Localnet: A Simple Recurrent Neural Network Model for Protein Secondary Structure Prediction Using Local Amino Acid Sequences Only

Shutong Yang
Ningbo University
Yuhong Wang
National Center for Advancing Translational Sciences
Kennie Cruz-Gutierrez
National Center for Advancing Translational Sciences
Fangling Wu (flwu16@fudan.edu.cn)
Ningbo University
Chuan-Fan Ding
Ningbo University

Research Article

Keywords: protein secondary structure prediction, recurrent neural network, long-term short-term memory cell, local amino acid sequences

Posted Date: January 7th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-139322/v1

Abstract

Background

Protein secondary structure prediction (PSSP) is important for protein structure modeling and design. Over the past a few years, deep learning models have shown promising results for PSSP. However, the current good performers for PSSP often require evolutionary information such as multiple sequence alignments and even real protein structures (templates), entire protein sequences, and amino acid property profiles.

Results

In this study, we used a fixed-size window of adjacent residues and only amino acid sequences, without any evolutionary information, as inputs, and developed a very simple, yet accurate RNN model: LocalNet. The accuracy for three states of secondary structures is as high as 85.15%, indicating that the local amino acid sequence itself contains enough information for PSSP, a well-known classical view. By comparing to other predictors, we also achieve an state-of-art accuracy on dataset of CASP11, CASP12 and CASP13.

Conclusion

The well-trained models are expected to have good applications in protein structure modeling and protein design. This model can be downloaded from https://github.com/lake-chao/protein-secondary-structure-prediction.

Background

Protein is a large organic molecule and acts as the main undertaker of life activities of all creation, and its function mainly depends on spatial structure through the correct conformational folding and structural transitions. Protein secondary structure forms first, and acts as the seed in determining how proteins fold [1, 2] and how fast they fold [3]. Three dimensional structures of over 150,000 proteins, mainly determined by experimental approaches such as X-ray crystallography and nuclear magnetic resonance spectroscopy, are available from Protein Database Bank (PDB) [4]. Even though having advanced at an ever-faster rate, most experimental approaches remain expensive, time-consuming, and insufficient [5]. Therefore, computational, fast and high-precision protein secondary structure prediction (PSSP) from amino acid sequences is of great significance for understanding the function of proteins in the field of bioinformatics [6–8].

The most common types of secondary structures are designated as three states (Q3) of helix, strand and coil with alphabets (H, E, C) according to defined rules [9]. And they are used as labels to evaluate the predictor's performance. Many researches of PSSP have been published for over the past 6 decades [10–13]. From the 1960s, PSSP relied on amino acid sequences only. Chou & Fasman (1974) [13] reported a short-range predictor with a Q3 accuracy of 77% (which was corrected later as 47%). Garnier (1978) [14] proposed an improved statistical algorithm and achieved an accuracy of 63%. A number of machine learning based algorithms, such as SIMPA [15], BSPSS [16], GOR-V [17] and PSIPRED [18], improved Q3

accuracy by about 10%. Since mid-1990s, additional input features, especially evolutionary information, have been utilized to improve PSSP, and these methods have become predominant today. SPIDER3 [19] utilized a recurrent neural network with long-term short-term memory cell (LSTM-BRNN) model in order to capture long-range interactions, and increased the accuracy of three-state secondary structure prediction to 84%. MUFOLD-SS [20] reported a Q3 accuracy of 84% with several parallel CNNs. Meanwhile, PORTER5 [21] employed an ensemble of BRNNs and achieved a Q3 accuracy of 84%. Short after that, PSRMS [22], DeepCNF [23], Ensemble of Contextnet [24], SecNet [25], 2DCNN-BLAST [26] pushed the accuracy limit to 84%-86%.

However, these models with good performance for PSSP often requires evolutionary information such as multiple sequence alignments and even real protein structures (templates), entire protein sequences, and amino acid property profiles as input. The SPIDER3-single [27], which was based upon entire amino acid sequences, only reported a Q3 accuracy of 72.5% using a deep neural network model of LSTM-BRNN. SPIDER3-single was inferior to its original model SPIDER3, and higher prediction accuracy seems to depend upon a good combination of evolutionary information and long-range interactions.

Deep Learning [28, 29] methods allow deep neural networks discovering the representations from raw data for specific tasks such as classification and pattern detection. Among various deep learning models, the most commonly used in bioinformatics is artificial neural network (ANN), composed of input layer, output layer, and hidden layer (Fig. 1a) and Recurrent neural network (RNN).

RNN has loops, which allow information to be passed from one step of the network to the next (Fig. 1b). In the past decade, RNN has had an incredible success in various problems such as speech recognition, language modeling, and translation, and one key to the successes is the use of the LSTM model, a special kind of RNN. LSTM model, introduced by Hochreiter & Schmidhuber, [30], includes the cell state (Fig. 1c), which is like a conveyor belt running through a sequence. LSTM allows information to flow along a sequence unchanged, enables linear interactions with each element in the sequence, and is thus able to capture and model long-range interactions.

RNN seems a natural architect for PSSP because it is created for and takes advantage of continuous sequence data. In this work, we designed a simple LSTM-RNN model, called LocalNet, for prediction of secondary structures. Unlike SPIDER3 and other high-performance models, we did not use any evolutionary information such as multiple sequence alignments and only used amino acid sequences from windows of fixed size as the input.

Results

Model optimization

In this study, the output from RNN is directly connected to the output layer. We experimented with various number of fully connected hidden layers and observed no noticeable improvement in accuracy and only worsened overfitting problem. For the LSTM cell in the RNN network, we tried various numbers of units and found that 32 units produced best accuracy with the minimum number of weights.

Multi-class cross-entropy loss is used as the cost function. For the ADAM optimizer, a learning rate of 0.001 produced a satisfactory result. The training over the training data set was terminated at 20 epochs due to noticeable overfitting after that, and it took about 18 minutes on the Linux workstation we used.

Here we plot the cost and accuracy versus epoch for the window size of 19 as an example. In Fig. 2 (right), the cost dropped significantly during the first 3 epochs, and followed by a gradual decrease. Correspondingly, the accuracy increased dramatically during the first 3 epochs followed by a gradual change. After 8 epochs, the accuracy on the training data set kept increasing, but the one on the validation data set started to degrade, apparently due to over-fitting. The optimal accuracy on validation data set is 0.836 at 8th epoch (Fig. 2 (left)). The models with the best accuracy on the validation data set were saved and used to benchmark the testing data set.

The finally minimized cost and all benchmark metrics numbers for the three data sets of training, validation and testing are given in Table 1.

Table 1
Summary of benchmarks for Q3 prediction models.

Window size	**7**	**9**	**11**	**13**	**15**	**17**	**19**	**21**
Optimal epochs	17	11	18	20	8	7	8	13
Optimized cost	3526.74	3146.50	2800.08	2561.26	2499.68	2411.42	2297.41	2149.44
Training data set accuracy	0.781	0.808	0.826	0.845	0.840	0.846	0.848	0.859
Validation data set accuracy	0.776	0.798	0.814	0.824	0.831	0.834	0.836	0.838
Test data set accuracy	0.774	0.800	0.815	0.827	0.834	0.837	0.845	0.843

As shown in Table 1, LocalNet performs generally better as window sizes increase. For validation data set, the best accuracy is reached at the window size of 21, and for test data set, the best window size is 19. We tried to extend the window size further and observed no significant improvement on prediction accuracy or even degraded performance on the validation data set. This implies that protein secondary structures are mainly determined by local sequences; long range interaction, as claimed in several literatures, does not seem necessary to achieve a good prediction accuracy.

Performance on three states of helix, strand and coil

We measured the performance of the optimal model for window size of 19 residues on CASP11[31], CASP12 [32] and CASP13[33] datasets, which contain 105, 96, and 125 domain sequences, respectively.

The performance of LocalNet is comparable among these four data sets (Fig. 3). Taken CASP11 as an example, the detailed prediction accuracies of Q3, H, E, and C are 85.0%, 92.6%, 82.2%, and 60.5%, respectively. Besides, the prediction accuracy of H is higher than 90% for CASP11, CASP12, CASP13, and Culled PDB.

Comparison of the recent predictors

We compared the performance between LocalNet and other state-of-the-art models on three independent datasets: CASP11, CASP12, and CASP13. All protein targets (template-based and free-modeling targets) were used to evaluate LocalNet and the results are listed in Table 2. For data sets of CASP11 and CASP13, LocalNet's accuracy (85.0%) is comparable to those of DCRNN, MUFOLD-SS and Ensemble of Contextnet. For CASP12, LocalNet performs worse than these three top performers with Q3 accuracy of 80.5%, but it is still better than Spider3, RaptorX and DeepProf.

DCRNN used both a deep convolutional and recurrent neural network with multiscale CNNs and three layers of BGRU, and it is much more complicated than LocalNet. DCRNN's input includes protein amino acid sequence, long-range contacts, sequence pattern, and other amino acid profiles. DCRNN's performance on the three CASP data sets is only marginally better than LocalNet, which only used a single RNN module and local amino acid sequences.

In terms of input, both SPIDER3-single and our model are based upon amino acid sequences only. The LSTM-BRNN structure of SPIDER3-single is similar to SPIDER3, but the accuracy is significantly lower. The authors contribute the accuracy of 72.5% to using the whole protein sequence as input and capturing long-range interactions between residues. LocalNet, having a much simpler structure than LSTM-BRNN, achieved better accuracy, and the sliding window strategy may account for the enhanced accuracy. By using a short window of amino acid sequence instead of the entire protein sequence as input, we are able to generate a much larger number of samples to train LocalNet. Sufficient sample size is particularly crucial for deep learning models to extract the functional relationship between variables.

For each feature and each dataset, the best three scores are marked in bold. Models which implement RNN or LSTM algorithms are marked italic. Empty cells represent predictions that were not reported. The Q3 accuracy is taken from the papers. [22, 34–36]

Table 2
Comparison of PSSP models of Q3 accuracy on CASP11-13.

Method/Algorithm	CASP11	CASP12	CASP13	year
PSIPRED[18]	80.7%	79.2%	80.7%	1999
JPRED4[34]	80.4%	78.5%	–	2015
DCRNN[35]	85.3%	–	–	2016
SPIDER3[19]	81.5%	79.8%	81.7%	2017
MUFOLD-SS[20]	85.2%	83.4%	79.6%	2018
CRRNN[36]	84.2%	82.6%	–	2018
Porter5[21]		–	82.9%	2018
RaptorX[37]	81.0%	78.6%	81.1%	2018
DeepCNF[23]	84.7%	82.1%	80.2%	2019
Ensemble of Contextnet[24]	–	82.7%	84.9%	2019
NetSurfP-2.0 (hhblits)[38]	–	82.4%	–	2019
DeepProf[39]	–	76.4%	–	2019
Bi-LSTM ensemble[40]	84.3%	–	–	2019
DNSS2[41]	–	–	82.2%	2019
2DCNN-BLAST[26]	81.5%	–	–	2020
LocalNet	85.0%	80.5%	85.0%	2020

Time performance

The processing of a single protein took on average 0.027 s (0.003 s-1.2 s) on NetSurfP-2.0 [37]. For DeepSeqVec [36], the average running time for a single protein was 0.08 with a minimum of 0.006 for the batch containing the shortest sequences (67 residues on average) and a maximum of 14.5 s (9860 residues on average). The only processing LocalNet needs is to break protein sequences into continuous fragments, and it took less than a mini second even for proteins of 9860 residues.

Discussions

In this study, we only used sliding windows of amino acids of fixed size on protein amino acid sequences, or fragments, as input and we did not utilize any information of evolution such as multiple sequence alignment and long-range interaction. For a window size of 19, LocalNet achieved a Q3 accuracy of 85.2%, comparable to other top performers. Our results are consistent with the traditional view that the secondary

structures are of local nature and mainly determined by local amino acid sequences.[5] Long-range contacts may have some impacts on protein secondary structure, but the impacts are likely insignificant.

Two factors may have contributed to LocalNet's excellent performance. First, even for a window size of 19 residues, the training data set consists of over 750,000 samples. For deep learning models of high dimension, sufficient training data size is crucial to reasonably approximate the unknown underlying mapping function from inputs to outputs. Generally, it is common knowledge that too little training data results in poor approximation. Second, as described above, in LocalNet, a fragment is considered to form H if the 5 consecutive residues at the center are assigned as H by DSSP, and a fragment is considered to form E or C if the 3 consecutive residues at the center are assigned as E or C, respectively by DSSP. These rules are consistent with well-known knowledge and helpful in removing noisy data; data quality is another crucial factor in building good deep learning models.

LocalNet's Q3 accuracy is comparable to those of other top performers; but LocalNet's accuracy for H is significantly higher than other models' and its accuracy for C is significantly lower. There are two reasonable explanations for these differences. First, among H, E, and C, only H is a relatively stable secondary structure. E is not a locally stabilized secondary structure; instead, it is stabilized by forming hydrogen bonds with distance residues on amino acid sequences. Unlike H, the geometry of E is irregular, and C is even more irregular and rarely stable. Second, other top performers generally use evolutionary information, and their better accuracy in prediction of C is likely derived from multiple sequence alignment. This explanation is also consistent with the observation that models utilizing templates have improved accuracy.

The generated high-quality model for PSSP in this study is expected to have good applications in protein structure modeling and protein design. For protein folding problems, for example, if alpha helix could be reliably predicted, it will significantly reduce the sampling space for locating the global free energy minimum. In protein engineering such as antibody engineering, amino acids need to be changed to improve protein's physical/chemical and other properties. An accurate model for secondary structure prediction is obviously helpful in guiding such design.

Conclusions

We developed a very simple, and yet highly accurate models for Q3 prediction using LSTM-RNN algorithms. The high accuracies show that local amino acid sequence itself contains sufficient information for secondary structure prediction without any homogenous information. The trained models are expected to have good applications in protein structure modeling and protein design, and they may also help in understanding protein folding mechanism.

Materials And Methods

Dataset and Hardware

12,358 protein X-ray structures in the precompiled culled PDB list [38] from PDB were selected for this study. This list used a cutoff of 30% amino acid sequence identity, and all proteins in this list have a resolution better than 2.0Å and a R-factor smaller than 2.5 (Table 3). We removed proteins with 40 residues or less and generated a refined list of 11,897 proteins. DSSP software [39] was used to assign proteins' Q3 secondary structures. Labels H, G, and I are assigned to class H; E and B to E; and S, T, and C to C. The 11,897 proteins were split into the training data set of 10,719 entries, test data set of 581 entries, and validation data set of 597 entries, randomly.

Table 3. Description of data sets.

Data sets	Number of entries
Culled PDB set [a]	12,358
Refined set [b]	11,897
Training set	10,719
Validation set	581
Testing set	597

a Total number of proteins from the Culled PDB list using 30% sequence identity cutoff. b Entries with >40 residues.

The neural network training and prediction were performed on a Dell PowerEdge R940xa server with four Intel Xeon Platinum 8160 processors (each with 24 cores), 3TB of RAM and four 16GB NVIDIA Tesla V100 graphic processing unit, installed with Ubuntu 16.04.6 distribution, python 3.5, CUDA driver version 10.0, cuDNN version 7.4, TensorRT 5.1 and TensorFlow 1.13.1. A python script was written to implement the neural network model and optimize the loss function.

Model

As illustrated in Figure 4, LocalNet starts with an RNN module, followed by one output classification layer. The LSTM cell consists of 32 units. The input to the RNN is the fragment sequence with each residue encoded by a one-hot vector. SoftMax layer is used as the classification layer, and the output layer consists of 3 nodes for H, E, and C.

Backpropagation is used for training the network [40]. Optimization of the loss function is carried out by mini-batch of a size 128 and the ADAM optimizer [41], which is implemented as tf.train.AdamOptimizer in the Tensorflow library [42].

Input features and preprocessing

In this study, we focused on Q3 prediction of H, E and C [43]. The Q3 accuracy was calculated by following equation:

$$Q_3 = 100\% \times \frac{\sum_{i=1}^{3} N_i}{N_{total}} \quad (1)$$

Where Ntotal is the total number of residues and Ni is the number of correctly predicted residues in state i [34].

For a residue with assigned secondary structure by DSSP, this residue and its neighbor residues are extracted from a protein sequence to form a data sample. To test the impact of window size on prediction accuracy, 8 window sizes, 7, 9, 11, 13, 15, 17, 19, and 21 were used. A fragment is considered to form H if the 5 consecutive residues at the center are assigned as H by DSSP, and a fragment is considered to form E or C if the 3 consecutive residues at the center are assigned as E or C, respectively, by DSSP. These rules are based upon known biochemistry and used to ensure data quality. A typical α helix contains about ten amino acids (about three turns) due to stabilizing interactions [44], and β sheets consist of typically 3 to 10 amino acids long with backbone in an extended conformation [45].

The sizes of the extracted samples of three data sets at different window sizes are given in Table 4.

Table 4. The number of three data sets at different window sizes.

21	19	17	15	13	11	9	7	Type	Data Set
460,589	469,844	479,350	489,083	498,852	508,545	517,841	526,194	H	
199,046	202,626	206,532	210,920	215,784	221,063	226,598	231,801	E	Training
91,929	93,287	94,707	96,121	97,656	99,242	100,897	102,719	C	
23,545	24,055	24,562	25,089	25,615	26,112	26,610	27,083	H	
10,994	11,180	11,387	11,605	11,864	12,149	12,440	12,714	E	Test
4,950	5,014	5,084	5,158	5,238	5,313	5,400	5,483	C	
25,692	26,196	26,717	27,241	27,756	28,267	28,752	29,186	H	
11,167	11,337	11,548	11,779	12,039	12,302	12,610	12,884	E	Validation
5,108	5,190	5,278	5,346	5,406	5,480	5,559	5,665	C	

Abbreviations

3D, 3-dimensional; PDB, Protein Database Bank; PSSP, protein secondary structure prediction; ANN, artificial neural network; RNN, recurrent neural network; LSTM, Long-Term Short-Term Memory; Q3: Three-state secondary structure per-residue accuracy; BRNN, Bayesian regularization neural networks

Declarations

Acknowledgements

Not applicable.

Authors Contributions

SY wrote the first draft and provided test data sets. YW trained the whole RNN model, evaluated models on different data sets and tasks and helped revise the manuscript. FW helped with discussions about the manuscript. KCG helped to manage the system. CFD improve the manuscript. All authors read and approved the final manuscript. All authors have given approval to the final version of the manuscript.

Funding

This work was supported by the National Natural Science Foundation of China [21773035, 21927805].

Availability of data and materials

The data and model can be downloaded from https://github.com/lake-chao/protein-secondary-structure-prediction.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Authors' information

[1] Zhejiang Provincial Key Laboratory of Advanced Mass Spectrometry and Molecular Analysis, Institute of Mass Spectrometry, School of Material Science and Chemical Engineering, Ningbo University, Ningbo, Zhejiang 315211, China.

[2] National Center for Advancing Translational Sciences, 9800 Medical Center Drive, Maryland 20850, USA.

[3] Department of Chemistry, Fudan University, Shanghai 200438, China.

References

1. Zhou YQ, Karplus M: Interpreting the folding kinetics of helical proteins. *Nature* 1999, 401(6751):400-403.
2. Ozkan SB, Wu GA, Chodera JD, Dill KA: Protein folding by zipping and assembly. *Proc Natl Acad Sci U S A* 2007, 104(29):11987-11992.
3. Plaxco KW, Simons KT, Baker D: Contact order, transition state placement and the refolding rates of single domain proteins. *J Mol Biol* 1998, 277(4):985-994.
4. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: The Protein Data Bank. *Nucleic Acids Res* 2000, 28(1):235-242.
5. Lupas AN: Review: Protein Secondary Structure Prediction Continues to Rise. *J Struct Biol* 2001, 134(2+3):204-218.
6. Zhu S, Liu Y: Protein Secondary Structure Online Server Predictive Evaluation. *Journal of Physics: Conference Series* 2019, 1237:052005.
7. Robson B, Mordasini T, Curioni A: Studies in the assessment of folding quality for protein modeling and structure prediction. *J Proteome Res* 2002, 1(2):115-133.
8. Robson B, Mordasini T, Curioni A: Studies in the assessment of folding quality for protein modeling and structure prediction. *Journal of Proteome Research* 2002, 1(2):115-133.
9. Kabsch W, Sander C: Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* 1983, 22(12):2577-2637.
10. Bettella F, Rasinski D, Knapp EW: Protein Secondary Structure Prediction with SPARROW. *Journal of Chemical Information and Modeling* 2012, 52(2):545-556.
11. Yaseen A, Li YH: Context-Based Features Enhance Protein Secondary Structure Prediction Accuracy. *Journal of Chemical Information and Modeling* 2014, 54(3):992-1002.
12. Kieslich CA, Smadbeck J, Khoury GA, Floudas CA: conSSert: Consensus SVM Model for Accurate Prediction of Ordered Secondary Structure. *Journal of Chemical Information and Modeling* 2016, 56(3):455-461.
13. Chou PY, Fasman GD: Prediction of protein conformation. *Biochemistry* 1974, 13(2):222-245.
14. Garnier J, Osguthorpe DJ, Robson B: Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol* 1978, 120(1):97-120.
15. Levin JM: Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng* 1997, 10(7):771-776.
16. Schmidler SC, Liu JS, Brutlag DL: Bayesian segmentation of protein secondary structure. *J Comput Biol* 2000, 7(1-2):233-248.
17. Kloczkowski A, Ting KL, Jernigan RL, Garnier J: Combining the GOR V algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins* 2002,

49(2):154-166.

18. Jones DT: Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices. *J Mol Biol* 1999, 17(2):195-202.
19. Heffernan R, Yang YD, Paliwal K, Zhou YQ: 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. *Bioinformatics* 2017, 33(18):2842-2849.
20. Fang C, Shang Y, Xu D: MUFOLD-SS: New deep inception-inside-inception networks for protein secondary structure prediction. *Proteins* 2018, 86(5):592-598.
21. Torrisi M, Kaleel M, Pollastri G: Porter 5: state-of-the-art ab initio prediction of protein secondary structure in 3 and 8 classes. *bioRxiv* 2018:289033.
22. Ma YM, Liu YH, Cheng JY: Protein Secondary Structure Prediction Based on Data Partition and Semi-Random Subspace Method. *Scientific Reports* 2018, 8.
23. Torrisi M, Kaleel M, Pollastri G: Deeper Profiles and Cascaded Recurrent and Convolutional Neural Networks for state-of-the-art Protein Secondary Structure Prediction. *Scientific Reports* 2019, 9.
24. Long SY, Tian P: Protein secondary structure prediction with context convolutional neural network. *Rsc Advances* 2019, 9(66):38391-38396.
25. Shapovalov M, Dunbrack RL, Vucetic S: Multifaceted analysis of training and testing convolutional neural networks for protein secondary structure prediction. *Plos One* 2020, 15(5).
26. Kumar P, Bankapur S, Patil N: An enhanced protein secondary structure prediction using deep learning framework on hybrid profile based features. *Applied Soft Computing* 2020, 86.
27. Heffernan R, Paliwal K, Lyons J, Singh J, Yang YD, Zhou YQ: Single-sequence-based prediction of protein secondary structures and solvent accessibility by deep whole-sequence learning. *Journal of Computational Chemistry* 2018, 39(26):2210-2216.
28. Ma JZ, Wang S, Wang ZY, Xu JB: Protein contact prediction by integrating joint evolutionary coupling analysis and supervised learning. *Bioinformatics* 2015, 31(21):3506-3513.
29. LeCun Y, Bengio Y, Hinton G: Deep learning. *Nature* 2015, 521(7553):436-444.
30. Hochreiter S, Schmidhuber J: Long short-term memory. *Neural Comput* 1997, 9(8):1735-1780.
31. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A: Critical assessment of methods of protein structure prediction: Progress and new directions in round XI. *Proteins* 2016, 84 Suppl 1(Suppl 1):4-14.
32. Moult J, Fidelis K, Kryshtafovych A, Schwede T, Tramontano A: Critical assessment of methods of protein structure prediction (CASP)-Round XII. *Proteins* 2018, 86 Suppl 1(Suppl 1):7-15.
33. Kryshtafovych A, Schwede T, Topf M, Fidelis K, Moult J: Critical assessment of methods of protein structure prediction (CASP)—Round XIII. *Proteins: Structure, Function, and Bioinformatics* 2019, 87(12):1011-1020.
34. Smolarczyk T, Roterman-Konieczna I, Stapor K: Protein Secondary Structure Prediction: A Review of Progress and Directions. *Current Bioinformatics* 2020, 15(2):90-107.

35. Hou J, Guo Z, Cheng J: DNSS2: improved ab initio protein secondary structure prediction using advanced deep learning architectures. *bioRxiv* 2019.
36. Heinzinger M, Elnaggar A, Wang Y, Dallago C, Nechaev D, Matthes F, Rost B: Modeling aspects of the language of life through transfer-learning protein sequences. *Bmc Bioinformatics* 2019, 20(1).
37. Klausen MS, Jespersen MC, Nielsen H, Jensen KK, Jurtz VI, Sonderby CK, Sommer MOA, Winther O, Nielsen M, Petersen B *et al*: NetSurfP-2.0: Improved prediction of protein structural features by integrated deep learning. *Proteins-Structure Function and Bioinformatics* 2019, 87(6):520-527.
38. Wang GL, Dunbrack RL: PISCES: a protein sequence culling server. *Bioinformatics* 2003, 19(12):1589-1591.
39. Joosten RP, te Beek TA, Krieger E, Hekkelman ML, Hooft RW, Schneider R, Sander C, Vriend G: A series of PDB related databases for everyday needs. *Nucleic Acids Res* 2011, 39(Database issue):D411-419.
40. Rumelhart DE, Hinton GE, Williams RJ: Learning Representations by Back-Propagating Errors. *Nature* 1986, 323(6088):533-536.
41. Kingma DP, Ba J: Adam: A Method for Stochastic Optimization. *Computer Science* 2014.
42. Tensorflow [www.tensorflow.org]
43. Kabsch W, Sander C: Dictionary of Protein Secondary Structure - Pattern-Recognition of Hydrogen-Bonded and Geometrical Features. *Biopolymers* 1983, 22(12):2577-2637.
44. Alpha-Helix (α-Helix). In: *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine.* Berlin, Heidelberg: Springer Berlin Heidelberg; 2006: 44-44.
45. Beta-Sheet Structure (β-Sheet Structure). In: *Encyclopedic Reference of Genomics and Proteomics in Molecular Medicine.* Berlin, Heidelberg: Springer Berlin Heidelberg; 2006: 134-134.

Figures

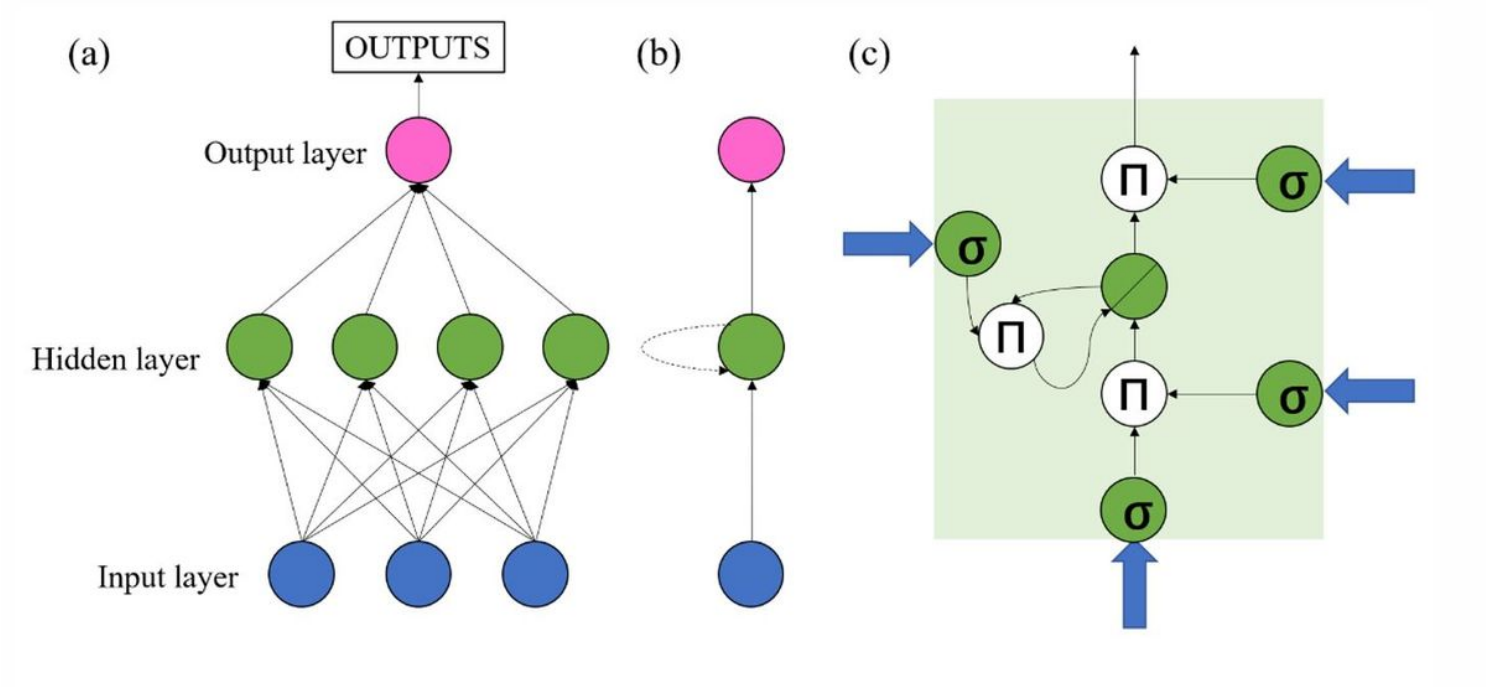


Figure 1

Illustration of ANN structure (a), RNN structure (b) with LSTM memory cell, which contains forget gate, input gate, output gate and cell state (c).

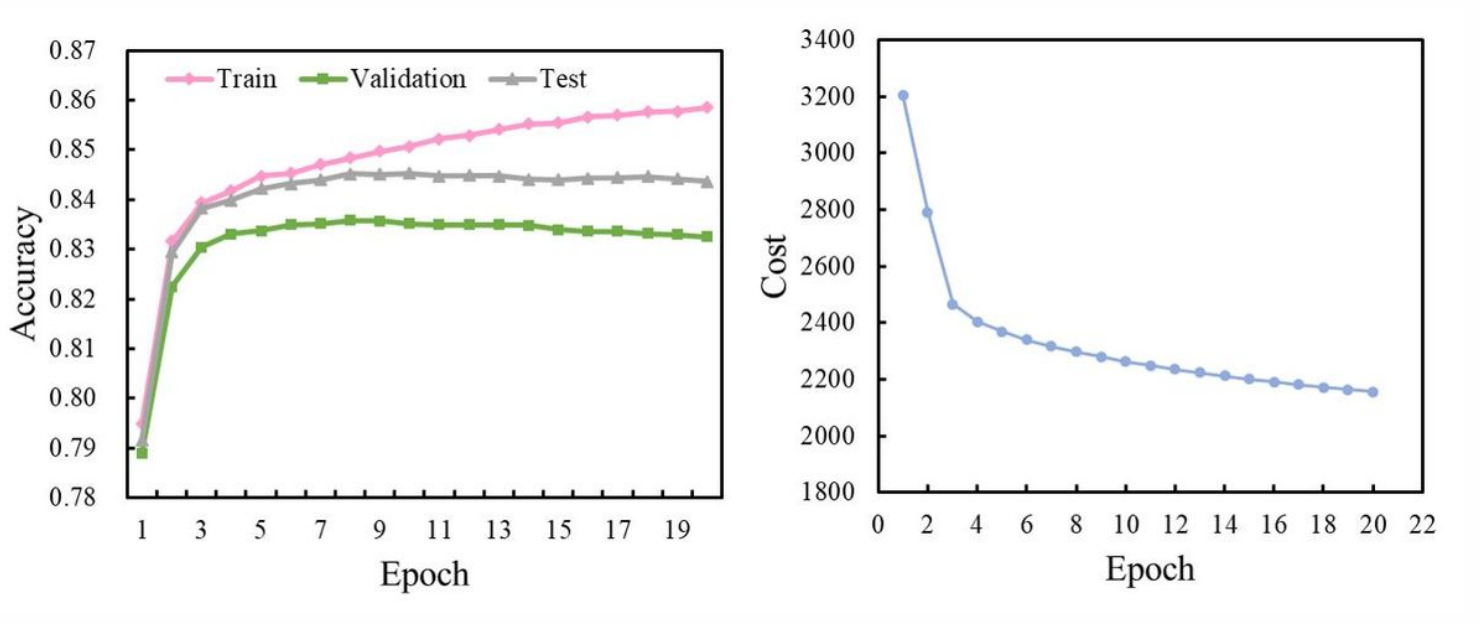
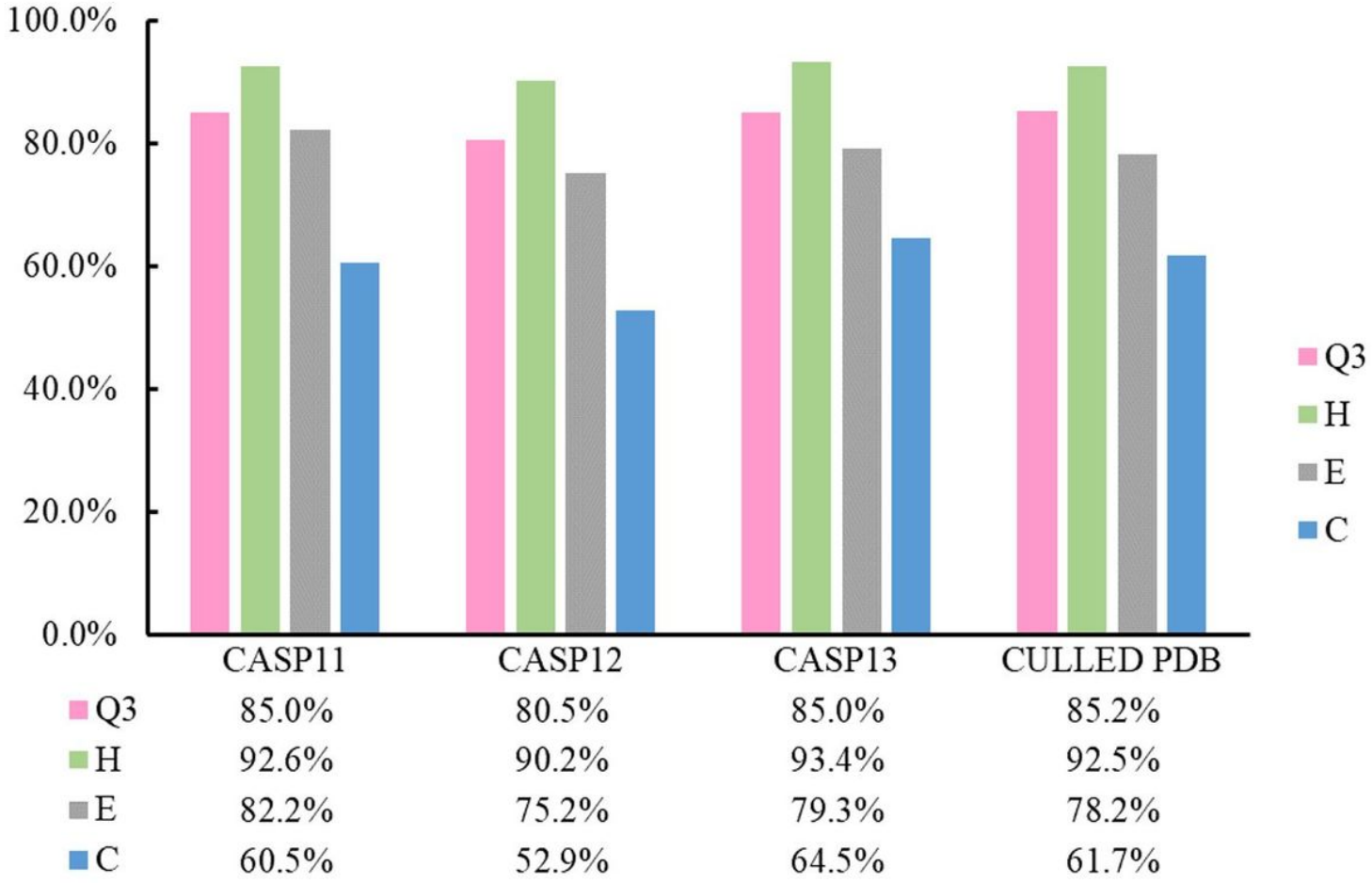


Figure 2

Accuracy (Left) and Cost (Right) for Q3 prediction model versus epoch. Window size is 19.



	CASP11	CASP12	CASP13	CULLED PDB
Q3	85.0%	80.5%	85.0%	85.2%
H	92.6%	90.2%	93.4%	92.5%
E	82.2%	75.2%	79.3%	78.2%
C	60.5%	52.9%	64.5%	61.7%

Figure 3

Prediction accuracy of Q3, H, E, and C for CASP11, CASP12, CASP13 and the culled PDB database.

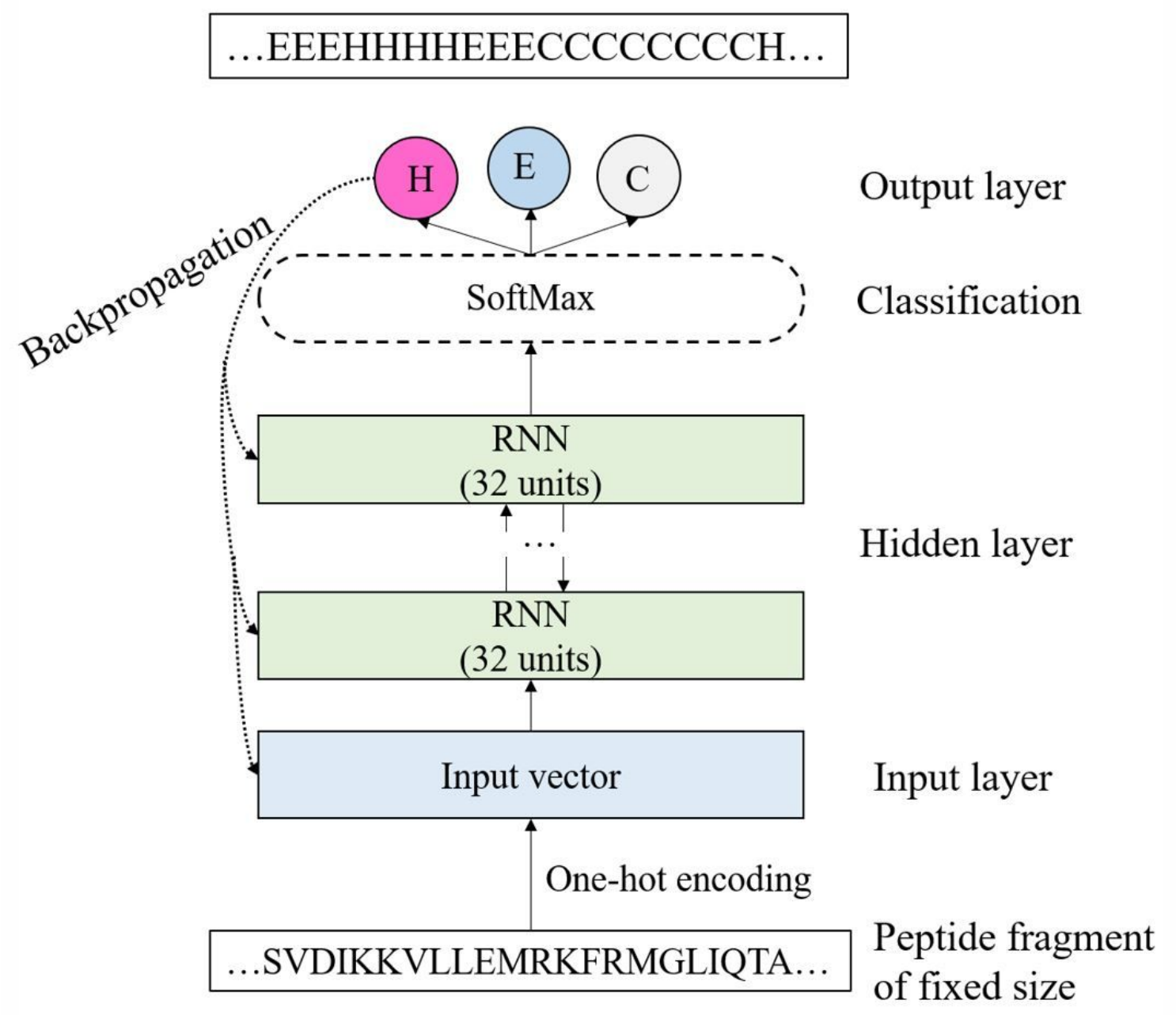


Figure 4

Diagram of the structure of LocalNet used in this study.