Bioinformatics, 33(18), 2017, 2842–2849
doi: 10.1093/bioinformatics/btx218
Advance Access Publication Date: 18 April 2017
Original Paper

OXFORD

Structural bioinformatics

Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility

Rhys Heffernan[1], Yuedong Yang[2,*], Kuldip Paliwal[1] and Yaoqi Zhou[2,*]

[1]Signal Processing Laboratory, Griffith University, Brisbane, QLD 4111, Australia and [2]Institute for Glycomics and School of Information and Communication Technology, Griffith University, Southport, QLD 4222, Australia

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on October 27, 2016; revised on January 20, 2017; editorial decision on April 8, 2017; accepted on April 15, 2017

Abstract

Motivation: The accuracy of predicting protein local and global structural properties such as secondary structure and solvent accessible surface area has been stagnant for many years because of the challenge of accounting for non-local interactions between amino acid residues that are close in three-dimensional structural space but far from each other in their sequence positions. All existing machine-learning techniques relied on a sliding window of 10–20 amino acid residues to capture some 'short to intermediate' non-local interactions. Here, we employed Long Short-Term Memory (LSTM) Bidirectional Recurrent Neural Networks (BRNNs) which are capable of capturing long range interactions without using a window.

Results: We showed that the application of LSTM-BRNN to the prediction of protein structural properties makes the most significant improvement for residues with the most long-range contacts ($|i-j| > 19$) over a previous window-based, deep-learning method SPIDER2. Capturing long-range interactions allows the accuracy of three-state secondary structure prediction to reach 84% and the correlation coefficient between predicted and actual solvent accessible surface areas to reach 0.80, plus a reduction of 5%, 10%, 5% and 10% in the mean absolute error for backbone ϕ, ψ, θ and τ angles, respectively, from SPIDER2. More significantly, 27% of 182724 40-residue models directly constructed from predicted Cα atom-based θ and τ have similar structures to their corresponding native structures (6Å RMSD or less), which is 3% better than models built by ϕ and ψ angles. We expect the method to be useful for assisting protein structure and function prediction.

Availability and implementation: The method is available as a SPIDER3 server and standalone package at http://sparks-lab.org.

Contact: yaoqi.zhou@griffith.edu.au or yuedong.yang@griffith.edu.au

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

1 Introduction

Proteins perform a wide variety of molecular functions, and most of those functions are determined by the protein's three-dimensional structures. The problem of predicting the three-dimensional structure of a protein, given only its linear sequence of residues, is crucially important due to the high cost of experimental protein structure determination and the low cost of DNA sequencing to obtain protein sequences. However, even with this importance it is a problem which has remained unsolved since its inception half a century ago (Gibson and Scheraga, 1967; Dill and MacCallum, 2012; Zhou *et al.*, 2011). Due to the challenge of predicting three-dimensional protein structure, the problem is commonly divided into smaller, more achievable problems, in the hopes that their solutions will ultimately lead to the solution for three-dimensional structure prediction. One type of these sub-problems is the prediction of one-dimensional structural properties of proteins, which can be represented as a one-dimensional vector along the protein sequence.

The most commonly predicted one-dimensional structure of proteins is secondary structure. Secondary structure prediction can be dated back to 1951 when Pauling and Corey proposed helical and sheet conformations of protein backbone based on hydrogen bonding patterns (Pauling *et al.*, 1951). Secondary structure is a course-grained descriptor of the local structure of the polypeptide backbone. Despite its long history, secondary structure prediction continues to be an active area of research as the accuracy of three-state prediction increases slowly from <70% to 82–84% (Rost, 2001; Zhou and Faraggi, 2010; Heffernan *et al.*, 2015; Wang *et al.*, 2016; Mirabello and Pollastri, 2013; Yaseen and Li, 2014), approaching the theoretical limit in the range of 88–90% (Rost, 2001).

Protein backbone structure can be described continuously by torsion angles ϕ and ψ. A number of methods have been developed to predict ϕ and ψ as both discrete (Kuang *et al.*, 2004; Kang *et al.*, 1993) and continuous (Wood and Hirst, 2005; Dor and Zhou, 2007; Xue *et al.*, 2008; Lyons *et al.*, 2014) labels. More recently, a method (Lyons *et al.*, 2014) was developed for predicting Cα atom-based θ and τ that describes the structure of four neighbouring residues, complement to the single residue structure described by ϕ and ψ.

Along with the previously mentioned local structure descriptors, a range of global solvent exposure descriptors are also important for understanding and predicting protein structure, function and interactions (Gilis and Rooman, 1997; Lee and Richards, 1971; Rost and Sander, 1994; Tuncbag *et al.*, 2009). These include solvent Accessible Surface Area (ASA), Contact Number (CN) and Half Sphere Exposure (HSE). CN is a count of the number of neighbouring residues in three-dimensional space, within a specific distance cut off. HSE extends the ideas of CN and adds directionality information by differentiating between the counts in a top and bottom half of the sphere (Hamelryck, 2005).

Earlier methods for ASA prediction were limited to the prediction of discrete states (Holbrook *et al.*, 1990; Rost and Sander, 1994; Pollastri *et al.*, 2002b) and have since progressed to continuous value predictions (Dor and Zhou, 2007; Garg *et al.*, 2005; Yuan and Huang, 2004; Ahmad *et al.*, 2003; Adamczak *et al.*, 2004; Heffernan *et al.*, 2015). A number of methods for CN and HSE prediction have also been developed (Kinjo and Nishikawa, 2006; Song *et al.*, 2008; Heffernan *et al.*, 2016)

The accuracy of all of these methods, however, was hampered by their inability to capture long-range, non-local interactions effectively. Non-local interactions refer to the interactions between residues that are close in the three-dimensional space, but far from each other in their sequence positions. Commonly used simple Artificial Neural Networks (ANNs) (Faraggi *et al.*, 2012), and Deep Neural Networks (DNNs) (Lyons *et al.*, 2014; Heffernan *et al.*, 2016, 2015), relied on their input windows of neighbouring residue information, which means that they are unable to fully learn the relationship between the residues in the whole sequence. Convolutional Neural Networks (CNNs) have been used to overcome the limitations of windowing by applying layers of shifting convolutional kernels across the input sequence (Wang *et al.*, 2016). The addition of each subsequent convolutional layer includes a window of information from the layer prior. This has the effect of growing the size of the window being applied to the input data, each time a layer is added. However, even though the effective window size increases at each layer, the total window is still finite and limited. DeepCNF for example uses five CNN layers, each with a window size of 11 (five on either side), this is only able to capture information from residues plus or minus 25 sequence positions away. Bidirectional Recurrent Neural Networks (BRNNs) (Schuster and Paliwal, 1997) can exploit both preceding and following dependencies across the whole sequence and have been used in a number of bioinformatic studies (Baldi *et al.*, 1999; Pollastri *et al.*, 2002a; Mirabello and Pollastri, 2013; Magnan and Baldi, 2014). Long Short-Term Memory (LSTM) cells were introduced to be able to learn both local and non-local intra-sequence relationships more efficiently (Hochreiter and Schmidhuber, 1997). In recent years, LSTM-based networks have enjoyed significant success in a wide variety of tasks that involve learning from sequences, from speech recognition (Amodei *et al.*, 2015), to natural language processing (Sundermeyer *et al.*, 2012), and handwriting recognition (Graves and Schmidhuber, 2009). Its recent application to protein intrinsic disorder prediction demonstrates its ability to capture non-local interactions in protein sequences (Hanson *et al.*, 2017).

In this paper we apply a BRNN, using LSTM cells, to the problem of prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. Using exactly the same datasets and iterative procedure used in previous works, (Lyons *et al.*, 2014; Heffernan *et al.*, 2016, 2015) we demonstrated significant improvement in all structural properties predicted (excluding contact-based properties), particularly for residues with more non-local contacts. This work highlights the importance of capturing non-local interactions in predicting protein one-dimensional structural properties.

2 Materials and methods

2.1 Datasets

To demonstrate the power of LSTM-BRNNs, we employed exactly same datasets as used in our previous studies (Lyons *et al.*, 2014; Heffernan *et al.*, 2015, 2016). The full dataset contains 5789 proteins with a sequence similarity cut off of 25% and X-ray resolution better than 2.0 Å. From the full set of data, 4590 proteins were randomly selected to be the training set (TR4590), and the remaining 1199 are used as the independent test set (TS1199). The TR4590 set is further split into a training and validation set (TR4131 and VAL459, respectively) so that the validation set can be used for early stopping during training. The TR4590 set is also split into 10 evenly sized folds for 10-fold cross validation, each of these folds also has designated train and validation sequences.

In addition, we obtained all the high-resolution (<3 Å) proteins deposited in 2016 (prior to September 16, 2016) that are below 30% sequence similarity among each other and to all proteins

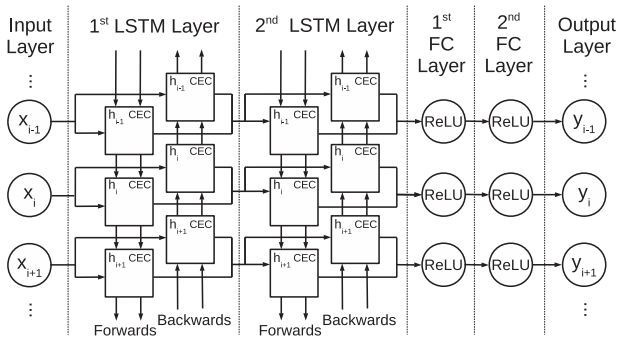

Fig. 1. Network architecture of LSTM-BRNN employed in all of the four iterations

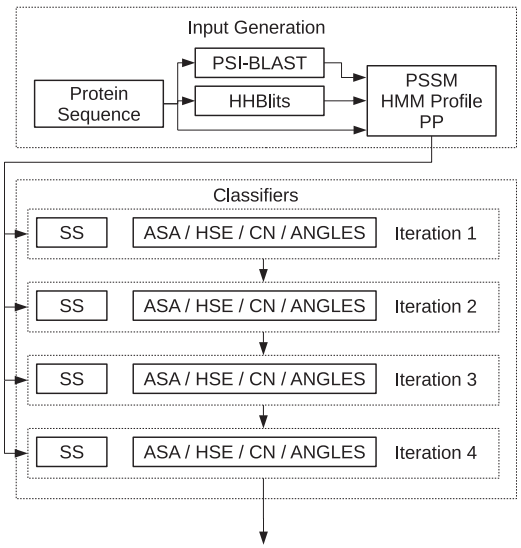

Fig. 2. System flowchart. Each of the four iterations contains two models (SS, and ASA/HSE/CN/ANGLES), for a total of eight LSTM-BRNN based models. The network architecture for each of the eight models are described by Figure 1

deposited prior to 2016. This constitutes a separate independent test set of 115 proteins.

2.2 Long short-term memory network bidirectional recurrent neural network

We employed a machine learning model called a Bidirectional Recurrent Neural Network (BRNN) (Schuster and Paliwal, 1997). As shown in Figure 1, the model uses two BRNN layers with 256 nodes per direction, followed by two more fully connected hidden layers with 1024 and 512 nodes, respectively. The BRNN layers use Long Short-Term Memory (LSTM) cells (Hochreiter and Schmidhuber, 1997) due to their ability to learn both distant and close intra-sequence dependencies. The fully connected nodes in the network each use Rectified Linear Unit (ReLU) activations. To reduce network overfitting, we utilize the dropout algorithm (Srivastava *et al.*, 2014) with a dropout ratio of 50% during training. The model was trained using Adam optimization (Kingma and Ba, 2014), a method for efficient stochastic optimization that is able to compute adaptive learning rates and requires little hyperparameter tuning. The validation set loss was used for early stopping during training to stop the networks from overfitting to the training data being used. We use Google's open-sourced TensorFlow library (Abadi *et al.*, 2016) to implement and train our networks. To speed up training, we use the GPU version of TensorFlow and train on an Nvidia GeForce Titan X GPU.

Two distinct networks are trained for each iteration of the model. For the secondary structure prediction the network uses the cross-entropy loss function which is well suited to classification problems (LeCun *et al.*, 2006). The ASA, ϕ, ψ, θ, τ, HSE and CN prediction network uses square loss which is well suited to regression problems (LeCun *et al.*, 2006). During training both of these networks mask out the loss contributed by any undefined labels, such as residues with no secondary structure assignment according to the program Dictionary of Secondary Structure of Proteins (DSSP) (Kabsch and Sander 1983).

2.3 Iterative learning

We employed an iterative procedure employed previously for sequence-based prediction of protein structural properties (Faraggi *et al.*, 2012; Lyons *et al.*, 2014; Heffernan *et al.*, 2016, 2015). As shown in Figure 2, the first set of predictors take seven representative physio-chemical properties (PP) of amino acids (Fauchère *et al.*, 1988), 20-dimensional Position Specific Substitution Matrices (PSSM) from PSI-BLAST (Altschul *et al.*, 1997), and 30-dimensional hidden Markov Model sequence profiles from HHBlits (Remmert *et al.*, 2012) per residue as input, and make predictions for SS, ASA, backbone angles, HSE and CN (iteration 1). These features were selected previously in the development of SPIDER for secondary structure and contact prediction (Lyons *et al.*, 2014; Heffernan *et al.*, 2016, 2015). Unlike previous systems utilizing neural networks, no window is needed as the entire sequence was input into the network. Both the long and short-range interactions are remembered by enforcing the constant error flow in the LSTM cell, regardless of sequence separation. The outputs from iteration 1 are then added to the PSSM, PP and HMM profile features, as inputs to a second set of predictors (iteration 2). This process is repeated two more times (iterations 3 and 4), for a total of four sets of networks as the result converges. Hereafter (and also for the available SPIDER3 server), results from iteration 4 were utilized. To avoid potential over training, we have kept the training and testing data completely separate. The TR4950 set is split into the same 10 folds for training every iteration of the network.

2.4 Outputs

For secondary structure labels, DSSP (Kabsch and Sander, 1983) specifies eight secondary structure states. Three helix shapes: 3_{10}-helix G, alpha-helix H, and pi-helix I; two strand types: beta-bridge B and beta-strand E; and three coil types: high curvature loop S, beta-turn T, and coil C. For this work, we predict three states by converting the DSSP assigned states G, H, and I to H; B and E to E; and S, T, and C to C. These secondary structure labels are calculated from their experimentally determined structures. The networks used to predict ASA, ϕ, ψ, θ, τ, HSE and CN have 14 total outputs, the first for ASA; the next eight nodes for $\sin(\phi)$, $\cos(\phi)$, $\sin(\psi)$, $\cos(\psi)$, $\sin(\theta)$, $\cos(\theta)$, $\sin(\tau)$ and $\cos(\tau)$, respectively; the next four for HSEα-up, HSEα-down, HSEβ-up, and HSEβ-down; and the final output node is for CN. Utilizing the sine and cosine functions for angles is to remove the effect of the angle's periodicity (Lyons *et al.*, 2014). The sine and cosine predictions are converted back to angles by the equation $\alpha = \tan^{-1}[\sin(\alpha)/\cos(\alpha)]$.

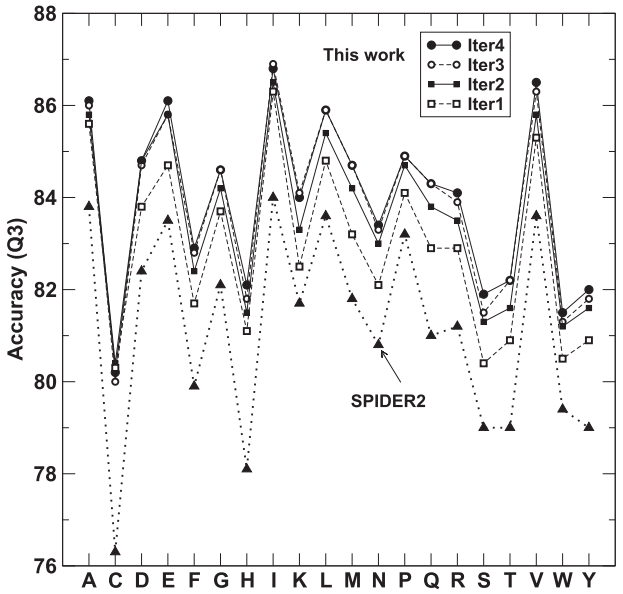

Fig. 3. The accuracy of three-state secondary structure prediction for individual amino acids at different iterations, compared to that of SPIDER2 for the independent test set (TS1199)

Table 1. The accuracy of this work in the 10-fold cross validation and independent test (TS1199) as compared to SPIDER2 trained iteratively on deep neural networks

	This Work		SPIDER2
Data	10-fold	TS1199	TS1199
SS (Q3) (%)	84.16	84.48	81.8
MAE: ϕ (°)	18.3	18.3	19.3
MAE: ψ (°)	27.0	27.0	30.0
MAE: θ (°)	7.74	7.64	8.13
MAE: τ (°)	29.2	28.9	32.4
ASA (CC)	0.788	0.795	0.759
CN (CC)	0.809	0.815	0.814
HSEβ-up	0.781	0.790	0.782
HSEβ-down	0.736	0.745	0.742
HSEα-up	0.783	0.792	0.784
HSEα-down	0.745	0.754	0.755

2.5 Performance measure

Secondary structure prediction accuracy is measured as the percentage of correctly classified residues in the database, and is known as Q3 for three state prediction. The Pearson Correlation Coefficient (PCC) and Mean Absolute Error (MAE) between true and predicted values are used as the accuracy measures for ASA (Faraggi *et al.*, 2012), HSE (Song *et al.*, 2008) and CN (Yuan, 2005; Heffernan *et al.*, 2016). The accuracy of ϕ, ψ, θ and τ angles are measured as the Mean Absolute Error (MAE) between true and predicted angles. To compensate for the periodicity in angles, the error in the prediction of residue i is defined as the smaller of d_i and $360 - d_i$, where d_i is $|A_i^{\text{pred}} - A_i^{\text{true}}|$. Because both ϕ and ψ have two peaks in their distributions, they can be split into two classes, where each class is defined as being closer to one of the two peaks. To evaluate large-angle errors the ϕ angles are split into one state from [0° to 150°] and a second state made up of angles from [(150° to 180°) and (−180° to 0°)]. The ψ angles are split into [−100° to 60°] for one state, and [(−180° to −100°) and (60° to 180°)] for the second.

For 10-fold cross validation, a full model (iteration 1–4) is trained for each fold. 10-fold validation results that are reported are the mean across each of the 10-folds. All other results in this work are from a single model, utilizing the TR4131 set for training data, and VAL459 as validation data.

3 Results

Figure 3 (see also Supplementary Table S1) compares the accuracy of secondary structure prediction at individual amino acid levels from this study at four iterations to that of SPIDER2. This is the result from the independent test set (TS1199). The accuracies of the first iteration are more accurate than the results from the final iteration of SPIDER2 for all 20 amino acid residues. The accuracies of individual amino acids are all higher than 80%. There is a strong correlation between SPIDER2 and the frequency of amino acid residues with a Pearson's Correlation Coefficient (PCC) of 0.71, indicating that higher abundance means higher accuracy. Similar strong correlation is observed with a PCC of 0.69 between the frequency and the accuracy at the forth iteration of this new method. A PCC of 0.97 between SPIDER2 and this work (the accuracy at the forth iteration), confirms a systematic improvement for all residues. The overall accuracy improves from 83.4% in the first iteration, 84.1% in the second iteration, 84.4% in the third iteration, to 84.5% in the forth iteration. The largest improvement is from the first to the second iteration. The improvement from the third to forth iteration is statistically significant with a P value of 0.002. Thus, we have set the number of iterations to four.

Table 1 compares the results from 10-fold cross validation (TR4590) and those from the large test set (TS1199). The two sets of training and test results are essentially identical across all structural properties ranging from secondary structures, angles, to contact numbers, indicating the robustness of the method. Table 1 also shows that all structure properties are significantly improved over SPIDER2 except contact numbers that were trained and tested on the same dataset with a deep neural network in three iterations. For the same test set, the new method has a 2.7% increase in accuracy of secondary structure prediction, 5%, 10%, 5% and 10% decrease in MAE for ϕ, ψ, θ and τ angles, respectively. Large errors in angles were also reduced. For the independent test set, the two-state accuracies for ϕ and ψ (see Methods for definition) are 96.8% and 88.6%, respectively. These are 0.2% and 1.8% absolute improvement over SPIDER2. The correlation coefficients, between predicted and actual ASA, improve from 0.76 to 0.80. Improvement in correlation coefficients for various contact numbers was smaller. This could be due to compensation of errors in predicting local and non-local contacts, as only the total number of contacts were predicted.

Figures 4 and 5 show the dependence of the accuracy of secondary structure prediction on the number of non-local and local contacts in a residue, respectively. Here, we defined non-local contacts as contacts between two residues that are more than or equal to 20 residues away in their sequence positions, but <8Å away in terms of their atomic distances between Cα atoms. We employed a sequence separation of 20 as a cutoff value because this is a typical sliding window size employed in secondary structure prediction. All residues were divided into 20 bins according to their non-local contacts. Each bin contains at least 12 500 residues. Despite the use of deep learning neural networks, the three-state accuracy of secondary structure prediction of SPIDER2 decreases nearly linearly when the number of non-local contacts is >5. This confirms the failure of commonly employed deep learning neural techniques in capturing

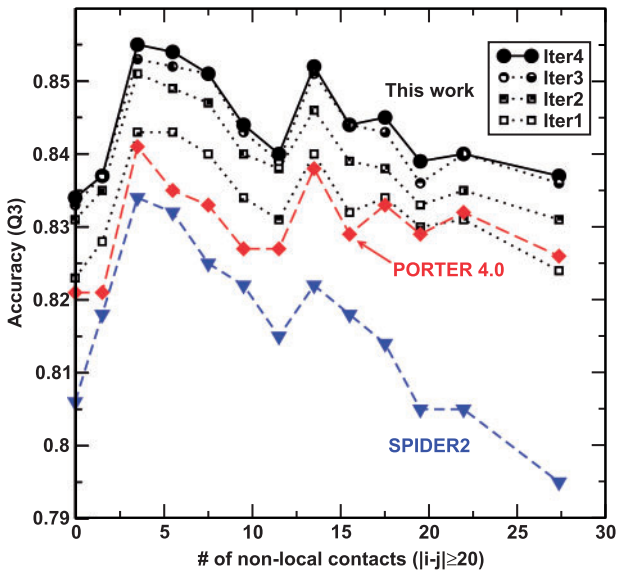

Fig. 4. The secondary structure accuracy as a function of number of non-local contacts ($|i-j| \geq 20$) for four iterations as compared to the SPIDER2 and PORTER 4.0 results for the independent test set (TS1199)

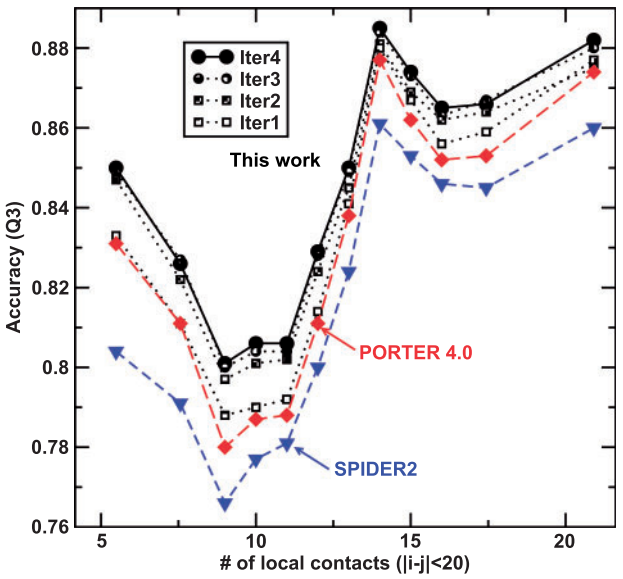

Fig. 5. The secondary structure accuracy as a function of number of local contacts ($|i-j| < 20$) for four iterations as compared to the SPIDER2 and PORTER 4.0 results for the independent test set (TS1199)

long-range interactions. In contrast, the three-state accuracy given by this work has a much slower decay when the number of non-local contacts is >5, as shown in Figure 4. The improvement in Q3 over SPIDER2 is about 2% for low numbers of non-local contacts, and 3–4% for high numbers of non-local contacts. This clearly indicates that capturing non-local interactions by LSTM-BRNNs is the key factor in the improvement. Interestingly, PORTER 4.0, a BRNN-based method, can also captures non-local interactions when it is compared to the result of iteration 1. It seems that LSTM-BRNN is better than BRNN in correctly detecting residues with intermediate numbers of non-local contacts (5–12). In addition, Figure 5 further shows that improving accuracy of residues with non-local interactions is also accompanied by improving accuracy of residues with local interactions.

Capturing non-local interactions is also the key for improving prediction of the non-local structural property: ASA. Figure 6 illustrates the dependence of the average MAE for ASA on non-local contacts in a residue. Only small improvement is observed when the number of non-local contacts is <12. As the number of non-local contacts further increases, the improvement becomes greater and greater, with the largest improvement at the largest number of non-local contacts numbers collected for statistics. Interestingly the forth iteration is able to improve ASA for residues with a larger number of non-local interactions.

The ultimate purpose of local angle prediction is to predict three-dimensional structures of proteins. Real-value prediction of angles allows us to construct backbones from predicted angles. Figure 7 plots the fractions of constructed models of which root-mean-squared distance from the corresponding native structure is less than a given value. The fraction of constructed three-dimensional structures with a similar conformation [RMSD $\leq$ 6Å (Reva *et al.*, 1998)] for all 182725 40-residue fragments in a dataset of 1199 proteins is 16.3% predicted by ϕ and ψ angles and 19.1% predicted by θ and τ angles by SPIDER2. By comparison, the LSTM-BRNN model achieved 23.7% and 27.3%, respectively. This is a significant 7–8% absolute improvement. We note that using θ and τ angles yields more accurate models. This is likely due to the fact that θ and τ angles have longer range information (3–4) residues than ϕ and ϕ

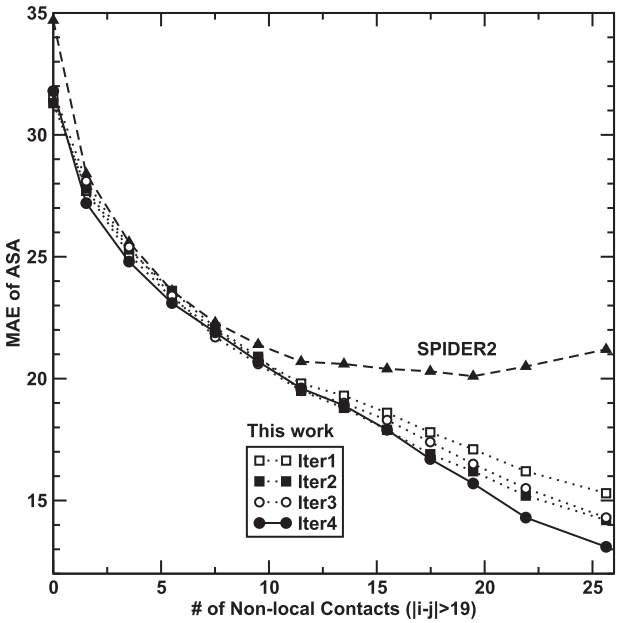

Fig. 6. The average mean absolute error of solvent accessible surface area as a function of average number of non-local contacts ($|i-j| > 19$) for four iterations as compared to the SPIDER2 result for the independent test set (TS1199)

angles (one residue). The inserts in Figure 7 shows one example of the improvement in model quality from 6Å RMSD by using ϕ/ψ angles to 2Å RMSD by using θ/τ. Better coil regions between the first strand and the first helix, as well as between the second strand and the last helix for the θ/τ-based model are observed.

It is of interest to compare the accuracy of predicted secondary structure with other state-of-the-art techniques. Table 2 compares the performance of this work with Jpred (Drozdetskiy *et al.*, 2015), SPINE-X (Faraggi *et al.*, 2012), PSIPRED 3.3 (Jones, 1999), SCORPION (Yaseen and Li, 2014), PORTER 4.0 (Mirabello and Pollastri, 2013), and DeepCNF (Wang *et al.*, 2016) using the TS115

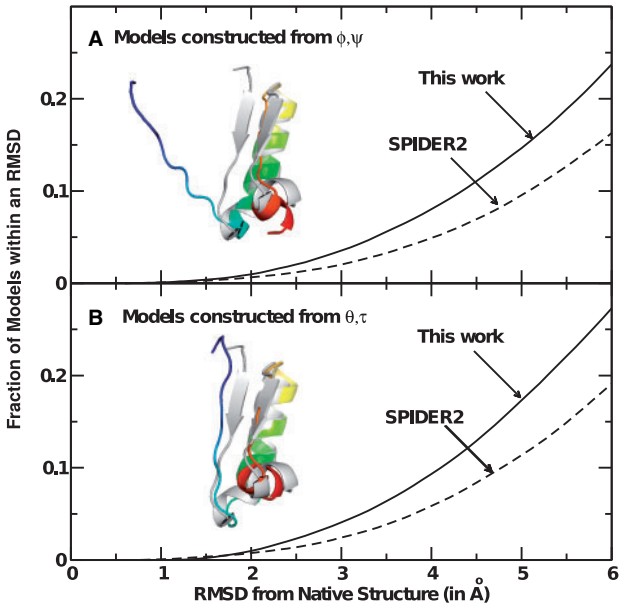

Fig. 7. The fraction of 40-residue structural fragments constructed from ϕ/ψ angles (**A**) or θ/τ angles (**B**) of which root-mean-squared distance from the corresponding native structure is less than a given value (X-axis) for all 182,724 40-residue fragments in the independent test set (TS1199). The result of the current study is compared to that of SPIDER2 as labelled. Inserts demonstrated one example of native structure (Residues 941 to 980 of carbamoyl phosphate synthesize or the chain A of PDB 1a9x shown in grey) in comparison to the model (shown in rainbow) constructed from predicted ϕ/ψ angles (A) and θ/τ angles (B) with RMSD = 6Å and 2Å, respectively.

Table 2. Method comparison using newly released structures (TS115)

Method	TS115	
	Q3 (%)	*P*-value[a]
Jpred 4 [b]	77.1	1.2e-8
SPINE-X	80.1	2.1e-14
PSIPRED 3.3	80.2	3.9e-10
SCORPION	81.7	6.4e-7
PORTER 4.0	82.0	1.3e-5
SPIDER2	81.9	1.3e-7
DeepCNF	82.3	0.0251
This work	83.9	–

[a]Paired *t*-test from this work.

[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

set of recently determined protein structures. The results of these methods were obtained recently in the review paper (Yang *et al.*, 2016). The Q3 accuracies are 77.1% for Jpred4, 80.1% for SPINE-X, 80.2% for PSIPRED 3.3, 81.7% for SCORPION, 81.9% for SPIDER2, 82.0% for PORTER 4.0, and 82.3% for DeepCNF, compared to 83.9% for this work. The differences between this work and all other methods are statistically significant ($P < 0.05$).

Table 3 shows the confusion matrices for SPIDER 3, PORTER 4.0 and DeepCNF results when tested with TS115 sequences. For the TS115 set this work achieves Q_3 83.9%, Q_C 83.3%, Q_E 76.1% and Q_H 88.1%. For comparison PORTER 4.0 achieves Q_3 82.0%, Q_C 80.8%, Q_E 75.5% and Q_H 86.2%; and DeepCNF achieves Q_3 82.3%, Q_C 84.2%, Q_E 77.6% and Q_H 82.9%. This indicates that

Table 3. Confusion matrix of this work, PORTER 4, and DeepCNF tested with TS115

Method		C (%)	E (%)	H (%)
This work	C	83.3	22.0	11.2
	E	6.2	76.1	0.7
	H	10.5	1.9	88.1
PORTER 4.0	C	80.8	22.6	12.8
	E	8.0	75.5	1.0
	H	11.2	1.8	86.2
DeepCNF	C	84.2	21.0	16.5
	E	7.7	77.6	0.6
	H	8.2	1.4	82.9

our method has the highest accuracy in helical prediction with slightly lower accuracy for coil and sheet prediction than DeepCNF and slightly higher accuracy for coil and sheet prediction than PORTER 4.0.

4 Discussion

This study applied Long Short-Term Memory based neural networks to the prediction of one-dimensional structural properties of proteins, ranging from secondary structure, backbone torsion angles, to solvent accessible surface area. LSTM-BRNNs were employed because of their demonstrated capability in capturing long-range interactions between distant events in a range of applications, from speech recognition (Xiong *et al.*, 2016), to natural language processing (Sundermeyer *et al.*, 2012), and handwriting recognition (Graves and Schmidhuber, 2009), as well as their successful application to protein intrinsic disorder prediction (Hanson *et al.*, 2017). Indeed, the new method was able to make a 3% improvement in secondary structure prediction in the large training set of 1199 proteins over SPIDER2, based on an identical training set. This overall 84% accuracy is further confirmed by an additional independent test set of 115 recently determined protein structures. This result is significantly better than the highest reported secondary-structure accuracy of DeepCNF, who employed a deep 5-layer convolutional neural field model (deep convoluational neural networks, followed by a conditional random field) and a larger training set of 5600 proteins.

What is more significant is perhaps the ability of achieving a correlation coefficient of 0.8 between predicted and actual solvent accessible surface areas. This is remarkable because the previous record of 0.76 by SPIDER2 was considered close to the theoretical limit. ASA prediction is challenging not only because it is a global structural property but because solvent accessible surface areas are not highly conserved. It was shown that the correlation coefficient of solvent accessibility between homologs is only 0.77 (Rost and Sander, 1994). Exceeding the limit posed by homologous structures indicates the role played by the real sequence information, such as physio-chemical properties of the actual protein sequence.

The ultimate test for the method developed here is whether or not it will be useful for predicting three-dimensional structures. We directly employed predicted angles to construct three-dimensional models of 40 residue fragments without using any energy functions for refinement or optimization. Two structures are considered as in a similar conformation if the RMSD between them is <6Å. We showed that 24% of structures built by ϕ and ϕ are similar to their native structures. More than one quarter (27%) are in a similar structure when θ and τ are employed. Such large numbers of correct prediction means that not only helical and sheet, but more

importantly, coil regions of many 40-residue fragments were predicted with high precision.

The challenge of accounting for non-local interactions was the biggest barrier that has limited the accuracy of predicted protein structural properties for many years. LSTM-BRNN successfully learns long-range interactions without the use of a fixed sliding window. Breaking the impasse by using LSTM-BRNNs, signals that the future of the protein folding problem, represented protein contact map predictions, may be solvable by employing a powerful machine-learning technique, trained on a large dataset. The work in this area is in progress. To further confirm the performance of SPIDER3, we have collected a larger independent test set for structures released between June, 2015 and February, 2017 with <30% sequence identity in between or to any proteins deposited before the time. The dataset contains 673 proteins, totally 171003 residues. SPIDER3 produces an Q3 accuracy of 83.74%, compared to 81.6% by DeepCNF (*P*-value<1E-99). This >2% difference is similar to the previous result in the smaller test set of 115 proteins.

Funding

This work was supported in part by National Health and Medical Research Council (1059775 and 1083450) of Australia to Y.Z. 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).

Conflict of Interest: none declared.

References

Abadi,M. *et al*. (2016) Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv:1603.04467*.

Adamczak,R. *et al*. (2004) Accurate prediction of solvent accessibility using neural networks-based regression. *Proteins*, **56**, 753–767.

Ahmad,S. *et al*. (2003) Real value prediction of solvent accessibility from amino acid sequence. *Proteins*, **50**, 629–635.

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.

Amodei,D. *et al*. (2015) Deep speech 2: End-to-end speech recognition in english and mandarin. *arXiv:1512.02595*.

Baldi,P. *et al*. (1999) Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, **15**, 937–946.

Dill,K.A., and MacCallum,J.L. (2012) The protein-folding problem, 50 years on. *Science*, **338**, 1042–1046.

Dor,O., and Zhou,Y. (2007) Real-SPINE: an integrated system of neural networks for real-value prediction of protein structural properties. *Protein*, **68**, 76–81.

Drozdetskiy,A. *et al*. (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res*., **43**, W389–W394.

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. Comput. Chem*., **33**, 259–267.

Fauchère,J.L. *et al*. (1988) Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Pept. Protein Res*., **32**, 269–278.

Garg,A. *et al*. (2005) Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins*, **61**, 318–324.

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.

Gilis,D., and Rooman,M. (1997) Predicting protein stability changes upon mutation using database-derived potentials: solvent accessibility determines the importance of local versus non-local interactions along the sequence. *J. Mol. Biol*., **272**, 276–290.

Graves,A., and Schmidhuber,J. (2009) Offline handwriting recognition with multidimensional recurrent neural networks. In: Koller D., Schuurmans D., Bengio Y., and Bottou L., editors, *Advances in Neural Information Processing Systems 21*. Curran Associates, Inc., Red Hook, NY, p.545–552.

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*. (2017) Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*, **33**, 685–692.

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

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

Holbrook,S.R. *et al*. (1990) Predicting surface exposure of amino acids from protein sequence. *Protein Eng*., **3**, 659–665.

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

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.

Kingma,D., and Ba,J. (2014) Adam: a method for stochastic optimization. *arXiv:1412.6980*.

Kinjo,A.R., and Nishikawa,K. (2006) CRNPRED: highly accurate prediction of one-dimensional protein structures by large-scale critical random networks. *BMC Bioinformatics*, **7**, 401.

Kuang,R. *et al*. (2004) Protein backbone angle prediction with machine learning approaches. *Bioinformatics*, **20**, 1612–1621.

LeCun,Y. *et al*. (2006) A tutorial on energy-based learning. In: Bakir, G., Hofman, T., Schölkopf, B., Smola, A., Taskar, B. (eds.) *Predicting Structured Data*. MIT Press, Cambridge.

Lee,B., and Richards,F.M. (1971) The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol*., **55**, 379–400.

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

Magnan,C.N., and Baldi,P. (2014) SSpro/ACCpro 5: almost perfect prediction of protein secondary structure and relative solvent accessibility using profiles, machine learning and structural similarity. *Bioinformatics*, **30**, 2592–2597.

Mirabello,C., and Pollastri,G. (2013) Porter, PaleAle 4.0: high-accuracy prediction of protein secondary structure and relative solvent accessibility. *Bioinformatics*, **29**, 2056–2058.

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.

Pollastri,G. *et al*. (2002a) Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins*, **47**, 228–235.

Pollastri,G. *et al*. (2002b) Prediction of coordination number and relative solvent accessibility in proteins. *Proteins*, **47**, 142–153.

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) Review: protein secondary structure prediction continues to rise. *J. Struct. Biol*., **134**, 204–218.

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

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

Song,J. *et al*. (2008) HSEpred: predict half-sphere exposure from protein sequences. *Bioinformatics*, **24**, 1489–1497.

Srivastava,N. *et al*. (2014) Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res*., **15**, 1929–1958.

Sundermeyer,M. *et al*. (2012) LSTM neural networks for language modeling. In: *Proceedings Interspeech*. p.194–197.

Tuncbag,N. *et al*. (2009) Identification of computational hot spots in protein interfaces: combining solvent accessibility and inter-residue potentials improves the accuracy. *Bioinformatics*, **25**, 1513–1520.

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

Wood,M.J., and Hirst,J.D. (2005) Protein secondary structure prediction with dihedral angles. *Proteins*, **59**, 476–481.

Xiong,W. *et al*. (2016) Achieving human parity in conversational speech recognition. *arXiv:1610.05256*.

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

Yang,Y. *et al*. (2016) Sixty-five years of long march in protein secondary structure prediction: the final stretch? *Brief. Bioinform*., DOI: 10.1093/bib/bbw129.

Yaseen,A., and Li,Y. (2014) Context-based features enhance protein secondary structure prediction accuracy. *J. Chem. Inf. Model*., **54**, 992–1002.

Yuan,Z. (2005) Better prediction of protein contact number using a support vector regression analysis of amino acid sequence. *BMC Bioinformatics*, **6**, 248.

Yuan,Z., and Huang,B. (2004) Prediction of protein accessible surface areas by support vector regression. *Proteins*, **57**, 558–564.

Zhou,Y., and Faraggi,E. (2010) Prediction of one-dimensional structural properties of proteins by integrated neural networks. In: Rangwala, R. and Karypis, G. (ed.) *Introduction to Protein Structure Prediction*, Chap. 4, John Wiley & Sons, Inc., Hoboken, NJ. p.45–74.

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