Sequence-Based Prediction of Multiple Lipid Modification Sites in Proteins by Integration of Pseaac and Statistical Moments

WAQAR HUSSAIN
F2016279019

SUPERVISED BY
DR. YASER DAANIAL KHAN

SCHOOL OF SYSTEMS AND TECHNOLOGY
UNIVERSITY OF MANAGEMENT AND TECHNOLOGY
LAHORE

2018

SEQUENCE-BASED PREDICTION OF MULTIPLE LIPID MODIFICATION SITES IN PROTEINS BY INTEGRATION OF PSEAAC AND STATISTICAL MOMENTS

By

WAQAR HUSSAIN
(F2016279019)

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF SCIENCE IN COMPUTER SCIENCE

SCHOOL OF SYSTEMS AND TECHNOLOGY
UNIVERSITY OF MANAGEMENT AND TECHNOLOGY
LAHORE

2018

SUPERVISED BY

DR. YASER DAANIAL KHAN

CO-SUPERVISED BY

DR. NOUMAN RASOOL

DECLARATION

I **Waqar Hussain** s/o **Altaf Hussain** ID: **F2016279019**, Session 2016-2018 hereby declare that the matter printed in the thesis titled "**Sequence-Based Prediction of Multiple Lipid Modification Sites in Proteins by Integration of PseAAC and Statistical Moments**" is my own work and has not yet been printed, published and submitted as research work, thesis or publication in any form in any university, research institution etc. in Pakistan or abroad.

Dated: **Waqar Hussain**

DEDICATION

I dedicate this thesis

to

my beloved Parents and Teachers, without their prayers

and efforts

I would have never been able

to

work hard and gain this achievement.

PREFACE

This thesis, entitled “Sequence-Based Prediction of Multiple Lipid Modification Sites in Proteins by Integration of PseAAC and Statistical Moments” is based on the prediction of multiple lipid modification sites in a protein.

Lipid modification of a protein is known for regulation of various physiological factors, such as protein-membrane interactions, protein-protein interactions, protein stabilization and enzymatic functionality. Identification of lipid modification sites through experimental mechanisms i.e. site-directed mutagenesis and high throughput mass spectrometry can be costly, labour associated and time-consuming. Therefore, it is observed that there is a dire need for the development of a computational method which can help in predicting the lipid modification sites in an efficient and accurate way. Due to the association of these lipid modification sites with various diseases, its timely prediction can help in diagnosing and controlling the associated fatal diseases. Chapter 1 of this thesis describes the introduction of the study. Chapter 2 describes the materials and methods of the study. Chapter 3 illustrates the results. Chapter 4 provides comparative analysis and discussion on results. Chapter 5 describes the webserver. Chapter 6 concludes this study. The dataset for NMG, SFC, SGC and SPC is provided in Appendix A, B, C and D, respectively.

The study and the proposed predictor are beneficial for scientists to predict lipid modification sites in protein in an accurate, efficient and cost-effective way.

ACKNOWLEDGEMENT

All praise to **Almighty Allah**, most tolerant and most humane who empowered me and has given me capacities to do some research work and to add to the honourable field of learning.

This thesis depicts the examination work embraced at School of Systems and Technology, University of Management and Technology, Lahore under the supervision of **Dr Yaser Daanial Khan** to whom I am very obligated for supervision, proposing the subject, entire time direction, support and recommendations. His recommendations, discussions, directions and remarks were consistently a source of inspiration and enthusiasm for me. He is always kind and prepared to help the students with their problems in research. I am very obliged to him for his supervision, support, knowledge sharing and valuable time.

I would like to express my special appreciation and thanks to my co-supervisor **Dr Nouman Rasool**, who have been a tremendous mentor to me. I would like to thank him for encouraging me in my research and for allowing me to grow. His advice, suggestions and encouragements have always been priceless in every field of my life.

Last but not least, I feel insufficiency in vocabulary to discover appropriate words to express my emotions for my **Parents** who raised and groomed me all through of my life, whose hands are constantly raised for prayers which made me effective in each field of my life. Their day and night support for me empowers me to join higher thoughts of life, taking care of all of the issues and to achieve my objectives.

Waqar Hussain

TABLE OF CONTENTS

DECLARATION i

DEDICATION ii

PREFACE iii

ACKNOWLEDGEMENT iv

LIST OF FIGURES vii

LIST OF TABLES viii

ABSTRACT ix

CHAPTER 1. INTRODUCTION 1

1.1 BACKGROUND 1

1.2 *N*-Myristoylation 1

1.3 Prenylation 4

1.4 Palmitoylation 4

1.5 PROBLEM STATEMENT 5

1.6 RELATED WORK 5

1.7 OBJECTIVES OF THE STUDY 6

CHAPTER 2. MATERIALS AND METHODS 8

2.1 BENCHMARK DATASET 8

2.1.1 Benchmark Dataset for N-Myristoyl Glycine 8

2.1.2 Benchmark Dataset for S-Farnesyl Cysteine 9

2.1.3 Benchmark Dataset for S-Geranylgeranyl Cysteine 10

2.1.4 Benchmark Dataset for S-Palmitoyl Cysteine 12

2.2 SAMPLE FORMULATION 13

2.2.1 Statistical Moments Calculation 13

2.2.2 Site Vicinity Vector (SVV) 15

2.2.3 Constructing Position Relative Incidence Matrix (PRIM) and Reverse

Position Relative Incidence Matrix (RPRIM).. 15

2.2.4 Frequency Matrix (FM) Determination 16
2.2.5 Computation of Accumulative Absolute Position Incidence Vector (AAPIV) and Reverse Accumulative Absolute Position Incidence Vector (RAAPIV) 17
2.3 PREDICTION ALGORITHM 17
CHAPTER 3. RESULTS 20
3.1 ESTIMATED ACCURACY 20
3.1.1 Metrics for accuracy estimation 20
3.1.2 Self-consistency Testing 21
3.1.3 Testing via 10-fold cross-validation 25
CHAPTER 4. COMPARATIVE ANALYSIS AND DISCUSSION 27
CHAPTER 5. DEVELOPMENT OF WEBSERVER 32
CHAPTER 6. CONCLUSION 33
REFERENCES 35
APPENDIX A 41
Positive dataset for NMG sites - 893 41
Negative dataset for NMG sites – 1093 43
APPENDIX B 46
Positive dataset for SFC sites – 90 46
Negative dataset for SFC sites – 100 46
APPENDIX C 47
Positive dataset for SGC sites – 74 47
Negative dataset for SGC sites – 100 47
APPENDIX D 48
Positive dataset for SPC sites – 436 48
Negative dataset for SPC sites – 500 49

LIST OF FIGURES

Figure 1.1: Myristate - A 14 carbon saturated fatty acid 2
Figure 1.2: N-myristoylation at (a) co-translational and (b) post-translational level ... 3
Figure 1.3: Attachment of Myristate with HIV-1 Matrix Protein. Blue chain represents the myristate, red represents the helices and yellow represents the random coil (PDB ID: 1UPH) 4
Figure 1.4: Graphical illustration of the 5-step rule 7
Figure 2.1: Flowchart for the methodology 8
Figure 2.2: Graphical representation of data flow in ANN 18
Figure 3.1: Self-consistency testing confusion matrix for NMyristoylG-PseAAC 21
Figure 3.2: Self consistency testing confusion matrix for SFarnesylC-PseAAC..... 22
Figure 3.3: Self consistency testing confusion matrix for SGeranylgeranylC-PseAAC 23
Figure 3.4: Self-consistency testing confusion matrix for SPalmitoylC-PseAAC..... 24
Figure 4.1: ROC curve for NMyristoylG-PseAAC vs. GPS-Lipid (N-Myristoylation) 28
Figure 4.2: ROC curve for SFarnesylC-PseAAC vs. GPS-Lipid (S-Farnesylation) .. 28
Figure 4.3: ROC curve for SGeranylgeranylC-PseAAC vs. GPS-Lipid (S-Geranylgeranylation) 29
Figure 4.4: ROC curve for SPalmitoylC-PseAAC vs. GPS-Lipid (S-Palmitoylation) 29
Figure 5.1: The graphical user interface of the prediction webserver developed in this study 32

LIST OF TABLES

Table 2.1: Amino acid indexing for Site Vicinity Vector 15
Table 3.1: Summarized self-consistency testing results for multiple lipid modification sites prediction..... 24
Table 3.2: 10-fold cross-validation results for multiple lipid modification sites predictors (Average of 10 folds) 25
Table 4.1: Comparative analysis of multiple lipid modification sites predictor proposed in this study and GPS-Lipid using benchmark dataset of the proposed study 30

ABSTRACT

Lipid modification of a protein, which can be co-translational or post-translational, is known for regulation of various physiological factors, such as protein-membrane interactions, protein-protein interactions, protein stabilization and enzymatic functionality. Due to the association of these lipid modification sites with various diseases, its timely prediction can help in diagnosing and controlling the associated fatal diseases. Here, we present a method for prediction of multiple lipid modification sites, in which we have incorporated PseAAC with statistical moments for the prediction. The aim of this study is to propose a new and more accurate predictor for lipid modification sites, based on the 5-step rule, to make it easier for the experimental scientists getting desired results. A benchmark dataset of 893 positive and 1093 negative samples for *N*MyristoylG-PseAAC, 90 positive and 100 negative samples for *S*FarnesylC-PseAAC, 74 positive and 100 negative samples for *S*GeranylgeranylCPseAAC, and 436 positive and 500 negative samples for *S*PalmitoylC -PseAAC, is collected and used in this study. For feature vector, various position and composition relative features along with the statistical moments are calculated. Later on, a back propagation neural network is trained using feature vectors and scaled conjugate gradient descent with adaptive learning is used as an optimizer. Self-consistency testing and 10-fold cross-validation are performed to evaluate the performance of predictors, using accuracy metrics. For self-consistency testing of *N*MyristoylG-PseAAC, 96.93% *Acc*, 97.09% *Sp*, 96.80% *Sn* and 0.94 *MCC* is observed, whereas, for 10-fold cross validation 94.41% *Acc*, 94.06% *Sp*, 94.70% *Sn* and 0.89 *MCC* is observed. For self-consistency testing of *S*FarnesylC-PseAAC, 95.79% *Acc*, 96.67% *Sp*, 95.00% *Sn* and 0.92 *MCC* is observed, whereas, for 10-fold cross validation 93.68% *Acc*, 95.56% *Sp*, 92.00% *Sn* and 0.87 *MCC* is observed. For self-consistency testing of *S*GeranylgeranylC-PseAAC, 95.91% *Acc*, 95.77% *Sp*, 96.00% *Sn* and 0.92 *MCC* is observed, whereas, for 10-fold cross validation 92.98% *Acc*, 92.96% *Sp*, 93.00% *Sn* and 0.86 *MCC* is observed. For self-consistency testing of *S*PalmitoylC-PseAAC, 98.08% *Acc*, 98.62% *Sp*, 97.60% *Sn* and 0.96 *MCC* is observed, whereas, for 10-fold cross validation 94.66% *Acc*, 96.79% *Sp*, 92.80% *Sn* and 0.89 *MCC* is observed. Thus the proposed predictor can help in predicting the targeted lipid modification sites in an efficient and accurate way.

CHAPTER 1. INTRODUCTION

1.1 BACKGROUND

Cellular membranes, which are comprised of lipids, provide a barrier and boundary for cell proliferation and survival. This illustrates that lipids are one of the essential molecules for biological systems (Jiang *et al.*, 2018). Protein-membrane interactions are also controlled and regulated using lipids as certain proteins have the ability to bind to specific lipid molecules (Eisenhaber *et al.*, 2003). Moreover, the most frequent interaction mechanism of lipid molecules with proteins is the covalent modification/attachment, known as lipidation in proteins (Blanden *et al.*, 2017).