Bioinformatics, 35(14), 2019, 2403–2410
doi: 10.1093/bioinformatics/bty1006
Advance Access Publication Date: 7 December 2018
Original Paper

Structural bioinformatics

Improving prediction of protein secondary structure, backbone angles, solvent accessibility and contact numbers by using predicted contact maps and an ensemble of recurrent and residual convolutional neural networks

Jack Hanson [1,*], Kuldip Paliwal[1], Thomas Litfin[2], Yuedong Yang [3] and Yaoqi Zhou [2,4,*]

[1]Signal Processing Laboratory, Griffith University, Brisbane, QLD 4122, Australia, [2]School of Information and Communication Technology, Griffith University, Gold Coast, QLD 4215, Australia, [3]School of Data and Computer Science, Sun-Yat Sen University, Guangzhou, Guangdong 510006, China and [4]Institute for Glycomics, Griffith University, Gold Coast, QLD 4215, Australia

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on August 13, 2018; revised on November 2, 2018; editorial decision on November 29, 2018; accepted on December 6, 2018

Abstract

Motivation: Sequence-based prediction of one dimensional structural properties of proteins has been a long-standing subproblem of protein structure prediction. Recently, prediction accuracy has been significantly improved due to the rapid expansion of protein sequence and structure libraries and advances in deep learning techniques, such as residual convolutional networks (ResNets) and Long-Short-Term Memory Cells in Bidirectional Recurrent Neural Networks (LSTM-BRNNs). Here we leverage an ensemble of LSTM-BRNN and ResNet models, together with predicted residue-residue contact maps, to continue the push towards the attainable limit of prediction for 3- and 8-state secondary structure, backbone angles (θ, τ, ϕ and ψ), half-sphere exposure, contact numbers and solvent accessible surface area (ASA).

Results: The new method, named SPOT-1D, achieves similar, high performance on a large validation set and test set ($\approx$1000 proteins in each set), suggesting robust performance for unseen data. For the large test set, it achieves 87% and 77% in 3- and 8-state secondary structure prediction and 0.82 and 0.86 in correlation coefficients between predicted and measured ASA and contact numbers, respectively. Comparison to current state-of-the-art techniques reveals substantial improvement in secondary structure and backbone angle prediction. In particular, 44% of 40-residue fragment structures constructed from predicted backbone Cα-based θ and τ angles are less than 6 Å root-mean-squared-distance from their native conformations, nearly 20% better than the next best. The method is expected to be useful for advancing protein structure and function prediction.

Availability and implementation: SPOT-1D and its data is available at: http://sparks-lab.org/.

Contact: jack.hanson@griffithuni.edu.au or yaoqi.zhou@griffith.edu.au

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Deriving a protein's structure from its sequence alone remains an unsolved problem since its inception over 50 years ago (Gibson and Scheraga, 1967; Yang *et al.*, 2018). The main challenge stems from the exorbitantly large conformational space of a protein chain and the lack of an accurate energy function to model the folding process (Zhou *et al.*, 2011). As a result, it is necessary to address simpler problems in both the prediction of one-dimensional structural properties, such as backbone secondary structure and sidechain solvent accessibility, and two-dimensional structural properties, such as residue-residue contact maps.

Backbone secondary structure was first described by Pauling *et al.* (1951) in their findings of helical and sheet hydrogen bonding patterns in a protein backbone. This has been refined into either 3 or 8 local conformational states (Kabsch and Sander, 1983). The accuracy of secondary structure prediction has risen from 70% (Rost and Sander, 1993) to the latest 85% (Fang *et al.*, 2018a), approaching the theoretical upper bounds of effective prediction accuracy of 88–90% (Rost, 2001; Yang *et al.*, 2018).

Recognizing secondary structure as a coarse-grained description of protein backbone, more recent efforts have been shifted to the prediction of continuously valued backbone torsion angles. Backbone angles ϕ and ψ are measurements of the residue-wise torsion (Ramachandran *et al.*, 1963), whereas angles θ and τ are spread over 3 (dihedral about $C\alpha_{i-1}$-$C\alpha_i$-$C\alpha_{i+1}$) and 4 (torsion about the $C\alpha_i$-$C\alpha_{i+1}$ bond) residues, respectively (Korkut and Hendrickson, 2009). These angles all form a complementary basis for local backbone structure, and have been predicted as both discrete states (Kang *et al.*, 1993) and continuous values (Faraggi *et al.*, 2012; Heffernan *et al.*, 2015, 2017; Lyons *et al.*, 2014; Xue *et al.*, 2008).

In addition to local backbone structural properties, global three-dimensional structures of proteins can be characterized by the residue solvent accessibility. The distinction between buried and exposed (i.e. low and high solvent accessibility, respectively) residues is important as active sites are typically located on the surface of a protein. The solvent-Accessible Surface Area (ASA) is one such descriptor which measures the exposure of a residue to solvent (water) in its folded state. Another such metric is the Contact Number (CN) of a residue in a protein, which is the count of spatially close residues within a distance cutoff to a target residue. The Half-Sphere Exposure (HSE) adds directionality to this measurement by splitting the spherical distance cutoff into two halves (Hamelryck, 2005). These descriptors have also been predicted into discrete states and as continuous values (Heffernan *et al.*, 2016; Lee and Richards, 1971; Rost and Sander, 1994).

Recent predictors of these one-dimensional structural descriptors have been dominated by the advances in deep learning. For example, NetSurfP-2.0 used a large Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) network in a Bidirectional Recurrent Neural Network (BRNN) (Schuster and Paliwal, 1997) to predict for 3- and 8-state secondary structure, ASA and backbone angles ϕ and ψ (Klausen *et al.*, 2018). MUFOLD-SS and MUFOLD-Angle utilized variants of Inception networks (Szegedy *et al.*, 2017) to predict for 3- and 8-state secondary structure, and backbone angles ϕ and ψ, respectively (Fang *et al.*, 2018a,b). Porter5 employed an ensemble of BRNN's to predict 3- and 8-state secondary structure (Torrisi *et al.*, 2018). PSRSM utilizes a large ensemble of Support Vector Machines (SVM's) (Vapnik, 1998) trained on various training objectives using protein length-based partitioning and random subspacing of their training data (Ma *et al.*, 2018). Deep conditional neural fields have been employed to predict for 3- and 8-state secondary structure in DeepCNF (Wang *et al.*, 2016a). RaptorX-Angle predicts real-valued ϕ and ψ angles by combining k-means clustering and deep residual convolutional neural networks (ResNets) (Gao *et al.*, 2018; He *et al.*, 2016). Our own work SPIDER3 utilized an iterative application of an LSTM-BRNN to predict 3-state secondary structure, ASA, HSE and four backbone torsion angles θ, τ, ϕ and ψ (Heffernan *et al.*, 2017, 2018). Current improvement in method performance is due to larger databases of protein sequences and structures and the enhanced ability of current methods to capture non-local interactions between residues which are structural but not sequential neighbors (Heffernan *et al.*, 2017). The latter was made possible by whole-sequence learning (without using sliding windows) of LSTM-BRNN's and ultra-deep ResNets.

All of the above new deep learning methods for predicting 1D structural properties have relied on the separate application of either vanilla or LSTM BRNN's, Inception networks, or deep ResNets. Different types of NN have different capability in capturing local and/or nonlocal interactions. Indeed, we have shown that using an ensemble of models based on LSTM-BRNN, ResNet and Fully-Connected (FC) NN allows a significant improvement in the prediction of residue-residue contact map (Hanson *et al.*, 2018), and proline and non-proline *cis*-isomers (Singh *et al.*, 2018) by the integrated learning of local and nonlocal interactions. Moreover, previous studies have shown the improvement in secondary structure prediction resulting from the input of native contact maps (Ceroni and Frasconi, 2004; Ceroni *et al.*, 2005) and by the inference of a contact map through beta sheet pairing (Chu *et al.*, 2006).

Thus, given recent advancements in contact map prediction through coupling correlated mutations and deep learning (Adhikari *et al.*, 2017; Hanson *et al.*, 2018; Wang *et al.*, 2017), it may be profitable to employ an ensemble of different machine learning models and utilize predicted contact maps as input for secondary structure and solvent accessibility prediction. In this paper, the above idea was utilized to develop a method called SPOT-1D by using 9 LSTM-BRNN- and ResNet-based models for the prediction of multiple one-dimensional structure properties (namely 3- and 8-state secondary structure, backbone torsion angles θ, τ, ϕ, ψ and solvent accessibility descriptors ASA, HSEα-up and -down and CN). We demonstrate that the new method achieves significant improvement in backbone structure in terms of secondary structures and torsion angles and leads to near 20% improvement in three-dimensional fragment-structure models constructed from predicted angles, in addition to achieving high performance for the accuracy of predicting contact numbers and ASA.

2 Materials and methods

2.1 Neural network

The model utilized in SPOT-1D follows the methodology employed in our previous papers for contact map and *cis*-isomer prediction (Hanson *et al.*, 2018; Singh *et al.*, 2018). In brief, we utilize an ensemble of LSTM-BRNN and ResNet hybrid models to identify and propagate short- and long-term dependencies throughout the sequence. The ensemble consists of nine models with varying hyper-parameters which provides a set of diverse learning paths. A full overview of the ensemble and the individual models is presented in Supplementary Section S1.

2.2 Input features

Our input features consisted of two evolutionary profiles from three iterations of PSI-BLAST (Altschul *et al.*, 1997) with default parameters and HHBlits (Remmert *et al.*, 2012) with default parameters, respectively, physicochemical properties of each amino acid and the predicted contact map information from SPOT-Contact (Hanson *et al.*, 2018). Each protein's Position Specific Scoring Matrix (PSSM) was generated by three iterations of PSI-BLAST (Altschul *et al.*, 1997) against the UniRef90 sequence database updated in April 2018, generating 20 substitution probabilities per sequence residue. The HMM profile was generated by HHblits v3.0.3 with the Uniprot sequence profile database from October 2017 (Mirdita *et al.*, 2017), and provides 20 residue substitution values along with 10 transition frequency and the number of effective homologous sequences. Seven physicochemical properties of each amino acid, such as Van der Waal's volume and polarizability, are obtained from Meiler *et al.* (2001). Thus, we have 57 base features as the input to our base-feature model.

The full-feature model contains additional features obtained by windowing the predicted contact information over the target residue's preceding and succeeding W_n pairwise contact predictions obtained from SPOT-Contact (Hanson *et al.*, 2018). When the window goes outside the contact map (i.e. $0 \leq i < W_n$ or $L - W_n \leq i < L$), the value for the undefined positions is set to 0. As SPOT-Contact only predicts contacts for residues that are greater than or equal to 3 in sequence position separation (i.e. $|i - j| \geq 3$), the feature size of the contact map is $2 \times W_n - 4$ for each residue. The windowed partition of the predicted contact map is shown in Supplementary Figure S1.

Thus, the full-feature model contains an additional $2 \times W_n - 4$ contact-map-based features. All features are standardized to have zero mean and unit variance at the input of the model according to the means and standard deviations of the training data. The window size was tuned as a hyperparameter in each of the models trained (Supplementary Table S1).

2.3 Outputs

For the classification model, we have eleven predicted outputs, providing independent prediction for both the 8-state and 3-state secondary structural elements. The 8-state labels and their one-letter representation as defined by the Dictionary of Secondary Structure of Proteins (DSSP) are: 3_{10}-helix G, α-helix H, π-helix I, β-bridge B, β-strand E, high-curvature loop S, β-turn T and coil C (Kabsch and Sander, 1983). This can be condensed to the 3-state labels of strand E (8 state B and E), helix H (8 state G, H and I) and coil (everything else). As shown in (Heffernan *et al.*, 2018), predicting these conformations independently provides superior accuracies than when SS3 is inferred from SS8.

The 12 regression outputs correspond to the ASA, HSEα-up and -down, CN, $\sin(\theta)$, $\cos(\theta)$, $\sin(\tau)$, $\cos(\tau)$, $\sin(\phi)$, $\cos(\phi)$, $\sin(\psi)$ and $\cos(\psi)$. The ASA is predicted as relative ASA (rASA) so that the prediction is not biased by larger nor smaller residues, and converted to absolute ASA at the output. We define the distance cutoff for CN and the HSEα metrics as 13 Å. θ is defined as the angle between three successive Cα atoms (Cα_{i-1}-Cα_i-Cα_{i+1}) and τ is defined as the torsion angle about the Cα_i-Cα_{i-1} bond. Both of the sin and cosine of the backbone angles are predicted to account for angle periodicity (Lyons *et al.*, 2014). The predicted angle is recovered at the output of the network by $\alpha = \tan^{-1}\left[\frac{\sin(\alpha)}{\cos(\alpha)}\right]$. The DSSP software was used to generate each protein's SS3, SS8, ASA and ϕ and ψ angles from their PDB files. The remaining structural properties were generated with an in-house program.

2.4 Datasets

The data used in these experiments is the same as our previous works (Hanson *et al.*, 2018; Singh *et al.*, 2018). In summary, we culled 12 450 proteins from the PISCES server (Wang and Dunbrack, 2003) on Feb 2017 with the constraints of high resolution (<2.5Å), an R-free <1 and a sequence identity cutoff of 25% according to BlastClust (Altschul *et al.*, 1997). 1250 proteins deposited after June 2015 were separated into a test set, leaving 11200 proteins which were randomly divided into a train set (10 200 proteins) and validation set (1000). As in SPOT-Contact we removed the proteins over 700 residues in the above training, validation and test sets for efficient calculations. This reduces our training, validation and independent test sets to 10 029, 983 and 1213 proteins, respectively. We should emphasize that SPOT-Contact employed the same validation and test sets but with a smaller training set of 7557 proteins, which is a subset of the 10 029 training proteins after removing the proteins with >300 residues for efficient training of SPOT-Contact. The same validation and test sets for SPOT-1D and SPOT-Contact minimize the possibility of over training with SPOT-Contact as input for SPOT-1D.

To facilitate a fair comparison to other methods, we further obtained structures from the PDB released between 01/01/2018 and 07/16/2018 and solved with resolution < 2.5Å and R-free < 0.25 to form a new independent test set. In order to minimize evaluation bias associated with partially overlapping training data, we removed proteins with >25% sequence identity to structures released prior to 2018. The dataset was also filtered to remove redundancy at a 25% sequence identity cutoff. Finally, 13 proteins were removed with length > 700 due to the limitations of some external predictors (MUFOLD and PSRSM), leaving 250 high-quality, non-redundant targets. For convenience, we denote two independent test sets as TEST2016 (1213 proteins) and TEST2018 (250 proteins) as they were deposited between June 2015 and Feb 2017 and between Jan 2018 and July 2018, respectively. For a measure of our predictor on a harder set, we employ the TEST-HARD set from Hanson *et al.* (2018), a subset of TEST2016 after removing proteins with a Blast E-value of <0.1 against our Train set. In addition, we utilize 20 available Template-Free Modelling (TFM) proteins from CASP12 (Schaarschmidt *et al.*, 2018) for independent testing of the available methods (CASP12). This set has < 25% sequence similarity to our training set.

2.5 Performance evaluation

We treat the outputs for the classification model as class probabilities due to the outputs being squeezed into two independent probability distributions for 3- and 8-state prediction by the softmax function. Thus, whichever node provides the highest probability in each distribution is selected as the corresponding label for that class. Hence, we can analyze the performance of the classification model by its accuracy in designating the correct class label for both 3-state (Q3) and 8-state (Q8) prediction. We also analyze the Segment Overlap Value (SOV) for secondary structure predictions as proposed by Zemla *et al.* (1999). We utilize the paired t-test to obtain a significance value *P* for comparison of protein-wise accuracies.

The regression models are analyzed in two ways. The Pearson Correlation Coefficient (CC) is used to analyze the regression model's performance for HSEα-up and -down, ASA and CN, whereas the Mean Absolute Error (MAE) is used for measuring the predicted angle and true angle discrepancies. The angle error is taken as the explementary of the error if the error is greater than 180° due to the periodicity of angles.

2.6 Method comparison

We compare against several secondary structure and solvent accessibility prediction methods recently released in the literature. We downloaded the standalone versions of MUFOLD-SS and MUFOLD-Angle (available at http://dslsrv8.cs.missouri.edu/ cf797/ MUFoldAngle/and http://dslsrv8.cs.missouri.edu/ cf797/MUFoldSS, respectively) (Fang *et al.*, 2018a,b), RaptorX-Angles (available at https://github.com/lacus2009/RaptorX-Angle) (Gao *et al.*, 2018) and SPIDER3 and SPIDER3-Single (available: http://sparks-lab.org/ yueyang/download/index.php) (Heffernan *et al.*, 2017, 2018). We utilized the online server for DeepCNF (Server URL: http://raptorx2.uchicago.edu/StructurePropertyPred/predict/) (Wang *et al.*, 2016a), NetSurfP-2.0 (Server URL: http://www.cbs.dtu.dk/services/ NetSurfP-2.0/), PSRSM (Server URL: http://210.44.144.20: 82/protein_PSRSM/default.aspx) (Ma *et al.*, 2018) and Porter5 (Server URL: http://distilldeep.ucd.ie/porter/) (Torrisi *et al.*, 2018). For brevity in the results, MUFOLD-SS and MUFOLD-Angle will both be referred to as MUFOLD, and DeepCNF and RaptorX-Angle will both be referred to as RaptorX.

3 Results

Table 1 shows the results for the validation (983 proteins) and TEST2016 sets by the final ensemble model SPOT-1D. Similar performance across both test and validation sets are observed for all ten predicted variables, indicating the robustness of the model trained for unseen data. For example, predicted 3-state secondary structures have an overall accuracy of 87.5% for the validation set, 87.2% for TEST2016 and 86% for TEST-HARD. The solvent accessibility prediction achieved correlation coefficients of 0.823, 0.816 and 0.791 between predicted and actual ASA values for the validation, TEST2016 and TEST-HARD sets, respectively. The backbone torsion angle ψ has mean absolution errors of 22.5, 23.3 and 25.0 degrees for the validation, TEST2016 and TEST-HARD sets, respectively.

As a comparison, we also listed the performance of our previous method SPIDER 3 for TEST2016. Large improvement is observed for all predicted structural properties. These include 2.5% increase in the accuracy of three-state secondary structure prediction (4.1% in SOV), 4% improvement in ASA correlation coefficients, and 11% reduction in ψ mean absolution errors. Improvement in angle prediction is the largest as the MAE values for all four angles (θ, τ, ϕ and ψ) are reduced by 11–14%. Table 1 also shows the results of SPOT-1D without input of predicted contact maps (SPOT-1D-base). Inputting predicted 2D information leads to 0.5% increase in the accuracy of three-state secondary structure prediction, 0.4% relative improvement in ASA correlation coefficients, and 2% reduction in ψ mean absolution errors. Thus, predicted 2D information provides useful incremental improvement.

SPOT-1D improves over SPIDER3 and SPOT-1D-base for all 20 amino acid residues. Supplementary Figure S2 compares the accuracy of 3-state secondary structure prediction of these three methods. SPOT-1D yields between 1.8 and 3.5% increase over SPIDER3 per residue. The largest improvement is for Tryptophan residues (W). Smaller but consistent improvement is observed between SPOT-1D and SPOT-1D-base. The overall trend of the prediction performance for each amino acid residue is the same for each method with a high correlation of 0.97.

The improvement of SPOT-1D over SPIDER3 is the smallest for fully exposed residues with the relative ASA (rASA) at about 100%. As shown in Supplementary Figure S3, more than 2% improvement of SPOT-1D over SPIDER3 is observed for rASA between 0 and 90%. Similarly, inputting predicted contact maps also makes the largest impact for partially exposed residues (rASA≈0.4). Fully exposed residues are easier to predict as these residues involve mostly local interactions and are dominated by coil residues (6.1, 0.6, 93.3% are helix, sheet and coil, respectively). Moreover, the >90% accuracy achieved by SPIDER3 at rASA≈1 is certainly difficult to improve on. Despite this, SPOT-1D manages to improve on SPIDER3 at rASA≈1 by nearly 1%.

It is of interest to know how much the above-described improvement in prediction performance is due to the use of an ensemble. Supplementary Table S2 compares the results of the single models described in Supplementary Table S1 and that of the final consensus. The use of an ensemble leads to 0.5 and 0.8% improvement in 3-state and 8-state secondary structure prediction, 2% relative improvement in correlation coefficient for ASA prediction, and 4% relative improvement in MAE of ψ angle prediction. The only exception is a 1.5 and 0.3% reduction in term of segment overlaps (SOV3 and SOV8), respectively. The ensemble may have slightly disrupted the segment-level consistency of a single model. Nevertheless, SOV3 was improved more than the three-state accuracy (4.1 versus 2.5%) from SPIDER3 to SPOT-1D.

To gauge the robustness of our proposed model, we performed 10-fold Cross Validation (CV) on our training set using one of our ensemble architectures (Model 0 in Supplementary Table S1). Supplementary Table S3 compares the result of 10-fold CV (alongside the means and standard deviations over the 10 folds) against Model 0's performance on the Validation set and on TEST2016. Essentially the same performance on the validation set and 10-fold CV as well as the low standard deviations over the 10-folds was observed for all predicted properties in the training set, confirming the robustness of the model trained.

Table 1. Performance of our proposed predictor (SPOT-1D) on the validation set alongside the performance of our previous LSTM-BRNN method SPIDER3 and SPOT-1D-base (SPOT-1D without contact map) on TEST2016 and TEST-HARD

Dataset	Model	SS3	SOV3	SS8	SOV8	ASA	HSEα-U	HSEα-D	CN	θ	τ	ϕ	ψ
Validation	SPOT-1D	87.54	80.81	77.60	75.58	0.8228	0.8225	0.7993	0.8631	6.78	24.54	16.27	22.53
TEST2016	SPIDER–3	84.66	75.62	–	–	0.7873	0.7744	0.7436	–	7.72	29.24	17.88	26.66
	SPOT-1D-base	86.67	79.52	76.03	73.88	0.8127	0.8139	0.7904	0.8532	7.02	26.16	16.52	23.67
	SPOT-1D	87.16	79.73	77.10	74.98	0.8158	0.8164	0.7938	0.8571	6.89	25.38	16.27	23.26
TEST-HARD	SPIDER–3	83.66	77.06	–	–	0.7586	0.7494	0.7072	–	7.82	29.93	18.15	28.05
	SPOT-1D-base	85.37	79.16	75.01	72.75	0.7864	0.7975	0.7642	0.8204	7.21	27.53	17.15	25.62
	SPOT-1D	85.99	79.80	76.20	74.21	0.7908	0.8020	0.7674	0.8254	7.04	26.56	16.84	25.01

Note: All angle predictions (θ, τ, ϕ, ψ) are measured in mean absolute error (MAE), 3-state (SS3, SOV3) and 8-state (SS8, SOV8) secondary structure predictions are measured in % accuracy, and the solvent accessibility metrics are measured in correlation coefficient (CC).

SPOT-1D is more accurate in part because SPOT-1D is trained on the larger dataset (10 029 proteins) than SPIDER 3 (4590 proteins). To assess the contribution of the larger training size without the complication of iterative training in SPIDER 3, we compare the first iteration of SPIDER-3 (SPIDER-3-Iter1) with an identical architecture trained with our larger dataset (SPIDER-3-Iter1-2018). The results of these two methods are also compared to those of our singular Model 0 from SPOT-1D (SPOT-1D-0) for the TEST2016 dataset in in Supplementary Table S4. The large database size (the difference between SPIDER3-Iter1-2018 and SPIDER3-Iter1) contributed to about 1.6% improvement in three-state secondary structure prediction and 2.2 degree reduction in τ. Our single model alone is 1.3% more accurate than SPIDER3-Iter1-2018 and another 1.7 degree reduction in τ, indicating that our improvement is more than due to the enlargement of our training database. Although iterative learning could improve SPIDER3-Iter1-2018 by another 1% (based on the performance difference between SPIDER3-Iter1 and SPIDER3), it would be still lower than our final SPOT-1D by about 1% for three-state secondary structure prediction. Thus, SPOT-1D without iterative learning improves the performance over SPIDER 3, even after accounting for the difference in training databases.

Table 2 compares the performance of several recently developed methods for 1D structural property prediction for new PDB structures in TEST2018. These methods include RaptorX (DeepCNF and RaptorX-angle) for secondary structure and angle prediction, PSRSM for secondary structure prediction, PORTER-5 for secondary structure prediction, MUFOLD for secondary and torsion angle prediction, and NetSurfP-2.0 for secondary structure, ASA and angle prediction. The second best performance method is NetSurfP-2.0 which has a comparable performance in ASA with SPOT-1D, but 0.8 and 0.4% lower accuracy in 3-state secondary structure prediction and SOV, respectively (1.6 and 2.2% in 8-state prediction), and more than 1 degree higher in mean absolute errors for both ϕ and ψ angle prediction. SPOT-1D's improvement is significant, obtaining *P*-values of <0.0022 against all external predictors for SS3 prediction, and $<3.7 \times 10^{-7}$ in SS8 prediction. Analyzing the 13 proteins of >700 residues originally excluded from TEST2018 leads to a reduction of performance for all methods applicable to long proteins as shown in Supplementary Table S5. For example, SPOT-1D performs 0.97% worse for SS3 prediction, 0.41 degrees higher for ϕ prediction, and 0.019 lower CC for ASA when analyzing only large proteins in our set. However, as this set is only 13 proteins, this is inconclusive as to the definite impact of long proteins on prediction. For completeness, we also show the class-wise SS prediction performance through a confusion matrix for 3- and 8-state predictions in Supplementary Tables S6 and S7, respectively. SPOT-1D is consistently the highest predictor for each element in SS3 and SS8 prediction, except for sheet prediction which is slightly lower than NetSurfP-2.0. The improvement over SPIDER3 is the largest for sheet prediction (6.6% increase).

Statistically significant improvement of SPOT-1D over other methods is evident when plotting the performance stratified by the number of effective number of homologous sequences, Neff, calculated by HHblits (Remmert *et al.*, 2012). Protein sequences with low Neff have few diverse homologous sequences in the sequence library. Figure 1 and Supplementary Figures S4 and S5 compares method performance for 3-state secondary structure, ASA and ψ angle prediction, respectively. All methods show a nearly monotonic decrease in performance for proteins with few sequence neighbors (low Neff), confirming the importance of evolutionary information in predicting one-dimensional structural properties (Hanson *et al.*, 2018; Heffernan *et al.*, 2018; Singh *et al.*, 2018; Wang *et al.*, 2016a). For example, the accuracy of 3-state secondary-structure prediction by SPOT-1D drops from 86% at Neff = 10 to 77% for Neff = 1. The overall performance of 86% by SPOT-1D for secondary structure prediction for TEST2018 indicates that the majority of recently solved structures have many known homologous sequences. Indeed, the average Neff for TEST2018 is 6.9. With a few exceptions noted below, SPOT-1D improves over other methods for every Neff bin, indicating the statistical significance of the improvement. One exception is the comparable performance of RaptorX (DeepCNF), Porter-5 and SPOT-1D in secondary structure prediction for sequences with few homologous sequences (Neff = 1). Another exception is that the overall comparable accuracy in ASA prediction by NetSurfP-2.0 and SPOT-1D.

Unlike multi-state secondary structure, one advantage of real-value prediction of torsion angles is that they can be employed to construct three-dimensional structural models for comparison with the corresponding actual native structures. Here, we constructed 33214 fragment structures of 40-residue sequence windows in TEST2018. Model structures were initialized by placing the first three atoms in the X-Z plane. The chain was incrementally extended using the predicted angles. Bond lengths and bond angles were fixed to residue-independent ideal values for the ϕ/ψ-based model with ω angle fixed to 180 degrees. The Cα-based θ/τ model assumed a fixed distance of 3.8Å between Cα positions. Discontinuous fragments with incomplete solved structures were excluded from the analysis. The structural difference between the model and its native structure is measured by root-mean-squared distance (RMSD). A 6 Å RMSD is considered as significant structural similarity (Reva *et al.*, 1998). As shown in Figure 2A, the fraction of fragment structures with RMSD < 6Å constructed by using predicted ϕ and ψ angles is 36%

Table 2. Test performance of several recently developed predictors alongside SPOT-1D on the latest PDB structures (TEST2018)

Predictor	SS3	SOV3	*P*SS3	SS8	SOV8	*P*SS8	ASA	HSEα-U	HSEα-D	CN	θ	τ	ϕ	ψ
SPIDER-3-Single	72.57	64.08	$<1 \times 10^{-10}$	59.81	57.86	$<1 \times 10^{-10}$	0.570	0.603	0.533	0.619	11.07	45.39	23.77	43.05
RaptorX	81.62	66.58	$<1 \times 10^{-10}$	70.43	65.66	$<1 \times 10^{-10}$	–	–	–	–	–	–	21.01	35.95
PSRSM	81.94	74.22	$<1 \times 10^{-10}$	–	–	–	–	–	–	–	–	–	–	–
SPIDER-3	83.84	73.89	$<1 \times 10^{-10}$	–	–	–	0.768	0.764	0.716	–	7.73	29.62	18.38	28.10
PORTER-5	84.10	74.04	$<1 \times 10^{-10}$	73.22	70.27	$<9.89 \times 10^{-9}$	–	–	–	–	–	–	–	–
MUFOLD	84.78	77.56	$<2.73 \times 10^{-8}$	73.66	71.34	$<2.15 \times 10^{-9}$	–	–	–	–	–	–	17.78	27.24
NetSurfP-2.0	85.31	78.58	$<2.20 \times 10^{-3}$	73.81	71.14	$<3.64 \times 10^{-7}$	0.801	–	–	–	–	–	17.90	26.63
SPOT-1D-base	85.66	78.77	$<1.08 \times 10^{-2}$	74.26	71.45	$<1.33 \times 10^{-4}$	0.799	0.812	0.775	0.837	7.03	26.86	17.15	25.41
SPOT-1D	86.18	79.00	–	75.41	73.30	–	0.803	0.814	0.779	0.841	6.91	25.94	16.89	24.87

Note: All classification accuracies are presented in % accuracy, except for the significance metrics *P*, all angle predictions (θ, τ, ϕ, ψ) are measured in mean absolute error (MAE), and the solvent accessibility metrics are measured in correlation coefficient (CC).

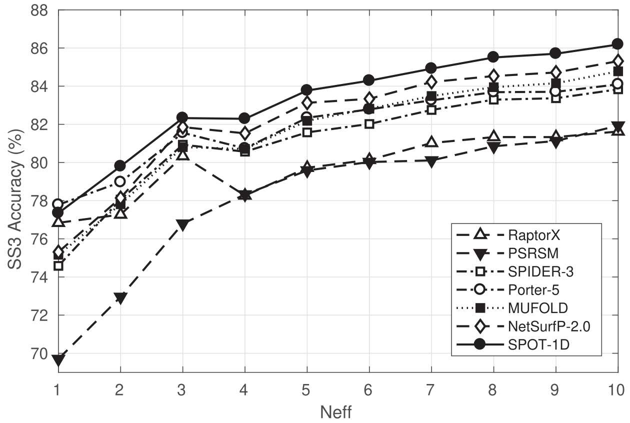

Fig. 1. The dependence of the accuracy of 3-state secondary structure prediction on the number of effective homologous sequences for the TEST2018 set (250 proteins). Note that the Neff of each protein is binned by rounding down it to the nearest integer

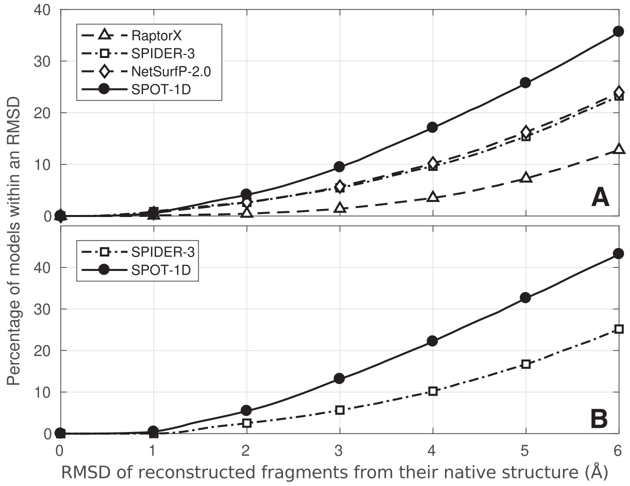

Fig. 2. The fraction of 40-residue fragments in the TEST2018 set whose models, constructed based on predicted angles, are below a given root-mean-squared distance (RMSD) from their corresponding native structures. (**A**) is based on ϕ/ψ predicted angles, and (**B**) is based on θ/τ angles

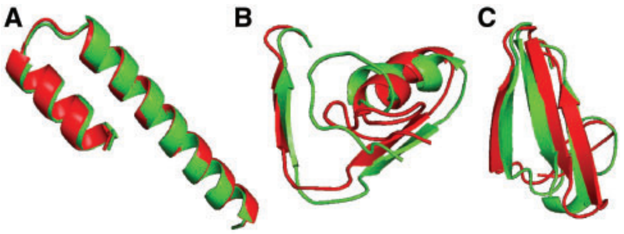

Fig. 3. The results of three 40-residue protein fragments: helical hairpin from residues 20 to 59 of protein 5N5EA (**A**), mixed helix and two sheets from residues 59 to 98 of protein 6FI2A, (**B**) and three-stranded antiparallel beta sheets from residues 326 to 365 of protein 6FQ3A (**C**). All fragments were reconstructed from the predicted θ/τ angles from SPOT-1D. The native structures are shown in green, and the reconstructed fragments in red (Color version of this figure is available at *Bioinformatics* online.)

by SPOT-1D, compared to 24% by NetSurfP-2.0 and SPIDER3. If fragment structures are constructed by the Cα-angle based θ and τ predicted by SPOT-1D, the fraction of fragment structures with RMSD < 6Å further increases to 44%, meaning that close of half of structures constructed have a reasonable structural similarity to their native structures, and an 18% improvement over SPIDER3 (Fig. 2B). As illustrative examples, Figure 3 shows accurately constructed models in helix bundle (A), mixed helix and sheet (B) and all-sheet fragments (C) with RMSD at 1.0Å, 4.3Å and 3.0Å, respectively. Accurate description of coil regions reflects the usefulness of angle prediction over secondary-structure prediction.

To further test our model, Supplementary Table S8 compares the results of the CASP12 dataset by different models. The performance for each predictor is considerably lower than that presented in Table 2, largely because these proteins are specifically hard cases which do not belong to folds present in any training data obtained before CASP12. SPOT-1D continues to outperform all tested methods (except for a comparable ASA to NetSurfP-2.0), although this performance improvement is not statistically significant largely due to the small sample size of 20 proteins.

4 Discussion

We have developed a new method for predicting one-dimensional structural properties of proteins based on an ensemble of different types of neural networks (LSTM-BRNN, ResNet and FC-NN) with predicted contact maps input from SPOT-contact. For a large independent test set of 1213 proteins (TEST2016), the method achieves unprecedented accuracy of 87 and 77% for 3-state and 8-state secondary structure prediction, respectively, and mean absolute errors of 16, 23, 7 and 25 degrees for ψ, ψ, θ and τ, respectively. It also provides high accuracy for ASA and HSE contact number prediction.

The significant improvement is due to several factors. Using the 3-state secondary structure as an example, predicted contact maps as a feature contributes 0.5% improvement in accuracy (Table 1). The employment of an ensemble of different types of neural networks contributes another 0.5% improvement (Supplementary Table S2). The remaining roughly 1% improvement over SPIDER3 comes from the large dataset (Supplementary Table S4). As the 3-state accuracy approaches its theoretical limit of 88–89% (Rost, 2001; Yang *et al.*, 2018), every bit of improvement is useful.

However, closing onto the theoretical limit for secondary structure prediction is true only for those protein sequences with many known sequence homologs. As Figure 1 shows, the accuracy for the secondary structure prediction drops to about 77% for sequences with few known homologs (Neff = 1). This is problematic because >90% known protein sequences have few homologous sequences (Ovchinnikov *et al.*, 2017). Thus, it is necessary to develop the methods based on information from single sequence only such as SPIDER3-single (Heffernan *et al.*, 2018). However, SPIDER3-single, using exactly the same neural network topology and training set as SPIDER 3, can only improve over SPIDER3 in secondary structure prediction by about 1% for proteins with Neff = 1. Thus, single-sequence-based prediction, the ultimate solution to the folding problem, remains challenging.

Interestingly, the performance of SPOT-1D is comparable to that of NetSurP-2.0 in ASA prediction (Supplementary Fig. S4). This occurred despite the significant improvement of SPOT-1D over NetSurP-2.0 in prediction of secondary structures (Fig. 1), angles (Supplementary Fig. S5) and construction of fragments from predicted angles (Fig. 2A). This could be a signal for reaching a possible theoretical limit for ASA prediction, because a CC of 0.8 or more between predicted and actual ASA (Tables 1 and 2) has already exceeded the 0.77 CC for the ASA between homologs (Rost and Sander, 1994). More studies by using other techniques are needed to confirm this bottleneck.

The most encouraging result of this paper is the large improvement of fragment structural accuracy. The 11–14% reduction of angle errors by SPOT-1D over SPIDER 3 (Table 1) leads to more than 10% increase in the fraction of sequences with RMSD < 6Å from native conformations. In fact, close to half (44%) of 33 214 40-residue fragments have $<$6Å RMSD when the fragments are constructed by predicted Cα-based angle and torsion angles. Thus, using SPOT-1D for angle prediction should significantly improve the accuracy of fragment libraries for de novo protein structure prediction (Wang *et al.*, 2016b). Direct use of predicted angles for sequence-to-structure alignment has also been proven useful for template-based structure prediction (Yang *et al.*, 2011).

It is of interest to know what causes some amino acid residues to be predicted easier than others, as this behavior seems to be consistent among different methods as shown in Supplementary Figure S2, and it has been shown previously (Heffernan *et al.*, 2017). One natural explanation is that different amino acids have different natural abundance in the training dataset and thus affect respective accuracy by different levels of training cases. Indeed, the correlation between amino acid abundance and the accuracy of secondary structure for individual amino acid has a reasonably strong correlation with a CC of 0.52. To further search for the underlying mechanism, we correlate the accuracy of secondary structure for individual amino acids to $>$550 one-dimensional structure properties collected in AAindex (Kawashima and Kanehisa, 2000). We found that the highest correlation (CC of 0.67) is to the distribution of amino acid residues in alpha-helices in thermophilic proteins. The propensity of an amino acid residue for forming helices in high temperature environment indicates its bias toward near-neighbor interactions. Thus, bias toward local interactions or helical propensity is another factor contributing easiness in secondary structure prediction as helices are the easiest to predict.

In the interest of profiling our method in terms of processing time, we have measured the time taken for each component of our SPOT-1D downloadable version. Supplementary Table S9 shows the time needed by our local machine for both a regular (PDB ID: 5ugwA) and long protein (PDB ID: 6ggyB) on both CPU and GPU. As can be seen in this table, the majority of processing time is spent on the PSSM generation, which is inhibited by the local machines disk read/write speed. However, if these files are readily available, they can be directly provided to the program, saving the bulk of processing time. Long proteins are also shown to take extensive time, especially for 2D analysis tools (SPOT-Contact, CCMpred, DCA). The use of CPU and GPU is shown to not make a major difference in time taken, as the speed increase introduced by GPU acceleration mainly comes during training. Protein 6ggyB is too large to be run on the GPU for SPOT-Contact. Compared to another readily available predictor MUFOLD (MUFOLD-SS and MUFOLD-Angle) for protein 5ugwA, SPOT-1D is only 2 min slower despite needing extra features. MUFOLD also requires both a PSSM and HHBlits profile. Finally, it should be noted that processing one file at a time is inefficient, as multiple sequences can be processed at once in the SPOT-1D suite, meaning each of the ensemble's models only needs to be loaded once.

Funding

This work was supported by Australia Research Council DP180102060 to Y.Z. and K.P., and in part by National Health and Medical Research Council (1121629) of Australia to Y.Z. We also gratefully acknowledge the use of the High Performance Computing Cluster 'Gowonda' to complete this research, and the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

Conflict of Interest: none declared.

References

Adhikari,B. *et al.* (2017) DNCON2: improved protein contact prediction using two-level deep convolutional neural networks. *Bioinformatics*, **1**, 7.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Ceroni,A. and Frasconi,P. (2004) On the role of long-range dependencies in learning protein secondary structure. In: *IEEE IJCNN*, Vol. **3**. IEEE, pp. 1899–1904.

Ceroni,A. *et al.* (2005) Learning protein secondary structure from sequential and relational data. *Neural Netw.*, **18**, 1029–1039.

Chu,W. *et al.* (2006) Bayesian segmental models with multiple sequence alignment profiles for protein secondary structure and contact map prediction. *IEEE ACM Trans. Comput. Biol.*, **3**, 98–113.

Fang,C. *et al.* (2018a) Mufold-ss: new deep inception-inside-inception networks for protein secondary structure prediction. *Proteins*, **86**, 592–598.

Fang,C. *et al.* (2018b) Prediction of protein backbone torsion angles using deep residual inception neural networks. *IEEE ACM Trans. Comput. Biol.*

Faraggi,E. *et al.* (2012) Spine x: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles. *J. Comp. Chem.*, **33**, 259–267.

Gao,Y. *et al.* (2018) Raptorx-angle: real-value prediction of protein backbone dihedral angles through a hybrid method of clustering and deep learning. *BMC Bioinformatics*, **19**, 100.

Gibson,K.D. and Scheraga,H.A. (1967) Minimization of polypeptide energy. i. preliminary structures of bovine pancreatic ribonuclease s-peptide. *Proc. Natl. Acad. Sci. USA*, **58**, 420–427.

Hamelryck,T. (2005) An amino acid has two sides: a new 2d measure provides a different view of solvent exposure. *Proteins*, **59**, 38–48.

Hanson,J. *et al.* (2018) Accurate prediction of protein contact maps by coupling residual two-dimensional bidirectional long short-term memory with convolutional neural networks. *Bioinformatics*, **34**, p4039–p4045.

He,K. *et al.* (2016) Identity mappings in deep residual networks. In: *Eur. Conf. Comp. Vis.* Springer, Amsterdam, The Netherlands, pp. 630–645.

Heffernan,R. *et al.* (2015) Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning. *Nat. Sci. Rep.*, **5**, 11476.

Heffernan,R. *et al.* (2016) Highly accurate sequence-based prediction of half-sphere exposures of amino acid residues in proteins. *Bioinformatics*, **32**, 843–849.

Heffernan,R. *et al.* (2017) Capturing non-local interactions by long short term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure. *Bioinformatics*, **33**, 2842–2849.

Heffernan,R. *et al.* (2018) Single-sequence-based prediction of protein secondary structure, backbone angles, solvent accessibility, half-sphere exposure, and contact number by long short-term memory bidirectional recurrent neural networks. *J. Comp. Chem*, **26**, 2210–2216.

Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.

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

Kang,H.S. *et al.* (1993) Estimation and use of protein backbone angle probabilities. *J. Mol. Biol.*, **229**, 448–460.

Kawashima,S. and Kanehisa,M. (2000) Aaindex: amino acid index database. *Nucleic Acids Res.*, **28**, 374.

Klausen,M.S. *et al.* (2018) Netsurfp-2.0: improved prediction of protein structural features by integrated deep learning. *bioRxiv*, 311209.

Korkut,A. and Hendrickson,W.A. (2009) A force field for virtual atom molecular mechanics of proteins. *Proc. Natl. Acad. Sci. USA*, **106**, 15667–15672.

Lee,B. and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.*, **55**, 379–IN4.

Lyons,J. *et al.* (2014) Predicting backbone cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *J. Comp. Chem.*, **35**, 2040–2046.

Ma,Y. *et al.* (2018) Protein secondary structure prediction based on data partition and semi-random subspace method. *Nat. Sci. Rep*, **8**, 9856.

Meiler,J. *et al.* (2001) Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Mol. Model*, **7**, 360–369.

Mirdita,M. *et al.* (2017) Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Res.*, **45**, D170–D176.

Ovchinnikov,S. *et al.* (2017) Protein structure determination using metagenome sequence data. *Science*, **355**, 294–298.

Pauling,L. *et al.* (1951) The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. USA*, **37**, 205–211.

Ramachandran,G. *et al.* (1963) Stereochemistry of polypeptide chain configurations. *J Mol. Biol.*, **7**, 95–99.

Remmert,M. *et al.* (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.

Reva,B.A. *et al.* (1998) What is the probability of a chance prediction of a protein structure with an rmsd of 6 å? *Fold. Des.*, **3**, 141–147.

Rost,B. (2001) Protein secondary structure prediction continues to rise. *J. Struct. Biol.*, **134**, 204–218.

Rost,B. and Sander,C. (1993) Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad Sci. USA*, **90**, 7558–7562.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.

Schaarschmidt,J. *et al.* (2018) Assessment of contact predictions in casp12: co-evolution and deep learning coming of age. *Proteins*, **86**, 51–66.

Schuster,M. and Paliwal,K.K. (1997) Bidirectional recurrent neural networks. *IEEE Trans. Signal Proc.*, **45**, 2673–2681.

Singh,J. *et al.* (2018) Detecting proline and non-proline cis isomers in protein structures from sequences using deep residual ensemble learning. *JCIM*, **58**, 2033–2042.

Szegedy,C. *et al.* (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: *AAAI*, Vol. **4**, p. 12.

Torrisi,M. *et al.* (2018) Porter 5: fast, state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. *bioRxiv*, 289033.

Vapnik,V.N. (1998) *Statistical Learning Theory*, Vol. **1**. Wiley, New York.

Wang,G. and Dunbrack,R.L. (2003) Pisces: a protein sequence culling server. *Bioinformatics*, **19**, 1589–1591.

Wang,S. *et al.* (2016a) Protein secondary structure prediction using deep convolutional neural fields. *Nat. Sci. Rep.*, **6**, 18962.

Wang,T. *et al.* (2016b) Lrfraglib: an effective algorithm to identify fragments for de novo protein structure prediction. *Bioinformatics*, **33**, 677–684.

Wang,S. *et al.* (2017) Accurate de novo prediction of protein contact map by ultra-deep learning model. *PLoS Comput. Biol.*, **13**, 1–34.

Xue,B. *et al.* (2008) Real-value prediction of backbone torsion angles. *Proteins*, **72**, 427–433.

Yang,Y. *et al.* (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. *Bioinformatics*, **27**, 2076–2082.

Yang,Y. *et al.* (2018) Sixty-five years of the long march in protein secondary structure prediction: the final stretch? *Brief. Bioinform.*, **19**, 482–494.

Zemla,A. *et al.* (1999) A modified definition of sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins*, **34**, 220–223.

Zhou,Y. *et al.* (2011) Trends in template/fragment-free protein structure prediction. *Theor. Chem. Acc.*, **128**, 3–16.