Sixty-five years of the long march in protein secondary structure prediction: the final stretch?

Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Jack Hanson, Kuldip Paliwal and Yaoqi Zhou

Corresponding author: Yaoqi Zhou, Institute for Glycomics, Griffith University, Parklands Drive, Southport, QLD 4222, Australia. Tel.: +61 (0)7 5552 8288; Fax +61 (0)7 5552 9040; E-mail: yaoqi.zhou@griffith.edu.au

Abstract

Protein secondary structure prediction began in 1951 when Pauling and Corey predicted helical and sheet conformations for protein polypeptide backbone even before the first protein structure was determined. Sixty-five years later, powerful new methods breathe new life into this field. The highest three-state accuracy without relying on structure templates is now at 82–84%, a number unthinkable just a few years ago. These improvements came from increasingly larger databases of protein sequences and structures for training, the use of template secondary structure information and more powerful deep learning techniques. As we are approaching to the theoretical limit of three-state prediction (88–90%), alternative to secondary structure prediction (prediction of backbone torsion angles and Cα-atom-based angles and torsion angles) not only has more room for further improvement but also allows direct prediction of three-dimensional fragment structures with constantly improved accuracy. About 20% of all 40-residue fragments in a database of 1199 non-redundant proteins have <6 Å root-mean-squared distance from the native conformations by SPIDER2. More powerful deep learning methods with improved capability of capturing long-range interactions begin to emerge as the next generation of techniques for secondary structure prediction. The time has come to finish off the final stretch of the long march towards protein secondary structure prediction.

Key words: secondary structure prediction, backbone structure prediction, torsion angle prediction, deep neural networks; machine learning
Introduction

Proteins are linear polymeric chains made of protein-specific sequences of 20 amino acid residue types. Proteins can perform a wide variety of molecular functions [1] ranging from molecular recognition, catalysis, molecular motors, to structural support, in part because they can fold into many different three-dimensional structural shapes [2, 3]. Thus, understanding how proteins function requires knowledge of their structures. Over 200 million protein sequences have been collected in Genbank, [4] while only ~100 000 protein structures have been deposited in Protein Data Bank, the central depository of all protein structures [5]. The gulf between the number of known sequences and the number of determined structures indicates that predicting protein structures computationally is the only practical solution, considering the high cost of protein structure determination (~$100 000 per protein [6]) and low cost of whole-genome sequencing (~$1000) [7]. However, it remains challenging to predict three-dimensional structures from protein sequences reliably without using known structures as templates [8, 9]. Thus, it is necessary to divide the protein structure prediction problem into smaller subproblems with the hope that the solutions to the subproblems will lead to the solution of the bigger problem (divide and conquer).

The most well-known subproblem of protein structure prediction is the prediction of protein secondary structure, or the local conformation of a protein’s polypeptide backbone. Solution of the secondary structure prediction problem is important in its own right because protein tertiary structures are classified into structural folds according to how secondary structure elements (helices and sheets) are packed and permuted [10, 11]. In other words, knowing secondary structures provides an approximate idea about overall structural categories. Moreover, secondary structure plays an important role in determining how proteins fold [12, 13] and how fast they fold [14]. As a result, the accuracy of protein secondary structure prediction directly impacts the accuracy of protein structure prediction (template-based or template-free) [15–18], prediction of solvent exposure of amino acid residues [19–21] and discrimination of structured from unstructured, intrinsically disordered protein regions [22, 23]. In particular, predicted probabilities of secondary structures in those intrinsically unstructured or disordered proteins (i.e. absence of unique tertiary structures) can provide clues for functional sites in their unstructured regions by binding or induced folding [24, 25]. In addition, because structures are more conserved than sequences and structures determine functions, predicted secondary structures are proven useful in protein sequence alignment [26, 27] and protein function prediction [28, 29]. Owing to the importance of secondary structure in protein structure stability and function, disease-causing mutations are often located in regions with secondary structures [30, 31]. As a result, predicted secondary structures are an important feature in the methods for discriminating disease-causing from neutral genetic variations (missense mutations [32] and small insertions/deletions (nonframeshifting [33] or frameshifting and nonsense mutations [34]).

Historically, secondary structure prediction predates the first protein structure (myoglobin) determined by X-ray crystallography in 1958 [35]. By analysing possible hydrogen-bonding patterns, Pauling and Corey [36, 37] proposed in 1951 that the dominant secondary structural motifs were $\alpha$-helices (hydrogen bonds between the i-th and the i + 4-th residues) and $\beta$-sheets (sequential hydrogen bonding between neighbouring segments rather than within a local segment). This hypothesis has since been confirmed by experimentally solved structures. Typically, high-resolution X-ray structures [38–40] were extracted to build training and test sets for secondary structure prediction, although other experimental techniques such as nuclear magnetic resonance (NMR) chemical shifts [41, 42], circular dichroism spectra [43, 44] and infrared spectroscopy [45, 46] have also been used to infer secondary structure and assess predicted ones. Despite long 65 years of history, secondary structure prediction continues to be highly active with steady improvement of accuracy till today. This raises the question of whether substantial further improvement in tackling this problem can be sustained. Because secondary structure prediction has been reviewed periodically [47–52], we will provide a discussion of only recent studies and the future prospects.

Secondary structure prediction continues with slowly rising accuracy

Figure 1 shows the number of publications per year and its accumulation since 1970s. These data were obtained by a title keyword search on protein secondary structure prediction in Web of Science on 11 August 2016, followed by manual inspection. Subject to the limitation of the keyword search, 266 methods were reported between 1973 and 2016. Despite the long history of method development, about five new methods have still been published every year since 2010.

Persistent interest in secondary structure prediction is largely because of the ability to make progress, albeit rather slowly, in improving the accuracy of secondary structure prediction. Accuracy of secondary structure prediction depends on how secondary structure is defined. The most commonly used standard is the secondary structure assignment method Dictionary of Secondary Structure of Proteins (DSSP) [53], which automatically assigns secondary structure into eight states according to hydrogen-bonding patterns. These eight states are often further simplified into three states of helix, sheet and coil. The most widely used convention is that helix is designated as $G (\alpha_1$ helix), $H (\alpha$-helix) and $I (\alpha$-helix); sheet as $B$ (isolated bridge) and $E$ (extended sheet); and all other states designated as a coil. Here, we focus on de novo secondary structure prediction in which methods were trained on labelled non-homologous sequences that have <25% sequence identity from each other. The reported three-state accuracy of secondary
structure prediction has gradually risen from 69.7% by PHD in 1993 [54], 76.5% by PSIPRED [55] in 1999, 80% by Structural Property prediction with Integrated Neural nEtwork (SPINE) [56] in 2007, 82% by Structural Property prediction with Integrated DEeP neurlal network 2 (SPIDER2) [57] in 2015, to 84% for several test data sets by Deep Convolution Neural Field network (DeepCNF) [58] in 2016. Although accuracies reported by different methods are not always directly comparable because of different data sets being used, there is a clear trend of a slow but steady improvement over the past 24 years.

What is the theoretical limit of secondary structure prediction?

The steady improvement re-raises the question regarding how much further we can go in this long march of protein secondary structure prediction. One limit imposed on secondary structure prediction is the somewhat arbitrary definition of three states. Ideal helices and sheets do not exist, and there are no clear boundaries between helix and coil nor sheet and coil states. It was shown that structural homologies differ by about 12% in secondary structure assignment [59] for those with >30% sequence identity [60]. This assignment inconsistency would limit the highest possible accuracy to about 88–90% [47].

Here, we re-examined the theoretical limit of secondary structure prediction by investigating the conservation of secondary structure among homologous proteins. The accuracy of state-of-the-art methods for secondary structure prediction relies on a sequence profile derived from multiple sequence alignment of homologous sequences, typically from position-specific substitution matrix (PSSM) calculated by Position-Specific Iterative Basic Local Alignment Search Tool (PSI-BLAST) [61]. Using PSSM implicitly assumes that homologous sequences have the same secondary structures. We used PSI-BLAST to scan homologous proteins in protein databank with E-value < 1E-3 against the 1199 proteins that were used as the test set for SPIDER for predicting Cx-based backbone angles [62]. We kept only the non-redundant (<30% sequence identity) high-resolution query chains (<3.0 Å) solved by X-ray crystallography for the analysis. The percentages of agreement on secondary structure between homologous sequence pairs were averaged for a given sequence identity (individual) or over all sequence identities (cumulative) higher than the given sequence identity. Figure 2 shows how the agreement between secondary structures changes as a function of sequence identity based on local sequence alignment. The agreement is >88% for >30% sequence identities, a commonly used cut-off for sequence homologies of identities (cumulative from high sequence identity).

The above limit, however, is most likely applicable to boundary regions, rather than internal regions of secondary structures. This is because the number of helices and sheets are more conserved than the lengths of helices and sheets. The internal regions of secondary structures are less likely subjected to assignment errors. This statement is confirmed by a detailed analysis of prediction errors at boundaries and in internal regions of helical, sheet and coil regions below.

Where does accuracy improvement come from?

Most early improvements in prediction were achieved by introducing better features. Early methods used features derived from single-residue properties [63–65]. This was followed by the inclusion of neighbouring residues within a window [68, 69, 71, 81], and later by a sequence evolution profile derived from multiple sequence alignment [75]. Sequence profiles, such as the PSSM from PSI-BLAST [61], contain conserved structural information across homologous sequences. Using sequence profiles was the main driving force for three-state accuracy going beyond 70% [54–56, 82]. In particular, SPINE [56] achieved 80% by
using classical neural networks trained on a large data set of 2640 non-redundant proteins with PSSM and physicochemical structural properties as its input.

Many recent methods improve accuracy by using actual secondary structures of homologous sequences (known as template-based approaches). Examples are HYPROSP [83], PROTEUS [84], MUPRED [85], DISTILL [86], GOR V [87], SPSSMpred [88], FLOPRED [89] and Protein Secondary Structure Prediction with Homology Analysis (SSpro) [90]. An accuracy of >90% can be achieved for some proteins by simply taking the secondary structures of their highly homologous sequences as shown in Figure 2. However, the majority of protein sequences do not have deposited structures of their homologous sequences. Thus, we focus on the methods trained and tested on non-redundant sets of <25% sequence identity.

Several methods were able to reach 81% or higher accuracy by simply updating sequences and structure databases and making minor algorithmic updates. Examples are PSIPRED 3.0 [91] and Jpred 3.0 [92]. Jpred4 was able to achieve 82% for a small representative set (150 sequences from 150 superfamilies not used in training) by training on 1338 proteins [93].

SPINE-X [94] improves over PSIPRED 3.2 by 0.7-2% through incorporating predicted backbone torsion angles and solvent accessibility iteratively. First, a neural network was used to predict secondary structure by using PSSM and a set of physicochemical properties of amino acids (PPAA). Then, PSSM and PPAA together with predicted secondary structures were used to predict solvent accessibility. Afterwards, PSSM and PPAA together with predicted secondary structures and predicted solvent accessibility were used to predict torsion angles. These predicted structural properties were used together with PSSM and PPAA to predict secondary structure for the second time. This procedure is repeated so that the secondary structure is predicted for a third time. This iterative method achieves 82% in 10-fold cross validation for a non-redundant data set of 2640 proteins (<25% sequence identity) and 81% on an independent test set of 1833 protein chains and 117 targets from the critical assessment of structure prediction (CASP9). Interestingly, SPINE-X is more accurate in predicting helical residues, whereas PSIPRED is more accurate in predicting coil residues.

Porter 4.0 [95] used two cascaded bidirectional recurrent neural networks: one for prediction and one for filtering. The method was trained and benchmarked by 5-fold cross-validation on a set of 7522 non-redundant proteins (<25% sequence identity) and a high-resolution subset of 2218 proteins (better than 2.5 Å). The three-state cross-validation accuracy is 82% for the full set and 81% for the high-resolution set. Independent testing and comparison with other methods was not available.

SeConDaRy structure Prediction (SCORPION) [96] used a data set of 166 633 high-resolution proteins at 50% sequence identity cut-off to collect triplet probabilities with a sequence separation of seven residues or less and calculated context-dependent pseudopotentials. Then, three separate neural networks (PSSM + pseudopotential, predicted secondary structure filter, and PSSM + modified pseudopotentials) were trained on 7987 chains at 25% sequence identity cut-off. Seven-fold cross-validation on 7987 chains achieves an accuracy of 82.7%, compared with 80.3% without the context-based pseudopotentials. It further shows that SCORPION is more accurate in predicting helical and sheet residues but less so in coil residues. The method had significantly better overall accuracy than several methods compared including Jpred, Porter and PSIPRED when independently tested by several small data sets.

SPIDER2 [57] applied deep neural networks to secondary structure prediction. A deep neural network refers to neural networks with more than two hidden layers [97]. Three layers were used in SPIDER2. SPIDER2, similar to SPINE-X, used an iterative improvement of secondary structure, backbone torsion angles and solvent accessibility at the same time in three iterations. The method achieves 81.8% for the independent test on 1199 high-resolution proteins (<2.0Å). When comparing SPIDER2 with SPINE-X, PSIPRED 3.3 and SCORPION in the independent test set of 1199 proteins as well as the CASP11 set, about >1% improvement over other methods was observed. SPIDER2 was not the first method that used deep neural networks for secondary structure prediction but is the most accurate according to reported accuracy [98, 99].

DeepCNF [58] is the first method using deep convolutional neural fields for secondary structure prediction. It stacks deep convolutional neural networks [100] with a conditional random-field model [101] on top as the output layer. The deep convolutional neural networks were used to capture the complex sequence–structure relation of longer sequence separation than a typical deep neural network, whereas the conditional random-field model allows detection of interdependence of secondary structure among neighbouring residues. It shows that five hidden layers and an 11-residue window were needed to achieve the best training. The method was trained on 5600 proteins with 25% sequence identity cut-off within itself and to several test sets including CASP targets. The accuracy of test sets ranges from 82.3 to 85.4%, which improves over several methods including SPINE-X, PSIPRED and Jpred by 1-4%, depending on the data sets.

The accuracy of state-of-the-art methods on the same independent test sets

To compare these state-of-the-art techniques by the same independent test, we downloaded on 20 September 2016 all protein structures determined by X-ray crystallography with resolution better than 3 Å and released after 1 January 2016. We used 3 Å resolution as a cut-off because we would like to have >100 structures, and a previous study showed that the accuracy of secondary structure prediction is only weakly dependent on X-ray resolution (0.1% difference from 2 to 3 Å cut-off) [56]. To provide a truly independent test set for all methods accepted before 2016 (including DeepCNF), we have removed sequences that have >30% identity to those released before 2016 according to CD-hit [102]. The final data set contains 115 proteins (TS115) with sequence lengths ranging from 43 to 1085 (the list is available at http://sparks-lab.org). Here, we have assumed that all published methods were not automatically re-trained with newly deposited proteins in 2016.

We conducted prediction of Jpred4, SCORPION, Porter4.0, PSIPRED 3.3, SPINE X, SPIDER2 and DeepCNF using their respective online servers except that a stand-alone version of SCORPION was used because the server was not working at the time of testing. As shown in Table 1, the three-state accuracies for these newly released targets are 77.1% by Jpred4, 80.1% by SPINE X, 80.2% by PSIPRED 3.3, 81.7% by SCORPION, 81.9% by SPIDER2, 82.0% by Porter 4.0 and 82.3% by DeepCNF. The overall accuracies for different methods are consistent with large-scale tests reported in respective studies. According to the performance in individual proteins, DeepCNF is statistically significantly different from all methods (P-value <0.05) except Porter 4.0, whereas SPIDER2 is not statistically significantly different.
from Porter 4.0, SCORPION and PSIPRED 3.3. Obviously larger independent data sets are needed to increase statistical power to separate these methods. A consensus prediction of top three (DeepCNF, Porter 4.0 and SPIDER2) or top five methods (DeepCNF, Porter 4.0, SPIDER2, SCORPION and PSIPRED3.3) achieved an overall accuracy of 83.5 and 83.6%, respectively, >82.3% by DeepCNF, but not statistically significantly different from DeepCNF. To confirm the weak dependence on X-ray resolution, we found that the accuracy of SPIDER2 is 81.9% for structures with <2.5 Å resolution, compared with 81.8% for structures with <3 Å resolution.

We also obtained the CASP12 (the protein targets from the 12th biannual meeting of Critical Assessment of Structure Prediction techniques) targets released in 2016 from http://www.predictioncenter.org/casp12/targetlist.cgi. There are 23 targets with structures accessible in the PDB. After removing the structures determined by the NMR technique and those low-resolution structures by X-ray crystallography (<3.5 Å), the final data set contains 15 targets. Their PDB and chain IDs are 4ympA, 5a7dB, 5a7dL, 5aatA, 5erfA, 5ja4AA, 5ja4B, 5jsVA, 5jsVB, 5jsVC, 5jmbA, 5jnaA, 5knpA and 5knaA. These structures are carefully selected to not be independent test because they were considered to be an independent test because they were carefully selected to not be homologous with any structures published before May 2016 by the CASP organizer. As shown in Table 1, the three-state accuracies for these 15 CASP12 targets are 75.1% by Jpred4, 76.9% by SPINE X, 78.6% by PSIPRED 3.3, 79.8% by Porter 4.0, 80.5% by SCORPION and 82.1% by DeepCNF. The overall relative accuracies for different methods are consistent with the larger data set of 115 proteins. However, because of the small test set, paired t-test suggests that only the differences between SPIDER2 and Jpred4 and between SPIDER2 and SPINE X are statistically significant with P-value <0.05. The difference between DeepCNF and SPIDER2 or between SCORPION and SPIDER2 is not statistically significant.

Jpred only predicts the sequence <800 residues. For TS115, there is one sequence (5hdtA) with 1085 residues. 5hdtA was divided into two chains with 800 residues and 285 residues, respectively.

Nevertheless, it has been shown that predicted secondary structures for CASP targets are usually more accurate than those derived from model structures predicted by various structure prediction techniques (template-based or template-free) [57, 94].

Where are the errors in secondary structure prediction?

Helices involve hydrogen bonds of sequence neighbours, whereas sheets are defined based on hydrogen bonds between amino acid residues that are not necessary sequence neighbours. As a result, helical residues are more accurately predicted than sheet residues as expected [47]. For example, the accuracy of SPIDER2 for the CASP11 data set is 86.2% for helix, 75.8% for sheet and 78.6% for coil [57]. One can also expect that the possibility of confusion between helical and sheet residues is lower than the confusion between helical and coil residues and between sheet and coil residues. Indeed, the confusion between helix and sheet residues is 1–2% and between coil and helix (or sheet) residues is 8–9% [56, 57, 94]. The larger confusion between coils and secondary structure elements raises the possibility that the boundaries (capping) of helices and sheets have larger errors. Several methods were developed to predict capping regions [103, 104].

Table 1. Method comparison based on Q3 using newly released structures (TS115) and CASP12 targets (15 proteins) for secondary structure prediction

<table>
<thead>
<tr>
<th>Data set Method</th>
<th>TS115</th>
<th>CASP12</th>
<th>Server location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Q3</td>
<td>P-value</td>
<td>Q3</td>
</tr>
<tr>
<td>Jpred4</td>
<td>0.771</td>
<td>0.0007</td>
<td>0.751</td>
</tr>
<tr>
<td>SPINE X</td>
<td>0.801</td>
<td>0.0002</td>
<td>0.769</td>
</tr>
<tr>
<td>PSIPRED 3.3</td>
<td>0.802</td>
<td>0.12</td>
<td>0.780</td>
</tr>
<tr>
<td>SCORPION</td>
<td>0.817</td>
<td>0.45</td>
<td>0.805</td>
</tr>
<tr>
<td>SPIDER2</td>
<td>0.819</td>
<td>NA</td>
<td>0.798</td>
</tr>
<tr>
<td>PORTER 4.0</td>
<td>0.820</td>
<td>0.17</td>
<td>0.798</td>
</tr>
<tr>
<td>DeepCNF</td>
<td>0.823</td>
<td>0.01</td>
<td>0.821</td>
</tr>
</tbody>
</table>

Note:
*Paired t-test from SPIDER2.

Table 2 analysed errors of predicted secondary structures in TS115 and CASP12 data sets by using SPIDER2 and DeepCNF as examples. Both showed low confusion between helix and sheet (<1%), followed by ~7% confusion between sheet and coil and ~10% confusion between helix and coil. Interestingly, there is significantly less confusion between sheets and coils than between helices and coils. DeepCNF is more accurate in discriminating coils from sheets and helices from sheets, whereas SPIDER2 is more accurate in separating helices from coils. We further confirmed that errors in helical and sheet boundaries are significantly larger than errors in the interior of a helix or a sheet. Here, we defined a helical/sheet residue as internal if its two nearest neighbouring residues are also helical/sheet residues and as a boundary if one or both of the nearest neighbours has a different secondary structural assignment. As shown in Table 2, about 10–11% helical errors are in helical internal regions but 38–43% are at boundaries. The same is true for sheet prediction with the largest errors (~38%) at boundaries. On the other hand, errors on coil residues less depend on their
boundaries, consistent with their flexibility in conformations. The overall accuracies (Q3) are 88% by SPIDER2 and 89% by DeepCNF in internal regions of secondary structures and 69% by SPIDER2 and 68% by DeepCNF at boundaries.

Another possible source of errors is the short chameleon sequences that have different types of secondary structure in different proteins [105, 106]. These sequences are implicated in amyloid-related diseases [107]. However, several studies [108–110] have indicated that chameleon sequences were predicted as accurately as other regions, likely because a typical sliding window (20 residues) is much longer than 10 residues, which is the longest length of chameleon sequences found [106]. This suggests that local interactions play the dominant role in determining the secondary structure of short chameleon sequences.

Hydrophobic interactions drive folding of soluble proteins [111] that have a hydrophobic core and hydrophilic surface. Helices and sheets buried deep inside a hydrophobic core have a similar folding environment as transmembrane helices [112, 113]. There is a possibility that machine learning techniques may confuse hydrophobic sheets inside protein cores with transmembrane helical segments and reduce the accuracy of beta-sheet prediction. But this possibility is low, as Table 2 shows that the confusion between helices and sheets is <1%.

To further examine this possibility, we used results of TS115 by SPIDER2 as an example. Figure 3A shows that the accuracies in fully buried regions of proteins are the highest for helices and sheets but the lowest for coils. The confusion between helices and sheets is nearly independent of solvent accessibility (Figure 3B). Thus, there is little evidence of confusion between transmembrane helices and hydrophobic sheets inside protein cores. On the other hand, exposed helices and sheets are more difficult to predict and easier to confuse with coil residues.

The largest source of errors, however, is likely because of the inability of current methods in incorporating the effect of non-local interactions. Non-local interactions are the interactions between residues that are close in three-dimensional space but far from each other in their respective sequence positions. Entropy densities of primary and secondary structure sequences suggested that local inter-sequence correlations contributed only one-fourth of the total information needed to determine the secondary structure [114]. Kihara showed that the accuracy of predicted helical residues and sheet residues is negatively correlated with residue contact order, or number of non-local contacts [115]. This is consistent with the fact that inputting native non-local contacts into an interaction-enriched bidirectional recurrent neural network improves secondary structure prediction accuracy from 79.9 to 84.6% [116, 117].

Table 2. Overall misclassification errors of H, E and C states and the errors in the internal and at the boundary region of secondary structure elements along with prediction accuracy in the internal and at boundary regions of secondary structures for newly released structures (TS115) by SPIDER2 and DeepCNF

<table>
<thead>
<tr>
<th>Method</th>
<th>SPIDER2 (%)</th>
<th>DeepCNF (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>H =&gt; E</td>
<td>0.96</td>
<td>0.52</td>
</tr>
<tr>
<td>H =&gt; C</td>
<td>9.5</td>
<td>10.1</td>
</tr>
<tr>
<td>E =&gt; C</td>
<td>7.6</td>
<td>7.0</td>
</tr>
<tr>
<td>H/E/C (internal)</td>
<td>9.7/14.5/14.3</td>
<td>11.0/11.9/10.9</td>
</tr>
<tr>
<td>H/E/C (boundary)</td>
<td>37.9/38.3/24.3</td>
<td>43.1/37.8/23.4</td>
</tr>
<tr>
<td>Q3 (internal)</td>
<td>87.9</td>
<td>88.9</td>
</tr>
<tr>
<td>Q3 (boundary)</td>
<td>68.6</td>
<td>67.9</td>
</tr>
</tbody>
</table>

Figure 3. The dependence of accuracies and misclassifications on solvent accessibility. The accuracy of predicting helices (QH), sheets (QE) and coils (QC) and the overall accuracy (Q3) (A) and the misclassifications of helices to coils, sheets, sheets to coils and helices to helices and sheets (B) as a function of solvent accessibility for TS115 by SPIDER2.

Figure 4. The dependence of accuracy on non-local contacts. The secondary structure accuracy as a function of the number of non-local contacts (|i−j| > 19) for the independent test set (TS1199) by SPINE X and SPIDER2.

Overriding predicted secondary structures has also led to improved model structures [118–120]. To further illustrate the dependence of secondary structure on non-local contacts, Figure 4 shows the dependence of Q3 as a function of number of non-local contacts (defined as |i−j| > 19 and the Cα-Cα distance <8 Å). We chose a cut-off of 19-residue separation because most secondary structure prediction used a window size of 10–20 residues. Because TS115 only has a few hundreds of residues at seven non-local contacts or more, we used the independent test set of 1199 high-resolution proteins from SPIDER [62]. Each statistical bin contains at least 12 500
residues. The three-state accuracy of secondary structure prediction of SPINE X as well as SPIDER2 decreases nearly linearly when the number of non-local contacts is >5. This happens despite that SPIDER2 used a deep learning neural network, confirming the failure of regular and deep neural networks in capturing long-range interactions.

**Alternatives to discrete three-state secondary structure prediction**

Although secondary structure prediction in three states is important, its limitation cannot be overlooked. This is because three secondary structure states are only a coarse-grained representation of the backbone structure with helical and sheet residues that often deviate significantly from standard helix and sheet conformations. In fact, DSSP, the most widely used secondary structure assignment program, defined eight states. This leads to several recent methods dedicated to eight-state prediction such as SSpro8 [121], RaptorXss8 [122], SCORPION [96, 123] and DeepCNF [58]. DeepCNF, in particular, pushed the eight-state accuracy to beyond 70%.

However, the boundaries between different states are somewhat arbitrarily defined. Even in three states, different secondary assignment techniques can differ by much as 15% [60]. Inconsistent assignment determines the theoretical limit of secondary structure prediction as discussed above. Moreover, predicted coil residues do not have a well-defined conformation.

Backbone structures can be defined continuously in three rotational (torsion) angles along the C-N (\(\omega\)), Cα-N (\(\psi\)) and Cα-C (\(\phi\)) bonds, respectively. \(\omega\) is approximately fixed at 180° for the common trans and 0° for the rare cis conformation because of rigid planar peptide groups. Thus, only two torsion angles per residue are needed for defining a backbone structure. The first continuous or real-value prediction of \(\phi\) and \(\psi\) was made by Xue et al. in 2008 [124]. The mean absolute errors (MAEs) were 25° for \(\phi\) and 38° for \(\psi\), respectively, based on 10-fold cross-validation of 2640 high-resolution non-redundant proteins.

The accuracy angle has been substantially improved over the years [94, 125, 126] with the highest accuracy reported by SPIDER2 (MAE = 19° for \(\phi\) and 30° for \(\psi\) using an independent test of 1199 high-resolution non-redundant proteins). Backbone torsion angles, however, are single-residue structural properties. By comparison, helical and sheet conformations implied hydrogen bonding between two residues that are separated in sequence positions (four for \(_{310}\) helix, five for \(\alpha\)-helix and undefined for sheet residues). It is desirable to have angles representing more than single residues. The peptide planar leads to the neighboring Cα atoms at about a fixed distance of around 3.8 Å from each other. Thus, the protein structure can also be uniquely represented by two angles between Cα_{i} - Cα_{i+1} (\(\theta\)) for three-residue coupling and a dihedral angle rotated about the Cα_{i} - Cα_{i+1} bond (\(\varphi\)) for four-residue coupling. Such angles have been widely used in coarse-grained modelling of dynamics, folding and assessment of protein structures [127–132]. This representation is complementary to \(\phi\) and \(\psi\) angles (single residue) and secondary structures (>3 residues).

**Table 3** compares the performance of SSpro8 and DeepCNF in eight-state performance for TS115 and CASP12 targets. SCORPION was not included because its server is not working at the time of testing. The accuracies according to Q8 are 68 and 69% for TS115 and CASP12, respectively, by SSpro8 and 72 and 73%, respectively, by DeepCNF. This table confirms that DeepCNF achieved >70% accuracy in eight-state prediction.

**Table 4** examines the accuracy of angle prediction according to MAE for different angles by SPINE X and SPIDER2. SPIDER2 yields a significant improvement in \(\phi\) and \(\psi\) over SPINE X MAE values for TS115 and CASP12 sets (18° for \(\phi\) and 28–29° for \(\psi\)) are slightly better than the reported MAE values (19° for \(\phi\) and 30° for \(\psi\) using an independent test of 1199 high-resolution non-redundant proteins).

 Backbone torsion angles, however, are single-residue structural properties. By comparison, helical and sheet conformations implied hydrogen bonding between two residues that are separated in sequence positions (four for \(_{310}\) helix, five for \(\alpha\)-helix and undefined for sheet residues). It is desirable to have angles representing more than single residues. The peptide planar leads to the neighboring Cα atoms at about a fixed distance of around 3.8 Å from each other. Thus, the protein structure can also be uniquely represented by two angles between Cα_{i} - Cα_{i+1} (\(\theta\)) for three-residue coupling and a dihedral angle rotated about the Cα_{i} - Cα_{i+1} bond (\(\varphi\)) for four-residue coupling. Such angles have been widely used in coarse-grained modelling of dynamics, folding and assessment of protein structures [127–132]. This representation is complementary to \(\phi\) and \(\psi\) angles (single residue) and secondary structures (>3 residues).

SPIDER [62] was the first method for sequence-based prediction of \(\theta\) and \(\varphi\) using an independent test of 1199 high-resolution non-redundant proteins. As **Table 4** shows, MAE values are 8° for \(\theta\) and 32° for \(\varphi\) using an independent test of 1199 high-resolution non-redundant proteins. The MAE values achieved by SPIDER2 are 8° for \(\theta\) and 32° for \(\varphi\) for both TS115 and CASP12 sets, confirming the reported accuracy.

The above results also raised a question about the potential theoretical limit for sequence-based prediction. Similar to secondary structures in **Figure 2**, using homologous sequences for predicting torsion angles would lead to 15.6° for \(\phi\) and 19.4° for \(\psi\) based on MAE for homologous sequences of >30% sequence identities, compared with 19° for \(\phi\) and 30° for \(\psi\) by SPIDER2. There is significant room for improving \(\psi\), in particular. Similarly, we found that MAE values for homologous sequences of >30% sequence identities are 5° for \(\theta\) and 15.5° for \(\varphi\), compared with 8° for \(\theta\) and 32° for \(\varphi\) by SPIDER2. Again, there is significant room for improving \(\varphi\). One nice feature regarding predicted angles is that it can be used to reconstruct the overall backbone structure [57], some of which can achieve low root-mean-squared distance (RMSD) without performing additional optimization and refinement. One example at 40 residues long is shown in **Figure 5**. The fraction of constructed three-dimensional structures with a correct fold (RMSD < 6 Å [133]) for all 182 725 40-residue fragments in a data set of 1199 proteins is small but highly significant (16.3% by predicted \(\varphi\) and \(\psi\) angles and 19.1% by predicted \(\theta\) and \(\varphi\) angles). It indicates room for further improvement in angle prediction, but is also in part because small angle shifts may lead to large changes in overall structures. It is of interest to note that more fractions of accurate structures are constructed by using predicted \(\theta\) and \(\varphi\) angles. This is likely because these two angles involve—three to four

### Table 3. Method comparison based on Q8 using newly released structures (TS115) and CASP12 targets (15 proteins) for eight-state secondary structure prediction

<table>
<thead>
<tr>
<th>Data set</th>
<th>Method</th>
<th>TS115</th>
<th>CASP12</th>
<th>Server location</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Q8</td>
<td>P-value*</td>
<td>Q8</td>
</tr>
<tr>
<td></td>
<td>SSpro8</td>
<td>0.68</td>
<td>3E-9</td>
<td>0.69</td>
</tr>
<tr>
<td></td>
<td>DeepCNF</td>
<td>0.72</td>
<td>NA</td>
<td>0.73</td>
</tr>
</tbody>
</table>

*Paired-t test from DeepCNF.
Table 4. Method comparison based on MAE using newly released structures (TS115) and CASP12 targets (15 proteins) for prediction of backbone angles ($\phi$, $\psi$, $\theta$ and $\chi$)

<table>
<thead>
<tr>
<th>Data set Method</th>
<th>TS115 $\phi$ (%)</th>
<th>TS115 $\psi$ (%)</th>
<th>CASP12 $\phi$ (%)</th>
<th>CASP12 $\psi$ (%)</th>
<th>Server location</th>
</tr>
</thead>
<tbody>
<tr>
<td>SPINE-X</td>
<td>19.4</td>
<td>32.9</td>
<td>19.1</td>
<td>33.2</td>
<td><a href="http://sparks-lab.org/SPINE-X/">http://sparks-lab.org/SPINE-X/</a></td>
</tr>
<tr>
<td>SPIDER2</td>
<td>18.2</td>
<td>29.3</td>
<td>17.8</td>
<td>28.1</td>
<td><a href="http://sparks-lab.org/server/SPIDER2/">http://sparks-lab.org/server/SPIDER2/</a></td>
</tr>
<tr>
<td>SPIDER2</td>
<td>7.89</td>
<td>30.8</td>
<td>8.31</td>
<td>31.1</td>
<td><a href="http://sparks-lab.org/server/SPIDER2/">http://sparks-lab.org/server/SPIDER2/</a></td>
</tr>
</tbody>
</table>

Figure 5. Direct prediction of three-dimensional structure by predicted angles. Structure (dark colour) constructed directly from $\phi/\psi$ angles compared with native structure (light colour) for residues 24–63 from PDB 5fdy chain A.

residues, compared with only one residue in the case of $\psi$ and $\psi$ angles.

Real-value prediction of angles, however, does not provide an estimate of errors in predicted values. This is different from multistate prediction where actual output values for different states can be normalized as probabilities of predicted states. Recently, Gao et al. [134] show that it is possible to predict the absolute errors of predicted local backbone angles with reasonable accuracy by deep neural networks with a Spearman correlation coefficient between predicted and actual errors at about 0.6. Predicted errors were found useful for model quality assessment.

Another source of errors in using predicted angles for model construction is the rare cis-conformation of proline and other residues ($\omega = 0^\circ$), which leads to a reduction of the distance between two neighbouring Cα atoms from 3.81 to 2.94 Å [135]. A few methods were developed for predicting proline conformations [136–140] with the highest reported accuracy at 72%. Unfortunately, the links to two recently published Web servers are no longer active for further examination. There is a clear lack of development in $\omega$, despite the importance of cis-trans conformational transitions in protein folding and function [141, 142].

All of the above methods for backbone structure prediction were based on secondary structures derived from three-dimensional structures determined by X-ray crystallography. However, proteins are dynamic, and secondary structures will be subjected to fluctuation in a solution, although they may be stabilized in a crystal environment [143]. The S2D method [144] opens a new avenue of secondary structure prediction by training on the probability distributions of secondary structure elements in disordered states derived from NMR chemical shifts for 2223 protein sequences. This technique is based on three separate single hidden layer feedforward neural networks, two of which used different window sizes for three-state prediction, while the remaining neural network was used to incorporate global secondary structure contents for final prediction. The method can also identify intrinsically disordered proteins. In addition, it yielded a three-state accuracy at about 79% for a data set of 1833 structured protein chains, validating the use of secondary structure probabilities derived from chemical shifts.

**Future perspective**

Secondary structure prediction has reached an accuracy (about 84%) that is close to the theoretical limit (~88%). Can we make the last stretch of this long-standing challenge? Recent accuracy improvements [91–93, 95] have resulted from increasingly larger sequence [4] and structural databases [5, 145], more sophisticated deep learning neural networks [57, 58], and the use of structural template information in whole [83–90] or in fragments [96, 123]. As the number of protein sequences, along with the number of solved structures, continues to rise exponentially, the secondary structure prediction accuracy is likely to continue its incremental improvement. One main obstacle, however, is the difficulty in capturing non-local interactions between those residues that are close in three-dimensional space but far from each other in their respective sequence positions [115]. All existing techniques have relied on window-based features to capture ‘non-local’ interactions limited to 10–30 amino acid residues apart. This is certainly not sufficient for medium to large proteins, in particular.

Non-local interactions can be described by residue-residue contact maps. Contact maps can be predicted from correlated mutations or evolution coupling [146–149], assuming that mutations are correlated because of close proximity in structure. This type of method is highly accurate for large protein families with thousands of sequences. A recent study further found that the codon-level information has added benefit for improving contact prediction [150]. Many machine learning techniques with or without evolutionary coupling techniques (for recent reviews, see [151, 152]) were also developed for contact map.
prediction, and their accuracy is assessed in biannual CASP meetings, with the best precision at about 30% [152, 153]. However, how to take advantage of these predicted non-local contacts (or structures) for improving secondary structure prediction remains to be further explored [118, 154], perhaps in the context of coupling of secondary and tertiary structure prediction [13, 120, 155].

In 1997, Hochreiter and Schmidhuber [156] showed that useful long-range interactions between a series of time-resolved events can be memorized by enforcing the constant error flow, regardless of the time lapse. This long short-term memory (LSTM) network was demonstrated by its ability to capture long-range dependencies in bidirectional recurrent neural network applications [157] for state-of-the-art interpretation and prediction, most notably in speech- and image-related problems [158, 159]. Such a network should be able to capture non-local interactions in a protein without using a sequence window. Indeed, the application of the LSTM network to protein disorder prediction has confirmed its improvement over regular neural networks in detecting structured and unstructured regions [160]. Our initial application (Heffernan et al., submitted) to protein secondary structure and backbone angle prediction (SPIDER3) has produced a three-state accuracy of 83.9% for the independent test set TS115 along with a 10% reduction of MAEs for $\phi$, $\psi$ and $\tau$ angles without expanding the training database. Moreover, 27% of 182 724 40-residue fragments built by predicted $\theta$ and $\tau$ angles are <6 Å RMSD away from their native conformations. More importantly, its improvement over SPIDER2 is the largest for residues with the highest number of non-local contacts ($|i-j|>19$). This result signals the emergence of the fourth-generation deep learning-based methods that are promising to complete the last stretch of long march in secondary (and therefore tertiary) structure prediction.

Key Points

- The accuracy of state-of-the-art three-state secondary structure prediction is at all time high of 82–84%.
- The improvement comes from large databases, the use of template and powerful deep learning techniques.
- Alternative to secondary structure prediction (backbone angle prediction) has more room for further improvement.
- Future is bright as next-generation deep learning techniques can remember long-range interactions.

Acknowledgements

The authors thank the Australian Research Council grant LE150100161 for infrastructure support. We also gratefully acknowledge the use of the High Performance Computing Cluster ‘Gowonda’ to complete this research. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (Q CIF).

Funding

The National Natural Science Foundation of China (grant numbers 61271378 and 61671107 to Y.Y. and J.W.), the Taishan Scholars Program of Shandong province of China, National Natural Science Foundation of China (grant number 61540025) and National Health and Medical Research Council of Australia (grant numbers 1059775 and 1083450 to Y.Z.). Specialized Research Fund for the Doctoral Program of Higher Education (SRFDP) (grant number 20130031120001 to J.G.)

References

7. Mardis ER. Anticipating the 1,000 dollar genome. Genome Biol 2006; 7:112.


