CSI-LSTM: A Web Server to Predict Protein Secondary Structure using Bidirectional Long Short Term Memory and NMR Chemical Shifts

Zhiwei Miao^{1,†}, Qianqian Wang^{1,†}, Xiongjie Xiao², Linhong Song², Xu Zhang², Conggang Li², Xin Zhou², Bin Jiang^{2,*}, Maili Liu^{2,*}

1. Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan, 430074, China
2. Key Laboratory of Magnetic Resonance in Biological Systems, State Key Laboratory of Magnetic Resonance and Atomic and Molecular Physics, National Center for Magnetic Resonance in Wuhan, Wuhan Institute of Physics and Mathematics, Innovation Academy for Precision Measurement Science and Technology, Chinese Academy of Sciences, Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology, Wuhan 430071, China; University of Chinese Academy of Sciences, Beijing 10049, China

[†] Authors with equal contribution

* Corresponding Authors: jbin@apm.ac.cn (Bin Jiang), ml.liu@apm.ac.cn (Maili Liu)

Abstract

The description and understanding of protein structure rely on secondary structure heavily. Secondary structure determination and prediction are widely used in protein structure related research. The secondary structure prediction methods based on NMR chemical shifts are convenient to use, so they are popular in protein NMR research. In recent years, there is significant improvement in deep neural network, which is consequently applied in many search fields. Here we proposed a deep neural network based on bidirectional long short term memory (biLSTM) to predict protein 3-state secondary structure using NMR chemical shifts of backbone nuclei. Compared with the existing methods of the same sort, the accuracy of the proposed method was improved. And a web server was built to provide secondary structure prediction service using this method.

Keywords:

Nuclear magnetic resonance, Chemical shifts, Protein secondary structure prediction, Deep learning, Long short term memory

1. Introduction

The structure of proteins is described at four levels of hierarchy[1]. A description of covalent bond linked amino acid residues in a polypeptide chain is primary structure of protein. Protein secondary structure is comprised of stable arrangements of amino acid residues in some specific structural patterns, and maintained by hydrogen bonds. Protein secondary structure can be described by 3-state category (helix, strand, and coil), or 8-state category (α -helix, π -helix, 3_{10} -helix, β -bridge, β -sheet, bend, turn, and loop)[2]. There are several protocols to transform the 8-state category to 3-state category [3]. Polypeptide chain folds up into 3-dimensional structure, which is tertiary structure. If a protein has two or more subunits, their spatial arrangement is quaternary structure.

Protein secondary structure is of significance to describe and understand protein tertiary structure[4]. Secondary structure provides rich structural information, thus its identification or prediction is widely used in protein 3D structure determination[5] and prediction[6], protein fold recognition[7], and intrinsically disordered protein (IDP) identification[8]. Because of its importance, protein secondary structure identification and prediction has long been attractive in protein research field. The first attempt of secondary prediction dated back to 1970s[9, 10]. Since then, a few of secondary structure prediction methods were developed. The methods based on the statistics of amino acids residues were firstly proposed[3, 10], while their 3-state secondary structure (Q_3) accuracy was less than 60%. As more sophisticated statistical methods were applied, such as artificial neural networks (ANN)[11] and machine learning[12], the prediction accuracy was increased, but still less than 65%[13]. Recent years, the prediction accuracy was raised up to more than 80% by introducing evolutionary information derived from multiple sequence alignment profile[13, 14] and advanced machine learning methods[3],

such as support vector machine[15] and deep neural networks[16]. It was reported that the accuracy limit of secondary structure may be 88%[17], due to the flexible protein structure and indistinct assignment in the joint regions between two states.

As for nuclear magnetic resonance (NMR) based protein structure determination, secondary structure is applied in protein structure generation and refinement[5, 18]. Traditionally, secondary structure is identified using inter-nucleic distance information provided by NOESY experiment. NMR chemical shifts contain rich structural information, and are much easier to obtain than NOEs, thus chemical shifts were explored to predict secondary structures as early in 1990s[19]. The first secondary structure prediction method based on chemical shifts is Chemical Shift Index or CSI, which is popular due to its fast speed, convenience, and fine accuracy[19-22]. Combined CSI protocol with more sophisticated statistical calculation or the sequence-based prediction program[23, 24] mentioned above, the Q3 accuracy can be increased. The representatives of this category are CSI 2.0[25] and TALOS-N[26]. Most of the previous mentioned methods provide 3-state classification, which is insufficient when applied in protein structure generation and refinement[20]. CSI 3.0 combines several prediction program to produce 11 types of secondary and super-secondary structures, and elevate prediction accuracy in further[20]. Both CSI 2.0 and 3.0 provide prediction service through online websites (<http://csi.wishartlab.com/cgi-bin/index.py>). Compared with sequence-based methods, NMR chemical shifts based methods could reach higher prediction accuracy[3, 20].

In contrary to traditional computing which depends on human instructions, artificial intelligence such as artificial neural network allows model modifies itself. Artificial neural network derived from the working pattern of neurons in brains, and comprise of units similar to biological neurons[27]. Deep neural network (DNN) refers to a network containing many neuron layers. Deep neural networks can express highly complicated models, thus they are successfully applied in many fields. In the procedure of model learning, the weights (parameters) of neural network are adjusted iteratively to improve the accuracy of result.

Recurrent neural network (RNN) is a class of artificial neural network, in which neural nodes are connected along ordered sequences, therefore RNN can extract information implied in ordered sequences[28]. Long short term memory (LSTM)[29] is a modified RNN architecture, which resolves the problems of gradient vanishment and information forgetting.

Apparently, peptide chain is one sort of ordered sequence. In this paper we proposed CSI-LSTM, a deep neural network based on bidirectional LSTM, to predict protein secondary structure using NMR chemical shifts. And a web site was developed to offer the structure prediction service. The model obtained Q3 accuracy 87.1% when it was applied on the test dataset. Compared with previous methods of the same sort, the prediction accuracy was improved.

2. Methods.

2.1 Dataset source

Protein secondary structures, which were prediction targets and data labels in model training, were obtained from DSSP database (<https://swift.cmbi.umcn.nl/gv/dssp/>). This database provided PDB IDs (protein's identification in Protein Data Bank), amino acid sequences, secondary structures, and other information related with secondary structures of over 157 thousand proteins for now. The model proposed here is to perform 3-state structure classification, while the secondary structures fall into 8 catalogues (H for α -helix, I for π -helix, G for 3_{10} -helix, B for β -bridge, E for β -sheet, S for bend, T for turn, and C for loop)[2, 3] in DSSP database. There are several methods to transform the 8-state secondary structure into 3-state (H for helix, E for strand, and C for coil) [3]. In this paper we used the following rule to transform 8-state catalogues to 3-state catalogues: transform 8-state catalogues H, G, I to 3-state catalogue H, transform 8-state catalogue E to 3-state catalogue E, and transform 8-state catalogues S, T, B, C to 3-state catalogue C.

NMR Chemical shift values of protein backbone, which were important features of data samples, were obtained from BMRB website (<https://bmrdb.io/>). A homemade python script was used to query whether

there is BMRB entry for a PDB ID. Considering there may be several entries in BMRB website for one PDB ID, which could introduce confusion in dataset preparation, only when one PDB ID was found to correspond only one BMRB entry, the assigned chemical shifts of that entry would be downloaded. In the following data check step, the protein data would be excluded, if their peptide sequences were broken or too short (shorter than 25 amino acid residues). The proportion of assigned chemical shifts of each protein was calculated. If the proportion of residues, whose assigned backbone nuclei (^1HN , $^1\text{H}_\alpha$, ^{15}N , ^{13}CO , $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$) were less than three, was more than 25% for one protein, the data of that protein was abandoned. Finally, there were 2860 proteins in dataset.

Even the proteins with too low assigned proportion were filtered out, there still were many unassigned backbone nuclei in the dataset. Those missing assigned chemical shifts were filled by the following procedure. Firstly, 3-residue segments were generated from the whole protein dataset as demonstrated in figure 1. Then the averaged backbone chemical shifts of middle residues were calculated from all the 3-residue segments with same sequence. Note that the missing chemical shifts were excluded from the average calculation. Thirdly, check each segment whether there is any missing backbone chemical shift in its middle residue, and fill the missing chemical shifts with the averaged values of the same segment sequence. Fourthly, there were still some 3-residue segments in which middle residues can't find the corresponding averaged chemical shifts. In that case, the missing chemical shifts were replaced with -100 ppm. Finally, those 3-residue segments were combined back into the original protein sequences.

Figure 1. Slice one protein sequence into 3-residue segments as demonstrated in this figure, where b stands for a virtual blank amino acid which was padded at two terminals of protein before slicing.

After missing chemical shifts were repaired, protein dataset was randomly separated into two parts, the sizes of which were 2574 and 286 respectively. The larger part was used in model training and validation, and the smaller one was used as test dataset. The PDB IDs and molecular weights of those proteins were listed in Supporting Information.

Each protein in training and validation dataset was firstly padded outside of both terminals with 5 blank amino-acid residues respectively, similar in figure 1. The so-called blank residue here is a virtual residue, whose backbone chemical shift values are all -100 ppm. Then, a 11-residue width sliding window was used to generate segments of 11 residues size. The 2574 proteins in training and validation dataset generated 259422 segments, of which 90% were randomly selected as training data, and the remained 10% were used as validation data. The 286 protein in test dataset generated 28531 segments. After training the proposed model should predict the 3-state secondary structures of the central residues in those segments.

2.2 Features and labels

Data features are critical to machine learning. The prediction model described here uses the following features.

(1) Amino-acid residue names. The amino-acid residue names are the basic input features in secondary structure prediction. Besides 20 standard amino acids, a virtual amino acid was designed to represent the virtual blank residue which were used in protein sequence padding, as stated before. The amino-acid names were one-hot coded[30] as sparse arrays with 21 elements, only one of which is non-zero.

(2) Chemical shifts of backbone nuclei (^1HN , $^1\text{H}_\alpha$, ^{15}N , ^{13}CO , $^{13}\text{C}_\alpha$, $^{13}\text{C}_\beta$) [25], which were download from BMRB website. The missing chemical

shifts were repaired as previously mentioned. Note that, even though glycine lack $^{13}\text{C}_\beta$ and proline lacks ^1HN intrinsically, those nonexistent chemical shifts were treated as common missing chemical shifts and repaired too. Chemical shifts need be standardized according to their mean value and standard deviation of the same type of nucleus.

(3) Physicochemical properties of amino acids. Apparently, these properties have obvious influence in the formation of secondary structure, so they have been used in secondary structure prediction since 1980s[31]. Seven physicochemical properties (steric parameter, polarizability, volume, hydrophobicity, isoelectric point, helix probability, sheet probability) were used as features here[32].

The training labels or prediction targets are the 3-state secondary structures of the central residues of the segments.

Figure 2. The architecture of CSI-LSTM for protein secondary structure prediction. The model is made up of two bidirectional LSTM (biLSTM) layers (dark blue) and two full connected (FC) layers (light blue). The neuron numbers of two FC layers are 200 and 3. The output layer consists of three neurons to predict the 3-state secondary structure of the central residue of the input 11-residue segment.

2.3 Model training

CSI-LSTM is a deep neural network based on bidirectional LSTM (biLSTM) for protein secondary structure prediction shown in figure 2. This network is made up of two biLSTM layers and two full connected[33] (FC) layers. The memory length of biLSTM is 128. The number of nodes of the two FC layers are 200 and 3. As stated previously, the sequence length of input data is 11, thus the size of output data of the second biLSTM is $128 \times 11 \times 2$ (memory length $\times$ sequence length $\times$ direction number), which is also the input size of the first FC layer. The node number of output layer is 3, which corresponds to three secondary structure types. The hyperparameters including biLSTM memory length and FC layer nodes were determined by experience and parameter tuning.

The model training was performed by a homemade PyTorch script run on a workstation, equipped with Intel Xeon E5-2630v2 CPU, 32GB ram, and Nvidia Geforce RTX 2080Ti GPU. The batch size was 4096, and epoch number was 300. Adaptive moment estimation (Adam)[34] algorithm was used in gradient decent optimization. The learning rate was 0.0005, and weight decay was 0.0005. When the model was trained on the training dataset, the loss value for training and validation datasets were calculated simultaneously. Every time when validation loss was lower than the previous lowest value, the model parameters were saved. The model parameter set with the lowest validation loss was chosen after model training.

2.4 Web server

A web site was built to provide protein secondary structure prediction using CSI-LSTM. The web site accepts protein information including primary sequence and NMR chemical shifts in NMR-STAR file format[35], and returns secondary structure prediction result. The web URL address is <http://www.proteindeeplearning.info/>.

3. Result and Discussion

Table S1 in Supporting Information listed the secondary structures

information of each type of residue in training and validation dataset. Some types of residues have very biased secondary structure distribution. For instance, there are 11634 proline residues in training and validation dataset, while only 754 proline residues are in sheet state. Serious biased structure distribution usually leads to bad prediction performance on rare state, because wrong prediction on rare state has small influence on loss function. To improve prediction performance in those cases, the segments of which the central residues were proline (P), aspartic acid (D), and glycine (G) residues in sheet state, were duplicated in training and validation dataset. The prediction performance of CSI-LSTM on training and validation dataset was also listed in Table S1. Totally, the accuracies for helix, sheet, and coil were 93.15%, 90.40%, and 89.63%, and the whole accuracy was 90.8%.

Table 1. Prediction accuracy (Acc.) of CSI-LSTM on the test dataset.

A.A. residue	Helix No.	Sheet No.	Coil No.	Residue No.	Acc. of helix	Acc. of sheet	Acc. of Coil	Residue Accuracy
A	827	287	886	2000	93.7%	79.8%	84.5%	87.7%
C	133	129	329	591	87.2%	81.4%	85.1%	84.8%
D	432	151	1028	1611	85.4%	76.2%	89.9%	87.4%
E	875	305	981	2161	94.4%	85.9%	81.1%	87.2%
F	315	319	327	961	93.7%	91.2%	75.2%	86.6%
G	235	209	1738	2182	73.6%	71.3%	93.7%	89.4%
H	163	121	506	790	84.1%	76.9%	90.9%	87.3%

I	485	520	484	1489	94.6%	87.3%	77.3%	86.4%
K	743	353	1013	2109	92.3%	80.5%	83.3%	86.0%
L	1023	500	906	2429	93.6%	90.6%	78.3%	87.2%
M	230	96	337	663	95.7%	82.3%	87.8%	89.7%
N	270	123	783	1176	78.9%	74.0%	92.6%	87.5%
P	148	71	1046	1265	69.6%	60.6%	92.7%	88.2%
Q	408	176	531	1115	90.2%	85.8%	80.4%	84.8%
R	478	251	630	1359	91.2%	80.9%	80.0%	84.1%
S	414	248	1412	2074	85.8%	75.4%	92.4%	89.0%
T	319	372	774	1465	84.3%	87.1%	85.1%	85.5%
V	555	691	612	1858	94.2%	92.6%	76.8%	87.9%
W	118	95	133	346	94.1%	84.2%	72.2%	83.0%
Y	248	303	336	887	92.3%	88.8%	77.1%	85.3%
Total	8419	5320	14792	28531	90.5%	84.6%	86.0%	87.1%

Table 1 listed the secondary structure information and CSI-LSTM prediction accuracies of each type of residue in test dataset. The whole

prediction accuracy was 87.1%, and the accuracies for helix, sheet and coil were 90.5%, 84.6%, and 86% respectively. Among all types of residues, CSI-LSTM has best performance on methionine (M) and glycine (G), with accuracies 89.7% and 89.4%. In spite of satisfactory prediction performance on glycine, it was shown that the prediction for helix and sheet states of glycine was not so good. The reason has been stated before that, glycine is rarely in helix or sheet state, thus wrong prediction for rare state has small impact on loss function. The similar circumstance also happened to proline (P). As mentioned early, the accuracy limit of secondary structure prediction is around 88% by rough estimate[3, 17]. Therefore, the 87.1% prediction accuracy on test dataset means this method is close to the theoretical limit.

To compare the prediction performance with the existing methods. The same test dataset was uploaded to CSI 2.0[25] and 3.0[20] prediction web server in NMR-STAR[35] file format. Note that, for unknown reason the CSI web servers didn't provided the prediction results for all uploaded proteins. The prediction accuracies of each protein from the three methods were listed in the excel file of Supporting Information, and were plot in figure 3. The orange and blue dots depict the results given by CSI-LSTM and CSI 2.0/3.0 respectively. From figure 3, the orange dots scatter higher and closer than blue ones, which means the proposed method provided better prediction result. The average and standard deviation values of protein secondary structure prediction accuracies were (81.7%, 83.7%, 87.2%) and (10.8%, 8.1%, 7.2%) for CSI 2.0, CSI 3.0, and CSI-LSTM on test dataset. Note that, the CSI-LSTM accuracy here (87.2%) was slightly different from the counter one (87.1%) in previous paragraph, because the average accuracy 87.2% came from statistics on prediction of proteins, while the average accuracy 87.1% came from statistics on prediction of 3-residue segments.

Figure 3. Accuracy comparison with CS 2.0 and 3.0. Every dot represents prediction accuracy of one protein in test dataset. The orange dots depict the results given by the proposed method, and the blue dots depict the results from CSI 2.0 or 3.0.

Commonly, assigned chemical shift values are usually complete for a protein. The missing proportion of assigned chemical shift values for the proteins in test datasets were calculated. The secondary prediction accuracies versus missing proportion of assigned chemical shifts were plot for three methods in figure S1. The Spearman correlation coefficients between accuracies and missing proportion were -0.039, -0.045, -0.114, with p values 0.516, 0.495, 0.0533, for CSI 2.0, CSI 3.0 and CSI-LSTM, which means all the three methods are robust to chemical shifts lack.

4. Summary

A deep neural network based on bidirectional long short term memory (biLSTM) was proposed to predict protein 3-state secondary structure using NMR chemical shifts of backbone nuclei. Compared with the previous methods of the same sort, the new method has better prediction accuracy. A web server was built to provide protein secondary structure prediction service using this method.

Supporting Information

The supporting information including the PDB IDs and molecular weights of proteins used as training and test data, the secondary structures information and prediction results of the training and validation dataset, the protein secondary structure accuracies by three methods. The supporting information is available free of charge on publications website.

Acknowledgements

This research was funded by National Key R&D Program of China (2018YFA0704002, 2018YFE0202300, 2017YFA0505400), National Natural Science Foundation of China (21735007, 21991080, 21921004), and CAS Key Research Program of Frontier Sciences (QYZDJ-SSW-SLH027).

References

- [1] D.L. Nelson, M.M. Cox, *Lehninger Principles of Biochemistry*, Seventh Edition, Macmillan, 2017.
- [2] W. Kabsch, C. Sander, DICTIONARY OF PROTEIN SECONDARY STRUCTURE - PATTERN-RECOGNITION OF HYDROGEN-BONDED AND GEOMETRICAL FEATURES, *Biopolymers*, 22 (1983) 2577-2637.
- [3] T. Smolarczyk, I. Roterman-Konieczna, K. Stapor, Protein Secondary Structure Prediction: A Review of Progress and Directions, *Current Bioinformatics*, 15 (2020) 90-107.
- [4] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, *Nucleic Acids Research*, 28 (2000) 235-242.
- [5] K. Wuthrich, *NMR with proteins and nucleic acids*, Wiley, 1991.
- [6] B. Kuhlman, P. Bradley, Advances in protein structure prediction and design, *Nat. Rev. Mol. Cell Biol.*, 20 (2019) 681-697.
- [7] J. Soding, A. Biegert, A.N. Lupas, The HHpred interactive server for protein homology detection and structure prediction, *Nucleic Acids Research*, 33 (2005)

W244-W248.

- [8] B. He, K.J. Wang, Y.L. Liu, B. Xue, V.N. Uversky, A.K. Dunker, Predicting intrinsic disorder in proteins: an overview, *Cell Research*, 19 (2009) 929-949.
- [9] M. Levitt, C. Chothia, STRUCTURAL PATTERNS IN GLOBULAR PROTEINS, *Nature*, 261 (1976) 552-558.
- [10] P.Y. Chou, G.D. Fasman, PREDICTION OF PROTEIN CONFORMATION, *Biochemistry*, 13 (1974) 222-245.
- [11] H. Bohr, J. Bohr, S. Brunak, R.M.J. Cotterill, B. Lautrup, L. Norskov, O.H. Olsen, S.B. Petersen, PROTEIN SECONDARY STRUCTURE AND HOMOLGY BY NEURAL NETWORKS - THE ALPHA-HELICES IN RHODOPSIN, *Febs Letters*, 241 (1988) 223-228.
- [12] S. Muggleton, R.D. King, M.J.E. Sternberg, PROTEIN SECONDARY STRUCTURE PREDICTION USING LOGIC-BASED MACHINE LEARNING, *Protein Engineering*, 5 (1992) 647-657.
- [13] Q. Jiang, X. Jin, S.J. Lee, S.W. Yao, Protein secondary structure prediction: A survey of the state of the art, *Journal of Molecular Graphics & Modelling*, 76 (2017) 379-402.
- [14] B. Rost, C. Sander, Improved prediction of protein secondary structure by use of sequence profiles and neural networks, *Proceedings of the National Academy of Sciences*, 90 (1993) 7558.
- [15] J.J. Ward, L.J. McGuffin, B.F. Buxton, D.T. Jones, Secondary structure prediction with support vector machines, *Bioinformatics*, 19 (2003) 1650-1655.
- [16] R. Heffernan, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, Y. Zhou, Improving prediction of secondary structure, local backbone angles and solvent accessible surface area of proteins by iterative deep learning, *Scientific Reports*, 5 (2015) 11476.
- [17] B. Rost, Review: Protein secondary structure prediction continues to rise, *J. Struct. Biol.*, 134 (2001) 204-218.
- [18] A.W. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C.L. Qin, A. Zidek, A.W.R. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D.T. Jones, D. Silver, K. Kavukcuoglu, D. Hassabis, Improved protein structure prediction using potentials from deep learning, *Nature*, 577 (2020) 706-+.
- [19] D.S. Wishart, B.D. Sykes, F.M. Richards, THE CHEMICAL-SHIFT INDEX - A FAST AND SIMPLE METHOD FOR THE ASSIGNMENT OF PROTEIN SECONDARY STRUCTURE THROUGH NMR-SPECTROSCOPY, *Biochemistry*, 31 (1992) 1647-1651.
- [20] N.E. Hafsa, D. Arndt, D.S. Wishart, CSI 3.0: a web server for identifying secondary and super-secondary structure in proteins using NMR chemical shifts, *Nucleic Acids Research*, 43 (2015) W370-W377.
- [21] H.R. Eghbalnia, L.Y. Wang, A. Bahrami, A. Assadi, J.L. Markley, Protein energetic conformational analysis from NMR chemical shifts (PECAN) and its use in determining secondary structural elements, *Journal of Biomolecular NMR*, 32 (2005) 71-81.
- [22] C.C. Wang, J.H. Chen, W.C. Lai, W.J. Chuang, 2DCSi: identification of protein

secondary structure and redox state using 2D cluster analysis of NMR chemical shifts, *Journal of Biomolecular NMR*, 38 (2007) 57-63.

[23] Y.J. Wang, O. Jardetzky, Probability-based protein secondary structure identification using combined NMR chemical-shift data, *Protein Science*, 11 (2002) 852-861.

[24] L.H. Hung, R. Samudrala, Accurate and automated classification of protein secondary structure with PsiCSI, *Protein Science*, 12 (2003) 288-295.

[25] N.E. Hafsa, D.S. Wishart, CSI 2.0: a significantly improved version of the Chemical Shift Index, *Journal of Biomolecular NMR*, 60 (2014) 131-146.

[26] Y. Shen, A. Bax, Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks, *Journal of Biomolecular NMR*, 56 (2013) 227-241.

[27] M. Minsky, Neural-analog networks and the brain model problem, Princeton University, in, Ph. D. Thesis, 1954.

[28] A. Tealab, Time series forecasting using artificial neural networks methodologies: A systematic review, *Future Computing and Informatics Journal*, 3 (2018) 334-340.

[29] S. Hochreiter, J. Schmidhuber, Long Short-Term Memory, *Neural Computation*, 9 (1997) 1735-1780.

[30] D.M. Harris, S.L. Harris, P. Prinz, T. Crawford, *Digital Design and Computer Architecture*, Morgan Kaufmann, 2019.

[31] M.J. Zvelebil, G.J. Barton, W.R. Taylor, M.J.E. Sternberg, PREDICTION OF PROTEIN SECONDARY STRUCTURE AND ACTIVE-SITES USING THE ALIGNMENT OF HOMOLOGOUS SEQUENCES, *Journal of Molecular Biology*, 195 (1987) 957-961.

[32] J. Meiler, M. Muller, A. Zeidler, F. Schmaschke, Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks, *Journal of Molecular Modeling*, 7 (2001) 360-369.

[33] Y. Bengio, Learning Deep Architectures for AI, *Foundations and Trends® in Machine Learning*, 2 (2009) 1-127.

[34] D.P. Kingma, J. Ba, Adam: A Method for Stochastic Optimization, in: *International Conference on Learning Representations*, 2014, pp. 1-13.

[35] E.L. Ulrich, K. Baskaran, H. Dashti, Y.E. Ioannidis, M. Livny, P.R. Romero, D. Maziuk, J.R. Wedell, H. Yao, H.R. Eghbalnia, J.C. Hoch, J.L. Markley, NMR-STAR: comprehensive ontology for representing, archiving and exchanging data from nuclear magnetic resonance spectroscopic experiments, *Journal of Biomolecular NMR*, 73 (2019) 5-9.