OXFORD

Briefings in Bioinformatics, 19(3), 2018, 482–494

doi: 10.1093/bib/bbw129
Advance Access Publication Date: 31 December 2016
Paper

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)75552 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

Yuedong Yang is a Research Fellow at Institute for Glycomics, Gold Coast Campus, Griffith University, Australia. His research focuses on algorithm development in structural bioinformatics, more specifically in the area of structure and function prediction of proteins and RNA.
Jianzhao Gao is a Lecturer at School of Mathematical Sciences and LPMC, Nankai University, Tianjin, China. He conducts research at the interface of mathematics and biology with a specific focus on machine learning and datamining of biological data.
Jihua Wang is Professor and Director of Shandong Provincial Key Laboratory of Biophysics, Institute of Biophysics, Dezhou University, China. He is interested in interpreting biological phenomena by using physical principles.
Rhys Heffernan is a PhD student in the Signal Processing Laboratory, Griffith University, Brisbane, Australia. His research focuses on using advanced signal-processing techniques to capture deeper meaning of biological data.
Jack Hanson is a PhD student in the Signal Processing Laboratory, Griffith University, Brisbane, Australia. His research activity involves implementation of modern machine learning techniques for improving protein structure prediction.
Kuldip Paliwal is a Professor in the Signal Processing Laboratory, Griffith University, Brisbane, Australia. His research activity covers signal processing, speech coding and recognition and machine learning.
Yaoqi Zhou is a Professor and Research Leader at Institute for Glycomics and School of Information and Communication Technology, Gold Coast Campus, Griffith University, Australia. He is also a visiting Professor at Dezhou University. His current research involves integration of computational prediction and experimental validation for small molecule and peptide drug discovery as well as protein structure and function prediction.
Submitted: 10 October 2016; **Received (in revised form):** 15 November 2016

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 permutated [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 α-helices (hydrogen bonds between the ith and the $i+4$th residues) and β-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.

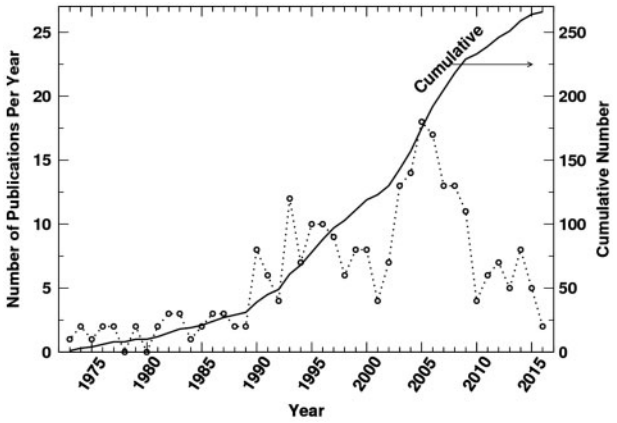

Figure 1. New methods continue in development for secondary structure prediction. The number of publications on protein secondary structure prediction per year and its cumulative increment.

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 (3_{10} helix), H (α-helix) and I (π-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 neuRal 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 Cα-based backbone angles [62]. We kept only the non-redundant (<90% 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 proteins. This result is consistent with the 88–90% limit suggested previously [59].

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.

Three generations of methods for improving secondary structure prediction

Secondary structure prediction techniques have been classified into three generations [47]. In the first generation, secondary

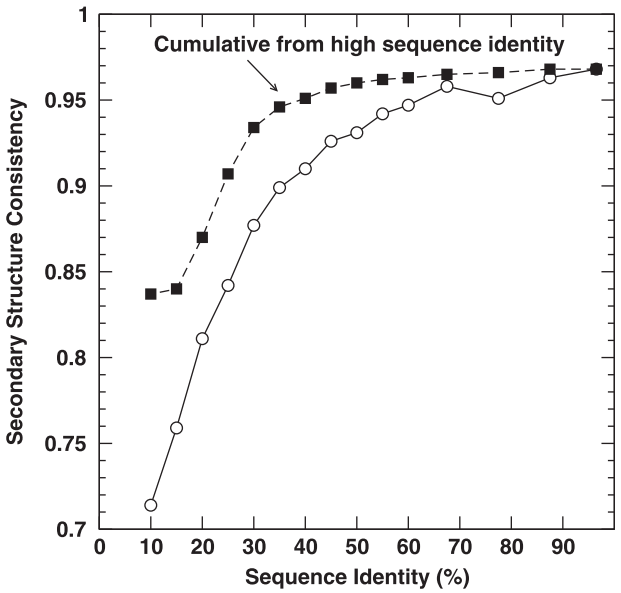

Figure 2. Conservation of secondary structure in homologous sequences. The average consistency on secondary structure of homologous sequences at a given sequence identity between two sequences compared or over all compared sequences above a given sequence identity (cumulative from high sequence identity).

structures were predicted from a protein sequence according to statistical propensities of amino acid residues towards a specific secondary structure element [63–65]. The most representative of this type of first-generation methods is the Chou–Fasman method [65], which combined propensities with heuristic rules. The second-generation methods, represented by the Garnier-Osguthorpe-Robson (GOR) method [66] and the Lim method [67], used a sliding window of neighbouring residues and various theoretical algorithms such as statistical information [66, 68, 69], graph theory [70], neural networks [71, 72], logic-based machine learning techniques [73] and nearest neighbouring methods [74]. The use of information from neighbouring residues is made possible as more protein structures became available to estimate pairwise, triplet or longer-segment frequencies. The third generation of techniques is characterized by using evolutionary information derived from alignment of multiple homologous sequences [54, 75]. During this period, new computational algorithms have been implemented. Examples are support vector machines [76, 77], Bayesian or hidden semi-Markov network [78, 79] and conditional random fields for combined prediction [80]. Out of these methods, neural-network-based models have been seen the highest reported accuracy [54–58].

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 physiochemical 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 physiochemical 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

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

Data set Method	TS115		CASP12		Server location
	Q3	*P*-value[a]	Q3	*P*-value[a]	
Jpred4	0.771[b]	0.0007	0.751	0.04	http://www.compbio.dundee.ac.uk/jpred4/index.html
SPINE X	0.801	0.0002	0.769	0.006	http://sparks-lab.org/SPINE-X/
PSIPRED 3.3	0.802	0.12	0.780	0.19	http://distillf.ucd.ie/porterpaleale/
SCORPION	0.817	0.45	0.805	0.44	Stand-alone version from http://hpcr.cs.odu.edu/c3scorpion/
SPIDER2	0.819	NA	0.798	NA	http://sparks-lab.org/server/SPIDER2/
PORTER 4.0	0.820	0.17	0.798	0.67	http://distillf.ucd.ie/porterpaleale/
DeepCNF	0.823	0.01	0.821	0.14	http://raptorx2.uchicago.edu/StructurePropertyPred/predict/

Note.
[a]Paired *t*-test from SPIDER2.
[b]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.

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 81.93% for structures with <2.5 Å resolution, compared with 81.87% 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, 5aotA, 5ereA, 5fjlA, 5j4aA, 5j4aB, 5j5vA, 5j5vB, 5j5vC, 5jmbA, 5jmuA, 5kkpA and 5ko9A. These structures are 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, 78.0% by PSIPRED 3.3, 79.8 % by Porter 4.0, 79.8% by SPIDER2, 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.

It is noted that the accuracy of 79.8–82.1% by Porter 4.0, SPIDER2 and DeepCNF in CASP is lower than the larger data set of 115 proteins as well as what was claimed for other test sets in their respectively original publications. This is in part because of the small test set of 15 targets. However, low performance for CASP targets has been observed earlier [57, 58, 94], largely because of the fact that remote homologies were removed by PSI-BLAST for CASP targets. New orphan sequences, which are highly dissimilar to existing sequences with known structures, likely have few sequence neighbours to yield effective sequence profiles and may have adopted less popular structural folds [11]. 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 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

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

Method	SPIDER2 (%)	DeepCNF (%)
H ⟺ E	0.96	0.52
H ⟺ C	9.5	10.1
E ⟺ C	7.6	7.0
H/E/C (internal)	9.7/14.5/14.3	11.0/11.9/10.9
H/E/C (boundary)	37.9/38.3/24.3	43.1/37.8/23.4
Q3 (internal)	87.9	88.9
Q3 (boundary)	68.6	67.9

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].

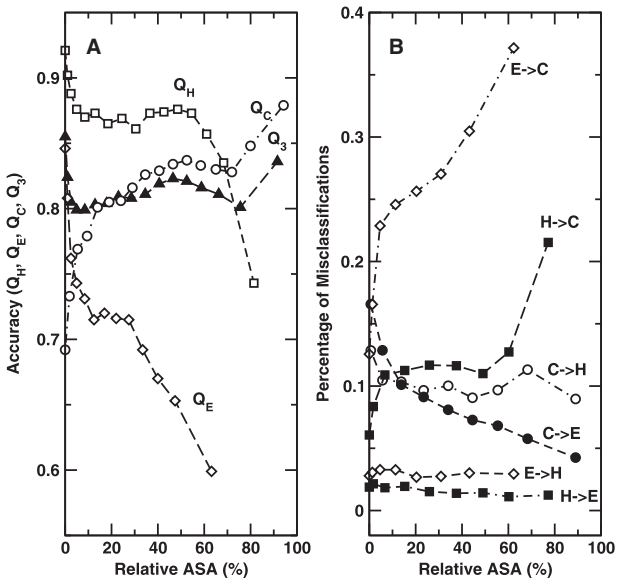

Figure 3. The dependence of accuracies and misclassifications on solvent accessibility. The accuracy of predicting helices (Q_H), sheets (Q_E) and coils (Q_C) and the overall accuracy (Q_3) (**A**) and the misclassifications of helices to coils, sheets, sheets to coils and helices and coils to helices and sheets (**B**) as a function of solvent accessibility for TS115 by SPIDER2.

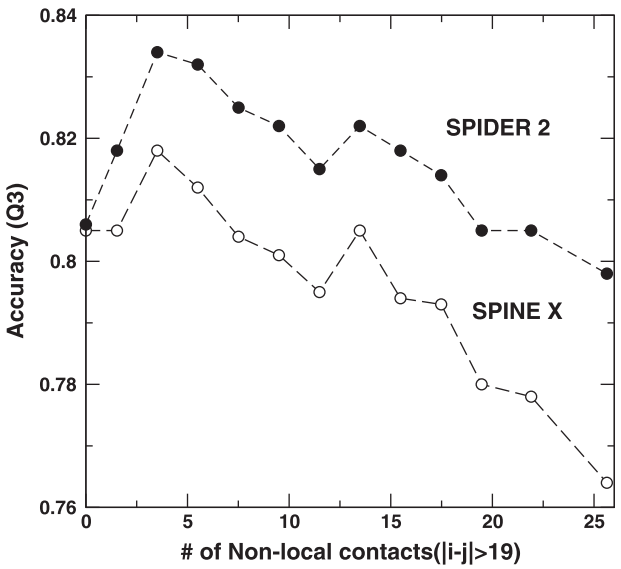

Figure 4. The dependence of accuracy on non-local contacts. The secondary structure accuracy as a function of the number of non-local contacts ($|i-j|>19$) for the independent test set (TS1199) by SPINE X and SPIDER2.

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 <8Å). 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

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

Data set Method	TS115		CASP12		Server location
	Q8	*P*-value[a]	Q8	*P*-value[a]	
SSPRO8	0.68	3E-9	0.69	0.014	http://scratch.proteomics.ics.uci.edu
DeepCNF	0.72	NA	0.73	NA	http://raptorx2.uchicago.edu/StructurePropertyPred/predict/

[a]Paired-t test from DeepCNF.

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%.

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.

However, the boundaries between different states are somewhat arbitrarily defined. Even in three states, different secondary assignment techniques can differ by as 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 (ω), Cα–N (ϕ) and Cα–C (ψ) bonds, respectively. ω 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 ϕ and ψ was made by Xue et al. in 2008 [124]. The mean absolute errors (MAEs) were 25° for ϕ and 38° for ψ, respectively, based on 10-fold cross-validation of 2640 high-resolution non-redundant proteins. The angle accuracy has been substantially improved over the years [94, 125, 126] with the highest accuracy reported by SPIDER2 (MAE = 19° for ϕ and 30° for ψ using an independent test of 1199 high-resolution non-redundant proteins) [57].

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 ϕ and ψ over SPINE X MAE values for TS115 and CASP12 sets (18° for ϕ and 28–29° for ψ) are slightly better than the reported MAE values (19° for ϕ and 30° for ψ 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 3_{10} helix, five for α-helix and undefined for sheet residues). It is desirable to have angles representing more than single residues. The peptide planar leads to the neighbouring 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 $C\alpha_{i-1} - C\alpha_i - C\alpha_{i+1}$ (θ) for three-residue coupling and a dihedral angle rotated about the $C\alpha_i - C\alpha_{i+1}$ bond (τ) 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 φ and ψ angles (single residue) and secondary structures (>3 residues). SPIDER [62] was the first method for sequence-based prediction of θ and τ and used a deep neural network. Its accuracy was subsequently improved by SPIDER2 with an iterative learning of multiple structural properties simultaneously [57]. The MAE values achieved by SPIDER2 are 8° for θ and 32° for τ using an independent test of 1199 high-resolution non-redundant proteins. As Table 4 shows, MAE values are 8° for θ and 31° for τ 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 φ and 19.4° for ψ based on MAE for homologous sequences of ≥30% sequence identities, compared with 19° for ϕ and 30° for ψ by SPIDER2. There is significant room for improving ψ, in particular. Similarly, we found that MAE values for homologous sequences of ≥30% sequence identities are 5° for θ and 15.5° for τ , compared with 8° for θ and 32° for τ by SPIDER2. Again, there is significant room for improving τ. 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 φ and ψ angles and 19.1% by predicted θ and τ 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 θ and τ angles. This is likely because these two angles involve—three to four

Table 4. Method comparison based on MAE using newly released structures (TS115) and CASP12 targets (15 proteins) for prediction of backbone angles (ϕ, ψ, θ and τ)

Data set Method	TS115		CASP12		Server location
	ϕ (°)	ψ (°)	ϕ (°)	ψ (°)	
SPINE-X	19.4	32.9	19.1	33.2	http://sparks-lab.org/SPINE-X/
SPIDER2	18.2	29.3	17.8	28.1	http://sparks-lab.org/server/SPIDER2/
	θ (°)	τ (°)	θ (°)	τ (°)	
SPIDER2	7.89	30.8	8.31	31.1	http://sparks-lab.org/server/SPIDER2/

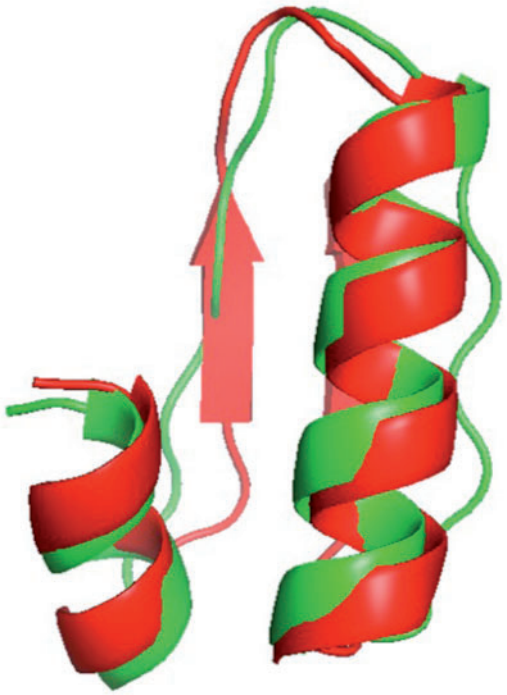

Figure 5. Direct prediction of three-dimensional structure by predicted angles. Structure (dark colour) constructed directly from φ/ψ angles compared with native structure (light colour) for residues 24–63 from PDB 5fdy chain A.

residues, compared with only one residue in the case of φ and ψ 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 ($\omega = 0°$), 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 Webservers are no longer active for further examination. There is a clear lack of development in ω, 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 φ, ψ and τ angles without expanding the training database. Moreover, 27% of 182 724 40-residue fragments built by predicted θ and τ 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 (QCIF).

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

1. Botstein D, Ashburner M, Ball CA, *et al*. Gene ontology: tool for the unification of biology, *Nat Genet* 2000;**25**:25–9.
2. Andreeva A, Howorth D, Chandonia JM, *et al*. Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res* 2008;**36**:D419–25.
3. Sillitoe I, Lewis TE, Cuff A, *et al*. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Res* 2015;**43**:D376–81.
4. Benson DA, Clark K, Karsch-Mizrachi I, *et al*. GenBank. *Nucleic Acids Res* 2015;**43**:D30–5.
5. Berman HM, Westbrook J, Feng Z, *et al*. The Protein Data Bank. *Nucleic Acids Res* 2000;**28**:235–42.
6. Terwilliger TC, Stuart D, Yokoyama S. Lessons from structural genomics. *Annu Rev Biophys* 2009;**38**:371–83.
7. Mardis ER. Anticipating the 1,000 dollar genome. *Genome Biol* 2006;**7**:112.
8. Zhou YQ, Duan Y, Yang YD, *et al*. Trends in template/fragment-free protein structure prediction. *Theor Chem Acc* 2011;**128**:3–16.
9. Tai CH, Bai H, Taylor TJ, *et al*. Assessment of template-free modeling in CASP10 and ROLL. *Proteins* 2014;**82(Suppl 2)**:57–83.
10. Murzin AG, Brenner SE, Hubbard T, *et al*. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 1995;**247**:536–40.
11. Dai L, Zhou Y. Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. *J Mol Biol* 2011;**408**:585–95.
12. Zhou Y, Karplus M. Interpreting the folding kinetics of helical proteins. *Nature* 1999;**401**:400–3.
13. Ozkan SB, Wu GA, Chodera JD, *et al*. Protein folding by zipping and assembly. *Proc Natl Acad Sci USA* 2007;**104**:11987–92.
14. Plaxco KW, Simons KT, Baker D. Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 1998;**277**:985–94.
15. Fischer D, Eisenberg D. Protein fold recognition using sequence-derived predictions. *Protein Sci* 1996;**5**:947–55.
16. Skolnick J, Kihara D, Zhang Y. Development and large scale benchmark testing of the PROSPECTOR_3 threading algorithm. *Proteins* 2004;**56**:502–18.
17. Rohl CA, Strauss CEM, Misura KMS, *et al*. Protein structure prediction using Rosetta. *Methods Enzymol* 2004;**383**:66–93.
18. Wu S, Skolnick J, Zhang Y. Ab initio modeling of small proteins by iterative TASSER simulations. *BMC Biol* 2007;**5**:17.
19. Ahmad S, Gromiha MM, Sarai A. Real value prediction of solvent accessibility from amino acid sequence. *Proteins* 2003;**50**:629–35.
20. Adamczak R, Porollo A, Meller J. Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins* 2004;**56**:753–67.
21. Heffernan R, Dehzangi A, Lyons J, *et al*. Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics* 2016;**32**: 843–9.

22. Schlessinger A, Rost B. Protein flexibility and rigidity predicted from sequence. *Proteins* 2005;**61**:115–26.
23. Radivojac P, Iakoucheva LM, Oldfield CJ, *et al*. Intrinsic disorder and functional proteomics. *Biophys J* 2007;**92**:1439–56.
24. Disfani FM, Hsu WL, Mizianty MJ, *et al*. MoRFpred, a computational tool for sequence-based prediction and characterization of short disorder-to-order transitioning binding regions in proteins. *Bioinformatics* 2012;**28**:i75–83.
25. Zhang T, Faraggi E, Li Z, *et al*. Intrinsically semi-disordered state and its role in induced folding and protein aggregation. *Cell Biochem Biophys* 2013;**67**:1193–205.
26. Zhou HY, Zhou YQ. SPEM: improving multiple sequence alignment with sequence profiles and predicted secondary structures. *Bioinformatics* 2005;**21**:3615–21.
27. Deng X, Cheng J. MSACompro: protein multiple sequence alignment using predicted secondary structure, solvent accessibility, and residue-residue contacts. *BMC Bioinformatics* 2011;**12**:472.
28. Godzik A, Jambon M, Friedberg I. Computational protein function prediction: Are we making progress? *Cell Mol Life Sci* 2007;**64**:2505–11.
29. Taherzadeh G, Zhou Y, Liew AW-C, *et al*. Sequence-based prediction of protein–carbohydrate binding sites using support vector machines. *J Chem Inf Model* 2016;**56**:2115–22.
30. Yue P, Li ZL, Moult J. Loss of protein structure stability as a major causative factor in monogenic disease. *J Mol Biol* 2005;**353**:459–73.
31. Khan S, Vihinen M. Spectrum of disease-causing mutations in protein secondary structures. *BMC Struct Biol* 2007;**7**:56.
32. Li B, Krishnan VG, Mort ME, *et al*. Automated inference of molecular mechanisms of disease from amino acid substitutions. *Bioinformatics* 2009;**25**:2744–50.
33. Zhao H, Yang Y, Lin H, *et al*. DDIG-in: Discriminating between disease-causing and neutral non-frameshifting micro-INDELs by support vector machines by means of integrated sequence- and structure-based features. *Genome Biol* 2013;**14**:R43.
34. Folkman L, Yang Y, Li Z, *et al*. DDIG-in: detecting disease-causing genetic variations due to frameshifting indels and nonsense mutations employing sequence and structural properties at nucleotide and protein levels. *Bioinformatics* 2015;**31**:1599–606.
35. Kendrew J, Bodo G, Dintzis H, *et al*. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 1958;**181**:662–6.
36. Pauling L, Corey RB, Branson HR. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc Natl Acad Sci USA* 1951;**37**:205–11.
37. Pauling L, Corey RB. Configurations of polypeptide chains with favored orientations around single bonds: two new pleated sheets. *Proc Natl Acad Sci USA* 1951;**37**:729–40.
38. Carugo O, Djinovic-Carugo K. Criteria to extract high-quality protein data bank subsets for structure users. *Methods Mol Biol* 2016;**1415**:139–52.
39. van Beusekom B, Perrakis A, Joosten RP. Data mining of macromolecular structures. *Methods Mol Biol* 2016;**1415**: 107–38.
40. Wang G, Dunbrack RL. Jr., PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res* 2005;**33**: W94–8.
41. Hafsa NE, Arndt D, Wishart DS. CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts. *Nucleic Acids Res* 2015;**43**:W370–7.
42. Li DW, Bruschweiler R. PPM_One: a static protein structure based chemical shift predictor. *J Biomol NMR* 2015;**62**:403–9.
43. Greenfield NJ. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* 2006;**1**:2876–90.
44. Micsonai A, Wien F, Kernya L, *et al*. Accurate secondary structure prediction and fold recognition for circular dichroism spectroscopy. *Proc Natl Acad Sci USA* 2015;**112**:E3095–103.
45. Dong A, Huang P, Caughey WS. Protein secondary structures in water from second-derivative amide I infrared spectra. *Biochemistry* 1990;**29**:3303–8.
46. Yang HY, Yang SN, Kong JL, *et al*. Obtaining information about protein secondary structures in aqueous solution using Fourier transform IR spectroscopy. *Nat Protoc* 2015; **10**:382–96.
47. Rost B. Review: protein secondary structure prediction continues to rise. *J Struct Biol* 2001;**134**:204–18.
48. Simossis VA, Heringa J. Integrating protein secondary structure prediction and multiple sequence alignment. *Curr Protein Pept Sci* 2004;**5**:249–66.
49. Heringa J. Computational methods for protein secondary structure prediction using multiple sequence alignments. *Curr Protein Pept Sci* 2000;**1**:273–301.
50. Yoo PD, Zhou BB, Zomaya AY. Machine learning techniques for protein secondary structure prediction: an overview and evaluation. *Curr Bioinform* 2008;**3**:74–86.
51. Zhou Y, Faraggi E, Prediction of one-dimensional structural properties of proteins by integrated neural network. In: H Rangwala, G Karypis (eds). *Protein Structure Prediction: Method and Algorithms*. Hoboken, NJ: Wiley, 2010, 44–74.
52. Pirovano W, Heringa J. Protein secondary structure prediction. *Methods Mol Biol* 2010;**609**:327–48.
53. Kabsch W, Sander C. Dictionary of protein structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983;**22**:2577–637.
54. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA* 1993;**90**:7558–62.
55. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. *J Mol Biol* 1999;**292**: 195–202.
56. Dor O, Zhou Y. Achieving 80% ten-fold cross-validated accuracy for secondary structure prediction by large-scale training. *Proteins* 2007;**66**:838–45.
57. Heffernan R, Paliwal K, Lyons J, *et al*. Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Sci Rep* 2015;**5**:11476.
58. Wang S, Peng J, Ma JZ, *et al*. Protein secondary structure prediction using deep convolutional neural fields. *Sci Rep* 2016;**6**:18962.
59. Rost B, Sander C, Schneider R. Redefining the goals of protein secondary structure prediction. *J Mol Biol* 1994;**235**: 13–26.
60. Zhang W, Dunker AK, Zhou YQ. Assessing secondary structure assignment of protein structures by using pairwise sequence-alignment benchmarks. *Proteins* 2008;**71**:61–7.
61. Altschul SF, Madden TL, Schaffer AA, *et al*. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997;**25**:3389–402.
62. Lyons J, Dehzangi A, Heffernan R, *et al*. Predicting backbone Calpha angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *J Comput Chem* 2014;**35**:2040–6.

63. Scheraga HA. Structural studies of ribonuclease.3. A model for the secondary and tertiary structure. *J Am Chem Soc* 1960;**82**:3847–52.
64. Finkelstein AV, Ptitsyn OB. Statistical analysis of correlation among amino acid residues in helical, beta-structural and non-regular regions of globular proteins. *J Mol Biol* 1971; **62**:613–24.
65. Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry* 1974;**13**:222–45.
66. Garnier J, Osguthorpe DJ, Robson B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 1978; **120**:97–120.
67. Lim VI. Algorithms for prediction of alpha-helical and beta-structural regions in globular proteins. *J Mol Biol* 1974; **88**:873–94.
68. Kabat EA, Wu TT. The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: comparison of predicted and experimental determination of -sheets in concanavalin A. *Proc Natl Acad Sci USA* 1973;**70**:1473–7.
69. Arnold GE, Dunker AK, Johns SJ, *et al*. Use of conditional probabilities for determining relationships between amino-acid-sequence and protein secondary structure. *Proteins* 1992;**12**:382–99.
70. Mitchell EM, Artymiuk PJ, Rice DW, *et al*. use of techniques derived from graph-theory to compare secondary structure motifs in proteins. *J Mol Biol* 1990;**212**:151–66.
71. Holley LH, Karplus M. Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA* 1989;**86**: 152–6.
72. Bohr H, Bohr J, Brunak S, *et al*. Protein secondary structure and homology by neural networks. The alpha-helices in rhodopsin. *Febs Lett* 1988;**241**:223–8.
73. Muggleton S, King RD, Sternberg MJE. Protein secondary structure prediction using logic-based machine learning. *Protein Eng* 1992;**5**:647–57.
74. Yi TM, Lander ES. Protein secondary structure prediction using nearest-neighbor methods. *J Mol Biol* 1993;**232**: 1117–29.
75. Zvelebil MJ, Barton GJ, Taylor WR, *et al*. Prediction of protein secondary structure and active-sites using the alignment of homologous sequences. *J Mol Biol* 1987;**195**: 957–61.
76. Hua S, Sun Z. A novel method of protein secondary structure prediction with high segment overlap measure: support vector machine approach. *J Mol Biol* 2001;**308**:397–407.
77. Ward JJ, McGuffin LJ, Buxton BF, *et al*. Secondary structure prediction with support vector machines. *Bioinformatics* 2003; **19**:1650–5.
78. Aydin Z, Altunbasak Y, Borodovsky M. Protein secondary structure prediction for a single-sequence using hidden semi-Markov models. *BMC Bioinformatics* 2006;**7**:178.
79. Yao XQ, Zhu H, She ZS. A dynamic Bayesian network approach to protein secondary structure prediction. *BMC Bioinformatics* 2008;**9**:49.
80. Liu Y, Carbonell J, Klein-Seetharaman J, *et al*. Comparison of probabilistic combination methods for protein secondary structure prediction. *Bioinformatics* 2004;**20**:3099–107.
81. Maxfield FR, Scheraga HA. Status of empirical methods for the prediction of protein backbone topography. *Biochemistry* 1976;**15**:5138–53.
82. Cuff JA, Barton GJ. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* 2000;**40**:502–11.
83. Lin HN, Chang JM, Wu KP, *et al*. HYPROSP II—a knowledge-based hybrid method for protein secondary structure prediction based on local prediction confidence. *Bioinformatics* 2005;**21**:3227–33.
84. Montgomerie S, Sundararaj S, Gallin WJ, *et al*. Improving the accuracy of protein secondary structure prediction using structural alignment. *BMC Bioinformatics* 2006;**7**:301.
85. Bondugula R, Xu D. MUPRED: A tool for bridging the gap between template based methods and sequence profile based methods for protein secondary structure prediction. *Proteins* 2007;**66**:664–70.
86. Pollastri G, Martin AJM, Mooney C, *et al*. Accurate prediction of protein secondary structure and solvent accessibility by consensus combiners of sequence and structure information. *BMC Bioinformatics* 2007;**8**:1–12.
87. Cheng HT, Sen TZ, Jernigan RL, *et al*. Consensus data mining (CDM) protein secondary structure prediction server: Combining GOR v and fragment database mining (FDM). *Bioinformatics* 2007;**23**:2628–30.
88. Li DP, Li TH, Cong PS, *et al*. A novel structural position-specific scoring matrix for the prediction of protein secondary structures. *Bioinformatics* 2012;**28**:32–9.
89. Saraswathi S, Fernandez-Martinez JL, Kolinski A, *et al*. Fast learning optimized prediction methodology (FLOPRED) for protein secondary structure prediction. *J Mol Model* 2012;**18**:4275–89.
90. Magnan CN, Baldi P. SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics* 2014;**30**:2592–7.
91. Buchan DWA, Ward SM, Lobley AE, *et al*. Protein annotation and modelling servers at University College London. *Nucleic Acids Res* 2010;**38**:W563–8.
92. Cole C, Barber JD, Barton GJ. The Jpred 3 secondary structure prediction server. *Nucleic Acids Res* 2008;**36**: W197–201.
93. Drozdetskiy A, Cole C, Procter J, *et al*. JPred4: a protein secondary structure prediction server. *Nucleic Acids Res* 2015;**43**:W389–94.
94. Faraggi E, Zhang T, Yang Y, *et al*. SPINE X: Improving protein secondary structure prediction by multi-step learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J Comput Chem* 2011;**33**: 259–63.
95. Mirabello C, Pollastri G. Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics* 2013;**29**:2056–8.
96. Yaseen A, Li YH. Context-based features enhance protein secondary structure prediction accuracy. *J Chem Inf Model* 2014;**54**:992–1002.
97. Hinton GE. Learning multiple layers of representation. *Trends Cogn Sci* 2007;**11**:428–34.
98. Qi YJ, Oja M, Weston J, *et al*. A unified multitask architecture for predicting local protein properties. *PLoS One* 2012;**7**: e32235.
99. Spencer M, Eickholt J, Cheng JL. A deep learning network approach to ab initio protein secondary structure prediction. *IEEE ACM Trans Comput Biol Bioinform* 2015;**12**:103–12.
100. Lee H, Grosse R, Ranganath R, *et al*. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In: *Proceedings of the 26th International Conference on Machine Learning*. Montreal, Canada, 2009.
101. Lafferty J, Mccallum A, Pereira F, Conditional random fields: probabilistic models for segmenting and labeling sequence

data. In: *18th International Conference on Machine Learning*. Morgan Kaufmann, San Francisco, CA, 2001, p. 282–9.
102. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006;**22**:1658–9.
103. Wilson CL, Boardman PE, Doig AJ, *et al*. Improved prediction for N-termini of alpha-helices using empirical information. *Proteins* 2004;**57**:322–30.
104. Midic U, Dunker AK, Obradovic Z. Improving protein secondary-structure prediction by predicting ends of secondary-structure segments. In: *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2005, pp. 490–97 (IEEE, Niagara Falls, Canada).
105. Mezei M. Chameleon sequences in the PDB. *Protein Eng* 1998;**11**:411–4.
106. Li W, Kinch LN, Karplus PA, *et al*. ChSeq: a database of chameleon sequences. *Protein Sci* 2015;**24**:1075–86.
107. Bahramali G, Goliaei B, Minuchehr Z, *et al*. Chameleon sequences in neurodegenerative diseases. *Biochem Biophys Res Commun* 2016;**472**:209–16.
108. Jacoboni I, Martelli PL, Fariselli P, *et al*. Predictions of protein segments with the same aminoacid sequence and different secondary structure: a benchmark for predictive methods. *Proteins* 2000;**41**:535–44.
109. Guo JT, Jaromczyk JW, Xu Y. Analysis of chameleon sequences and their implications in biological processes. *Proteins* 2007;**67**:548–58.
110. Ghozlane A, Joseph AP, Bornot A, *et al*. Analysis of protein chameleon sequence characteristics. *Bioinformation* 2009;**3**: 367–9.
111. Dill KA, Bromberg S, Yue K, *et al*. Principles of protein folding–a perspective from simple exact models. *Protein Sci* 1995;**4**:561–602.
112. Rees DC, Eisenberg D. Turning a reference inside-out: commentary on an article by Stevens and Arkin entitled: "Are membrane proteins 'inside-out' proteins?" (Proteins 1999;36:135-143). *Proteins* 2000;**38**:121–2.
113. Stevens TJ, Arkin IT. Are membrane proteins "inside-out" proteins? *Proteins* 1999;**36**:135–43.
114. Crooks GE, Brenner SE. Protein secondary structure: entropy, correlations and prediction. *Bioinformatics* 2004;**20**:1603–11.
115. Kihara D. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci* 2005;**14**:1955–63.
116. Ceroni A, Frasconi P. On the role of long-range dependencies in learning protein secondary structure. In: *2004 IEEE International Joint Conference on Neural Networks, Vols 1–4, Proceedings* 2004, p. 1899–1904 (IEEE, Budapest).
117. Ceroni A, Frasconi P, Pollastri G. Learning protein secondary structure from sequential and relational data. *Neural Netw* 2005;**18**:1029–39.
118. Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci USA* 2003;**100**: 12105–10.
119. Kinch L, Yong Shi S, Cong Q, *et al*. CASP9 assessment of free modeling target predictions. *Proteins* 2011;**79(Suppl 10)**: 59–73.
120. DeBartolo J, Colubri A, Jha AK, *et al*. Mimicking the folding pathway to improve homology-free protein structure prediction. *Proc Natl Acad Sci USA* 2009;**106**:3734–9.
121. Pollastri G, Przybylski D, Rost B, *et al*. Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins* 2002;**47**:228–35.
122. Wang Z, Zhao F, Peng J, *et al*. Protein 8-class secondary structure prediction using conditional neural fields. *Proteomics* 2011;**11**:3786–92.
123. Yaseen A, Li YH. Template-based C8-SCORPION: a protein 8-state secondary structure prediction method using structural information and context-based features. *BMC Bioinformatics* 2014;**15(Suppl 8)**:1–8.
124. Xue B, Dor O, Faraggi E, *et al*. Real-value prediction of backbone torsion angles. *Proteins* 2008;**72**:427–33.
125. Faraggi E, Xue B, Zhou Y. Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins* 2009;**74**: 847–56.
126. Singh H, Singh S, Raghava GP. Evaluation of protein dihedral angle prediction methods. *PLoS One* 2014;**9**:e105667.
127. Korkut A, Hendrickson WA. A force field for virtual atom molecular mechanics of proteins. *Proc Natl Acad Sci USA* 2009;**106**:15667–72.
128. Zhou Y, Karplus M. Folding thermodynamics of a model three-helix-bundle protein. *Proc Natl Acad Sci USA* 1997; **94**:14429–32.
129. Kihara D, Lu H, Kolinski A, *et al*. TOUCHSTONE: an ab initio protein structure prediction method that uses threading-based tertiary restraints. *Proc Natl Acad Sci USA* 2001;**98**: 10125–30.
130. Liwo A, He Y, Scheraga HA. Coarse-grained force field: general folding theory. *Phys Chem Chem Phys* 2011;**13**:16890–901.
131. Flocco MM, Mowbray SL. C alpha-based torsion angles: a simple tool to analyze protein conformational changes. *Protein Sci* 1995;**4**:2118–22.
132. Kleywegt GJ. Validation of protein models from Calpha coordinates alone. *J Mol Biol* 1997;**273**:371–6.
133. Reva BA, Finkelstein AV, Skolnick J. What is the probability of a chance prediction of a protein structure with an RMSD of 6 angstrom? *Fold Des* 1998;**3**:141–7.
134. Gao J, Yang Y, Zhou Y. Predicting the errors of predicted local backbone angles and non-local solvent-accessibilities of proteins by deep neural networks. *Bioinformatics* 2016, doi: 10.1093/bioinformatics/btw549.
135. Touw WG, Joosten RP, Vriend G. Detection of trans-cis flips and peptide-plane flips in protein structures. *Acta Crystallogr D Struct Biol* 2015;**71**:1604–14.
136. Frommel C, Preissner R. Prediction of prolyl residues in Cis-conformation in protein structures on the basis of the amino-acid-sequence. *FEBS Lett* 1990;**277**:159–63.
137. Song JN, Burrage K, Yuan Z, *et al*. Prediction of cis/trans isomerization in proteins using PSI-BLAST profiles and secondary structure information. *BMC Bioinformatics* 2006;**7**: 124.1-124.13.
138. Exarchos KP, Exarchos TP, Papaloukas C, *et al*. Detection of discriminative sequence patterns in the neighborhood of proline cis peptide bonds and their functional annotation. *BMC Bioinformatics* 2009;**10**(1):14.
139. Exarchos KP, Papaloukas C, Exarchos TP, *et al*. Prediction of cis/trans isomerization using feature selection and support vector machines. *J Biomed Inf* 2009;**42**:140–9.
140. Exarchos KP, Exarchos TP, Papaloukas C, *et al*. PBOND: web server for the prediction of proline and non-proline cis/trans isomerization. *Genomics Proteomics Bioinformatics* 2009;**7**: 138–42.
141. Pal D, Chakrabarti P. Cis peptide bonds in proteins: Residues involved, their conformations, interactions and locations. *J Mol Biol* 1999;**294**:271–88.

142. Dugave C, Demange L. Cis-trans isomerization of organic molecules and biomolecules: Implications and applications. *Chem Rev* 2003;**103**:2475–532.
143. Acharya KR, Lloyd MD. The advantages and limitations of protein crystal structures. *Trends Pharmacol Sci* 2005; **26**:10–4.
144. Sormanni P, Camilloni C, Fariselli P, *et al*. The s2D method: simultaneous sequence-based prediction of the statistical populations of ordered and disordered regions in proteins. *J Mol Biol* 2015;**427**:982–96.
145. Abriata LA. Structural database resources for biological macromolecules. *Brief Bioinform* 2016, in press. [Epub ahead of print]
146. Gobel U, Sander C, Schneider R, *et al*. Correlated mutations and residue contacts in proteins. *Proteins* 1994;**18**: 309–17.
147. Marks DS, Colwell LJ, Sheridan R, *et al*. Protein 3D structure computed from evolutionary sequence variation. *PLoS One* 2011;**6**:e28766.
148. Hopf TA, Colwell LJ, Sheridan R, *et al*. Three-dimensional structures of membrane proteins from genomic sequencing. *Cell* 2012;**149**:1607–21.
149. Morcos F, Pagnini A, Lunt B, *et al*. Estimation of residue-residue coevolution using direct coupling analysis identifies many native contacts across a large number of domain families. *Biophysical Journal* 2012;**102**:250A.
150. Jacob E, Unger R, Horovitz A. Codon-level information improves predictions of inter-residue contacts in proteins by correlated mutation analysis. *Elife* 2015;**4**: e08932.
151. Xie J, Ding W, Chen L, *et al*. Advances in protein contact map prediction based on machine learning. *Med Chem* 2015;**11**: 265–70.
152. Wuyun Q, Zheng W, Peng Z, *et al*. A large-scale comparative assessment of methods for residue–residue contact prediction. *Brief Bioinform* 2016, doi: https://doi.org/10.1093/bib/bbw106.
153. Monastyrskyy B, D'Andrea D, Fidelis K, *et al*. New encouraging developments in contact prediction: assessment of the CASP11 results. *Proteins* 2016;**84(Suppl 1)**:131–44.
154. Chu W, Ghahramani Z, Podtelezhnikov A, *et al*. Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2006;**3**:98–113.
155. Toth-Petroczy A, Palmedo P, Ingraham J, *et al*. Structured states of disordered proteins from genomic sequences. *Cell* 2016;**167**:158–70.e112.
156. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.
157. Schuster M, Paliwal KK. Bidirectional recurrent neural networks. *IEEE Trans Signal Process* 1997;**45**:2673–81.
158. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005;**18**:602–10.
159. Vinyals O, Toshev A, Bengio S, *et al*. Show and tell: a neural image caption generator. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, p. 3156–64 (IEEE, Boston, Massachusetts).
160. Hanson J, Yang Y, Paliwal K, *et al*. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics* 2016, in press.