Content

1. Bioinformatics databases
2. Sequence alignment and database searching
3. Phylogenic tree and multiple sequence alignment
4. Protein structure alignment
5. Protein secondary structure prediction
6. Protein tertiary structure prediction

Protein secondary structure prediction

杨建益

Email: yangjy@nankai.edu.cn

Webpage: http://yanglab.nankai.edu.cn/

Course: http://yanglab.nankai.edu.cn/teaching/bioinformatics/

Office: 数学科学学院，419室

Content

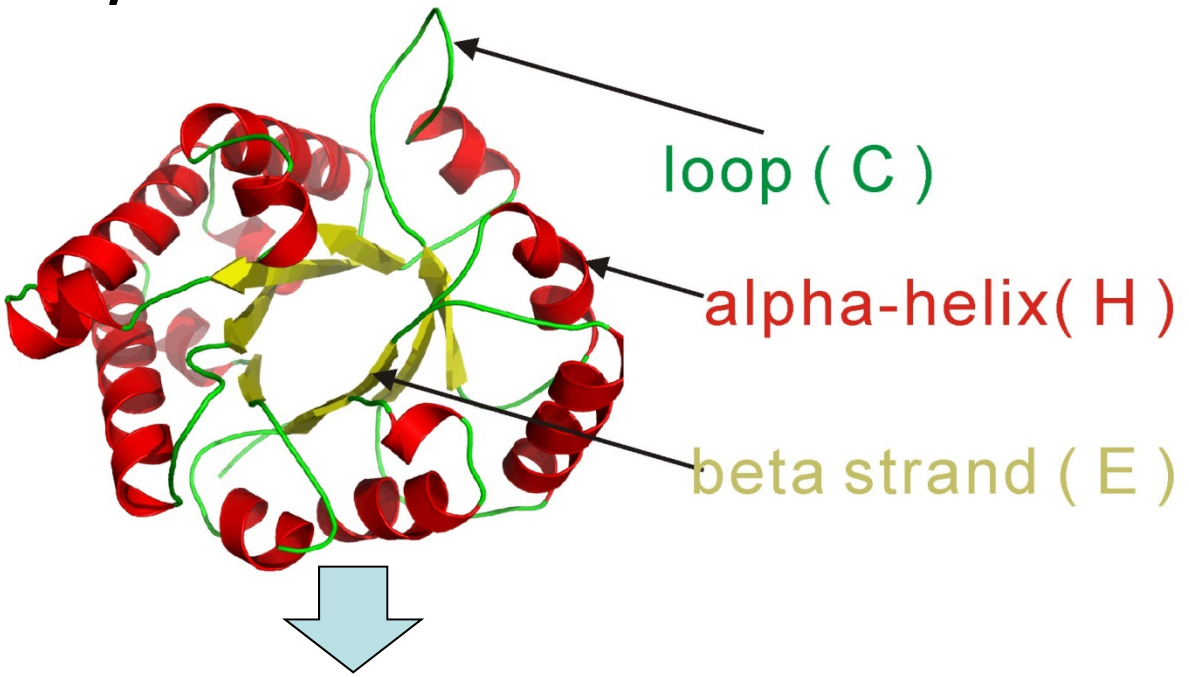
1. What is secondary structure?
2. Methods for predicting secondary structure
    a. PSIPRED
    b. Deep learning-based

What is secondary structure?

1, Primary structure: amino acid sequence (1D)

MVLEEGEWQLVLHVWAKVEADVAGHGQDILIRLFKEHPETLEKFDRVE
EAIIHVLHERHPGNFGADAQGAMNK

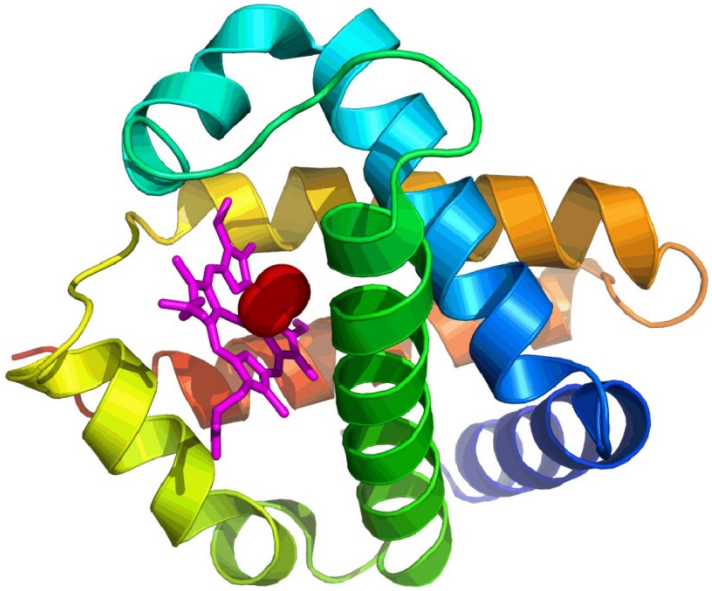
2, Secondary structure



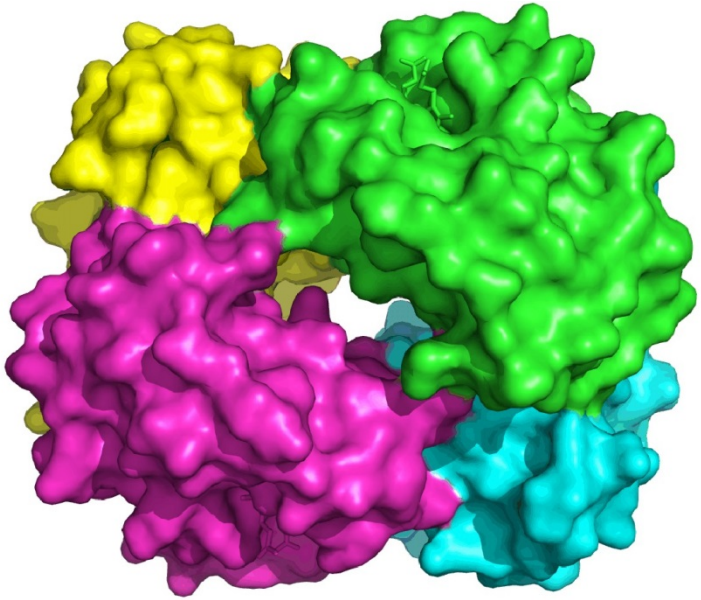
HHHHHHCCCEEEEECCCCCCCCCCHHHHHHHHHHCCCCEEE
EECCCCHHHHHHHHHHCCCCEEEECCCCCHHHHHHHH

What is secondary structure?

3, Tertiary structure



4, Quaternary structure



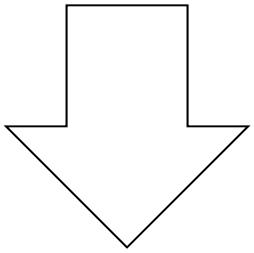
Content

1. What is secondary structure?
2. Methods for predicting secondary structure
    a. PSIPRED
    b. Deep learning-based

Secondary structure prediction

Problem:

MVLEEGEWQLVLHVWAKVEADVAGHGQDILIRLFKEHPETLEKFDRVE
EAIIHVLHERHPGNFGADAQGAMNK…



How to predict SS from amino acid sequence?

HHHHHHCCCEEEEECCCCCCCCCCHHHHHHHHHHCCCCEEEEECCCC
HHHHHHHHHHCCCCEEEECCCCCHHHHHHHH…

Can be solved with machine learning algorithms!

The machine learning framework

$$y = f(\mathbf{x})$$

output prediction function feature

- **Training:** given a *training set* of labeled examples $\{(\mathbf{x}_1,y_1), \ldots, (\mathbf{x}_N,y_N)\}$, estimate the prediction function f by minimizing the prediction error on the training set
- **Testing:** apply f to an un-seen *test example* $\mathbf{x}$ and output the predicted value $y = f(\mathbf{x})$

Briefings in Bioinformatics, 2016, 1–13

doi: 10.1093/bib/bbw129
Paper

Sixty-five years of the long march in protein secondary structure prediction: the final stretch?

Yuedong Yang, Jianzhao Gao, Jihua Wang, Rhys Heffernan, Kuldip Paliwal and Yaoqi Zhou

Corresponding author: Yaoqi Zhou, Institute for Glycomics, Griffith University, Parklands Drive, Southport, QLD 4222 28
Fax +61 (0)7 5552 9040; E-mail: yaoqi.zhou@griffith.edu.au

Abstract

Protein secondary structure prediction began in 1951 when Pauling and Corey predicted helical and sheet conformations for protein polypeptide backbone even before the first protein structure was determined. Sixty-five years later, powerful new methods breathe new life into this field. The highest three-state accuracy without relying on structure templates is now at 82–84%, a number unthinkable just a few years ago. These improvements came from increasingly larger databases of protein sequences and structures for training, the use of template secondary structure information and more powerful deep learning techniques. As we are approaching to the theoretical limit of three-state prediction (88–90%), alternative to second-

PSIPRED

Protein secondary structure prediction based on position-specific scoring matrices1

DT Jones - Journal of molecular biology, 1999 - Elsevier

Abstract A two-stage neural network has been used to predict protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST. Despite the simplicity and convenience of the approach used, the results are found to be superior to those produced by other methods, including the popular PHD method according to our own benchmarking results and the results from the recent Critical Assessment of Techniques for Protein Structure Prediction experiment (CASP3), where the method was evaluated by ...

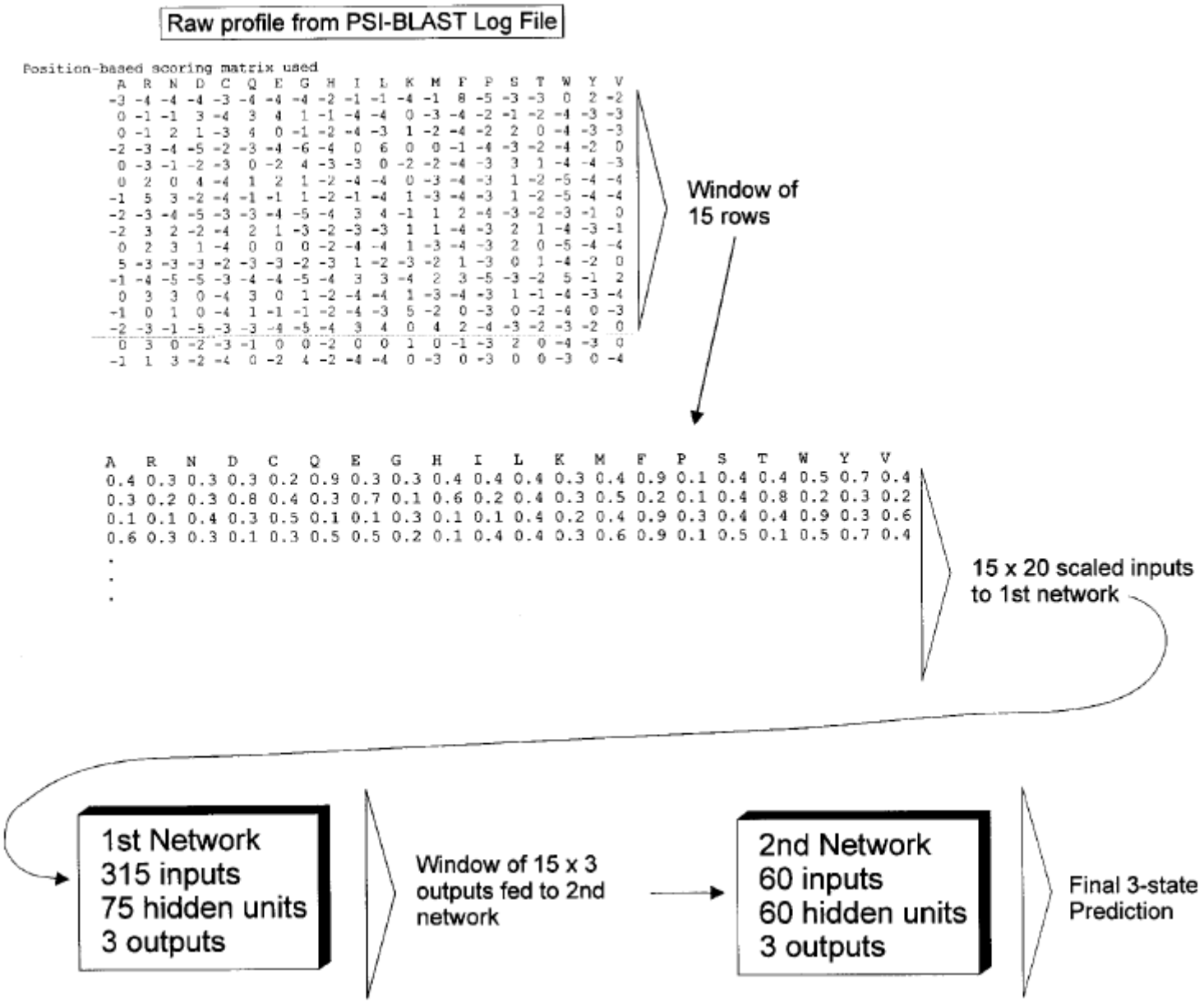
☆ 99 被引用次数：4872 相关文章 所有 29 个版本

Q3 accuracy: 80%

Professor David Jones

Welcome to my home page at University College London. I am currently Professor of Bioinformatics and Head of the Bioinformatics Group in the Department of Computer Science. I am also Director of the Bloomsbury Centre for Bioinformatics, which is a joint Research Centre between UCL and Birkbeck College and which also provides bioinformatics training and support services to biomedical researchers. My appointment is held jointly with the Department of Structural and Molecular Biology, although all mail should be addressed to the Computer Science Dept. as shown below.

PSIPRED


Raw profile from PSI-BLAST Log File
Position-based scoring matrix used
A R N D C Q E G H I L K M F P S T W Y V
-3 -4 -4 -4 -3 -4 -4 -4 -2 -1 -1 -4 -1 8 -5 -3 -3 0 2 -2
0 -1 -1 3 -4 3 4 1 -1 -4 -4 0 -3 -4 -2 -1 -2 -4 -3 -3
0 -1 2 1 -3 4 0 -1 -2 -4 -3 1 -2 -4 -2 2 0 -4 -3 -3
-2 -3 -4 -5 -2 -3 -4 -6 -4 0 6 0 0 -1 -4 -3 -2 -4 -2 0
0 -3 -1 -2 -3 0 -2 4 -3 -3 0 -2 -2 -4 -3 3 1 -4 -4 -3
0 2 0 4 -4 1 2 1 -2 -4 -4 0 -3 -4 -3 1 -2 -5 -4 -4
-1 5 3 -2 -4 -1 -1 1 -2 -1 -4 1 -3 -4 -3 1 -2 -5 -4 -4
-2 -3 -4 -5 -3 -3 -4 -5 -4 3 4 -1 1 2 -4 -3 -2 -3 -1 0
-2 3 2 -2 -4 2 1 -3 -2 -3 -3 1 1 -4 -3 2 1 -4 -3 -1
0 2 3 1 -4 0 0 0 -2 -4 -4 1 -3 -4 -3 2 0 -5 -4 -4
5 -3 -3 -3 -2 -3 -3 -2 -3 1 -2 -3 -2 1 -3 0 1 -4 -2 0
-1 -4 -5 -5 -3 -4 -4 -5 -4 3 3 -4 2 3 -5 -3 -2 5 -1 2
0 3 3 0 -4 3 0 1 -2 -4 -4 1 -3 -4 -3 1 -1 -4 -3 -4
-1 0 1 0 -4 1 -1 -1 -2 -4 -3 5 -2 0 -3 0 -2 -6 0 -3
-2 -3 -1 -5 -3 -3 -4 -5 -4 3 4 0 4 2 -4 -3 -2 -3 -2 0
0 3 0 -2 -3 -1 0 0 -2 0 0 1 0 -1 -3 2 0 -4 -3 0
-1 1 3 -2 -4 0 -2 4 -2 -4 -4 0 -3 0 -3 0 0 -3 0 -4
Window of 15 rows
A R N D C Q E G H I L K M F P S T W Y V
0.4 0.3 0.3 0.3 0.2 0.9 0.3 0.3 0.4 0.4 0.4 0.3 0.4 0.9 0.1 0.4 0.4 0.5 0.7 0.4
0.3 0.2 0.3 0.8 0.4 0.3 0.7 0.1 0.6 0.2 0.4 0.3 0.5 0.2 0.1 0.4 0.8 0.2 0.3 0.2
0.1 0.1 0.4 0.3 0.5 0.1 0.1 0.3 0.1 0.1 0.4 0.2 0.4 0.9 0.3 0.4 0.4 0.9 0.3 0.6
0.6 0.3 0.3 0.1 0.3 0.5 0.5 0.2 0.1 0.4 0.4 0.3 0.6 0.9 0.1 0.5 0.1 0.5 0.7 0.4
15 x 20 scaled inputs to 1st network
1st Network
315 inputs
75 hidden units
3 outputs
Window of 15 x 3 outputs fed to 2nd network
2nd Network
60 inputs
60 hidden units
3 outputs
Final 3-state Prediction

Deep learning-based methods

DeepCNF
Q3=~84%

SCIENTIFIC REPORTS

OPEN

Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields

Received: 28 June 2015
Accepted: 26 November 2015
Published: 11 January 2016

Sheng Wang[1,2], Jian Peng[3], Jianzhu Ma[1] & Jinbo Xu[1]

Protein secondary structure (SS) prediction is important for studying protein structure and function. When only the sequence (profile) information is used as input feature, currently the best predictors can obtain ~80% Q3 accuracy, which has not been improved in the past decade. Here we present DeepCNF

Deep learning-based methods

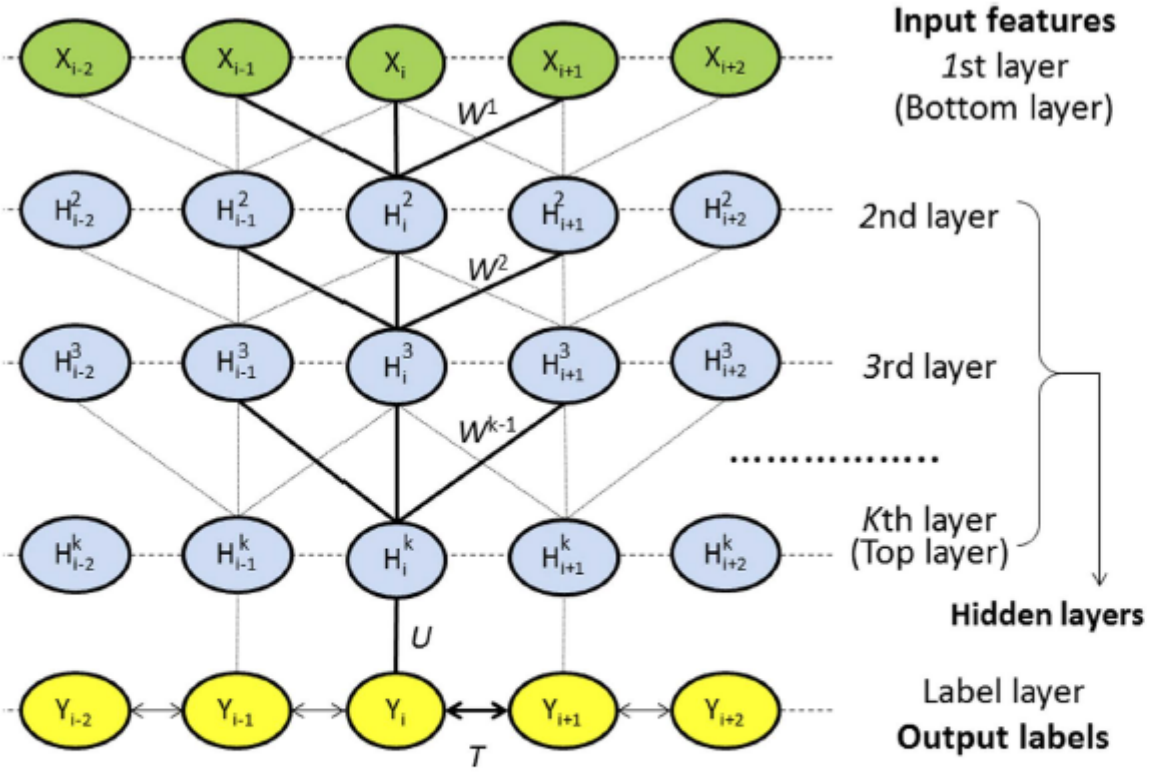


Figure 2. The architecture of DeepCNF, where i is the residue index and X_i the associated input features, H^k represents the k-th hidden layer, and Y is the output label. All the layers from the 1st to the top layer form a deep convolutional neural network (DCNN) with parameter $W^k\{k = 1, 2, \ldots, K\}$. The top layer and the label layer form a conditional random field (CRF) with U and T being the model parameters. U is the parameter used to connect the top layer to the label layer, and T is used to model correlation among adjacent residues.

Deep learning-based methods

SPIDER 3.0
Q3=84%

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

Structural bioinformatics

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

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

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

Deep learning-based methods

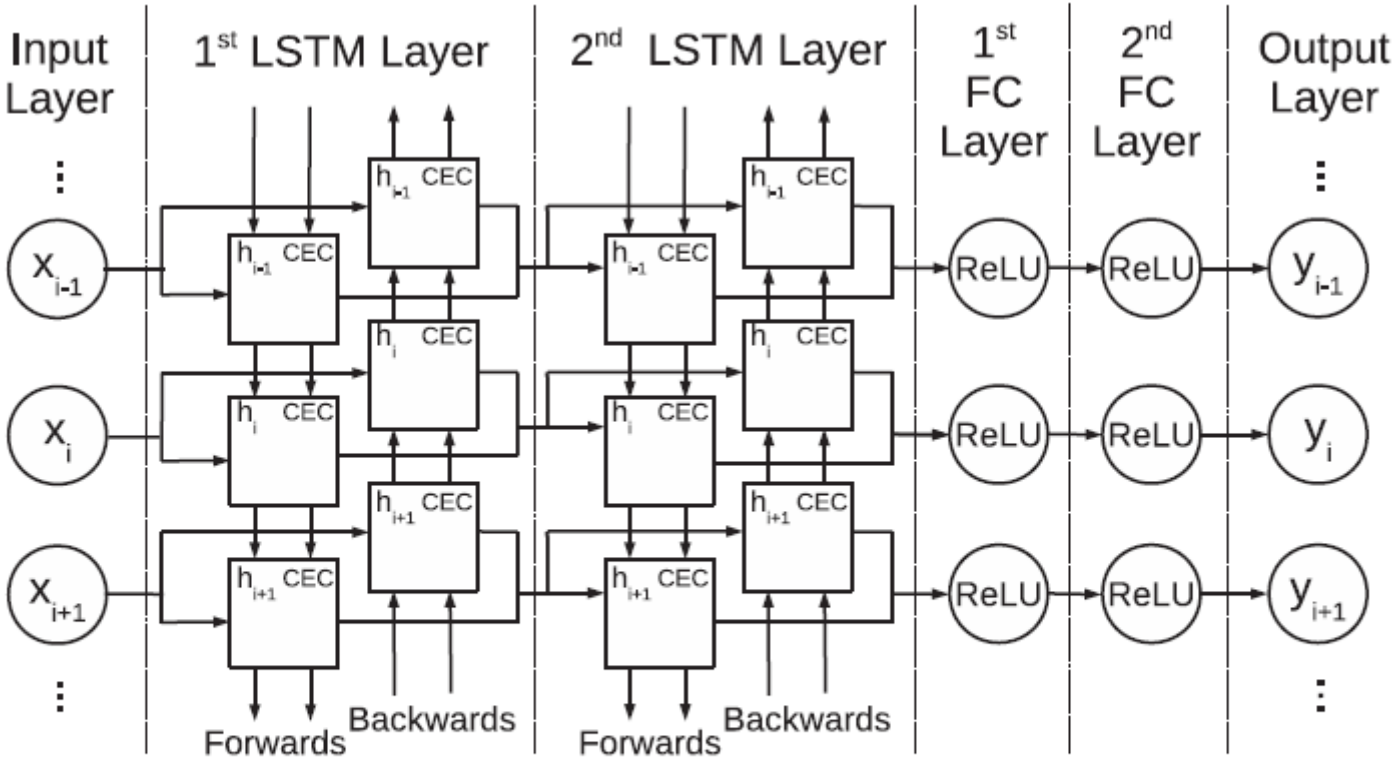


Fig. 1. Network architecture of LSTM-BRNN employed in all of the four iterations

Content

1. Bioinformatics databases
2. Sequence alignment and database searching
3. Phylogenic tree and multiple sequence alignment
4. Protein structure alignment
5. Protein secondary structure prediction
6. Protein tertiary structure prediction
7. Protein function prediction

Papers to read

Threading

J. U. Bowie, R. Luthy, D. Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. **Science**. (1991) 253:164-170.

MODELLER

A. Sali, T.L. Blundell. Comparative protein modelling by satisfaction of spatial restraints. J. Mol. Biol. 234, 779-815, 1993.

HHsearch

Söding J (2005). Protein homology detection by HMM-HMM comparison. Bioinformatics. 21 (7): 951–960.

Papers to read

Rosetta

K. T. Simons, C. Kooperberg, E. Huang, D. Baker. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. **J Mol Biol**. 1997 Apr 25;268(1):209-25.

I-TASSER

Yang et al, The I-TASSER Suite: protein structure and function prediction, Nature Methods, 12: 7-8 (2015).