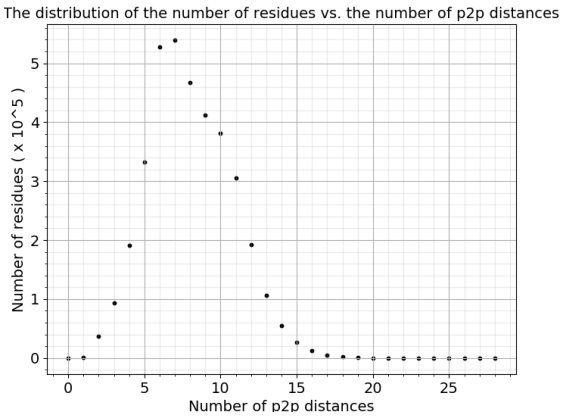
Table S11: **The probabilities for the six SEE types for the six residues Y73-K78 in 6a02, two residues R191-R192 in 1iv8 and the three residues A217-N219 in 4qpw**. DSSP assigns all of the six residues Y73-K78 in 6a02 to E while P2PSSE assigns them to $\mathcal{B}$, $\mathcal{T}$ and $\mathcal{U}$. DSSP assigns two residues R191-R192 in 1iv8 to E while P2PSSE assigns them to $\mathcal{U}$s. Both of them have $p(E)$s as their second largest probabilities. P2PSSE assigns all the three residues A217-D218-N219 to $\mathcal{E}$s while DSSP assigns them to U. The largest probability for each residue is in bold face while the second largest in *italics*.

Method	Helix	Sheet	Loop
P2PSSE	$\mathcal{H}, G$	$\mathcal{E}, B$	$\mathcal{T}, U$
DSSP	H, G, I	B, E	T, S, C
STRIDE	H, G, I	B, E	T, C
P-SEA	a	b	c
KAKSI	H	b	“,”
PROSS	H	E	T, P
SENGO	H, G, I	E, e	B, b, P, p, O
PCASSO	H	E	C

Table S12: **The regrouping into three SSE types, helix, sheet and loop, of the assignments by eight programs**. Each item in the table is the original SSE type notations used by the corresponding program.

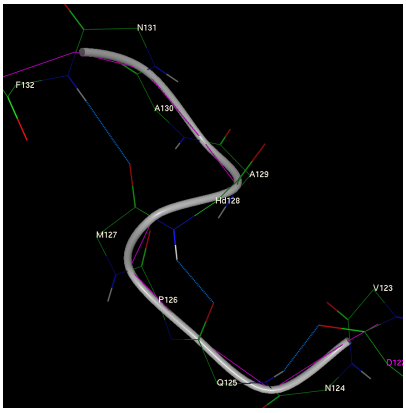


(a) $T_{pp} = 6.0$Å

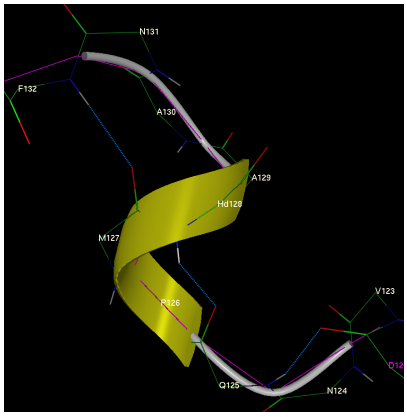


(b) $T_{pp} = 8.0$Å

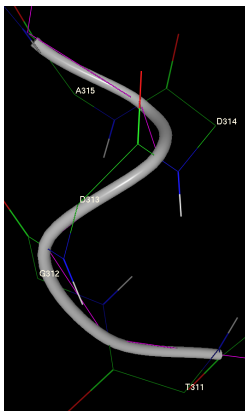
Figure 1



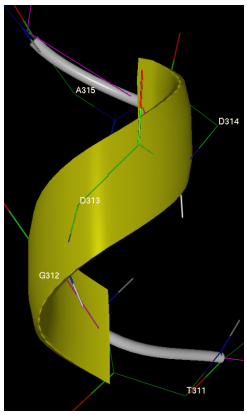
(a) Cartoon (1ah7)



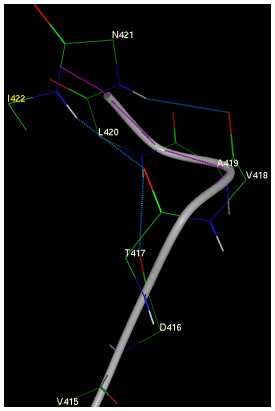
(b) Helix (1ah7)



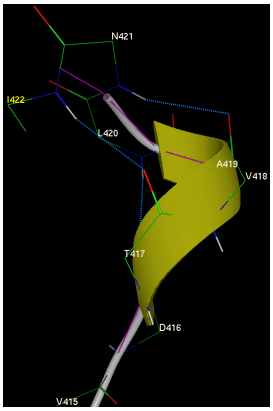
(c) Cartoon (5zxm)



(d) Helix (5zxm)

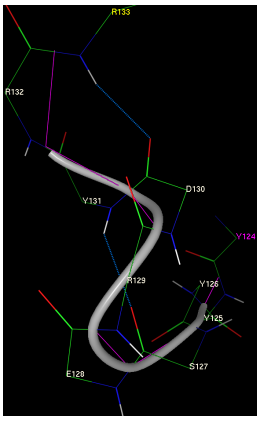


(e) Cartoon (5n13)

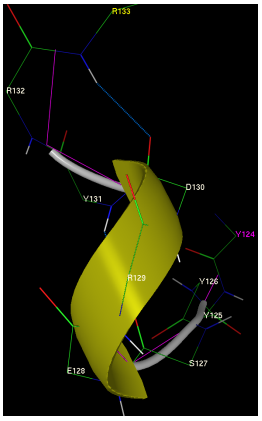


(f) Helix (5n13)

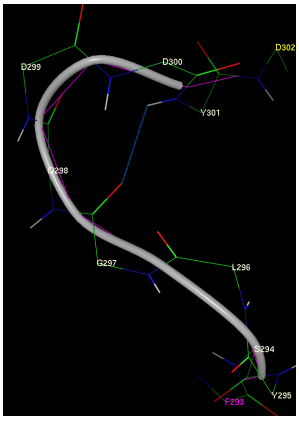
Figure 2



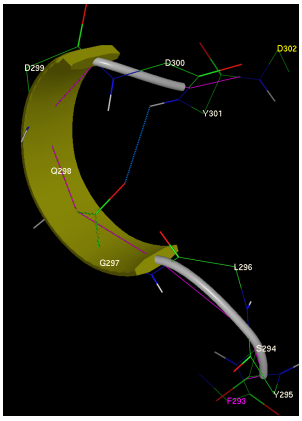
(a) Cartoon (5tkm)



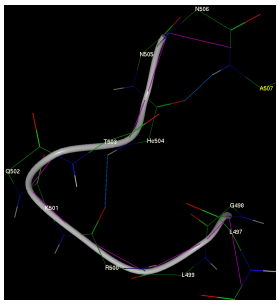
(b) Helix (5tkm)



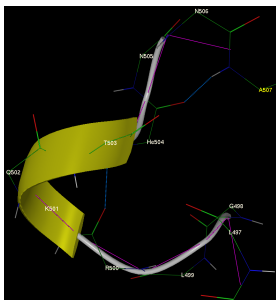
(c) Cartoon (5mx9)



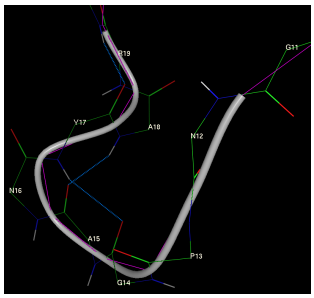
(d) Helix (5mx9)



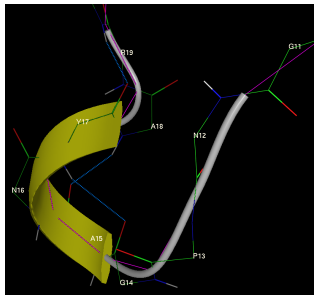
(e) Cartoon (1cza)



(f) Helix (1cza)

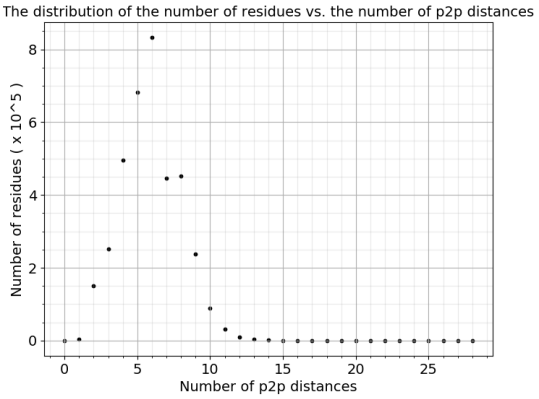


(g) Cartoon (2z2i)

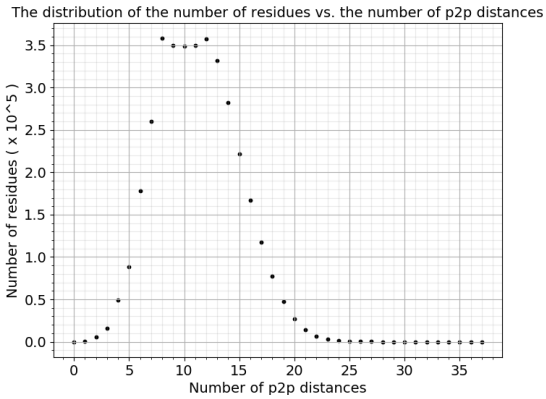


(h) Helix (2z2i)

Figure 3

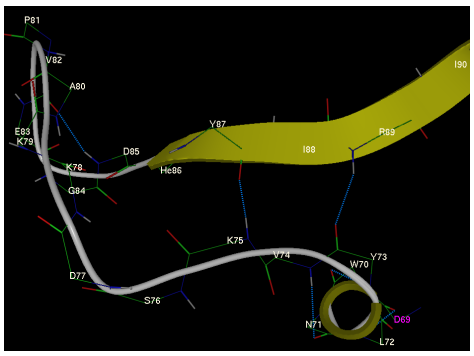


(a) $T_{pp} = 7.0$Å

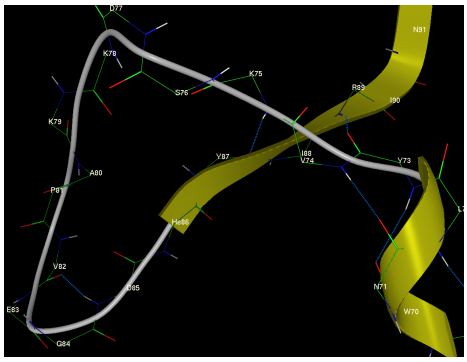


(b) $T_{pp} = 9.0$Å

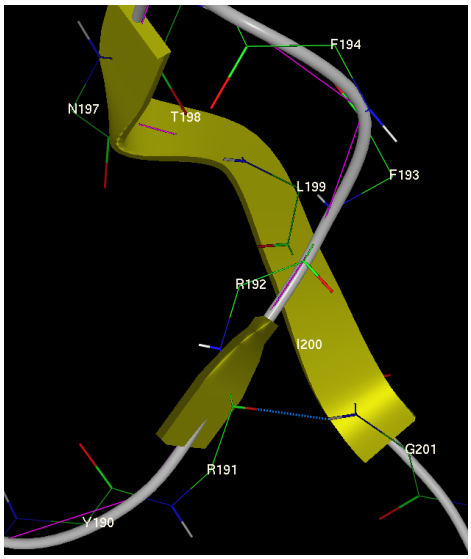
Figure S1



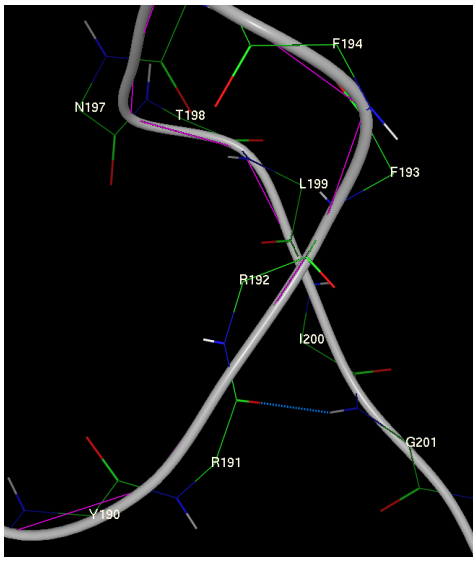
(a) Plane view (6a02)



(b) Side view (6a02)

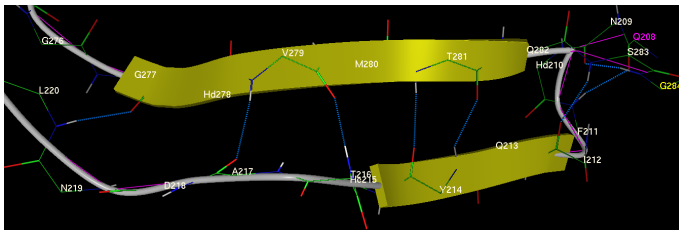


(c) Plane view (1iv8)

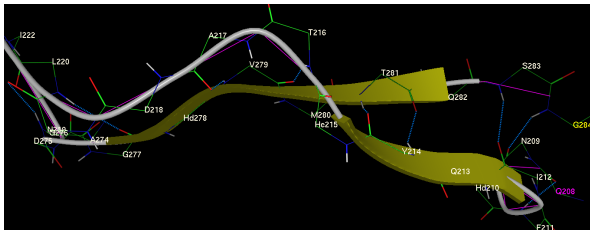


(d) Side view (1iv8)

Figure S2



(a) Plane view



(b) Side view

Figure S3