A Machine Learning Approach To Protein Fold Recognition

Guided By : Mr. Uday Nayak

Project Team
Ashwin Fernandes
Kavya Kotian
Chiraag Limaye
Karan Manghi

Introduction

- Proteins

- key functional units in living organisms involved in many biological processes in the cell.

What is Protein Folding?

- The physical process by which a protein chain acquires its native 3-dimensional structure.
- A protein's three dimensional structure largely determines its function. The process for identifying these structurally similar proteins is called fold recognition

Need for the system

- Several experimental methods like X-ray crystallography, NMR spectroscopy, and electron microscopy have been used to determine protein structure.
- Due to the significant cost and time required for using those methods :

  - Number of proteins with known structure(s) << number of known protein sequences
- Compared with the traditional experimental methods, machine learning-based methods offer better advantages in terms of robustness and reliable performance.

Problem Statement

To train neural networks to predict if a protein pair belongs to the same structural fold.

System Design

- Recurrent Neural Network with three layers.
- Input layer :
  - 84 nodes (84 feature scores)
- Output layer :
  - Provides a binary classification output(0 or 1)
- Sigmoidal function used as the activation function.
- Weights assigned as per the significance of a particular feature

System Design

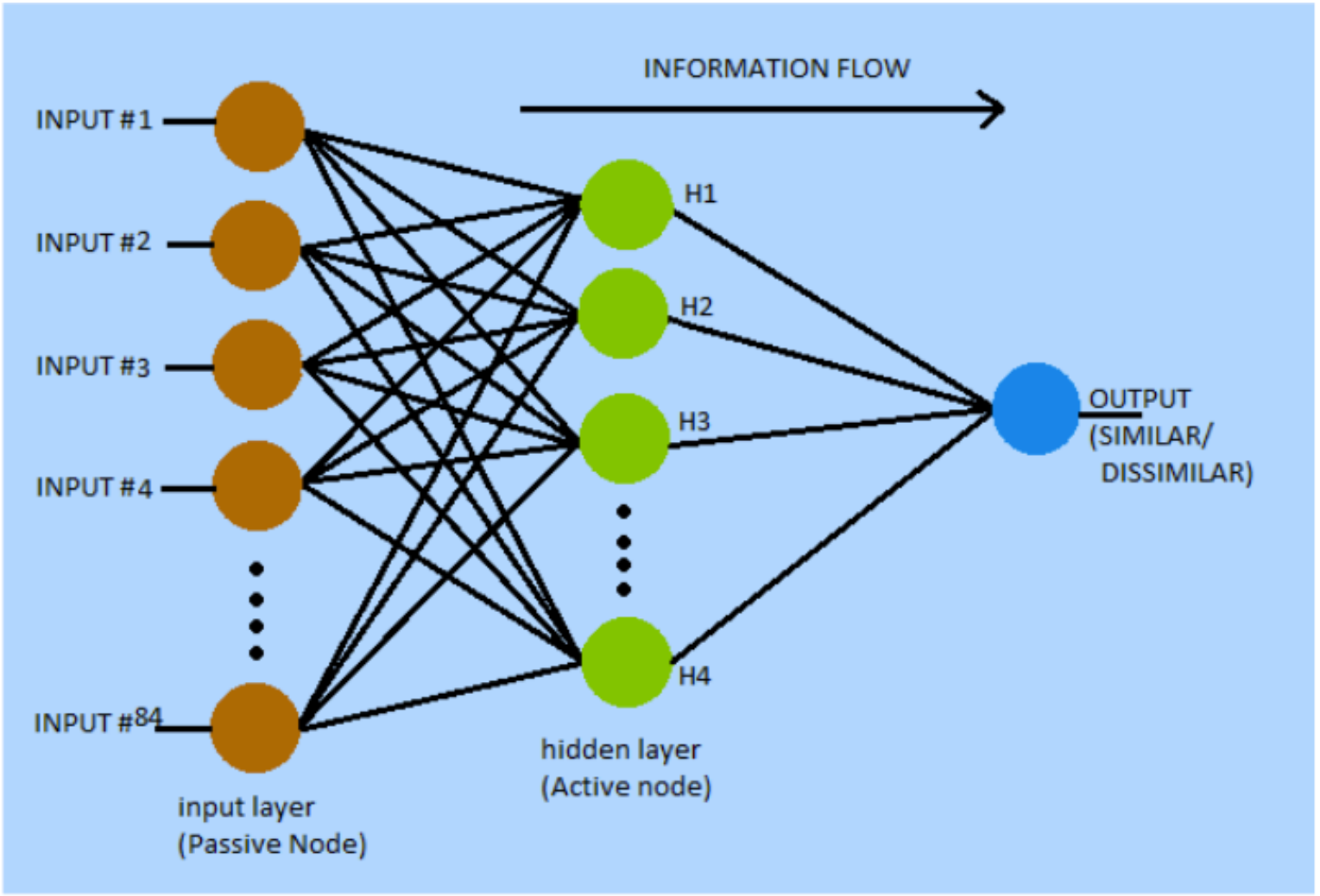

Figure 2.1: Neural Network

Implementation

- **Dataset**
  - Protein pairs having less than 40% sequence identity in the training set
  - Pairwise similarity calculated on basis of 5 types of sequence alignment tools
  - Dataset size : 3.8 MB

- Input data in text file format.Class labels specified as 1 or 0.
- Data normalized by min_max function.
- Cross validation : Data randomly split into testing and training set

Implementation

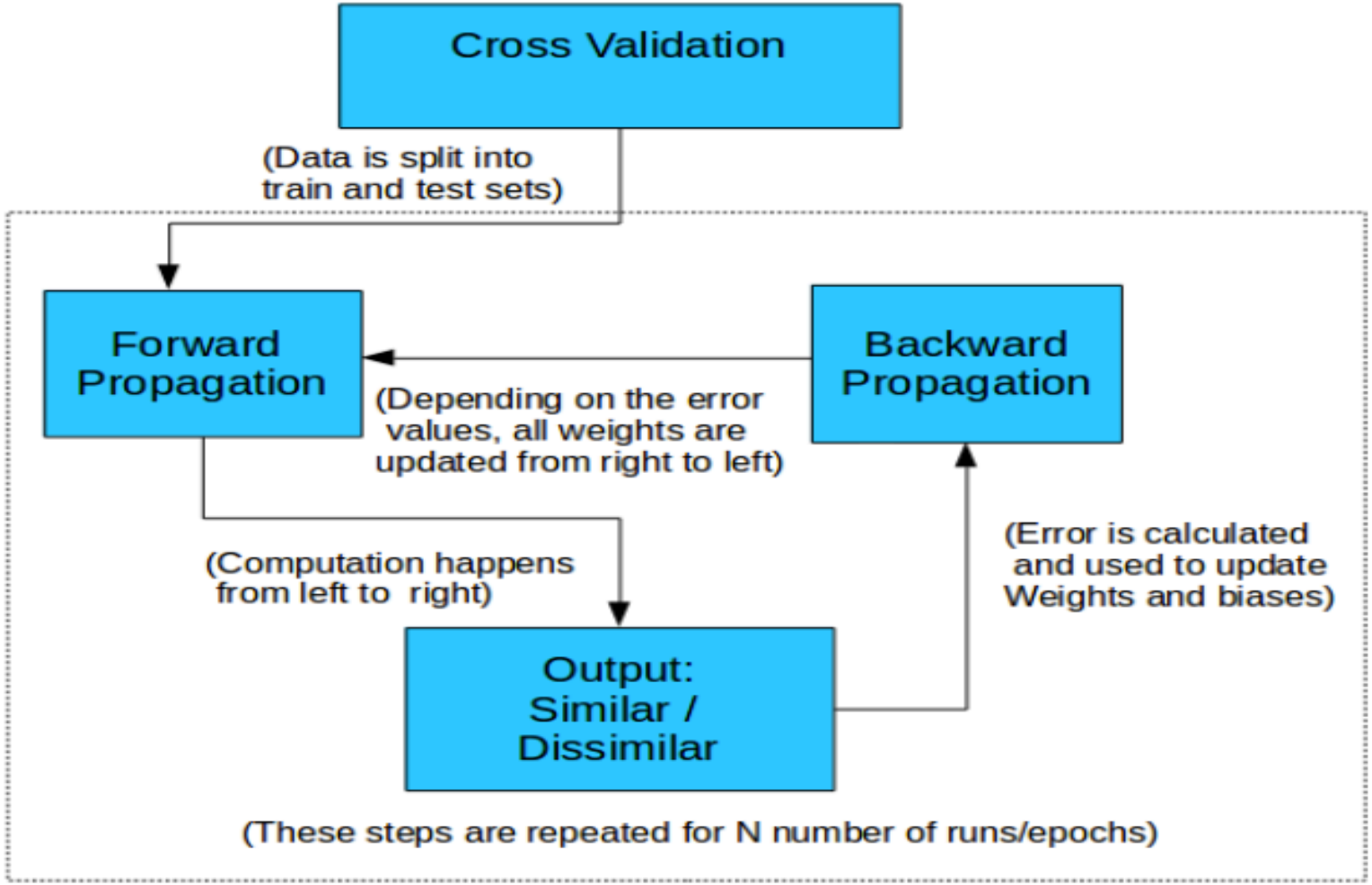

Figure 3.1: Block diagram of the Neural Network Implementation

Results

- The neural network is trained and tested with the help of a dataset of size 3.8MB
- Given the 84 feature scores for any two proteins our system can predict if the two protein contain similar folds.
- The existing software FOLDpro takes 54 input features into consideration.
- On conducting literature survey, it was observed that 30 more features contribute to the result inorder to increase accuracy. Thus our system considers 84 features cited by the latest discovery.

Conclusion & Future Work

- In this work , we have successfully implemented a Neural network to find whether a given protein pair belongs to the same fold.

- The accuracy of the system varies with the size of the dataset , the learning rate & the no of runs.

- **Future Scope :**
  - Implementing the data pre-processing module.
  - Using other machine learning methods to handle the Protein Fold Recognition problem better.

References

[1] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne; *The Protein Data Bank. Nucleic Acids Res* 2000; 28 (1): 235-242. doi: 10.1093/nar/28.1.235

[2] Leyi Wei and Quan Zou;*Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition*;School of Computer Science and Technology, Tianjin University, Tianjin 300354

[3] Jianlin Cheng, Pierre Baldi; *A machine learning information retrieval approach to protein fold recognition*; *Bioinformatics* 2006; 22(12):1456-1463.doi:10.1093/bioinformatics/btl102

[4] Stephen F. Altschul, Thomas L. Madden, Alejandro A. Scha ̈ffer, Jinghui Zhang, Zheng Zhang, Webb Miller, David J. Lipman; *BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res* 1997; 25 (17): 3389-3402. doi: 10.1093/nar/25.17.3389

[5] BLAST http://en.wikipedia.org/wiki/BLAST , last modified on 12 April 2017, at 02:26
[6] Jo, T. et al. *Homology Modeling of an Algal Membrane Protein, Heterosigma Akashiwo Na+ –ATPase. Membrane* 35(2), 80–85 (2010).
[7] Taeho Jo, Jie Hou, Jesse Eickholt Jianlin Cheng; *Improving Protein Fold Recognition by Deep Learning Networks*; SCIENTIFIC REPORT 5, ARTICLE NUMBER: 17573 (2015)
[8] Baker, D. Centenary Award and Sir Frederick Gowland Hopkins Memorial Lecture. *Biochemical Society Transactions* 42(2),225–229 (2014).
[9] Pressman, Roger (2010). *Software Engineering: A Practitioner's Approach. Boston: McGraw Hill*. pp.41–42. ISBN9780073375977.
[10] Murzin, A. G., Brenner, S. E., Hubbard, T. Chothia, C. *SCOP: a structural classification of proteins database for the investigation of sequences and structures.* Journal of molecular biology 247, 536–540 (1995).

[11] Matt Spencer, Jesse Eickholt, and Jianlin Cheng; *A Deep Learning Network Approach to ab initio*;IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL.12, NO. 1, JANUARY/FEBRUARY 2015

[12] Timothy K. Lee,Tuan Nguyen;*Protein Family Classification with Neural Networks*, STANFORD UNIVERSITY

[13] Helen M. Berman, John Westbrook, Zukang Feng, Gary Gilliland, T. N. Bhat, Helge Weissig, Ilya N. Shindyalov, Philip E. Bourne; *The Protein Data Bank. Nucleic Acids Res* 2000; 28 (1): 235-242. doi: 10.1093/nar/28.1.235

[14] Leyi Wei and Quan Zou;*Recent Progress in Machine Learning-Based Methods for Protein Fold Recognition*;School of Computer Science and Technology, Tianjin University, Tianjin 300354