In the name of Allah, the Compassionate,
The Merciful

Prediction of Beta Lactamase using Position and Composition Variant Features and Neural Networks

By

Muhammad Adeel Ashraf (15025050002)

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

March 2017

Supervised by:

Dr. Yaser Daanial Khan

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

Muhammad Adeel Ashraf 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

March 2017

.....	.....
Supervisor	*Signature*
.....	.....
External Examiner	*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

Muhammad Adeel Ashraf

Dedication

This thesis is dedicated to my Sweet parents

Their proficient

Guidance and splendid inspiration had

Positively influenced my life and made me capable

Of attaining this degree

Successfully.

Table of Contents

List of Figures:

List of Tables

Preface

This thesis aids in medical field to identify beta-lactamase from non-beta-lactamase sequences with its respective class using position and composition variant features. For this purpose, a prediction model is established.

Chapter 1 entitled “Introduction” discusses the role of enzymes, antibiotics, beta-lactamases and its classification, and related work. Chapter 2 entitled “Materials and Methods” discusses the data collection of sequences, Position Relative Incidence Matrix (PRIM), Reverse Position Relative Incidence Matrix (RPRIM), frequency matrix, Accumulative Absolute Position Incidence Vector (AAPIV), and Reverse Accumulative Absolute Position Incidence Vector (RAAPIV) for feature extraction of protein sequences, and neural network. Chapter 3 entitled “Results and Discussion” discusses the numerous validation tests as compared to other existing methods for beta lactamase with its respective class and non-beta-lactamase. Chapter 4 entitled “Conclusion” discusses that proposed method is better than the previous methods and works effectively. Appendix contains the complete data set on which the prediction model is constructed.

From the discussion in this thesis I believed that it is possible for one to get a better understanding of beta-lactamase prediction model. This research endeavors to empower in medical field.

Acknowledgement

All praise to Almighty Allah alone, the Omnipotent, and the most Compassionate. His prophet “MUHAMMAD” (SAW), the most perfect and exalted among and of ever born on the surface of the earth, who is forever torch of guideless and knowledge for humanity as a whole.

I feel great pleasure in expressing my heartiest gratitude to Supervisor Dr. Yaser Daanial Khan (Assistant Professor of Computer Science Department) University of Management and Technology, Lahore Campus for kind behavior, valuable suggestions, worth and keen supervision, scholarly criticisms, sympathetic attitude towards completion of this thesis. His kind behavior and attitude during period of thesis work is unforgettable. I am thankful to Dr. Nouman Rasool for their Kind guidance and support in research work. I express my sincere thanks to all respectable teachers, faculty members of computer science department, and Information Processing Center (IPC) of University of Management and Technology, Lahore.

I feel great happiness in expressing my thanks to my family and special thanks to Naveed Ashraf, Sehrish Batool, and Hafsa Adeel for their love and support. Prayers of my family are a treasure for my life.

This dissertation report is a proof of sincerity of those who helped me during this thesis work. So I submit my earnest thanks again to all of them for their encouragement and moral support. Last but not least I pray for health, happiness, and prosperity of all the participants.

Muhammad Adeel Ashraf

15025050002

ABSTRACT

β-lactamase produced by different bacteria confers resistance against β-lactam containing drugs. The gene encoding β-lactamase is plasmid borne and can easily be transferred from one bacterium to another during conjugation. By such transformations, the recipient also acquires resistance against the drugs of β-lactam family. β-lactam antibiotics play an important role in clinical treatment of diseases like skin infections, soft tissue infections, urinary tract infections and bronchitis. Furthermore, based on the physical structure of beta-lactamases, they can be classified into four classes namely A, B, C, and D. Class A, C and D include enzymes that hydrolyze their substrate by forming an acyl enzyme through active site serine while class B are metallo-enzymes that utilize at least one active site zinc ion to hydrolyze. This Thesis presents a computationally intelligent technique formulated to predict whether a given protein belongs to beta-lactamase or non-beta-lactamase. If the given sequence is beta-lactamase then identify beta-lactamase to its respective class (Class A, B, C, D). The computational model uses primary amino acid sequence structure as its input. Various metrics are derived from the primary structure to form a feature vector. Experimentally determined data of positive and negative beta-lactamases are collected and transformed into a feature vector space. An adaptive training algorithm is used to train a back propagation neural network for prediction purposes. Rigorous verification and validation tests are performed and metrics are collected to evaluate the authenticity of the proposed model. To differentiate beta-lactamase from non-beta-lactamase with an accuracy of 89.1% and Mathew correlation coefficient of 0.7823. Tests also reveal that the proposed model yields of 94.1%, 95.2%, 93.8%, 97.7%, and 90.4% along with Mathew correlation coefficient of 0.64, 0.75, 0.84, 0.73, and 0.81 for classes A, B, C, D and non-beta-lactamase class respectively.

Chapter 1:

Introduction

1.1 Introduction:

Proteins are the most abundant organic compounds found in living cell and form about 50% of dry weight of living body. They are present in all types of cell and all parts of cell. Chemically speaking, proteins are long chains of amino acids. Composition of amino acids includes carbon, nitrogen, oxygen and hydrogen. These elements are arranged in special arrangement in which main carbon is called alpha carbon to whom is attached a hydrogen atom, an amino group ($-NH_2$), carboxyl group an –R group (this can be a hydrogen atom, $-CH_3$ group etc.

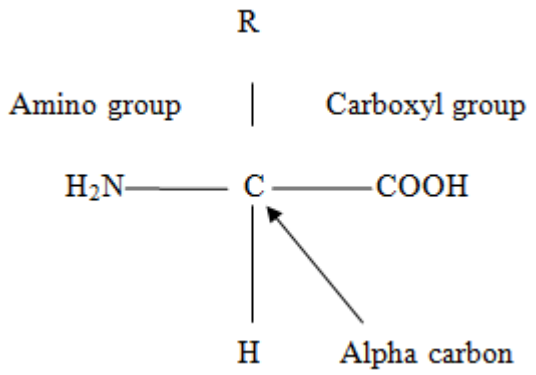

Figure 1: General Formula for Amino Acids

One amino acid is attached to other amino acid through peptide bond to form polypeptide chain. Polypeptide chain s may contain a few to 3000 or even more amino acids in proteins. About 170 types of amino acids are present in cells and tissues. Of these almost 25 are found in proteins. Most of proteins are however made up of 20 amino acids. These amino acids include: alanine, asparagine's, aspartic acid, arginine, glutamic acid, cysteine, glutamine, isoleucine, lysine, methionine, leucine, glycine-histidine, phenylalanine, serine, proline, threonine, tyrosine, tryptophan, and valine [1].