PROTEIN SECONDARY STRUCTURE PREDICTION USING DEEP LEARNING TECHNIQUES

डीप लर्निंग तकनीक के उपयोग से प्रोटीन की द्वितीयक संरचना का अनुमान

Thesis submitted in partial fulfillment
For the Award of Degree of
DOCTOR OF PHILOSOPHY

by

ASHISH KUMAR SHARMA
आशीष कुमार शर्मा

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY
(BANARAS HINDU UNIVERSITY) VARANASI – 221 005

Roll No: 17071511 **February 2021**

CERTIFICATE

It is certified that the work contained in the thesis titled "**Protein Secondary Structure Prediction using Deep Learning Techniques**" by **ASHISH KUMAR SHARMA**, has been carried out under my supervision and that this work has not been submitted elsewhere for a degree.

It is further certified that the student has fulfilled all the requirements of Comprehensive, Candidacy and SOTA.

Signature of Supervisor

Prof. Rajeev Srivastava
Department of Computer Science and Engineering
Indian Institute of Technology (BHU), Varanasi

DECLARATION BY THE CANDIDATE

I, **ASHISH KUMAR SHARMA,** certify that the work embodied in this thesis is my own bonafide work and carried out by me under the supervision of **Prof. Rajeev Srivastava** from **December-2017 to January-2021**, at the **Department of Computer Science and Engineering**, Indian Institute of Technology (BHU), Varanasi. The matter embodied in this thesis has not been submitted for the award of any other degree/diploma. I declare that I have faithfully acknowledged and given credits to the research workers wherever their works have been cited in my work in this thesis. I further declare that I have not willfully lifted up any other's work, paragraphs, text, data, results, etc., reported in journals, books, magazines, reports dissertations, theses, etc., or available at websites and included them in this thesis and cited as my own work.

Date: Signature of the Student

Place: Varanasi (ASHISH KUMAR SHARMA)

CERTIFICATE BY THE SUPERVISOR

It is certified that the above statement made by the student is correct to the best of my/our knowledge.

Signature of Supervisor

(Prof. Rajeev Srivastava)

Department of Computer Science and Engineering

Indian Institute of Technology (BHU), Varanasi

Signature of Head of the Department

COPYRIGHT TRANSFER CERTIFICATE

Title of the Thesis: Protein Secondary Structure Prediction using Deep Learning Techniques

Name of the Student: ASHISH KUMAR SHARMA

Copyright Transfer

The undersigned hereby assigns to the Indian Institute of Technology (Banaras Hindu University), Varanasi all rights under copyright that may exist in and for the above thesis submitted for the award of the DOCTOR OF PHILOSOPHY.

Date:

Place: Varanasi (Ashish Kumar Sharma)

Note: However, the author may reproduce or authorize others to reproduce material extracted verbatim from the thesis or derivative of the thesis for author's personal use provided that the source and the Institute's copyright notice are indicated.

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my deep sense of gratitude to all who helped me directly or indirectly during this thesis work. Firstly, I would like to thank my supervisor, **Prof. Rajeev Srivastava**, for being a great mentor and the best adviser I could ever have. His advice, encouragement and critics are source of innovative ideas, inspiration are causes behind the successful completion of this Thesis work. The confidence shown on me by him was the biggest source of inspiration for me. It has been a privilege working with him from several years. I am highly obliged to all the faculty members of Computer Science and Engineering Department for their support and encouragement. I express my sincere thanks to the Prof. K. K. Shukla, Prof. A. K. Tripathi, Prof. R.B. Mishra, Prof. S.K. Singh, Dr. Bhaskar Biswas of the department of Computer Science and Engineering, Prof. Subir Das, Department of Mathematical Sciences and Prof Ranjit Mahanty, Department of Electrical Engineering, IIT (BHU), for providing continuous support, encouragement and advice. I express my sincere thanks to all the Professors, Deans, office staff, supporting staff and PhD Research Scholars of Indian Institute of Technology (BHU) Varanasi India. I express my gratitude to Director, Registrars, Deans, Heads, and Student Alumni of the Indian Institute of Technology (BHU) Varanasi. My memory of study period at IIT (BHU) can never be complete without mentioning my fellow research scholars. Special thanks to Dr. Nagendra Pratap Singh, Dr. Arvind Tiwari, Dr. Vibhav Prakash Singh, Dr. Kuntesh K Jani, Dr. Gargi Srivastava, Dr. Roshan Singh, Dr. Tribikram Pradhan, Mr. Sushant Pandey, Mr. Ashwini Singh, Mr. Nigmendra Pratap Yadav, Mr. Ankit Jaiswal, Mr. Santosh Kumar Tripathy, Ms. Pratishtha Verma, Mrs. Divya Singh for their great help and cooperation.

I extend special thanks to the non-teaching staff in the Department, particularly, Mr. Manoj Singh, Mr. Ravi Kumar Bharati, Mr. Prakhar Kumar, Mr. Ritesh Singh and Mr. Shubham Pandey for their consistent support. My parents, Late. Neelam Sharma and Shree Kuwar Prasad Sharma, who gave me the power and brain to work out on this research and their blessings at every level made me to see this success. Words are insufficient to express my profound sense of gratitude to my loving wife Mrs. Khushaboo Sharma, as well as my loving parents-in-law Smt. Champak Lata Rai and Shree Ram Anant Rai whose encouragement and blessings gave me physical and moral strength throughout my career as well in the present research. I extend my thanks to my brothers Mr. Devesh Kumar Sharma and Mr Ashutosh Rai, and my sisters Mrs. Vibha Singh and Mrs. Ankita Singh, my great daughter Shaivya Sharma who is my parts of inspiration. Finally, I would like to wind up by paying my heartfelt thanks and prayers to the Almighty for his unbound love and grace.

Ashish Kumar Sharma

TABLE OF CONTENT

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
ASA	Accessible Surface Area
Bi-GRU	Bidirectional Gated Recurrent Unit
Bi-LSTM	Bidirectional Long Short Term Memory
Bi-RNN	Bidirectional Recurrent Neural Network
CNN	Convolutional Neural Network
CPU	Central Processing Unit
DL	Deep Learning
DNN	Deep Neural Network
DSSP	Define Secondary Structure of Protein
GOR	Garnier Osguthorpe-Robson
GPU	Graphics Processing Unit
GRU	Gated Recurrent Unit
HMM	Hidden Markov Model
LSTM	Long Short Term Memory
ML	Machine Learning
MSA	Multiple Sequence Alignment
NLP	Natural Language Processing
NMR	Nuclear Magnetic Resonance
NN	Neural Network
PSSM	Position Specific Scoring Matrix
TPR	True Positive Rate

TNR	True Negative Rate
FPR	False Positive Rate
FNR	False Negative Rate
PSSP	Protein Secondary Structure Prediction
RNN	Recurrent Neural Network
SVM	Support Vector Machine
RF	Random Forest
PseAAC	Pseudo Acid Composition

LIST OF SYMBOLS

B	Beta Bridge
b	biases
C	Coil Regions
E	Beta Strand
F	Activation Function
f_t	Forget Gate
G	310-helix
H	Alpha Helix
h_t	State Vector
I	pi-helix
i_t	Input Gate
K	Secondary Structure Classes
L	Length of Amino Acid Sequence
o_t	Output Gate
Q3	Three State Per Residue Accuracy
Q8	eight State Per Residue Accuracy
S	loop
T	Beta Turn
tanh	Tangent Function
V	Size of n-gram words vocabulary
W	Weight Matrix
x_t	Input Vector

Z	Probability Vector
Θ	Hadamard Product
Σ	Sigmoid Function

PREFACE

Protein is an essential molecule for defining the structure and function of body cells. However, protein sequences have enough information for determining their structure and function. There is a large gap between the number of protein sequences and the number of protein structures. Protein secondary structure prediction from its amino acid sequence is an important problem in molecular biology. Protein structural information is critical for understanding their function, but many protein sequences do not have their structural information. The large gap between protein sequence and structure motivates computational methods and, in particular, deep learning methods predicting structural information from the protein primary sequences.

In this thesis, we focused on predicting protein secondary structure from the primary sequences. Protein primary sequences represented using subsequence-based representations of amino acid sequences. The dense embedding vectors of primary sequences are trained over a large dataset of primary sequences. These dense embedding vectors combined with a bidirectional long short term memory network to capture the contextual information of amino acid residue for predicting secondary structure.

Firstly the thesis work focuses on primary sequence representation for deep learning methods. The primary sequences are segmented into twenty amino acid characters.

The dense embedding vector of amino acid residues combined with deep learning techniques for protein secondary structure prediction. Further, the primary sequences are a bag of character n-gram words. The different sizes of n-gram words computed and the dense vector of character n-gram words used to capture the local contextual

information for secondary structure prediction. The variable-length character n-gram words dense embedding vector used for secondary structure prediction to overcome the drawback of fixed size character n-gram word. The recurrent neural network used for secondary structure prediction, the bidirectional long short-term memory network, better capture the long contextual information.

Further protein secondary structure prediction mapped as a sequence to sequence modeling. The amino acid residues are labeled with secondary structure elements. The state vectors for encoder and decoder are computed using long short term memory network.