Thesis Report

Prediction of disulfide bonding sites in proteins using position and composition relative feature vectors

By

Mehreen Jamil (14024050003)

A Dissertation Submitted For the Degree of Master of Science in Computer Science in the Department of Computer Science, Faculty of Science, University of Management and Technology, Lahore

April 2017

Supervised by:

Dr. Yaser Daanial Khan

Co-Supervisor

Dr. Nouman Rasool

Copyright©2017 Mehreen Jamil

Certificate of Approval

It is certified that the research work presented in this thesis entitled "Prediction of Beta - Lactamase using Position and Composition Variant Features and Neural Networks" was conducted by

Mehreen Jamil under the supervision of Dr. Yaser Daanial Khan.

No part of this thesis has been submitted anywhere else for any other degree.

This thesis is submitted to Department of Computer Science for the partial fulfillment of the requirement for the degree of Master of Science in Computer Science

at the

Department of Computer Science, University of Management and Technology,

Lahore, Pakistan

April 2017

……	……
Supervisor	*Signature*
……	……
Director Graduate Studies	*Signature*
……	……
COD	*Signature*
……	……
Dean	*Signature*

Declaration

I declare that this dissertation is my own original work. Where collaborations with other researchers are involved, or materials generated by other researchers are included, the parties and/or materials are acknowledged or are explicitly referenced as appropriate.

This work is being submitted for the degree of Master of Science in Computer Science at the University of Management and Technology. This thesis has not been submitted to any other university or institution for any other degree or examination.

_____ _____

Date Signature

Mehreen Jamil

Dedication

I wholeheartedly thank and dedicate all my success to each one of them.

- Parents – For their unconditional support.
- Prof.Yasir Daniyal Khan – For support and advisory he provided for the successful execution of my thesis .

Table of Contents .....

List of Figures:

List of Tables:

Preface

This work helps to find out the proper disulfide bonds of a protein by using statistical features of residual values of amino acids .In order to predict disulfide bonding, a mathematical model is established. Chapter 1 labeled as “Introduction” discusses the importance of disulfide bonds of a protein’s binding and stability. Related work is also discussed in chapter 1 .Chapter 2 labeled as “Materials and Methods” describing the data collection sources, preprocessing, statistical moments, feature Vector and Position Relative Incidence Matrix, Reverse Position Relative Incidence Matrix, Frequency Matrix, Reverse Accumulative Absolute Position Incidence Vector and Accumulative Absolute Position Incidence vector. Neural Network training is also discussed in it. Chapter 3 labeled as “Results and Discussion” discuss the ROC and confusion matrix of proposed model. Different testing on both accumulative and independent data set are described in this section. Comparison of proposed system with other existing techniques also discussed in chapter 3. Chapter 4 labeled as “Conclusion” describes that current work is more accurate and efficient than others to predict the disulfide bonds of protein. Positive and Negative data sets with related accession id, are showed in Appendix section .

Acknowledgement

In this book complete design and documentation of my project is given. This is a mathematical model to predict disulfide bonds of a protein. Important source codes and related information is also provided so that this can be used for further improvement of this product.

I would like to put on record, my appreciation and gratitude to all who have rendered their support and input. Without them, it would not have been possible for me to shape this study.

I have received immense guidance from my Advisor Dr. Yaser Daanial Khan (Assistant Professor of Computer Science Department) and Dr. Nouman Rasool. I would therefore like to convey my sincere gratitude to them for their continuous help, support and time during the entire course of the project.

Especial thanks to all our teachers at UMT for enabling me to achieve my goals and fulfill my ambitions.

I am sincerely thankful to my family and special thanks to Sohaib Atif for his never ending love and support.

Information Processing Center (IPC) Staff for enduring me all these years and to anyone and everyone who has played some part in what I am today.

ABSTRACT

The presence of disulfide bonds in a protein confers an additional stability to protein against various threats. The formation of correct disulfide bonds between cysteine residues ensures proper folding of the protein during in vivo and in vitro folding process. Oxidation of these bonds may disturb the proper biological activity of a protein. Not all cysteines in a protein are involved in the formation of disulfide bonds,therefore prediction of accurate disulfide bonds is crucial for structural and functional relationship of a protein. Many neurodegenerative diseases are caused by the improper formation of disulfide bonds in the nervous system .The determination of fallacious S-S interaction is crucial for correct diagnosis of these disorders. In this study a novel method is used to predict the intra molecular disulfide bonding accurately using context based information. The surrounding amino acids of the cysteine, involved in disulfide bond, play a vital role in making disulfide bonds and used as feature vectors. The proposed method uses context-based data to calculate statistical moments. Statistical moments are very important as they are very sensitive regarding to position of data sequences. For prediction of intra molecular disulfide bonds, these moments are combined together to train neural networks. 10-fold validation on accumulative dataset gives us 87.52% accurate result. Estimation of accurate disulfide bonding against independent data set for 5-fold and 10-fold result in 82.4% and 88% respectively. The overall accuracy of system is 86.3% to sensitivity value 82.4% and specificity 93%.

Introduction

1.1Introduction:

Proteins are major constituents of a cell and perform a variety of functions inside and outside the living organisms. The backbone of proteins is made up of strong peptide bonds between amino acids. Additional stability of a protein is acquired by making intra molecular ionic interactions, hydrogen bond, van der Waals forces and disulfide bonds. Disulfide bond is a covalent linkage between two cysteine amino acids of a protein and is considered a strong bond after peptide link. These bonds are formulated in endoplasmic reticulum (ER) and help in attaining proper shape and strength by proteins. These bonds result in conformational changes in protein by decreasing entropy of unfolded state of a protein [36]. It is reported that 15% of human protein has dissolved bonds to perform proper function. A large fraction of secreted proteins have these bonds in order to resist the cellular environmental changes [37].