Single-Sequence-Based Prediction of Protein Secondary Structures and Solvent Accessibility by Deep Whole-Sequence Learning

Rhys Heffernan,[a] Kuldip Paliwal,[a] James Lyons,[a] Jaswinder Singh,[a] Yuedong Yang,[b] and Yaoqi Zhou,*[c]

Predicting protein structure from sequence alone is challenging. Thus, the majority of methods for protein structure prediction rely on evolutionary information from multiple sequence alignments. In previous work we showed that Long Short-Term Bidirectional Recurrent Neural Networks (LSTM-BRNNs) improved over regular neural networks by better capturing intra-sequence dependencies. Here we show a single-sequence-based prediction method employing LSTM-BRNNs (SPIDER3-Single), that consistently achieves Q3 accuracy of 72.5%, and correlation coefficient of 0.67 between predicted and actual solvent accessible surface area. Moreover, it yields reasonably accurate prediction of eight-state secondary structure, main-chain angles (backbone ϕ and ψ torsion angles and Cα-atom-based θ and τ angles), half-sphere exposure, and contact number. The method is more accurate than the corresponding evolutionary-based method for proteins with few sequence homologs, and computationally efficient for large-scale screening of protein-structural properties. It is available as an option in the SPIDER3 server, and a standalone version for download, at http://sparks-lab.org. © 2018 Wiley Periodicals, Inc.

DOI:10.1002/jcc.25534

Introduction

The protein folding paradigm states that a protein's structure is determined solely by its sequence. This is a fundamentally important concept, as protein structures provide the key to their functional mechanisms. Experimental techniques for solving protein structures, such as X-ray crystallography, nuclear magnetic resonance, and cryo-electron microscopy are time-consuming, labor-intensive, and costly; leading to a situation where the rate at which new proteins are discovered far outpaces our ability to experimentally determine their structures. This adds an urgency to finding a method to computationally predict a protein's structure, from its primary sequence alone. Despite its importance, the protein folding problem has remained unsolved[1–3] since its inception over half a century ago[4]. This is due to the difficulty in formulating an accurate energy function for the solvent-mediated interaction between amino acid residues.[2]

While it is challenging to predict a protein's structure directly from its sequence alone, accurate structure prediction can now be made using restraints derived from correlated mutations located from Multiple Sequence Alignments (MSAs) of homologous sequences, if a large number of homologous sequences are known.[5] Similarly, the accuracy of predicting protein secondary structure, an important subproblem of protein structure prediction, increased from approximately 60% by early single-sequence-based techniques[6] to beyond 70% with the introduction of evolutionary information from MSA,[7] and to 82%–84% with the latest deep long-range learning techniques also with evolutionary information as the key input.[3,8–10]

However, the majority of proteins (>90%) have few, if any, known homologous sequences.[5] In these cases, evolutionary information is limited or non-existent, and poor prediction accuracy is expected. It is quite possible that inaccurate evolutionary information might reduce the accuracy of prediction. Indeed, a recent single-sequence-based prediction of solvent Accessible Surface Area (ASA by ASAquick) is more accurate than evolution-profile trained methods for those proteins with few homologous sequences.[11] Thus, it is possible that one can simply improve prediction accuracy by the alternative use of single-sequence and evolution-based methods, depending on the size of the homologous sequence cluster for a given protein.

Moreover, such single-sequence-based prediction is computationally efficient because greater than 99% of the computational time is spent on generating evolutionary sequence profiles. Increasingly inexpensive sequencing techniques have led to an exponential increase in the number of known sequences. As a result, the computational time requirement for finding sequence profiles is continuing to increase. For

[a] R. Heffernan, K. Paliwal, J. Lyons, J. Singh
Signal Processing Laboratory, Griffith University, Brisbane, QLD, 4111, Australia
[b] Y. Yang
School of Data and Computer Science, Sun Yet-Sen University, Guangzhou, China
[c] Y. Zhou
Institute for Glycomics and School of Information and Communication Technology, Griffith University, Southport, QLD, 4222, Australia
E-mail: yaoqi.zhou@griffith.edu.au

Contract grant sponsor: This work was supported by Australian 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.

example, PSSM generation, by PSI-BLAST,[12] can take in the region of 30 min for a short protein (around 100 resides) up to multiple hours for a longer sequence (around 1000 residues). More importantly, a single-sequence-based prediction directly addresses the original subproblem of protein structure prediction: how far can we push the accuracy of predicting protein secondary structure from its sequence alone?

Although secondary structure prediction is dominated by methods based on evolutionary information, progress has been make in single-sequence-based prediction. SIMPA is a method employing a nearest neighbor model, which reported a Q3 of 67.7%.[13] BSPSS is a Bayesian solution which succeeds in incorporating non local information and reports a Q3 of 68.8%.[14] Kloczkowshi et al. presented a single sequence version of the GOR V method which reports a Q3 accuracy of 67.5%.[15] IPSSP extends upon BSPSS with a number of changes including iterative training and improved residue dependency model reporting a Q3 of 70.3%.[16] Bidargaddi et al. combined BSPSS with a Neural Network (NN) to reach a Q3 of 71%.[17] Bouziane et al. used an ensemble of Support Vector Machine (SVM) and NN models to achieve Q3 results around 66%.[18] The popular PSIpred also maintains a single sequence version (hereon referred to as PSIpred-Single) which achieves results around 70%.[19] All of these methods, however, have not yet taken advantage of recent computational advancements in deep learning.[20–22]

Recently, we employed Long Short-Term Memory (LSTM) Bidirectional Recurrent Neural Networks (BRNNs) to predict several one-dimensional structural properties of proteins by iterative learning.[9] LSTM networks take the whole protein sequence as input, rather than a sliding window, to capture long-range interactions between residues that are close in three-dimensional space, but far from each other in sequence positions. This allows the method (SPIDER3) to achieve the highest reported accuracy for several one-dimensional structural properties such as secondary structure (~84% Q3), solvent accessibility (with a correlation coefficient between predicted and actual solvent accessibility at 0.8), and backbone torsion angles (e.g., a mean absolute error of 27° for ψ).

In this article, we examine if LSTM-BRNNs can make more accurate single-sequence prediction of secondary structure than existing techniques, by using the same iterative deep learning technique as SPIDER3. Unlike previous single-sequence-based methods, this method (named SPIDER3-Single) will not only predict secondary structure in three states, but also in all eight states. Moreover, we will predict four backbone torsion angles (ϕ, ψ, θ, and τ), solvent accessibility, Half Sphere Exposure (HSE), and Contact Number (CN). We show that SPIDER3-Single is able to achieve reasonably accurate prediction for all structural properties predicted.

Methodology

Datasets

Our dataset was obtained by downloading 30% non-redundant sequences from cullpdb in February 2017, with a resolution of less than 2.5 Å, and *R*-factor less than 1.0. Of these, we removed sequences with a length of less than 30 residues, a similarity greater than 25% according to BlastClust,[12] and any sequences with incomplete information. This resulted in 12,442 sequences, which were split into 11192 submitted before and 1250 after June 2015 (TS1250). The 11,192 sequences submitted before June 2015 were used for training and evaluation by separating into a training set of 9993 proteins called TR9993, and a test set of 1199 proteins called TS1199. The TR9993 set was used for 10-fold cross validation, in which the set was divided into 10 sets with each employed in turn as the test set and the remainder as the training set. We obtained a more difficult subset from TS1250 by further excluding potential homologous proteins from the training set with a PSI-Blast *E*-value cutoff of 0.1. This subset, labeled TS-Hard, has 280 sequence.

Input features

Each single sequence is represented by a one-hot vector of size $20 \times L$, where L is the length of the protein chain. Although single-sequence derived features such as physiochemical properties[8,9] and the BLOSUM matrix[23] can also be used, we found that these did not offer any improvement. This is likely because these features are simply linear weight matrices that are learnable by the neural network employed here. Thus, only the one-hot vector is employed.

Outputs

The secondary structure assignment program Define Secondary Structure of Proteins (DSSP) specifies eight secondary structure states, comprised of two helix, two strand, and three coil states.[24] The three helix states are: 3_{10}-helix (G), alpha-helix (H), and pi-helix (I); the strand states are: beta-bridge (B) and beta-strand (E); and the three coil types are high curvature loop (S), beta-turn (T), and coil (C). These eight states are converted into a three-state problem using the following conversion: G, H, and I to H; B and E to E; and S, T, and C to C. We designed our network to independently predict 8 and 3 state secondary structure for a total of 11 output nodes.

In addition to secondary structure, our method also predicts Accessible Surface Area (ASA), ϕ, ψ, θ, τ, Half Sphere Exposure (HSE), and Contact Number (CN). ASA is a measure of the level of exposure of each residue within the protein to solvent (water). Active sites of proteins are often located on their surface, therefore knowing the level of exposure of each residue can provide insight as to where that activity might occur. ϕ and ψ are two of the backbone torsion angles, along with ω, that describe a protein's local backbone structure. ω is not predicted here because it is usually at 180 due to the planarity of the peptide bond. θ and τ angles describe the orientation between neighboring residues according to Cα atoms. Specifically θ is the angle between $C\alpha_{i-1} - C\alpha_i - C\alpha_{i+1}$, and τ is the dihedral angle rotated about the $C\alpha_i - C\alpha_{i+1}$ vector.[25] CN is the count of the number of residues within a distance cutoff, in three-dimensional space, of a given residue. HSE extends the idea of CN by adding directionality information and differentiating

between counts in a top and bottom half of the sphere.[26] Hamelryck described two methods for defining the plane separating the upper and lower hemisphere, HSEα based on the neighboring Cα–Cα directional vector and HSEβ based on the Cα–Cβ directional vector. In this work we use the HSEα definition. For CN and HSE, residue distance is defined as the distance between Cα atoms with a 13 Å cutoff.

This leads to another 12 outputs: the first for ASA; the next eight nodes for sin(ϕ), cos(ϕ), sin(ψ), cos(ψ), sin(θ), cos(θ), sin(τ), and cos(τ), respectively; the next two for HSEα-up, HSEα-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.[25] The sine and cosine predictions are converted back to angles by the equation α = tan − 1[sin (α)/cos(α)].

Model details

SPIDER3-Single utilizes the same LSTM-BRNN networks that we demonstrated in our previous evolutionary-profile based work, SPIDER3.[9] Briefly, we use two BRNN layers with 256 nodes per direction, per layer; followed by two fully-connected hidden layers with 1024 and 512 nodes, respectively (Figure 1). In the BRNN layers, we employ LSTM cells for their ability to learn both distant and close intra-sequence dependencies (Figure 2).[27] LSTM cells can remember both the long and short-ranage interactions by enforcing the constant error flow regardless of sequence seperation, thus permitting the input of the entire protein sequence. The implementation of the network was done using Google's Tensorflow library, using the CUDA version on an Nvidia GeForce Titan X GPU to speed up training.[28]

The input to the model is simply an $L \times 20$ matrix of one-hot feature vectors, where L is the sequence length. This input is then provided to one of two distinct networks, one for the prediction of secondary structure, 3 and 8 states separately, and one for the remainder of the structural properties: ASA, HSEα, CN, ϕ, ψ, θ, and τ. As shown in Figure 1, these predictions are split in such a way that the classification and regression problems are performed by two separate networks. This allows each of those networks to have different loss functions that are better suited to the task.[29] For the classification of secondary structure, the network uses cross-entropy loss, while the regression performed for the rest of the predictions are better served by a square loss.

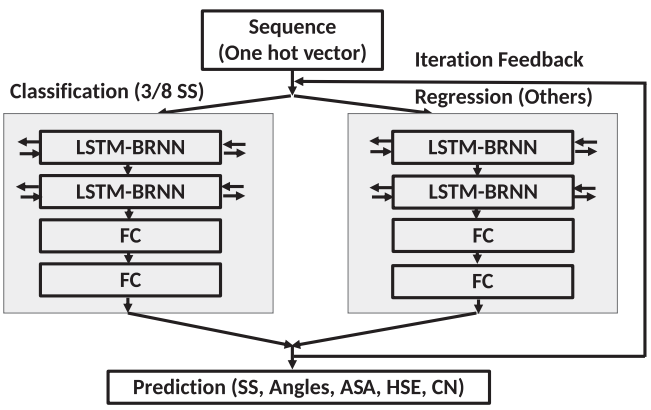

Figure 1. Overview of model.

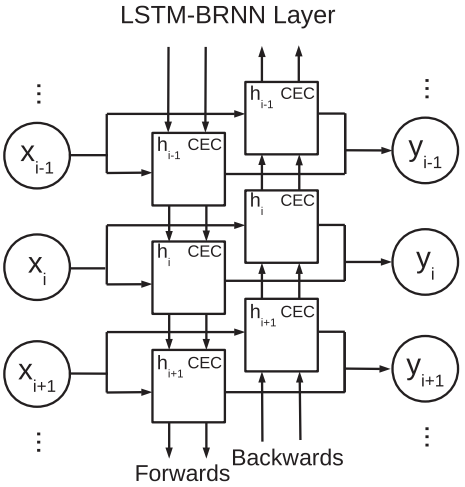

Figure 2. LSTM-BRNN layer architecture connected with the contant error carousel (CEC).

During training both of these networks mask any contribution to the loss made by any undefined labels. These undefined labels include residues with no secondary structure assignment according to the program DSSP,[24] or residues with no defined ϕ, ψ, θ, or τ. In these instances, the residue itself is not ignored, only any missing labels.

Iterative based methods have previously been shown to have an increased accuracy over using the same models without iterations.[8,9,25,30,31] SPIDER3 employed four iterations. Here the networks employed the one-hot vectors as the only input, and secondary structure or ASA,HSE,CN, ϕ, ψ, θ, and τ as the output in the first iteration. Then, the output of the first iteration, from both networks (classification and regression), is appended to the one-hot features as the input to a second pair of networks, which predict the same outputs again (Figure 1). This process of adding one iteration's output to the following iteration's input is repeated for as long as the results continue to improve.

It should be noted that care needs to be taken when using this iterative procedure so as to not test the network on data that it has already seen during training. This is an issue because to be able to train a subsequent iteration, we must have predictions (i.e., outputs from the proceeding iteration) for the training data. If one was to simply pass the training data through the first iteration networks, those predictions would be artificially high because the network has seen that data during training. To overcome this potential overtraining issue, the training set is split into 10 folds, and 10 different networks are trained. Each network is trained on nine of the folds, and tested on the remaining one. In this way we are able to make predictions for each of those folds one at a time, using networks that have not been trained with that data.

For SPIDER3-Single we stop after the third iteration. However, unless specifically mentioned we will only report the results from iteration 2 because of the insignificant difference between the second and third iterations for the majority of outputs.

Performance measure

Secondary structure prediction accuracy is defined as the percentage of residues for which the state is correctly predicted; Q3 for three state prediction, and Q8 for eight state prediction. For ASA,[30] CN,[31,32] and HSE[33] the Pearson Correlation Coefficient (PCC) between the true and predicted values are calculated and reported. The accuracy of ϕ, ψ, θ, and τ angles are reported as the Mean Absolute Error (MAE) between the true and predicted angle vales. The MAE of the angles is defined as the smaller of α_i and 360 − α_i, where α_i is $|\alpha_i^{\text{pred}} - \alpha_i^{\text{true}}|$, to account for the periodicity of the angles.

Results

Table 1 shows the 10-fold cross validation results of the network. It can be seen that all of the results improve between the first and second iterations, but plateau by the third. Specifically the third iteration's results degrade all except ϕ, ψ, θ, and τ where there was no change for ϕ and θ, and a minor MAE improvement (0.1) for ϕ and τ. Thus, here and hereafter, we will only report the result of the second iteration.

In Table 1 we note that three-state secondary structure converted from 8-state prediction is not as accurate as direct prediction in three states. Thus, hereafter we will report the result from the direct prediction of three state only for three-state secondary structure.

Table 2 examines the method's performance across different datasets. For 10-fold cross validation and two large independent test sets (TS1199 and TS1250), the performance is essentially the same. For example, Q3 varies from 72.4% to 72.6%, Q8 from 59.8% to 60.1%, CC for ASA from 0.66 to 0.67, MAE for ψ from 43.5 to 43.8 degrees. The consistency of this performance indicates SPIDER3-Single can provide equally accurate predictions for unseen data.

Table 2 also shows the result for TS-Hard, the subset of TS1250 after removing any potential homologs to the training set using PSI-Blast according to an *E*-value cutoff of 0.1. Interestingly, the overall performance is better for some structural properties. For example, Q3 increases from 72.5% to 73.2% although the CC of ASA decreases slightly from 0.67 to 0.66. By comparison, evolution-profile-based SPIDER3 performs worse on TS-Hard than TS1250 (Table 2) where Q3 decreases from 84.3% for TS1250 to 81.9% for TS-Hard. Table 3 shows the mean accuracy of three state secondary structure prediction across each protein in the data set, along with the standard deviation of those accuracies, and the *p*-value of a paired *t*-test between the SPIDER3-Single predictions and that of SPIDER3 and PSIpred-Single. The *p*-values show that the results of SPIDER3-Single are significantly different from those of the other two methods. This suggests that SPIDER3-Single is less dependent how many homologous sequences a protein has, as TS-Hard has a higher proportion of proteins with fewer homologous sequences.

To confirm the above possibility, Figure 3 plots Q3 as a function of the number of effective homologous sequences (Neff), a parameter used in HHblits to measure the size of homologous sequence cluster.[34] Indeed, SPIDER3 has a systematic reduction in accuracy as Neff decreases whereas Q3 for SPIDER3-Single is mostly independent of Neff. Interestingly, Q3 for SPIDER3-Single is higher than Q3 for SPIDER3 for Neff less than 1.5. This confirms that a lack of homologous sequences is detrimental to the accuracy of evolutionary-profile based methods.

Figure 3 also shows the result of PSIpred-Single. Similar to SPIDER3-Single, it has a Q3 mostly independent of Neff. SPIDER3-Single consistently outperforms PSIpred-Single across the entire range of Neff with an overall 3% improvement (from 69.6% to 72.5%) in Q3.

Similar results are also observed for ASA. As shown in Figure 4 SPIDER3-Single outperforms SPIDER3 for Neff less than 1.5 and is mostly independent of Neff whereas SPIDER3 performs the best for proteins with a higher Neff. Figure 4 also compares SPIDER3-single with the single-sequence method ASAquick. SPIDER3-Single consistently outperforms ASAquick across the entire range of Neff in correlation coefficients.

The improved performance of SPIDER3-Single over SPIDER3 for Neff less than 1.5 is also observed for all other commonly predicted one-dimensional structural properties (backbone angles, HSEα-up, and contact numbers) except HSEα-down. These results are shown in Supporting Information Figures S1–S5. This performance difference suggests a possible consensus

Table 1. Accuracy of the 10-fold cross validation by SPIDER3-Single in three iterations with the best performance highlighted, according to the fraction of residues in correctly predicted three and eight state (Q3 and Q8), Pearson Correlation Coefficients (CC), and Mean Absolute Error (MAE).

	it. 1	it. 2	it. 3
Q3	71.81%	**72.42%**	**72.36%**
Q3 from 8 state	70.28%	**71.08%**	**70.96%**
Q8	59.30%	**60.07%**	**59.97%**
ASA (CC)	0.666	**0.670**	**0.669**
HSEα-up (CC)	0.611	**0.612**	**0.610**
HSEα-down (CC)	0.551	**0.566**	**0.562**
CN (CC)	0.638	**0.643**	**0.640**
ϕ (MAE)	24.6	24.3	**24.2**
ψ (MAE)	44.5	**43.8**	**43.8**
θ (MAE)	11.0	10.7	**10.6**
τ (MAE)	46.6	**45.8**	**45.8**

Table 2. SPIDER3-Single performance across different datasets according to fraction of residues in correctly predicted three and eight states (Q3 and Q8), Pearson Correlation Coefficients (CC), and Mean Absolute Error (MAE).

	10-fold	TS1199	TS1250	TS-Hard
Q3	72.42%	72.56%	72.52%	73.24%
Q8	60.07%	60.11%	59.80%	61.72%
ASA (CC)	0.670	0.671	0.666	0.661
HSEα-up (CC)	0.612	0.612	0.606	0.604
HSEα-down (CC)	0.566	0.568	0.565	0.559
CN (CC)	0.644	0.643	0.638	0.636
ϕ (MAE)	24.3	24.5	24.1	23.4
ψ (MAE)	43.8	43.5	43.5	42.4
θ (MAE)	10.7	11.3	11.3	11.1
τ (MAE)	45.8	45.8	46.0	44.7

Table 3. Performance of secondary structure prediction by SPIDER3, PSIpred-Single, and SPIDER3-Single, on two datasets at the residue level.

	TS1250	TS-Hard
SPIDER3	84.31%	81.87%
PSIpred-Single	69.61%	70.21%
SPIDER3-Single	72.52%	73.24%

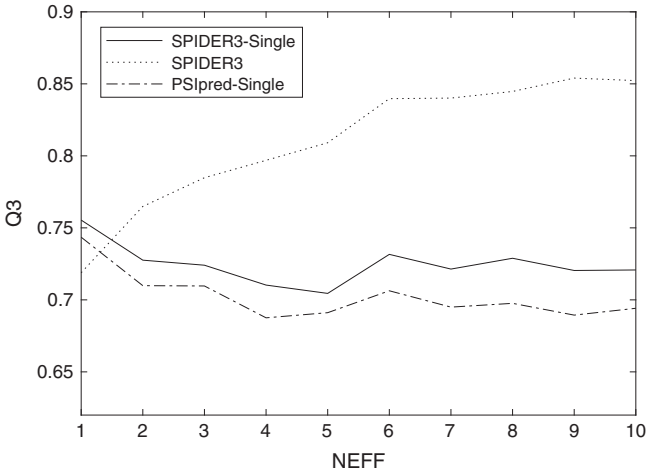

Figure 3. Accuracy of predicted secondary structure (Q3) given by SPIDER3, SPIDER3-Single, and PSIpred-Single as labeled as a function of the number of effective homologous sequences (Neff). Neff values were binned by rounding to their nearest integer value.

prediction technique: employing SPIDER3 for Neff greater than 1.5 and SPIDER3-Single for Neff ≤ 1.5. This method combination leads a minor improvement of Q3 from 84.3 from SPIDER3 to 84.4% by the consensus for TS1250 but a larger improvement from 81.9% to 82.1% for TS-Hard. Obviously, the level of improvement strongly depends on the fraction of sequences with Neff ≤ 1.5.

The significant improvement of SPIDER3-Single over previous single-sequence methods observed in Figures 3 and 4 is likely due to the ability of LSTM-BRNN to better account for inter-residue dependencies between sequentially close and distant residues. To confirm this hypothesis, Figures 5 and 6 compare the performance of secondary structure prediction by SPIDER3-Single and by PSIpred-Single as a function of number of local (|(i–j)|< 20) and long-range (|(i–j)|≥ 20) contacts, respectively, of a residue. For surface residues with few contacts (short or long-range), SPIDER3-Single and PSIpred-Single have a smaller performance difference (~2%). As the number of contacts between short and long-range sequentially separated residues increases, the difference in performance increases to 5%–6%.

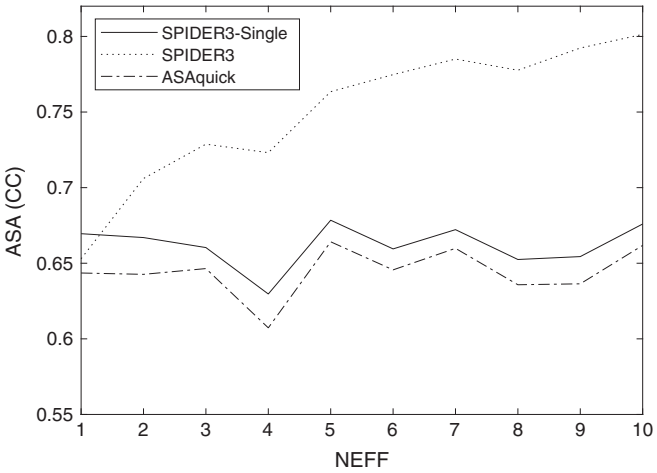

Figure 4. Accuracy of predicted ASA (measured by correlation coefficients between predicted and actual ASA) given by SPIIDER3, SPIDER3-Single, and ASAquick as labeled as a function of the number of effective homologous sequences (Neff). Neff values were binned by rounding to their nearest integer value.

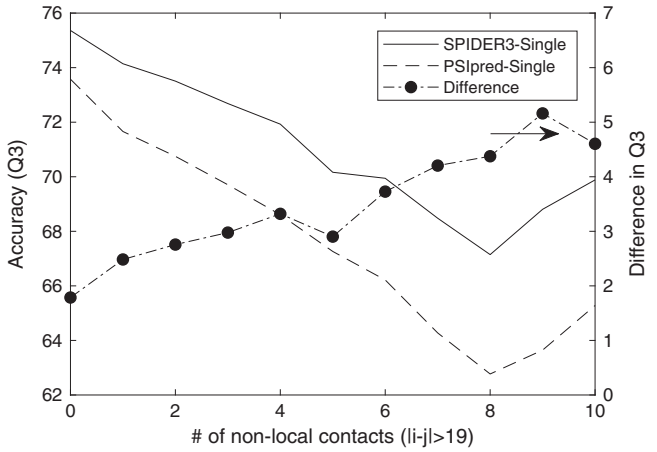

Figure 5. Performance comparison between SPIDER3-Single and PSIpred-Single for secondary structure prediction (Q3) on TS1250, as a function of residues with different number of non-local (long-range) contacts, along with the difference in Q3 (SPIDER3-Single minus PSIpred-Single) with its y-axis to the right.

These results demonstrate the power of LSTM-BRNN over the regular NN used in PSIpred-Single.

Figure 7 compares SPIDER3-Single and SPIDER3 in accuracy of the secondary structure prediction for each residue type. The overall accuracy of SPIDER3 is higher than that of SPIDER3-Single for all residue types. There is a strong correlation (CC = 0.98) between two sets of the accuracy. This

Table 4. Performance in secondary structure prediction by SPIDER3, PSIpred-Single, and SPIDER3-Single, on two datasets according to mean and standard devation at the protein level.

	TS1250			TS-Hard		
	Mean	Std	*p*-value	Mean	Std	*p*-value
SPIDER3	84.74%	0.0717	2.67e-201	82.19%	0.0874	5.75e-21
PSIpred-Single	70.78%	0.0901	7.65e-15	71.36%	0.0955	3.55e-04
SPIDER3-Single	73.65%	0.0935		74.36%	0.1015	

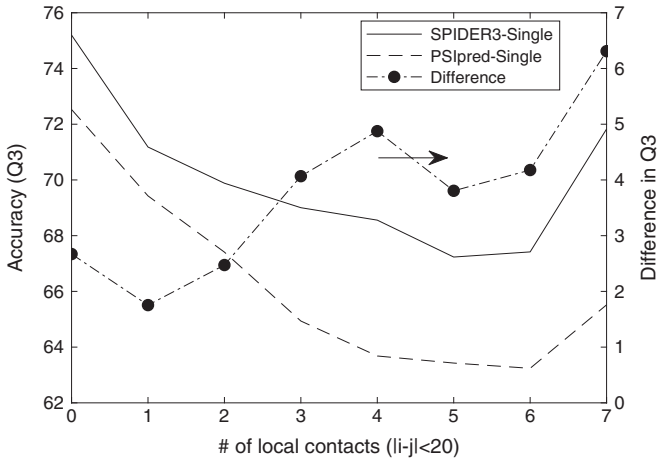

Figure 6. Performance comparison between SPIDER3-Single and PSIpred-Single for secondary structure prediction (Q3) on TS1250, as a function of residues with different number of local (short-range) contacts, along with the difference in Q3 (SPIDER3-Single minus PSIpred-Single) with its y-axis to the right.

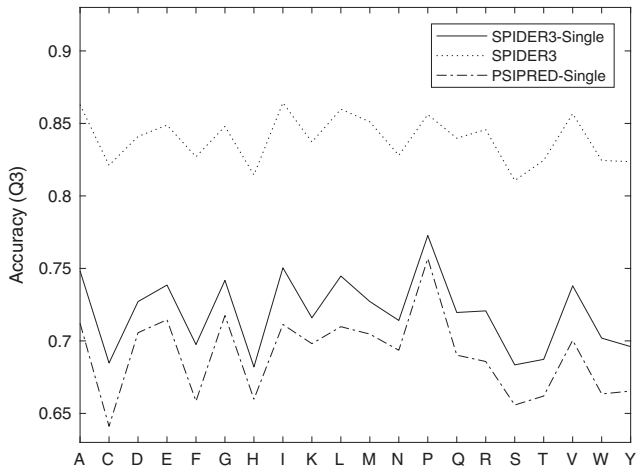

Figure 7. Accuracy of secondary structure prediction (Q3) for individual amino acids for SPIDER3-Single and SPIDER3 on the TS1250 dataset. Single letter codes were used for amino acid residues.

suggests that the predictability of each residue type is intrinsic and independent of evolution. One such intrinsic property is abundance of amino acid residue types in protein sequences, where there is a moderate correlation between prediction results and residue abundance (CC = 0.56 for SPIDER3 and CC = 0.53 for SPIDER3-Single). Interestingly, evolution reduces the fluctuation in accuracy between residue types, where the standard deviations are 0.0167 for SPIDER3 and 0.0261 for SPIDER3-Single.

Conclusions

We have developed a new single-sequence-based method for predicting several one-dimensional structural properties. This method employs the same network architecture as the previously developed SPIDER3, except that SPIDER3 utilizes evolutionary profiles generated from multiple sequence alignment. We showed that the single-sequence technique results in a more accurate prediction than the evolution-based technique when few homologous sequences are available for producing evolutionary profiles. As the majority of proteins have few homologous sequences, this computationally efficient method is expected to be useful for screening analysis of secondary structure and solvent accessibility in large scale prediction.

Acknowledgments

This work was supported by Australian 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. This research/project has also been undertaken with the aid of the research cloud resources provided by the Queensland Cyber Infrastructure Foundation (QCIF).

Keywords: secondary structure prediction · solvent accessibility prediction · contact prediction · protein structure prediction · backbone angles

How to cite this article: RhysHeffernan, K. Paliwal, J. Lyons, J. Singh, Y. Yang, Y. Zhou. *J. Comput. Chem.* **2018**, 9999, 1–7. DOI: 10.1002/jcc.25534

Additional Supporting Information may be found in the online version of this article.

[1] K. Dill, J. MacCallum, Science **2012**, 338, 1042.
[2] Y. Zhou, Y. Duan, Y. Yang, E. Faraggi, H. Lei, Theor. Chem. Acc. **2011**, 128, 3.
[3] Y. Yang, J. Gao, J. Wang, R. Heffernan, J. Hanson, K. Paliwal, Y. Zhou, Brief. Bioinform. **2018**, 19, 482.
[4] K. Gibson, H. Scheraga, Proc. Natl. Acad. Sci. U. S. A. **1967**, 58, 420.
[5] S. Ovchinnikov, H. Park, N. Varghese, P.-S. Huang, G. Pavlopoulos, D. Kim, H. Kamisetty, N. Kyrpides, D. Baker, Science **2017**, 355, 294.
[6] P. Chou, G. Fasman, Biochemistry **1974**, 13, 222.
[7] M. J. Zvelebil, G. J. Barton, W. R. Taylor, M. J. Sternberg, J. Mol. Biol. **1987**, 195, 957.
[8] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Sci. Rep. **2015**, 5, 11476.
[9] R. Heffernan, Y. Yang, K. Paliwal, Y. Zhou, Bioinformatics **2017**, 33, 2842.
[10] S. Wang, J. Peng, J. Ma, J. Xu, Sci. Rep. **2016**, 6, 18962.
[11] E. Faraggi, Y. Zhou, A. Kloczkowski, Proteins: Struct. Funct. Bioinf. **2014**, 82, 3170.
[12] S. Altschul, T. Madden, A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. Lipman, Nucleic Acids Res. **1997**, 25, 3389.
[13] J. Levin, Prot. Eng., Des. Sel. **1997**, 10, 771.
[14] S. Schmidler, J. Liu, D. Brutlag, J. Comput. Biol. **2000**, 7, 233.
[15] A. Kloczkowski, K.-L. Ting, R. Jernigan, J. Garnier, Prot. Struct. Funct. Bioinf. **2002**, 49, 154.
[16] Z. Aydin, Y. Altunbasak, M. Borodovsky, BMC Bioinform. **2006**, 7, 178.
[17] N. Bidargaddi, M. Chetty, J. Kamruzzaman, Neurocomputing **2009**, 72, 3943.
[18] H. Bouziane, B. Messabih, A. Chouarfia, Soft Comput. **2015**, 19, 1663.
[19] D. Jones, J. Mol. Biol. **1999**, 292, 195.
[20] Y. LeCun, Y. Bengio, G. Hinton, Nature **2015**, 521, 436.
[21] D. Kingma, J. Ba, *arXiv:1412.6980*, **2014**.
[22] M. Sundermeyer, R. Schlüter, H. Ney, Proc. INTERSPEECH, International Speech Communication Association (ISCA), Baixas, France, **2012**, p. 194.
[23] S. Henikoff, J. Henikoff, Proc. Natl. Acad. Sci. U. S. A. **1992**, 89, 10915.
[24] W. Kabsch, C. Sander, Biopolymers **1983**, 22, 2577.

[25] J. Lyons, A. Dehzangi, R. Heffernan, A. Sharma, K. Paliwal, A. Sattar, Y. Zhou, Y. Yang, J. Comput. Chem. **2014**, 35, 2040.
[26] T. Hamelryck, Prot.: Struct. Funct. Bioinf. **2005**, 59, 38.
[27] S. Hochreiter, J. Schmidhuber, Neural Comput. **1997**, 9, 1735.
[28] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, *arXiv:1603.04467*, **2016**.
[29] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, F. Huang, Predicting Structured Data, MIT Press, Boston, **2006**.
[30] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, Y. Zhou, J. Comput. Chem. **2012**, 33, 259.
[31] R. Heffernan, A. Dehzangi, J. Lyons, K. Paliwal, A. Sharma, J. Wang, A. Sattar, Y. Zhou, Y. Yang, Bioinformatics **2016**, 32, 843.
[32] Z. Yuan, BMC Bioinform. **2005**, 6, 248.
[33] J. Song, H. Tan, K. Takemoto, T. Akutsu, Bioinformatics **2008**, 24, 1489.
[34] M. Remmert, A. Biegert, A. Hauser, J. Söding, Nat. Met. **2012**, 9, 173.

Received: 12 February 2018
Revised: 11 May 2018
Accepted: 14 June 2018
Published online on 2 February 2018