Prediction of Protein Solubility in *Escherichia coli* and Experimental Verification

By

Shahid Mehmood

14003140056

Advisor:

Dr. Nouman Rasool

Department of Chemistry

School of science

University of Management and Technology,

Lahore Pakistan

2017

Prediction of Protein Solubility in *Escherichia coli* and Experimental Verification

Submitted to University of management and technology Lahore

In partial fulfillment of the requirement

for the award of degree of

MS

CHEMISTRY

BY

SHAHID MEHMOOD

ID

14003140056

Session:2014-2016

In the name of ALLAH the beneficent, the merciful Read‟ Thy lord is most honorable most benevolent, who taught (to write) by pen; He taught man that which he knew not.

(Sura AL-ALAQ 30:3-5)AL- QURAN

DECLARATION

I Shahid Mehmood S/O Abdul Rauf ID: 14003140056 Session 2014-2016 here by declare that the matter printed in the thesis titled " **Prediction of protein solubility in *Escherichia coli* and experimental verification"** is my own work and has not 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: Shahid Mehmood

( )

RESEARCH COMPLETION CERTIFICATE

Certified that the research work contained in this thesis titled "**Prediction of protein solubility in *Escherichia coli* and experimental verification"** has been carried out and completed by Shahid Mehmood , ID:14003140056. The quantum and quality of the work contained in this thesis is adequate for the award of degree of MS /M.phil.

---------- ----------

Supervisor
Dr. Nouman Rasool
Assistant Professor
Department of Chemistry,
UMT, Lahore.

External examiner
Dr. Deeba
Department of Biological Science
FCU, Lahore .

---------- ----------

Dr. Sammia Shahid
Chairperson,
Department of Chemistry,
UMT, Lahore.

Dr. Muhammad Azhar Iqbal
Dean
School of Science,
UMT, Lahore.

This thesis is dedicated to my beloved father

Abdul Rauf

ABSTRACT

Soluble protein in proper concentration is very important for different experimental studies. Solubility of protein can be estimated by the sequence of amino acids in protein. The solubility of protein is important for biophysical and structural development. To achieve the soluble protein in high concentration is a major challenge. The protein which are heterologous expressed are often insoluble and their solubilization is highly trial and error process with low success rate. Although very highly overexpression in inclusion body is some time desirable which result in clean protein. A new method is develop which will predict the solubility of protein on overexpression in *E.coli*. This method use four classifier named as Multilayer Perceptron, Decision Tree, Random Forest, Bayes Classifier. Theses classifier were trained for the prediction of recombinant protein solubility. Many features are used by this method such as canonical variable (CV),Surrounding hydrophobicity, Solubility index composition, Intrinsic aggregation

propensity, Intrinsic Z-scores for aggregation, = tripeptide score, AI = aliphatic index, II= instability index, Fn= frequency of occurrence of Asn, Ft = frequency of occurrence of Thr, Fy= frequency of occurrence of Tyr. It is very simple and easy method for the prediction of recombinant protein solubility. To evaluate the validity of this method test is performed. For this purpose dataset consist of 1500 proteins, out of which 1000 are soluble and 500 are insoluble. Each classifier was trained for the prediction of 450 protein sequences. This method will predict the protein solubility with greater accuracy of about 95.9%. The accuracy of this method is also compared with the previous work or methods. Results shows that this method has more accuracy and precision then other previous works.

ACKNOWLEDGEMENT

Thanks to **ALLLAAH** , who pulled us through the times, when every stone was turned against us, guide us in the light and darkness and show the right path in the right direction, provided us with the opportunity, courage and ability to complete this humble contribution towards knowledge.

All respect for the **Holy Prophet Muhammad (PBUH**), who has sent as greatest educator for the guidance forever and whose teaching enables to see what is beautiful, to know what is true and to love what is good.

I express my deepest gratitude to my research supervisor, **Dr. Nouman Rasool** for encouragement, guidance, unfailing, patience, masterly advice and inspiring attitude, without which this study would not be taken.

I would like to thank **Dr. Sajid Mahmood**, whose guidance greatly helped me in the completion of my project. I thank to him for giving me encouragement and organizing the necessary infrastructure for my project to be possible.

And a special thanks to **Dr. Samia Shahid,** whose kind attitude always encouraged me.

I also thanks to **Waqar** for helping me in executing this project.

I would like this prospect to extend my admiration and appreciation to my research fellow Iram for her cordial harmonization.

I really under high obligation for expressing my best wishes for my friends Ali who encourage me and provide nice company and love

This acknowledgment would be incomplete if i do not pay my sincere and hearty thanks to my cherished and loving parents for their sacrifices, prayers, and affection without which it would have been just a dream to achieve any goal.

Table of Contents

List of Tables

List of Figures

List of symbols and abbreviation

%	Percentage
	Amino group
COOH	Carboxyle group
CCs	Charged regions or charge clusters
PCCs	Positive charge clusters
R	Alkyl group
PI	Isoelectric point
M	Molar
RSA	Relevant solvent accessiblity
ASA	Accessible surface area
COG	General composition of sequence
GSVM	Granular support vector machine
GRAVY	Grand average of hydrophaticity index
GI	Gain information
Rf	Relief factor
SVM	Support vector machine
GR	Gain ratio
CS	Chi squared
OR	One rule

SU	Symmetrical uncertainty
MeOH	Methanol
EMIM-Cl	1-Ethyl-3-methylimidazolium chloride

Chapter # 1: Introduction

1. INTRODUCTION

1.1. Protein

Protein holds central position in the functioning and structure of living organism. The proteins are responsible all physical and chemical function of life. Protein is a Greek word which mean first. It is about 55-58% of cell and is the major organic compound in all plant and animal cells. The proteins are large polymers (macromolecules) which are synthesized from amino acids. At present 300 amino acids are known and from these 300 amino acids only 20 amino acids are in L configuration which are building block of protein. The protein is made up of carbon, hydrogen, oxygen and Sulphur they are found in % approximation as carbon =50-55 hydrogen=6-8 oxygen =20-23 nitrogen =15-18 and Sulphur =0-4. These proteins are fundamental structure of body of living organism. The muscles, nail, heir and skin in animal are made up of protein. While on the other hand the seed of plant are also made up of proteins in the function prospectus. The most important different kind of enzymes which are protein in nature are responsible for different metabolic processes in cells.