بسم الله الرحمن الرحيم

∞∞∞

تم رفع هذه الرسالة بواسطة / حسام الدين محمد مغربي

بقسم التوثيق الإلكتروني بمركز الشبكات وتكنولوجيا المعلومات دون أدنى مسئولية عن محتوى هذه الرسالة.

ملاحظات : لا يوجد

Ain Shams University
Faculty of Computer and Information Sciences
Information Systems Department
Bioinformatics Program

Artificial Intelligence Approach for Protein Sequence Analysis

A thesis submitted as partial fulfillment of the requirements for the degree of Master's in Information Systems, Bioinformatics program

By

Farida Alaaeldin Mostafa Mohamed

Teaching Assistant at Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

Under the Supervision of

Prof. Dr. Nagwa Badr
Professor at Information Systems Department,
Dean of Faculty of Computer and Information Sciences,
Ain Shams University

Prof. Dr. Rasha Ismail
Professor at Information Systems Department,
Vice Dean for Graduate Studies and Research,
Faculty of Computer and Information Sciences,
Ain Shams University

Dr. Yasmine Afify
Lecturer at Information Systems Department,
Faculty of Computer and Information Sciences,
Ain Shams University

2022
Cairo

Acknowledgment

The assistance of my supervisors made this thesis achievable. I would want to thank Dr. Nagwa Badr, for her genuineness and support, which I will never forget. She is the icon of leadership and the ideal role model. Dr. Rasha Ismail has been an inspiration to me while I worked through this master's degree. This thesis would not have been possible without the help of Dr. Yasmine Afify, who guided me through the research process from the beginning and helped me gain a better grasp of the subject.

I am grateful for the tremendous possibilities my supervisors provided for me to progress professionally and for the extraordinary experiences they planned for me.

I am grateful for my parents' unwavering love and support, which keeps me driven and self-assured. My achievements and success are due to their belief in me. My gratitude goes out to my siblings, who remind me of the important things in life and are always encouraging in my experiences.

Finally, thank you to my friends for continually listening to me complain and chat things out, for cracking laughs when things grew too serious, and for making things better so that I could pursue a master’s degree.

Abstract

Protein sequence analysis helps in the prediction of protein functions. The objective of this thesis is to propose new deep learning models that are capable of classifying proteins based on their features extracted in either 1D or 3D and investigate the impact of data variations using 3D features on the deep learning-based protein sequence classification.

Regarding the 1D features, different protein descriptors were used and decomposed into modified feature descriptors using Empirical Mode Decomposition that were not employed in protein studies. Uniquely, we introduced using Convolutional Neural Network to learn and classify protein diseases. A dataset of 1563 protein sequences was classified into 3 different disease classes: AIDS, Tumor suppressor, and Proto-oncogene.

Results showed a significant increase in the performance of the Convolutional Neural Network model using modified feature descriptor over Support Vector Machine using rbf kernel function by 23.3% in accuracy. CTDT modified feature descriptors improved the deep learning model results by 19.5%, 39.6%, 23.3%, 29.9%, 24.3%, and 31.2% in AUC, MCC, accuracy, F1- score, recall, and precision, evaluation metrices respectively.

Regarding the 3D features, uniquely five feature extraction groups were utilized to create 3D features with two sizes (7x7x7 and 9x9x9). Three datasets are employed in the assessment, which are different in their sorts, sizes, and balance state namely, Disease and two Phage Virion Proteins datasets.

Results showed that the 7x7x7 feature matrix has a positive correlation between its dimensions, which has positive impact on the results reaching 71% in PVP-Balanced and 86% in disease dataset. Using the sum of the first three Intrinsic Mode Function components had a better impact than using the first component improving accuracy to 86.6% for disease dataset. The dataset size had a significant positive impact on training the Convolutional Neural Network model reaching 84%.

List of Publications

[1] Farida Alaaeldin Mostafa, Yasmine Mohamed Afify, Rasha Mohamed Ismail, Nagwa Lotfy Badr, "Protein Deep Learning Classification Using 3D Features," In 2021 Tenth International Conference on Intelligent Computing and Information Systems (ICICIS), 2021, pp. 462-466, DOI: 10.1109/ICICIS52592.2021.9694247

[2] Farida Alaaeldin Mostafa, Yasmine Mohamed Afify, Rasha Mohamed Ismail, Nagwa Lotfy Badr, “Deep Learning Model for Protein Disease Classification” In Current Bioinformatics, 2022; 17(3), pp. 245-253, DOI: 10.2174/1574893616666211108094205

[3] Farida A. Mostafa, Yasmine M. Afify, Rasha Ismail, Nagwa Badr. “Uncovering The Effects of Data Variation on Protein Sequence Classification Using Deep Learning” In International Journal of Intelligent Computing and Information Sciences, 2022; 22(2): 112-125. DOI: 10.21608/ijicis.2022.123177.1168

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AAC	Amino Acid Composition
ASMF	Adaptive Signal Model Fingerprinting
AUC	Area Under the Curve
CNN	Convolutional Neural Network
CTDC	C/T/D Composition
CTDT	C/T/D Transition
CTriad	Conjoint Triad
DFT	Discrete Fourier Transform
DWT	Discrete Wavelet Transform
ECG	ElectroCardioGram
EMD	Empirical Mode Decomposition
FN	False Negative
FP	False Positive
GAAC	Grouped Amino Acid Composition
IMFs	Intrinsic Mode Functions
MCC	Matthews Correlation Coefficient
MSA	Multiple Sequence Alignment
MSE	Mean Square Error
NADH	Nicotinamide Adenine Dinucleotide
ND5	Nicotinamide Adenine Dinucleotide dehydrogenase 5
ND6	Nicotinamide Adenine Dinucleotide dehydrogenase 6
Poly	Polynomial
PRD	Percent Root mean square Difference
PVP	Phage Virion Proteins

PVPred-SCM	Phage Virion Prediction-Scoring Card Method
RBF	Radial Basis Function
ReLU	Rectified Linear activation function
RQA	Recurrence Quantification Analysis
SCM	Scoring Card Method
SNRimp	improved Signal to Noise Ratio
SVM	Support Vector Machine
SVM-PCD	Support Vector Machine Physicochemical Distributions
SVM-RQA	Support Vector Machine Recurrence Quantification
TN	True Negative
TP	True Positive

Chapter 1

Introduction