Protein Remote Homology Detection Based on Deep Convolutional Neural Network

Yu Wang
Xi'an Jiaotong University
JunPeng Bao (baojp@mail.xjtu.edu.cn)
Xi'an Jiaotong University https://orcid.org/0000-0002-3866-5011
Fangfang Huang
Xi'an Jiaotong University
Jianqiang Du
Xi'an Jiaotong University
Yongfeng Li
People's Hospital of Lishui City

Research article

Keywords: Deep Learning, Resnet, Inception, Remote Homology Protein Detection

Posted Date: October 1st, 2019

DOI: https://doi.org/10.21203/rs.2.15388/v1

Abstract

Background: Protein remote homology detection has long received great attention in the field of bioinformatics, but the property of low protein sequence similarities heavily influences the accuracy of detection. Recently, such deep learning methods as LSTM have been adopted to deal with the problem. However, LSTM-based models will consume much time during the training process because of their cyclic connection mechanism and such problem will become more serious when dealing with long protein sequences.

Results: In this paper, we propose a CNN-based network, called ConvRes, to address the aforementioned shortcomings of existing methods in this field, which combines a variant Inception and Resnet block. Experimental results show that (1) this CNN-based network can classify the family of the remote homology proteins with comparable precision to the existing state-of-art method (ProDec-BLSTM) on the SCOP benchmark dataset. (2) ConvRes consumes less time with using only 15000 seconds whereas ProDec-BLSTM taking 150000 seconds.

Conclusion: This paper showcases that our proposed ConvRes network outperforms other existing models with regard to detecting remote homology proteins. The experimental results prove ConvRes network to be a viable and efficient model for remote homology protein detection. In the future work, we will improve the performance of ConvRes network by using other dataset and explore new representations to adapt for variable-length sequences.

Background

Protein remote homology detection plays an indispensable part in the field of bioinformatics. The detection and classification of remote homology protein sequences are of vital significance to understand the biological process because these proteins share similar structures and functions[1,2]. In the past decades, varieties of technologies and algorithms have been developed and designed for solving the aforementioned problem.

Many alignment-based methods including BLAST[3], FASTA[4], UCLUST[5], CD-HIT[6], profile alignment[7–12] and HMM alignment methods[13–15] have been proposed to compute the similarity of protein sequences. These methods are based on sequences alignments, consequently generating a similarity score. However, the performance of these methods has been restricted because of the low protein sequence similarities of these remote homology proteins.

Traditional machine learning methods have been successfully applied to pattern recognition by using the given fixed features as input. Inspired by this, some researchers proposed discriminative methods for protein remote homology detection, which trains a classifier based on positive and negative samples and then classifies these protein sequences at the prediction stage. Several kinds of kernels have been applied in the research such as LA kernel[16], motif kernel[17,] and mismatch kernel[18]. In addition, other

research which combines the physicochemical property to improve the accuracy of detecting the representation of protein[19–22] continues to emerge. However, the classification performance of these methods largely relies on the fixed features extracted by priori knowledge.

Compared with traditional machine learning, deep learning technologies can automatically capture the patterns of input data without priori knowledge. Several architectures of deep learning technologies, including Convolutional Neural Network (CNN) and Recurrent Neural Networks (RNN) have shown their merits on feature extraction and representation especially in the field of image[23–25] and Nature Language Processing (NLP)[26]. Meanwhile, there are varieties of deep learning-based methods successfully applied to bioinformatics in recent years such as protein classification[27], protein structure prediction[28,29], and protein subcellular localization[30,31]. Biological sequences can be considered as a special language and some researchers used RNN, especially Long Short-Term Memory (LSTM), to process biological sequences and find motifs among different sequences. LSTM has also been applied in the detection of remote homology proteins, among which ProDec-BLSTM[35] achieves the best performance by using a bidirectional LSTM (BLSTM) as classifier. ProDec-BLSTM converts initial protein sequences into pseudo proteins, and encodes these sequences by one-hot technology. Then, it adopts BLSTM as a classifier to recognize the family of the input sequences. However, training a neural network by BLSTM would consume much time because of the cyclic connections and this problem becomes more challenging when dealing with long protein sequences. By contrast, focused on the local sequence pattern, CNN can be promising to deal with biological sequences.

This paper proposes a ConvRes model, a CNN-based deep neural network, and proves it to be an efficient remote homology protein detector. Compared with the existing 10 related methods, the proposed model gains the state-of-art performance (evaluated by AUROC) on SCOP benchmark dataset. Furthermore, the cost of training time by this model is greatly reduced in contrast to the existing state-of-art method (ProDec-BLSTM).

Methods

2.1 SCOP Benchmark Dataset

The Structural Classification of Proteins (SCOP) database has been widely used to evaluate the performance of various methods on protein classification such as in ProDec-BLSTM. In this work, SCOP1.67 dataset is thus used (the same as ProDec-BLSTM) and it is accessible online.

Positive and negative samples of training and testing data are randomly selected for each of the 102 families contained in our dataset, with the average of 9077 sequences in each training dataset. There are 507,119 different sequences in total in this dataset, of which the minimum length is 13 and maximum length is 1264. The sequences with their length shorter than 400bp account for 96% of the dataset. Hence, the sequence length is constrained to 400bp in this study, which means that sequences with their length over 400bp will be correspondingly controlled at 400^{th} bp.

2.2 Sequences Representation

Since the physiological properties of protein rely on the physiological properties of amino acids, this study uses physiological properties of aminos acids to denote protein sequences. *Table 1* demonstrates all 12 types of physicochemical properties[37] of amino acids in our work, including chemical composition of the side chain (P1), polar requirement (P2), hydropathy index (P3), isoelectric point (P4), molecular volume (P5), polarity (P6), aromaticity (P7), aliphaticity (P8), hydrogenation (P9), Hydroxythiolation (P10), pK1(-COOH) (P11) and pK2() (P12).

2.3 Deep Neural Network Architecture Combined Inception and Resnet

This section illustrates our proposed ConvRes (shown in *Fig.1)*,, which combines a variant Inception and a Resnet Block. Input data are fed into a variant Inception block, aiming to extract abstract features of protein sequences by using various kernel sizes. The features of protein sequences can be enhanced after the Inception block because different kernel sizes can be seen as different window sizes according to protein sequences. Then, Resnet block is employed as a detector by using the aforementioned features as input. Finally, this architecture will recognize whether the input sequence belongs to a certain family. More details will be clarified in the following subsections.

2.3.1 1-D Inception Block

Inception network is a frequently used structure in the field of Convolutional Neural Network (CNN), which extracts features by several kernels with different sizes. More abstract features can be received through the Inception network even if the objective possesses different sizes in the set of pictures. As for biological sequences, the window size plays a vitally important role on the accuracy of classification. However, no previous studies could help stipulate the optimal window size. So this paper adopts a variant Inception, called 1-D Inception block, combining the Inception structure with 1-dimentional convolution (shown in *Fig.1)*.. Since the input of the Inception block is [Due to technical limitations, this equation is only available as a download in the supplemental files section], the output of this block can be described as follows,

[Due to technical limitations, this equation is only available as a download in the supplemental files section.] (1)

where *Conv1D* ( )represents 1-dimentional convolutional option of different filter size *f,k* stands for the number of different kernels in this block, and + means concatenate operation.

The enhanced features extracted from this block will be concatenated by channels, and sent to the following Resnet classifier to generate the final result.

2.3.2 Resnet classifier

Deep residual network (Resnet) is a highly configured edition of conventional CNN, which is formed by several convolutional layers and a residual operation between every two layers. Resnet solves the problem of gradient vanishing to some degree because of the residual operations, thus achieving better performance than conventional CNN.

The input feature ofconvolutional layer is [Due to technical limitations, this equation is only available as a download in the supplemental files section], so the input of layer is as follows,

[Due to technical limitations, this equation is only available as a download in the supplemental files section.] (2)

where*Conv* ( ) represents convolutional option of i_{th} layer, and + stands for concatenate operation.

This paper employs 18 layers of Resnet as the classifier for remote homology protein detection. This Resnet classifier contains an independent convolutional layer followed by a max-pooling layer, and 4 residual blocks (with 2 convolutional layers in each block) followed by an average-pooling layer and a full connection layer. The concatenated features extracted by the 1-D Inception block will be sent to this Resnet classifier, in which each layer uses the extracted features and initial input of the previous layer as input and provides feature extraction with a higher-level abstraction. The final dense layers will recognize whether the input sequence belongs to the current family or not.

2.3.3 Implementation details

This network is implemented by using *Keras 2.2.4* with the backend of *TensorFlow 1.9.0.* Six of different kernel sizes are adopted in our 1-D Inception block, which are set to 1, 3, 5, 9, 15, 21 respectively. The parameters used in Resnet block are the same as standard Resnet–18[36]. For each protein family, such binary classification network is trained and tested respectively. Each model is optimized by training for 150 epochs.

2.3.4 Performance evaluation

In this paper, the area under the receiver operating characteristic (AUROC) is used to evaluate the performance of our method and the existing methods. Receiver Operating Characteristic (ROC) curve is plotted by employing the true positive rate as *x* axis and the false positive rate as *y* axis according to different classification threshold. AUROC refers to the area under ROC plot, whose score is between 0 and 1.The stronger and better performance the classification achieves, the closer the AUROC score is to 1.

Result

This section compares our ConvRes model with several other related methods including PSI-BLAST[7], LA-kernel[16], GPkernel[17], GPextended[17], GPboost[17], Mismatch[18], SVM-Pairwise[32], eMOTIF[33], LSTM[34], and ProDec-BLSTM[35]. The results are shown in *Fig.2,* which indicates that ConvRes is comparable to ProDec-BLSTM and outperforms any other related methods.

As described in S*ection 1,* ProDec-BLSTM model includes two essential parts, which are pseudo proteins processing and a BLSTM classifier. To further evaluate the performance of our ConvRes model and the ProDec-BLSTM model, this work compares the training time of BLSTM (removes pseudo protein processing of ProDec-BLSTM) and ConvRes model on 16 families respectively. The result (shown in *Fig.3)* showcases that it is much quicker to train the ConvRes model than the BLSTM model (nearly 10 times). For a protein family, BLSTM takes about 150000s to train for 150 epochs, while ConvRes costs only 15000s. It is obvious that the CNN framework operates much faster than the BLSTM framework. Moreover, ProDec-BLSTM contains the processing of pseudo proteins using PSI-BLAST to generate PSSM, which will also consume lots of time. So our model requires much less training time with the performance comparable to ProDec-BLSTM.

Conclusion

This study proposes a CNN-based network that combines Inception and Resnet block to detect remote homology proteins. The proposed network can precisely classify proteins into the specific family that they belong to. Experimental results show that ConvRes achieves the top performance in comparison to other related methods on the SCOP benchmark dataset. Furthermore, this model saves much time than that of the existing state-of-art method (ProDec-BLSTM), which benefits from the local pattern detection properties of CNN. Furthermore, different window sizes in Inception block also enhance the features of protein sequences.

In the future work, we will improve the performance of ConvRes network by using other dataset and explore new representations to adapt for variable-length sequences.

Declarations

Acknowledgements

Not applicable.

Funding

Not applicable.

Availability of data and materials

The SCOP benchmark dataset supporting the conclusion of this article was published in [34], which is available on http://www.bioinf.jku.at/software/LSTM_protein/.

Authors' contributions

YW carried out remote homology detection studies and drafted the manuscript. JPB proposed the main idea of this study and guided the work. FFH worked together to complete experiments. JQD assisted in offering suggestions for biological knowledge. YFL assisted in providing suggestions for medical knowledge. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interest

The authors declare that they have no competing interests.

Reference

1. Chen J, Guo M, Wang X, et al. A comprehensive review and comparison of different computational methods for protein remote homology detection. Briefings in bioinformatics, 2016, 19(2): 231–244.
2. Wei L, Zou Q. Recent progress in machine learning-based methods for protein fold recognition. International journal of molecular sciences, 2016, 17(12): 2118.
3. Lobo. Basic Local Alignment Search Tool (BLAST). Journal of Molecular Biology, 2012, 215:403–410.
4. Lipman D J, Pearson W R. Rapid and sensitive protein similarity searches. Science, 1985, 227:1435–1441.
5. Edgar R C. Search and clustering orders of magnitude faster than BLAST. Bioinformatics, 2010, 26:2460.
6. Li W. Fast Program for Clustering and Comparing Large Sets of Protein or Nucleotide Sequences. Springer US, 2015.
7. Altschul S F, Madden T L, Schäffer A A, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids research, 1997, 25(17): 3389–3402.

8. Margelevičius M, Laganeckas M, Venclovas Č. COMA server for protein distant homology search. Bioinformatics, 2010, 26(15): 1905–1906.
9. Jaroszewski L, Li Z, Cai X, et al. FFAS server: novel features and applications. Nucleic acids research, 2011, 39(suppl_2): W38-W44.
10. Sadreyev R, Grishin N. COMPASS: a tool for comparison of multiple protein alignments with assessment of statistical significance. Journal of molecular biology, 2003, 326(1): 317–336.
11. Yang Y, Faraggi E, Zhao H, et al. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics, 2011, 27(15): 2076–2082.
12. Yan K, Xu Y, Fang X, et al. Protein fold recognition based on sparse representation based classification. Artificial intelligence in medicine, 2017, 79: 1–8.
13. Finn R D, Clements J, Eddy S R. HMMER web server: interactive sequence similarity searching. Nucleic acids research, 2011, 39(suppl_2): W29-W37.
14. Remmert M, Biegert A, Hauser A, et al. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature methods, 2012, 9(2): 173.
15. Karplus K, Barrett C, Hughey R. Hidden Markov models for detecting remote protein homologies. Bioinformatics (Oxford, England), 1998, 14(10): 846–856.
16. Saigo H, Vert JP, Ueda N, Akutsu T. Protein homology detection using string alignment kernels. Bioinformatics. 2004;20(11):1682–9.
17. Håndstad T, Hestnes A J H, Sætrom P. Motif kernel generated by genetic programming improves remote homology and fold detection. BMC bioinformatics, 2007, 8(1): 23.
18. Leslie C S, Eskin E, Cohen A, et al. Mismatch string kernels for discriminative protein classification. Bioinformatics, 2004, 20(4): 467–476.
19. Yang Y, Tantoso E, Li K B. Remote protein homology detection using recurrence quantification analysis and amino acid physicochemical properties. Journal of Theoretical Biology, 2008, 252(1): 145–154.
20. Webb-Robertson B J M, Ratuiste K G, Oehmen C S. Physicochemical property distributions for accurate and rapid pairwise protein homology detection. BMC bioinformatics, 2010, 11(1): 145.
21. Liu B, Wang X, Chen Q, et al. Using amino acid physicochemical distance transformation for fast protein remote homology detection. PloS one, 2012, 7(9): e46633.
22. Liu B, Chen J, Wang X. Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis. Molecular Genetics and Genomics, 2015, 290(5): 1919–1931.
23. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks//Advances in neural information processing systems. 2012: 1097–1105.
24. Li H, Lin Z, Shen X, et al. A convolutional neural network cascade for face detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 5325–5334.

25. Gatys L A, Ecker A S, Bethge M. Image style transfer using convolutional neural networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2414–2423.
26. Young T, Hazarika D, Poria S, et al. Recent trends in deep learning based natural language processing. ieee Computational intelligenCe magazine, 2018, 13(3): 55–75.
27. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Briefings in bioinformatics, 2017, 18(5): 851–869.
28. Spencer M, Eickholt J, Cheng J. A deep learning network approach to ab initio protein secondary structure prediction. IEEE/ACM transactions on computational biology and bioinformatics (TCBB), 2015, 12(1): 103–112.
29. Wang S, Peng J, Ma J, et al. Protein secondary structure prediction using deep convolutional neural fields. Scientific reports, 2016, 6: 18962.
30. Sønderby S K, Sønderby C K, Nielsen H, et al. Convolutional LSTM networks for subcellular localization of proteins. International Conference on Algorithms for Computational Biology. Springer, Cham, 2015: 68–80.
31. Almagro Armenteros J J, Sønderby C K, Sønderby S K, et al. DeepLoc: prediction of protein subcellular localization using deep learning. Bioinformatics, 2017, 33(21): 3387–3395.
32. Liao L, Noble W S. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. Journal of computational biology, 2003, 10(6): 857–868.
33. Ben-Hur A, Brutlag D. Remote homology detection: a motif based approach. Bioinformatics, 2003, 19(suppl_1): i26-i33.
34. Hochreiter S, Heusel M, Obermayer K. Fast model-based protein homology detection without alignment. Bioinformatics, 2007, 23(14): 1728–1736.
35. Li S, Chen J, Liu B. Protein remote homology detection based on bidirectional long short-term memory. BMC bioinformatics, 2017, 18(1): 443.
36. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770–778.
37. Zhao-Hui Q, Meng-Zhe J, Su-Li L, et al. A protein mapping method based on physicochemical properties and dimension reduction. Computers in Biology & Medicine, 2015, 57:1–7.

Table 1

Table 1. Physicochemical properties of amino acids

Amino acids	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
A	0	7.0	1.8	6.00	31	8.1	-0.11	0.239	0.33	-0.062	2.34	9.69
C	2.75	4.8	2.5	5.07	55	5.5	-0.184	0.22	0.074	0.38	1.71	10.78
D	1.38	13.0	-3.5	2.77	54	13.0	-0.285	0.171	-0.371	-0.079	2.09	9.82
E	0.92	12.5	-3.5	3.22	83	12.3	-0.067	0.187	-0.254	-0.184	2.19	9.67
F	0	5.0	2.8	5.48	132	5.2	0.438	0.234	0.011	0.074	1.83	9.13
G	0.74	7.9	-0.4	5.97	3	9.0	-0.073	0.16	0.37	-0.017	2.34	9.60
H	0.58	8.4	-3.2	7.59	96	10.4	0.32	0.205	-0.078	0.056	1.82	9.17
I	0	4.9	4.5	6.02	111	5.2	0.001	0.273	0.149	-0.309	2.36	9.68
K	0.33	10.1	-3.9	9.74	119	11.3	0.049	0.228	-0.075	-0.371	2.18	8.95
L	0	4.9	3.8	5.98	111	4.9	-0.008	0.281	0.129	-0.264	2.36	9.60
M	0	5.3	1.9	5.74	105	5.7	-0.041	0.253	-0.092	0.077	2.28	9.21
N	1.33	10.0	-3.5	5.41	56	11.6	-0.136	0.249	-0.233	0.166	2.02	8.80
P	0.39	6.6	-1.6	6.30	32.5	8.0	-0.016	0.165	0.37	-0.036	1.99	10.60
Q	0.89	8.6	-3.5	5.65	85	10.5	-0.246	0.26	-0.409	-0.025	2.17	9.13
R	0.65	9.1	-4.5	10.76	124	10.5	0.079	0.211	-0.176	-0.167	2.17	9.04
S	1.42	7.5	-0.8	5.68	32	9.2	-0.153	0.236	0.022	0.47	2.21	9.15
T	0.71	6.6	-0.7	6.16	61	8.6	-0.208	0.213	0.136	0.348	2.63	10.43
V	0	5.6	4.2	5.96	84	5.9	-0.155	0.255	0.245	0.212	2.32	9.62
W	0.13	5.2	-0.9	5.89	170	5.4	0.493	0.183	0.011	0.05	2.38	9.39
Y	0.20	5.4	-1.3	5.66	136	6.2	0.381	0.193	-0.138	0.22	2.20	9.11

Figures

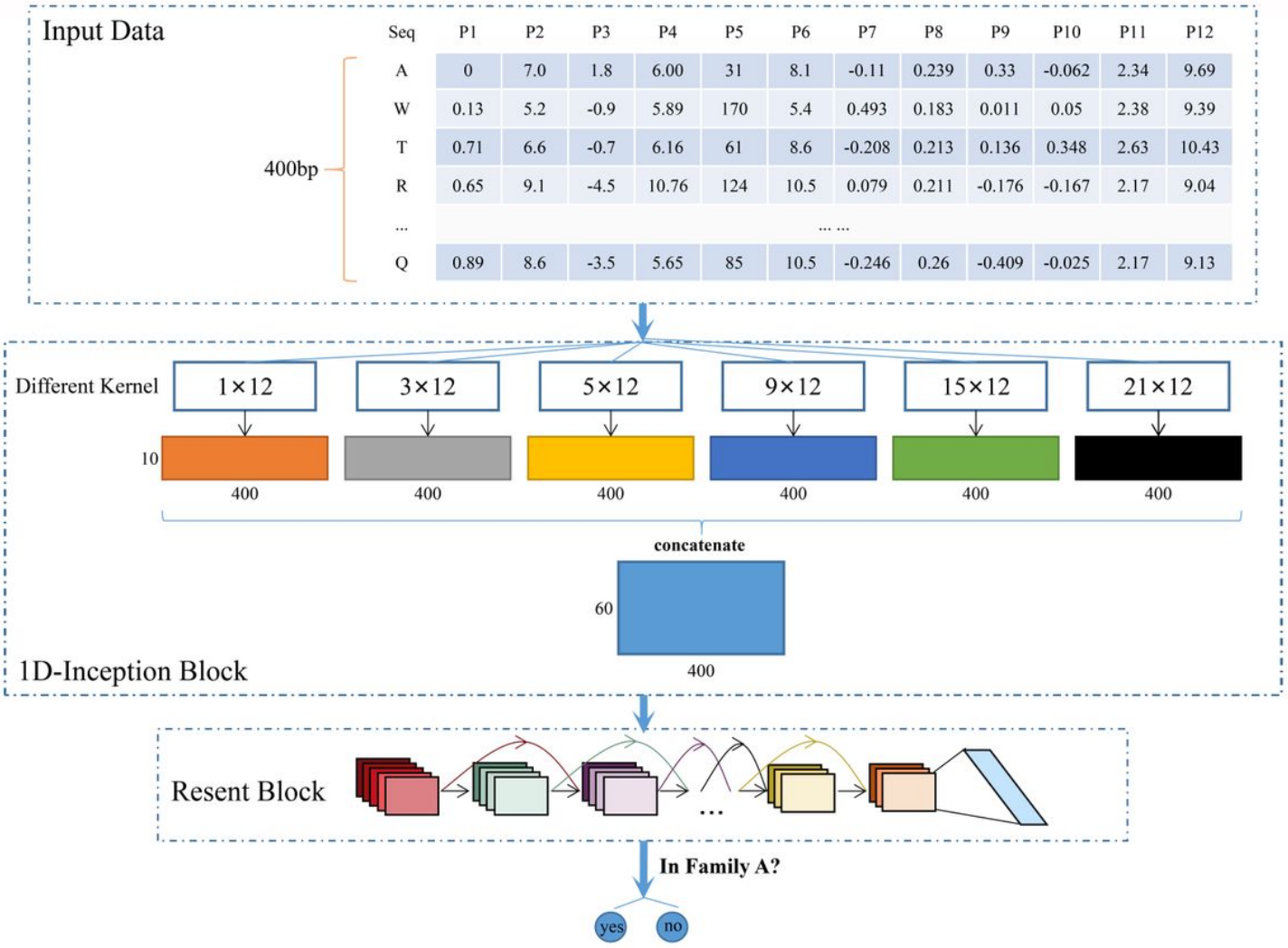

Figure 1

The architecture of our work.

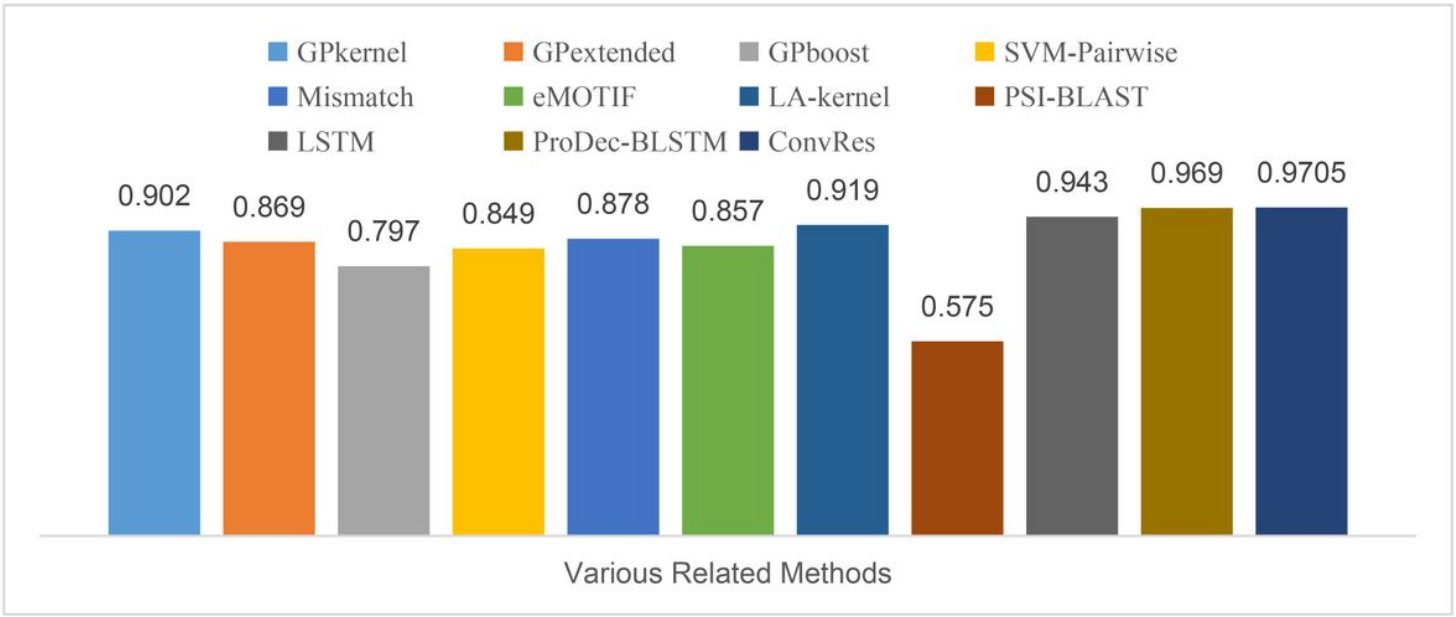

Figure 2

Performance of several related methods evaluated by mean AUROC.

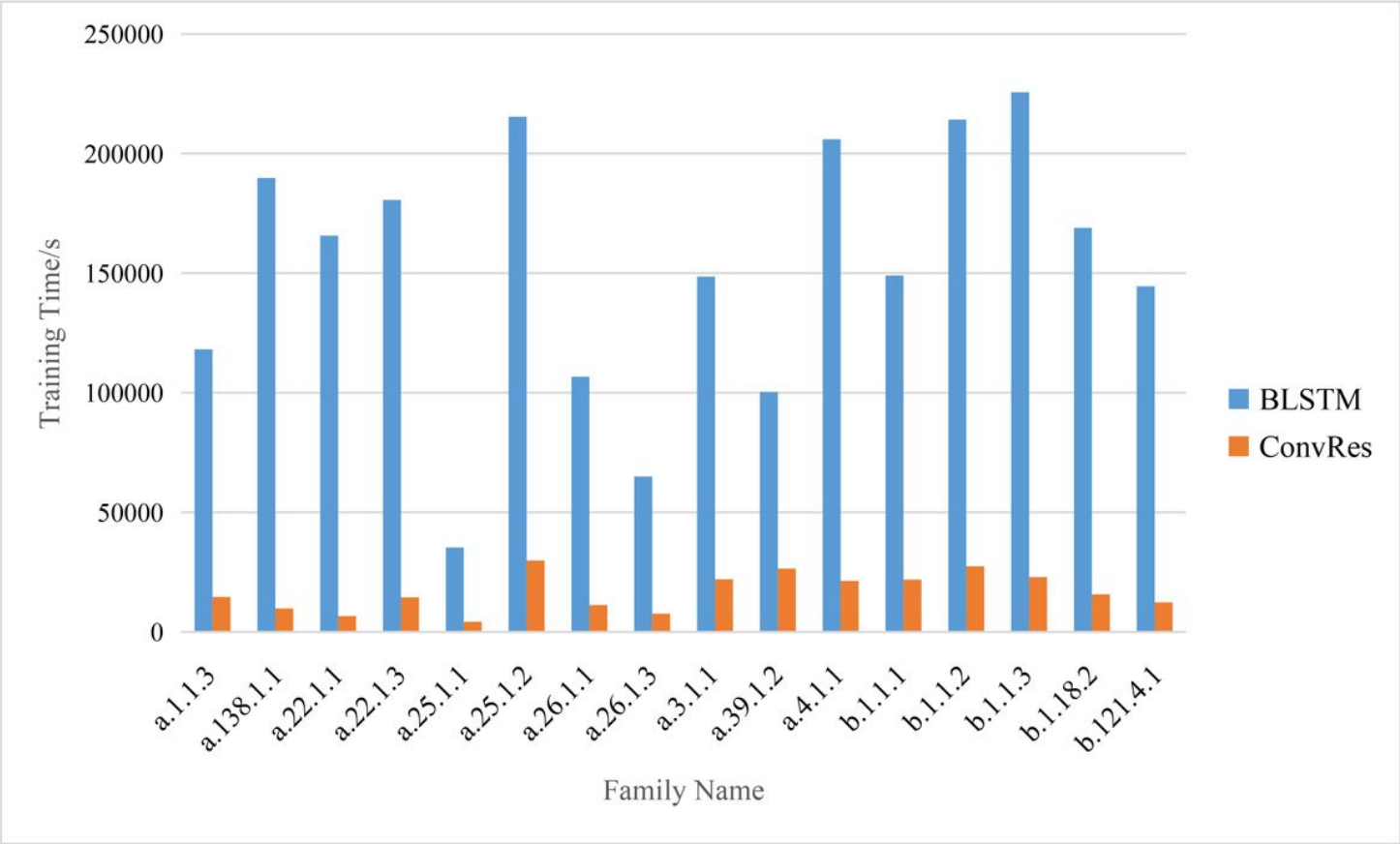

Figure 3

Training time of BLSTM and ConvRes model.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- eq0.5.jpg
- eq1.5.jpg
- eq1.jpg
- eq2.jpg