



# **FINAL YAER PROJECT REPORT**

## **Protein Secondary Structure Prediction**

**By**

**Bushra Fazal**

**Noorlhuda**

**Bachelors of Software Engineering**

**Bahria University (Karachi Campus)**

**2008**

## ACKNOWLEDGMENT

At the outset, it is our duty to thank Almighty ALLAH for giving us immense strength, good knowledge, great power for doing things in our life and to accomplish our project on time. We also like to thank our parents who prayed for our success in every part of life.

The journey for creating this project has been exhilarating one and it wouldn't be possible without the help of friends and many other professionals. This project would not be complete without the help of our Supervisor Mr. Twaha Ahmed Minai. We would also like to thank Mr. Rauf Shams and Mr. Usman Waheed whose expertise and understanding added very much effectness to our experience. They provided us with direction and technical support and helped in successful completion of the project. We doubt that we will ever be able to convey our appreciation fully, but we owe them our gratitude.

We would also like to thank Miss Madiha Saeed Rizvi of HEJ Karachi University for her special contributions as she us attain understanding of the domain knowledge as it is a vast subject and without her help we would never have understood the complexities of the subject.

We would also like to thank our Project Coordinator Mr. Abid Karim for his cooperation.

Without all the above mentioned people seeing this project through to completion would be a very difficult task for us and we are thankful for their support and their invaluable time.

## **ABSTRACT**

Finding structure of a protein sequence is an important field in Biochemistry. As it is a long and costly endeavor finding techniques and procedures to assist biochemist in their efforts to predict the structure of protein. The study of structure of protein is important as structure very closely related to the function that a protein performs. Thus finding the structure is the first step in finding the function of a protein.

Our final year project consists of three phases first selection of technique to predict protein structure. For that we reviewed various techniques but then we decided on using the already available data to provide to produce results that have more reliability as they are extracted from already reviewed protein structures. We establish a relationship between structure and sequence of protein through their families and on the basis of this relationship we predict structure.

Second phase was to collect data from various sources on internet according to our needs and create a dataset that fulfilled our requirements.

And third was to create a protein structure viewer and a tool to generate prediction.

# Table of Contents

|            |   |    |
|------------|---|----|
| Chapter 1. | Introduction .....                                  | 5  |
| 1.1        | Protein Structure Prediction .....                  | 5  |
| 1.2        | Brief introduction of domain .....                  | 6  |
| 1.3        | Database.....                                       | 7  |
| 1.4        | Introduction to project .....                       | 7  |
| 1.5        | System overview.....                                | 9  |
| 1.6        | Report Organization.....                            | 9  |
| Chapter 2. | Background and Literature review .....              | 12 |
| 2.1        | Domain Overview.....                                | 12 |
| 2.1.1      | Protein.....  | 12 |
| 2.1.2      | Amino Acid .....                                    | 13 |
| 2.1.3      | Synthesis .....                                     | 14 |
| 2.1.4      | Peptide bond formation.....                         | 16 |
| 2.1.5      | Structure of proteins .....                         | 17 |
| 2.1.6      | Motifs, domains and folds in protein structure..... | 20 |
| 2.1.7      | Protein Sequence Family .....                       | 22 |
| 2.1.8      | DNA.....  | 22 |
| 2.1.9      | RNA.....  | 23 |
| 2.1.10     | Transcription factor .....                          | 24 |
| 2.1.11     | Biological roles.....                               | 24 |
| 2.1.12     | Regulation of transcription factor activity .....   | 26 |
| 2.2        | Literature Review .....                             | 28 |
| Chapter 3. | Aim and Statement of Problem .....                  | 33 |
| 3.1        | Problem Description: .....                          | 33 |
| 3.2        | Solution to the Problem .....                       | 33 |
| Chapter 4. | Analysis and Design .....                           | 34 |
| 4.1        | How the idea Originate? .....                       | 34 |
| 4.2        | Requirement Specification.....                      | 35 |
| 4.2.1      | Functional Specification .....                      | 35 |
| 4.2.2      | Hardware Requirement.....                           | 35 |
| 4.2.3      | Software Requirement .....                          | 36 |

|            |                                   |    |
|------------|-----------------------------------|----|
| 4.3        | Project Plan .....                | 36 |
| 4.4        | Design Diagrams .....             | 39 |
| 4.4.1      | System block diagram .....        | 39 |
| 4.4.2      | Flowchart .....                   | 40 |
| 4.4.3      | Use case Diagram .....            | 41 |
| 4.4.4      | Design Use cases .....            | 42 |
| 4.4.5      | Block Diagram Implementation..... | 44 |
| 4.4.6      | ERD .....                         | 45 |
| 4.4.7      | Class Diagram.....                | 46 |
| Chapter 5. | Implementation.....               | 47 |
| 5.1        | Local Alignment .....             | 47 |
| 5.2        | Global Alignment .....            | 51 |
| 5.3        | Interfaces.....                   | 56 |
| 5.3.1      | Splash screen. ....               | 56 |
| 5.3.2      | Main page .....                   | 57 |
| 5.3.3      | Uniprot View .....                | 58 |
| 5.3.4      | Structure view.....               | 59 |
| 5.3.5      | Manage Database.....              | 60 |
| 5.3.6      | Protein insert.....               | 61 |
| 5.3.7      | Prediction interface.....         | 62 |
| Chapter 6. | Testing.....                      | 63 |
| 6.1        | Test plan.....                    | 63 |
| 6.2        | White Box Testing.....            | 65 |
| 6.2.1      | Flowchart .....                   | 68 |
| 6.2.2      | Flow graph.....                   | 69 |
| 6.2.3      | Cyclomatic Complexity .....       | 70 |
| 6.2.4      | Paths.....                        | 70 |
| 6.2.5      | Test Cases.....                   | 71 |
| 6.3        | Black Box Testing .....           | 72 |
| 6.3.1      | Interface 1 .....                 | 72 |
| 6.3.2      | Interface 2.....                  | 79 |
| Chapter 7. | Result.....                       | 84 |
| Chapter 8. | Conclusion.....                   | 85 |

|  |     |
|--|-----|
| Chapter 9. Future Works .....                                    | 86  |
| References .....   | 89  |
| Appendix A List of standard amino acid.....                      | 91  |
| Appendix B Information about amino acids .....                   | 92  |
| Appendix C Transcription Factors of Ecoli.....                   | 95  |
| Appendix D Protein Families in Ecoli Transcription Factors ..... | 108 |
| I) Activators.....   | 108 |
| II) Repressors.....  | 115 |
| Appendix E Uniprot File Format.....                              | 126 |
| Appendix F PDB File Format .....                                 | 127 |
| I) Secondary Structure Section.....                              | 127 |
| II) Coordinate Section.....                                      | 133 |
| Glossary.....  | 138 |