Analysis and Identification of Protein-Ligand Interacting Residues Using Computational Approaches

THESIS

Submitted in partial fulfillment
of the requirements for the degree of
DOCTOR OF PHILOSOPHY

by

P. PRIYADARSHINI PAI

Under the Supervision of
Dr. Sukanta Mondal

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI
2017

BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE, PILANI

CERTIFICATE

This is to certify that the thesis entitled **"Analysis and Identification of Protein-Ligand Interacting Residues Using Computational Approaches"** and submitted by **P. PRIYADARSHINI PAI** ID No **2012PHXF007G** for award of Ph.D. of the Institute embodies original work done by her under my supervision.

Sukanta Mondal

Signature of the Supervisor

DR. SUKANTA MONDAL

Assistant Professor

Date: 22-Apr-2017

Dedicated to my loving family and friends.

Acknowledgements

> "One often meets their destiny on the road they take to avoid it!" - Master Oogway, Kung-Fu Panda.

Had it not been for my unconventional thought process and ideologies in life, this research direction would not have involved me, not even remotely! However, now that I have reached so far, I look back and find several people who have contributed to my journey.

I would like to thank

first of all, my mentor and thesis supervisor, Dr. Sukanta Mondal, the leader of the Annotate Biomolecules Computationally (ABC) group, for advising me and being there for me in all circumstances. He is by far, one of the most awesome persons I have met. In the years that I have worked with him, I have learnt a lot from his composed, pleasant and problem-solving attitude towards science and life, in general. With a lot of patience, he taught me the basics of bioinformatics and computational biology, and helped me perform well, by giving examples of genuine commitment to research, intellectual honestly and persistence. From little knowledge of programming to participating in an international league of women in computing, Dr. Mondal, showed constant encouragement, faith in my abilities and helped me look beyond the conventional boundaries, by offering endless scientific discussions, teaching techniques and much more. I sincerely cannot thank him enough for the philosophical guidance and many other little and big things that he has done for me. Being his first Ph.D. student has been nothing short of a doting parent's first child. Workplace could not have been better without his contagious drive for exploring possibilities.

other members of the ABC group, who I was directly or indirectly involved with.They helped me learn teamwork. Special mention to Kaustubh and Gurdeep, for it was with their initial impetus that our first research article was published; to Sriranjini, with whom I have my first, first-author article; to Rohit and Vishnu C, who never gave up on the study findings we had together; to Saurav Raj, a thesis student I partly mentored, for he asked me questions and more questions, because of which I learnt much more than I knew; to Tirtharaj, a friend and colleague, who added another dimension to our work scenario with his extreme zest for numbers; and newer members Reshma and Chitra.

my Doctoral Advisory Committee (DAC) Members, Prof. Dibakar Chakrabarty and Dr. Kundan Kumar, for broadening my perspectives on scientific research. I am indebted to them for their kind participation, review, encouragement and constructive suggestions in shaping this thesis work effectively.

the Doctoral Research Committee (DRC), currently headed by Prof. Utpal Roy and its constituting faculty members. I am also obliged to the former convenors and members for their active participation in reviewing my work and support for all the formal proceedings of this work.

the Associate Dean, Academic Research Division (ARD), Prof. Prasanta Kumar Das, for his infectious enthusiasm towards science and constant efforts to raise our standards and stretch beyond limits. I am also thankful to his team including Mr. Pratap Behera for having put in so much of their time and efforts in helping us through various formalities.

the Head of Department (HOD) of Biological Sciences, Prof. Judith Braganca, for her exceptional leadership and friendly nature. I am grateful to her and all the former HODs (Prof. Meenal Kowshik and Prof. Utpal Roy) for providing me an opportunity to work for this thesis and also contribute in departmental work along with other faculty members and research scholars, to improve upon my teaching and interpersonal skills. I acknowledge their contribution and that of non-teaching staff including Mrs. Kamna Upadhyay, Mr. Mahadeo Shetkar, Mr. Mahaling Lamani and Mrs. Reshma for their support.

the institute altogether including the Chancellor: Dr. K. M. Birla; Vice-Chancellors: Prof. S. Bhattacharya (current) and Prof. B. N. Jain (former); Directors: Prof. Raghurama (current), Prof. S. Punnekat (former), Late Prof. S. K. Aggrawal (former) and Prof. K. E. Raman (former acting Director); Prof. S. D Manjare (former-in-charge, RCEDD), Prof. A. P. Koley (Instruction Division), Prof. N. Goveas (Academic Registration and Counseling); Admissions and Student Welfare Division; finance section for giving me not only with the opportunity to pursue doctoral studies in this prestigious institute, but also, proving me with infrastructure and fellowship for the purpose. I am also obliged to *Dr. Saby John K and Prof. Meenakshi Raman*, for having mentored me initially.

Special thanks also goes to all my friends in the campus without whom five years would not have passed by just like the planes flying above our campus to the airport. I am especially lucky to have a nice bond with Dr. Mondal's wife Swati and son Aayush, who always made me feel at-home, away from home. Our outings, meetings, get-togethers, parties were always fun-filled.

Last but not the least, I would like to thank my parents *Nataraj and Anita*, and my brother *Deepak*, for loving me so much, believing in my abilities and providing me their support always. Thanks also to my friends *Wasim Ansari, Ramesh Bishnoi, Vishnu G U, Yogesh Joshi, Sanita Ravindranath, Madhusudhan Gowda and Himanshu Gupta* for their encouragement, cheer and madness.

Besides these people, I would like to thank the amazing individuals contributing to open source scientific community efforts such as RCSB, UniProtKB, PyMOL,LaTeX, Python, *etc*. which have made works like this possible.
I can not remember how may people have made a difference and added value to my 'I'. Not that their contribution was not worthy of mention. Some of them came across my memory line and some did not. There will always be scope of improvement here too. Nevertheless, I thank everyone directly and indirectly for having this here.

Priyadarshini P. Pai
March 2017

Abstract

Proteins are key players in biochemical processes playing important roles as enzymes, structural components, messengers and transporters within the living organisms. They perform their function by interacting with ligands which include other proteins, peptides, nucleic acids, carbohydrates, vitamins, metals, *etc*. Certain residues in the protein architecture govern these functions by leading to the formation of a structural neighbourhood or site, capable of facilitating various interactions. Elucidation of how these biomolecules interact with each other can help decipher their mechanism of action, which may in turn, unveil various functional aspects with biomedical implications such as manifestation of diseases. Therefore, proteins and their interaction biology have been in focus of scientific investigations for their promising potential as disease biomarkers, in ligand-mediated functional regulation for disease management and more recently, artificially designed enzymes have also come to fore.

With the advent of high-throughput technology, there has been an increasing availability of experimental data concerning protein interactions. Given the importance of gaining insights into the functioning of various proteins, the need for multi-faceted characterisation has been rising. Various approaches have been proposed for analysis and identification of protein interactions, including unveiling of the crucial residues, domains and folds in their sequence and structure. These are broadly based on similarity transfer and statistical or machine learning techniques. The statistical or machine learning rely on pattern recognition, essentially associating patterns in experimental observations with functions. Wide diversity in the nature and occurrence of interacting residues along the proteins, within and across the families, add several complexities in deciphering important biological traits that can be used for discrimination purposes. Nevertheless scientists are relentlessly contributing to the development of computational approaches using novel perspectives.

In the overall scenario, as sequence information is more abundant, it is desirable that there be sequence-based simple-yet-highly efficient computational approaches. When as minimal information as the sequence is available, there are various challenges and concerns associated with prediction of protein properties or interactions. This thesis presents relevant contributions for identification of the protein architecture involved in ligand interactions using computational efforts. Studies revolve around objectives of analysis and identification of protein ligand interacting residues which have discrete and scanty occurrence such as in enzyme catalytic residues, exploring the scope of using an ensemble approach to enhance the performance and investigating non-parametric probabilistic approaches for residues occurring more or less continuously, such as in nucleic acid interactions.

It addresses issues of data imbalance and considers fundamental aspects of protein interactions to identify protein-interacting residues using supervised machine learning and statistical approach with advantages as above-mentioned. Analytical studies to gain an overview of the nature of various types of protein-ligand interactions involved in enzyme catalysis, mannose-interaction and nucleic-acid interactions have been described in various chapters using benchmarked non-redundant recent protein data. Biological properties such as evolutionary conservation and biochemical nature of the proteins in the interaction neighbourhood at sequence-levels, have been investigated for the development of approaches along with evolving tools such as support vector machines and random forests. These have been reported to have effectively contributed in unraveling various aspects of protein function, with powerful discriminative potential.

Based on study findings, novel methodologies involving selective features and robust discriminative function powered by domain-knowledge driven post-processing filters, local neighbourhood based ensembles, combinations with structural insights and conditional probability based perspective on local occurrence are proposed. The predictions obtained by each of the developed approaches are shown to significantly add to the comprehensive understanding of the prediction scenario either by offering complementarily comparable with or better performances than the state-of-art. They are made available to users as a software of Python codes requiring as minimal as sequence information. It is hoped that these contributions boost protein based experimental studies and eventually also aid in biomolecular design for industrial production and therapeutic applications.

Contents

Acknowledgements **iv**

Abstract **vii**

List of Figures **xiii**

List of Tables **xvii**

List of Abbreviations **xx**

1 Introduction **1**

1.1 Proteins, their interactions and biological significance 1

1.2 Function annotation comes of age: low to high resolution 3

1.3 Challenges associated with identification: Gaps in existing research 7

1.4 Motivation 8

1.5 Aim and objectives 9

1.6 Thesis outline 9

2 PINGU: PredIction of eNzyme catalytic residues usinG seqUence information **11**

2.1 Introduction 11

2.2 Materials and Methods 14

2.2.1 Datasets 14

2.2.2 Encoding of biological properties as features 16

2.2.3 Feature selection 19

2.2.4 Supervised machine learning based discriminative functions 19

2.2.5 Prediction performance assessment 22

2.3 Results and Discussion 24

2.3.1 Biochemical diversity of enzymes 24

2.3.2 Choice of classification models 27

2.3.3 Incrementally challenging predefined settings for independent testing . 30
2.3.4 Predictions in real scenario 30
2.3.5 Application of enzyme domain knowledge in post-processing 34
2.3.6 Case study 36
2.3.7 Software availability 37
2.4 Conclusion 37

3 MOWGLI: prediction of protein-MannOse interacting residues With ensemble classifiers usinG evoLutionary Information 39
3.1 Introduction 39
3.2 Materials and Methods 41
3.2.1 Datasets 42
3.2.2 Amino acid composition 44
3.2.3 Features 44
3.2.4 Prediction engine and performance assessment 45
3.3 Results and discussion 47
3.3.1 Dataset analysis: insights into residues interacting with mannose and its variants 47
3.3.2 Base-classifer and ensemble performance 47
3.3.3 Comparison with state-of-art 53
3.3.4 Case study 59
3.3.5 Software availability 61
3.4 Conclusion 62

4 ROBBY: pRediction Of Biologically relevant small molecule Binding residues on enzYmes 63
4.1 Introduction 63
4.2 Materials and Methods 65
4.2.1 Datasets 65
4.2.2 Features 67
4.2.3 Prediction engine and performance assessment 69
4.3 Results and discussion 69
4.3.1 Prediction insights with sequence information alone 69
4.3.2 Enhancing success rates, validity and applicability using structure information 73
4.3.3 Case study 75

4.3.4 Software availability 76
4.4 Conclusion 78

5 DORAEMON: conDitiOnal pRobabilistic Approach for idEntification of aMino acids interacting with ribONucleic acids 79
5.1 Introduction 79
5.2 Materials and Methods 82
5.2.1 Using local occurrence in a conditional probability-based perspective . 82
5.2.2 Implementation on benchmark datasets 86
5.2.3 Performance assessment 87
5.3 Results and discussion 88
5.3.1 Position-specific probability estimation and discrimination 88
5.3.2 Comparison with state-of-art RNA interacting residue predictors 92
5.3.3 Case study 93
5.3.4 Software availability 94
5.4 Conclusion 94

6 DORAMI: conDitiOnal pRobAbility based prediction Model of protein-deoxyribonucleic acid Interacting residues 95
6.1 Introduction 95
6.2 Materials and Methods 97
6.2.1 Datasets 97
6.2.2 Implementation of probabilistic approach and performance assessment 97
6.3 Results and Discussion 99
6.3.1 DNA-binding residue prediction 99
6.3.2 Cross-predictions 99
6.3.3 Comparative analysis of prediction using combined datasets 100
6.3.4 Case study 101
6.3.5 Software availability 101
6.4 Conclusion 101

7 Conclusion 103

Future scope and directions 107

References 109

Appendix I: Amino acid representations 135

Appendix II: List of Publications **137**

Appendix III: List of conferences and workshops **138**

Brief Biography of the candidate **140**

Biodata of the Supervisor **141**

List of Figures

1.1 An overview of protein sequence structure information. (A) Crystal structure (PDB code: *2vnvA* of BCLA Lectin from *Burkholderia cenocepacia* homodimer in complex with alpha-methyl-mannoside at 1.7 Å(B) First four residues of the protein's N-terminal are represented in 'ball and stick'. The nitrogen atoms are coloured in blue, the oxygen atoms in red and carbon atoms in cyan. C_α carbons are denoted and peptide bonds pointed out. (C) One of the protein chains are represented in single letter code depicting the sequence information. 2

1.2 An overview of protein interactions. (A) Crystal structure of the N-terminal exonuclease domain of the Epsilon Subunit of *E. coli* DNA Polymerase III at pH 8.5. The catalytic residues of this enzyme are shown in 'ball and stick' in the structure and highlighted in upper case in the sequence. (B) Crystal structure of the Poly(A)-binding protein (PABP)-binding site of eIF4G in complex with RRM1-2 of PABP and poly(A). This is an RNA binding protein whose interacting residues are highlighted in red in the structure and in uppercase in sequence. 4

2.1 Schematic workflow of a typical supervised machine learning approach. 15

2.2 An illustration of performance assessment using confusion matrix. 23

2.3 Number of catalytic residues per enzyme (chain) in the datasets. 25

2.4 Type of amino acids constituting the catalytic residues of enzymes. 26

2.5 Sequence lengths (number of amino acids in a chain) of enzymes. 26

2.6 Non-catalytic residues per catalytic residue in enzymes. 27

2.7 Selected number of features showing the best training performance on Dset650. 28

2.8 PINGU prediction performance of in real scenario on 60 diverse enzymes. . . . 31

2.9 Exploring preference of amino acids during PINGU predictions in real scenario. Abbreviations tpr (Sensitivity): true positive rate; fpr (1- Specificity): false positive rate. 36

2.10 Prediction performance of PINGU on 4-hydroxyproline betaine 2-epimerase. The alphabets in upper case indicate PINGU predictions; alphabets underlined are predicted ligand binding residues; and those highlighted are catalytic residues. 37

3.1 The overall study workflow where an ensemble architecture is built from created training dataset and optimised for prediction purposes. 46

3.2 Frequency of occurrence of the mannose (A) interacting residues (MIR) and (B) non-interacting (Non-MIR) in the datasets. 48

3.3 Prediction performance of RF (top panel) and SVM (bottom panel) based ensemble classifiers. 51

3.4 Comparative analysis of ensemble and non-ensemble classification approaches using RF (top panel) and SVM (bottom panel). 52

3.5 Prediction performance of MOWGLI (this work) in comparison with state-of-art mannose-specific predictor PreMieR (Agarwal et al., 2011) and general carbohydrate binding site predictor CBS-PSSM (Malik and Ahmad, 2007). 58

3.6 Case study. Prediction scenario for protein (A) Mannose-specific agglutinin and (B) Glucoamylase. The interacting residues are denoted by uppercase. The prediction marked by star, underline, tilde, and highlight are obtained by MOWGLI (this work), PreMieR (Agarwal et al., 2011), CBS-PSSM (Malik and Ahmad, 2007) and COACH (Yang et al., 2013b) respectively. 60

3.7 Schematic prediction scenario for a user provided query mannose-interacting protein 61

4.1 An overview of ROBBY development. 66

4.2 Assessment of prediction performance using (A) condensed PSSM with two different classifiers on Dset233 and (B) smoothened neighborhood information with best model. 71

4.3 Interaction residues of Uridylate kinase (PDB code: 3ek6A) from the Gram-negative plant pathogen *Xanthomonas campestris*. (A) Hexameric biological unit (B) LigASite mapped on one protein chain in blue and ROBBY in green (C) Sequence information: uppercase denoting interacting residues, underline representing ROBBY predictions, open circles denoting putative ligand binding residues using HOMCOS. 74

4.4 Case study. (A) biological assembly with highlighted ligand binding region in blue. Predictions obtained using ROBBY and MetaPocket 2.0 along with the experimental observation for (B) Alcohol Dehydrogenase (PDB code: *1a4uA*) and (C) Sensor Kinase PhoQ (PDB code: *3cgzA*). Sequence information provided with interacting residues in uppercase, ROBBY predictions underlined, MetaPocket 2.0 predictions denoted by star and their overlap is highlighted. . . 77

5.1 Bayesian network depicted for a protein fragment of length w where h is the hypothesis tested to know if the fragment is interacting or not given the probability of local occurrence of amino acids in that region. 87

5.2 Case study. (A) The structure of YaeJ (PDB code 4dh9Y) with Y chain shown in surface and interacting RNA is shown in cartoon. (B) The sequence of YaeJ. (A and B) The interacting regions are highlighted in red in structure and upper case in sequence, whereas the non-interacting regions are marked in green in structure and lower case in sequence. DORAEMON prediction is underlined, and FastRNABindR is highlighted. 93

6.1 Case study: Crystal Structure of regulatory protein NolR from *Sinorhizobium fredii* in complex with DNA (Lee et al., 2014) Color code: TP=Green, FN=Cyan, FP=Red, TN=Grey 102

List of Tables

1.1 Examples of recent general approaches for protein-ligand interaction prediction 5

1.2 Examples of recent specific approaches for protein-ligand interaction prediction 6

2.1 Ten-fold cross-validation results on Dset650 29

2.2 Summary of selected features for independent on Dtestset200 29

2.3 PINGU prediction on Independent test-dataset Dtestset200. 30

2.4 Prediction performance of PINGU on 60 diverse enzymes in real scenario. . . . 32

2.5 Prediction performance upon application of post-processing filter in real scenario on 60 diverse enzymes. 34

3.1 List of mannose and its variants used in the development of MOWGLI. The three letter representation are as per the PDB code. 43

3.2 Training prediction performance using different feature types. 49

3.3 Prediction performance of base-classifiers on an average. Each base-classifier obtained using 10-fold cross-validation on Dset128. 50

3.4 Prediction performance of MOWGLI as compared to state-of-art. The parameters which could not be assessed have been labelled as 'NA' 55

3.5 Prediction performance of PreMieR (Agarwal et al., 2011) 56

3.6 Prediction performance of CBS-PSSM (Malik and Ahmad, 2007) 57

4.1 Training performance on Dset233 71

4.2 Testing performance of various models on Dtestset78 72

4.3 Testing performance of various models on Dtestset78 75

5.1 Performance assessment measures for discrimination based on probability estimation from CE3213 89

5.2 Performance assessment measures across various distance cut-offs (definition) for dataset CE3213 90

5.3 Performance assessment measures across various window sizes (w) for dataset CE3213 91

5.4 Performance assessment measures across various fragment similarities for dataset CE3213 91

5.5 Independent testing of DORAEMON on RB122 92

5.6 Comparison with state-of-art on New_R15 dataset 92

6.1 Comparative analysis on New_D31 dataset 99

6.2 Performance with various definition 99

6.3 Prediction on New_D31 by means of CPT developed using RNA binding dataset (Pai et al., 2017) 100

6.4 Comparative analysis of prediction on New_R15 by means of CPT developed using DNA binding dataset Dset831 and DORAEMON for the optimised definition 5.0 Å. 100

6.5 Comparative analysis of prediction on New_R15 and New_D31 by means of CPT using combined datasets (selected definition 5.0 Å). 101

List of Abbreviations

Abbreviation	Full form
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid
PDB	Protein Data Bank
CSA	Catalytic Site Atlas
PSSM	Position Specific Scoring Matrix
SVM	Support vector machines
LLR	$l1$-regularized logistic regression
RF	Random forests
RBFN	Radial basis function networks
NnCS	Number of non-catalytic residue per catalytic residue
MIR	Mannose-interacting residue
Non-MIR	Non Mannose-interacting residue
CPT	Conditional Probability Table
TP	True Positive
FN	False Negative
FP	False Positive
TN	True Negative
SN	Sensitivity
PR	Precision
SP	Specificity
AC	Accuracy
MCC	Matthews correlation coefficient
FM	F-measure

Chapter 1

Introduction

Nature is perhaps one of the best known teacher. Not only has it provided philosophical answers to living beings, but to mankind, it has also become an ever attractive subject of intellectual interest, for example in understanding how biomolecules work. Attempts to understand how these biological molecules interact and play their functional roles are described in the following chapters. As a prelude to the contributions, a brief introduction to the overall context is provided below.

1.1 Proteins, their interactions and biological significance

Proteins, which we commonly know as components of daily food, are at deeper levels, key players in the central dogma of life. They have important roles in the cell such as structural (cytoskeletal), mechanical (muscle), biochemical (enzymes) and cell signaling (hormones), eventually contributing to cell growth, differentiation, maintenance and degeneration (Nelson et al., 2008). Certain residues in the protein sequences govern these functions by leading to the formation of a structural neighbourhood or site, capable of facilitating various interactions. Any disruptions in protein functions at sequence or structural level may lead to undesirable gain or loss of function, eventually manifesting as diseases (Gonzalez and Kann, 2012). Thus, gaining insights into various types of protein-ligand binding sites and their constituting residues, is an important step in the elucidation of protein function, which are involved in different cellular processes (Roche et al., 2015). A snapshot of the protein sequence and structure is shown in

Figure 1.1.

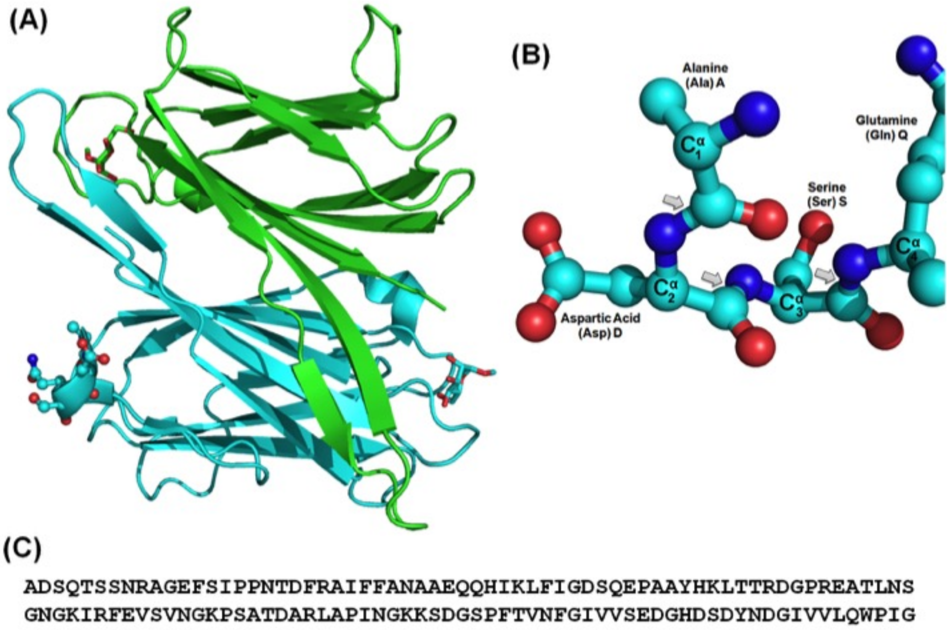

(A)
(B)
Alanine (Ala) A
Glutamine (Gln) Q
Serine (Ser) S
Aspartic Acid (Asp) D
C_1^α
C_2^α
C_3^α
C_4^α
(C)
ADSQTSSNRAGEFSIPPNTDFRAIFFANAAEQQHIKLFIGDSQEPAAYHKLTTRDGPREATLNS
GNGKIRFEVSVNGKPSATDARLAPINGKKSDGSPFTVNFGIVVSEDGHDSDYNDGIVVLQWPIG

Figure 1.1: An overview of protein sequence structure information. (A) Crystal structure (PDB code: *2vnvA* of BCLA Lectin from *Burkholderia cenocepacia* homodimer in complex with alpha-methyl-mannoside at 1.7 Å(B) First four residues of the protein's N-terminal are represented in 'ball and stick'. The nitrogen atoms are coloured in blue, the oxygen atoms in red and carbon atoms in cyan. C_α carbons are denoted and peptide bonds pointed out. (C) One of the protein chains are represented in single letter code depicting the sequence information.

Depending upon the function, proteins interact with a variety of other biomolecules including other proteins, nucleic acids, carbohydrates, metals, *etc*, (Darnell et al., 1990), using covalent and non-covalent interactions (Nelson et al., 2008). To facilitate such interactions, amino acids with required biochemical properties occur in the binding site, offering the required environment. Collisions between the protein and ligand are marked in this environment where molecular diffusion plays a decisive role (Schiavo et al., 2012) leading to conformational changes and complex formation. The mechanisms of interactions cover a broad spectrum of such energy-associated binding events (Kastritis and Bonvin, 2013) and can aid in gaining an understanding of protein function (Perozzo et al., 2004). An overview of protein interactions at sequence and structure levels are shown in Figure 1.2.

As proteins are ubiquitous, efforts towards understanding their roles have gathered large impetus over several years. A wide range of experimental methods are available for gaining multifaceted relevant insights including high-throughput methods such as two-hybrid systems (Rolland et al., 2014), mass spectrometry (Morris et al., 2014), phage display (Blikstad and Ivarsson, 2015), protein chip technology (Blikstad and Ivarsson, 2015), X-ray crystallography (Zheng et al., 2015), Nuclear Magnetic Resonance (NMR) (Ardenkjaer-Larsen et al., 2015), *etc*. With growing technology, vast amounts of sequence and structural information have become available, resulting in an increasing need for characterisation. Attempts to expedite characterisation have been made using computational techniques have been gathering impetus (Hu et al., 2016b). These are described below.

1.2 Function annotation comes of age: low to high resolution

Over many years, experimental studies have provided information on protein sequence, structure and associated functional aspects. These have formed the basis of a large number of computational methods of identification and characterisation. Some of the approaches that have been recently reported for protein-ligand interactions are mentioned in Table 1.1. These methods use information as sequence, structure or their combination using various techniques such as evolutionary trace (Lichtarge et al., 1996), similarity transfer (Yang et al., 2013b) and statistical or machine learning (Du et al., 2016).

They are either general predictors (Capra et al., 2009; Ming et al., 2017; Nagl et al., 1999) (Table 1.1) or targeted at specific ligand binding such as proteins (Yugandhar and Gromiha, 2017), nucleic-acids (Yan and Kurgan, 2017), heme (Xiong et al., 2012), metals (Lin et al., 2005), vitamins (Yu et al., 2014), *etc* (Table 1.2) . Besides these, there are also methods that combine multiple algorithms to achieve better predictions such as MetaPocket 2.0 (Zhang et al., 2011) and COACH (Yang et al., 2013b).

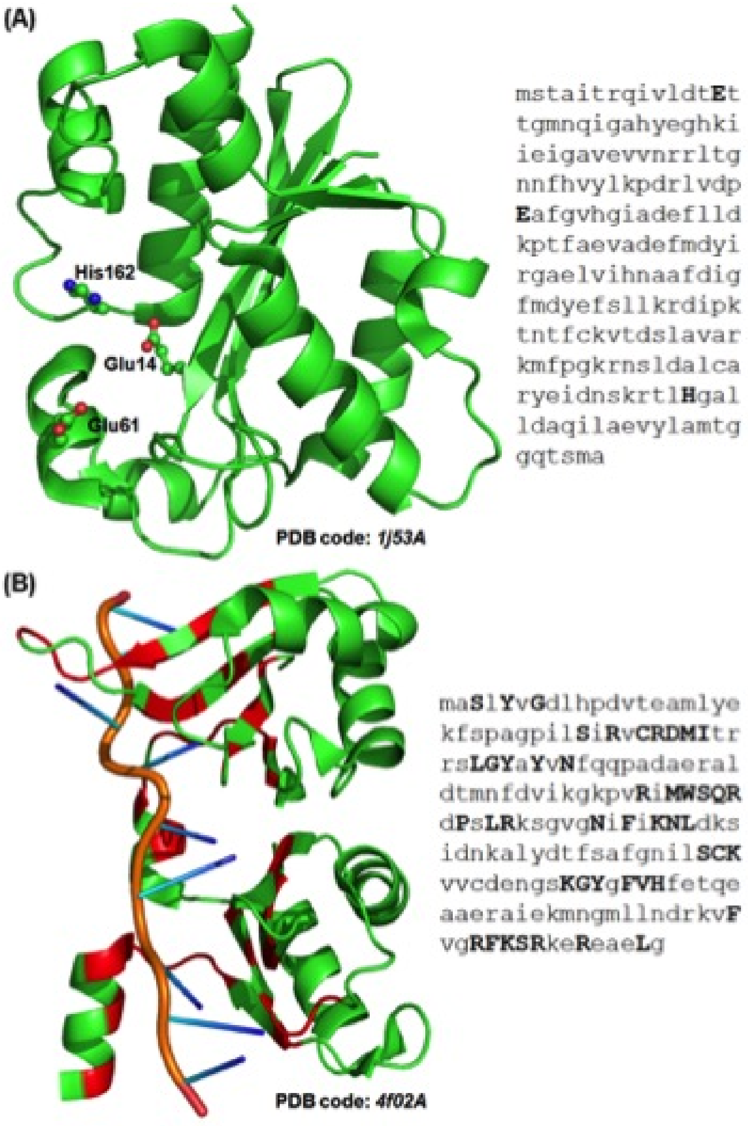


Figure 1.2: An overview of protein interactions. (A) Crystal structure of the N-terminal exonuclease domain of the Epsilon Subunit of *E. coli* DNA Polymerase III at pH 8.5. The catalytic residues of this enzyme are shown in ’ball and stick’ in the structure and highlighted in upper case in the sequence. (B) Crystal structure of the Poly(A)-binding protein (PABP)-binding site of eIF4G in complex with RRM1-2 of PABP and poly(A). This is an RNA binding protein whose interacting residues are highlighted in red in the structure and in uppercase in sequence.

Table 1.1: Examples of recent general approaches for protein-ligand interaction prediction

Approach description	**Reference**
GALAXY webservers: predicts putative ligands and protein-peptide complexes	(Heo et al., 2016)
TargetCom: combination of template-based and template-free methods	(Hu et al., 2016b)
LPIcom: web server for analysis, comparison and prediction of sites	(Singh et al., 2016)
PL-PatchSurfer: molecular local-surface based method	(Hu et al., 2014)
PRANK: probability-based putative pocket prioritisation	(Krivak and Hoksza, 2015)
SVM model based on the chemical-protein interactions from STITCH	(Zhao et al., 2014)
AFAL: profiles amino acids surrounding ligands in proteins	(Arenas-Salinas et al., 2014)
LigandRFs: sequence based random forest ensemble	(Chen et al., 2014)
SVM-based method with statistical depth function to define negative samples	(Wang et al., 2013)
eFindSite: meta-threading and machine-learning based method using auxiliary ligands	(Brylinski and Feinstein, 2013)
FunFOLD2 server: integrates protein-ligand binding site and quality assessment	(Roche et al., 2013)
PLB-SAVE: extracts geometrical construct of solid angles from surface atoms	(Lo et al., 2013)
AutoMap: analyses protein-ligand recognition using multiple ligand binding modes	(Agostino et al., 2013)
ETB-Viterbi: long-distance information and decoding in hidden Markov models	(Kern et al., 2013)
LISE: uses ligand-interacting and site-enriched protein triangles	(Xie et al., 2013)
S4MPLE: sampler for multiple protein-ligand entities	(Hoffer and Horvath, 2012)
Quantum.Ligand.Dock: docking with quantum entanglement refinement	(Kantardjiev, 2012)
COFACTOR: low resolution structural models for global alignment and local refinement	(Roy and Zhang, 2012)
COACH: complementary substructure comparison and sequence profile alignment	(Yang et al., 2013b)
PRL-Dock: docking based on hydrogen bond matching and probabilistic relaxation labeling	(Wu et al., 2012)
Multipositional correlations with graph theoretic clustering and kernel CCA	(Gonzalez et al., 2012)

Table 1.2: Examples of recent specific approaches for protein-ligand interaction prediction

Approach description	**Reference**
Catalytic residue and sites	
CRHunter: integrating multifacted information for catalytic residues	(Sun et al., 2016)
GASS: genetic algorithm bases active site search	(Izidoro et al., 2014)
EXIA2: based on the special side chain orientation of catalytic residues	(Lu et al., 2014)
Allosteric residue and sites	
AlloPred: allosteric sites, perturbation of normal modes alongside pocket descriptors	(Greener and Sternberg, 2015)
PARS: allosteric sites, protein dynamics and structural conservation based	(Panjkovich and Daura, 2014)
RNA-interacting residue and sites	
FastRNABindR: for protein-RNA interface residues, PSSM profiles based on 1% of the UniRef100	(Yasser et al., 2016)
PredRBR:Accurate Prediction of RNA-Binding Residues in proteins using Gradient Tree Boosting	(Liu et al., 2016)
RBscore&NBench: for nucleic acid binding residues, large-scale benchmarking database	(Miao and Westhof, 2016)
STarMIR: for microRNA binding sites, logistic models using CLIP data	(Rennie et al., 2014)
DNA-interacting residue and sites	
Evolution-based DNA-binding residue predictor using dynamic query-driven learning	(Chai et al., 2016)
SNBRFinder: Hybrid sequence based feature- and template- algorithm	(Yang et al., 2015)
DBSI: Structure-based SVM model and visualization	(Zhu et al., 2013)
Metal-interacting residue and sites	
Prediction of Metal Ion-Binding Sites in Proteins Using the Fragment Transformation Method	(Lu et al., 2012)
GRE4Zn: zinc binding sites, geometric restriction based models	(Liu et al., 2014)
nucleotide-interacting residue and sites	
TargetATPsite: ATP-binding sites, residue evolution image sparse representation and ensemble	(Yu et al., 2013)
GTP binding sites: radial basis function networks and significant amino acid pairs	(Ou et al., 2016)
NAD- and FAD-binding sites in proteins using the fragment transformation method	(Lu et al., 2015)
sugar-interacting residue and sites	
PreMieR: identification of mannose interacting residues using local composition	(Agarwal et al., 2011)
vitamin-interacting residue and sites	
Protein-vitamin binding residues using multiple heterogeneous subspace SVMs ensemble	(Yu et al., 2014)
heme-interacting residue and sites	
HEMEsPred: Structure-based Fast-adaptive Ensemble Learning Scheme	(Zhang et al., 2016)

The prediction scenario has been studied at various resolutions. Right from identifying whether a protein interacts with another biomolecule to which regions are involved in the protein and its counterpart, have been attempted with a focus on low to medium resolution prediction (Du et al., 2016; Zhao et al., 2013a). However, reviews suggest that each of these approaches have their advantages and limitations for use in the prediction scenario (Du et al., 2016; Roche et al., 2015). The evolutionary trace methods largely depends on evolutionary conservation to assign scores to various residues for their functional importance (Lichtarge et al., 1996). But at sequence levels, not all residues that are conserved interact with ligands. As a result, even template-based or similarity transfer methods show limitations in characterising proteins which have similar sequence or structure but differ in function or vice-versa. In such cases, *de novo* methods, which rely on pattern information, are particularly useful, as they associate functions based on patterns available for experimental observations.

Different types of discriminating functions have been used in the computational approach development such as classification, regression and hybrid methods (Yan and Kurgan, 2017).These methods use biological properties such as protein topology, binding energies, evolutionary conservation, frequency of occurrence and biochemical nature to encode information of the protein interaction architecture (Du et al., 2016; Roche et al., 2015). Although statistical or machine learning approaches have their own limitations in terms of finding optimal trade-off in the prediction scenario, because of their inherent advantages of generalisation and applicability in novel characterisation, many attempts for uncovering functionally diverse protein interactions have been made. However these are also mired with challenges offering ample avenues for relevant research and development.

1.3 Challenges associated with identification: Gaps in existing research

Many challenges lurk ahead of computational identification approaches, in terms of development, right from collecting reliable data to making the approach more suitable for experimental applications (Jacobson et al., 2014). One of the major concerns is the fact that for a given protein, the number of residues involved in interactions vary considerably (Bartlett et al., 2002; Khazanov and Carlson, 2013). They are governed by evolution and conserved depending upon the importance of their functional roles. Studies have revealed that not all functionally important residues are evolutionarily conserved to same extents and not all evolutionarily conserved residues are equally important for function (Jensen, 1976; Tawfik, 2010). Some ligand interactions are surface-based, whereas others are buried in deep pockets (Konc and Janezic, 2007), which means, the innate architectural environment required also vary biochemically. Further, issues of intermediate conformational changes in the proteins also exist which need careful considerations. In a nutshell, the diversity in the nature and involvement of protein residues for interactions with ligands is large (Chakrabarti and Lanczycki, 2007). This makes devising a simple-yet-widely applicable assumption for identification of protein-interactions at various resolutions - a complicated task. Often the associated inherent complexities reflect in the approaches as false predictions. Scientists have been venturing into finding an assumption or discriminative function such that the trade-off between true and false predictions leads to achieving a more or less accurate prediction scenario for real-time applications. Despite several milestones, there is still ample scope of research in this area which can be enriched by devising novel perspectives based on domain knowledge and application of robust learning.

1.4 Motivation

The vast variation of shape, sizes, and composition of protein-ligand binding sites and the ligands they bind, makes devising a general prediction method very challenging. Since the residue composition of a ligand binding site determines the interactions, their accurate identification can aid in learning more about the ensuing protein-ligand binding events and mechanisms involved. From a broader perspective, understanding general composition of these sites is of great importance for large-scale protein function annotation (Khazanov and Carlson, 2013). As there has been an increasing availability of information associated with protein interactions such as their sequence, structure, binding affinities, conformational changes, involved geometry, regulation and biomedical implications, widely-encompassing computational efforts can aid in adding novel perspectives to the protein-ligand prediction scenario. This thesis is a result of a strong motivation to look beyond convention and develop perspectives for problem-based learning. It aims to present explorations for protein-ligand interacting residue identification using supervised machine learning and probabilistic perspectives powered by domain knowledge, to eventually aid in understanding diseases and for applications in industry and therapeutics. In order to achieve the targeted goal, three fundamental objectives were laid down as mentioned below (in the next section).

1.5 Aim and objectives

Having identified a subject to be studied motivated by the eventual potential impact it can have in drug-design and protein-engineering based applications, the following objectives were laid down to perform various studies and utilise the findings arising thereof in making meaningful interpretations for identification of protein-ligand interaction sites computationally.

Objectives:

- Analysis and identification of key residues in protein-ligand interactions that have discrete and scanty occurrence such as in enzyme catalytic residues.
- Exploring ensemble architecture for achieving enhanced prediction of protein-ligand interacting residues.
- Use of non-parametric probabilistic approach for protein-ligand interacting residues that occur more or less continuously such as in nucleic acid interactions.

1.6 Thesis outline

This chapter (***Chapter one***), provides an introduction to the issue of protein-ligand interacting residue identification using computational approaches. Following chapters specifically address various challenges associated with protein-ligand interacting residue identification. ***Chapter two*** has studies pertaining identification of ligand interacting residues directly involved in the catalytic reaction, known as catalytic residues.The scope of achieving highly sensitive-yet-precise identification in real-time cases is explored. The challenge associated with this type of prediction scenario is the scanty occurrence of catalytic residues in the entire length of the proteins and this chapter offers novel perspectives to address this issue. ***Chapter three*** highlights how drawing consensus of protein-interaction neighbourhood information can serve in achieving enhanced prediction of protein-mannose interacting residues. It shows why generalisation in supervised machine learning based approaches is not a straight forward process and how knowledge guided choice or design of prediction architecture can play a crucial role in materialising better applicability. ***Chapter four*** shows how varied number and type of neighbourhood information when included in ensemble architecture can help in prediction. Additionally, the scope of enhancing the prediction scenario using predicted structural insights is also provided. ***Chapter five*** deals with prediction of another biologically very important class of ligands, *i.e.*, RNA. The number of interacting residues depend on the type of RNAs and can vary based on numerous biological factors. Devising a generally applicable algorithm with a considerable prediction power despite the wide range of evolutionary and biochemical requirements associated with nucleic acid binding proteins is undoubtedly challenging. This chapter illustrates how challenges associated with specific-yet-broad class of ligand interactions can be addressed using a non-parametric approach emphasising on local amino acid occurrence. ***Chapter six*** addresses identification of protein-DNA interactions using the same conditional probability perspective and presents issues of cross-prediction within nucleic acids, joint prediction and scope of improvement in the overall nucleic acid interacting residue prediction scenario. ***Chapter seven*** has conclusive remarks of all these studies, summary of findings, challenges associated and directions for further research and development.

Chapter 2

PINGU: PredIction of eNzyme catalytic residues usinG seqUence information

This chapter has details of studies performed for sequence-based identification of ligand binding residues that are directly involved in biochemical catalysis. The primary question that motivated these investigations was whether the challenges in relevant identification based on various approaches can be supplemented using domain knowledge and made more suitable for experimental validation.

2.1 Introduction

Enzymes, the catalysts of biological systems, are marvellous molecular devices that determine the patterns of chemical transformations (Berg et al., 2015). They play important roles in catalysing biochemical reactions governing processes essential for life. For examples, DNA polymerase handles DNA replication at nuclear level, amino acyl tRNA synthetases facilitate translation of messenger RNA to protein, polypeptide N-acetyl galactosaminyl transferase helps in addition of N-acetyl-galactosamine to serine or threonine residues in O-linked glycosylation, *i.e.*, post-translational modifications at the cellular level, *etc* (Nelson et al., 2008). In order to perform various catalytic functions, the enzyme architecture is bestowed upon with certain distinct properties facilitating binding or interaction with substrates, cofactors and water molecules

(Bartlett et al., 2002). The amino acids constituting the catalytic architecture are usually organised and evolutionarily conserved and are known as catalytic residues (Furnham et al., 2014; Nagano, 2005). So certain mutations are tolerated during the evolutionary process, while certain others may lead to disruption in their naturally permissible states, eventually leading to loss or gain of activity manifesting into biochemical disorders (Goldberg, 1992) such as those related with growth (Visser, 1988), diabetes (Leahy, 2005), kidney (El Dib et al., 2013), neurological function (Ross and Tabrizi, 2011), cancer (Vinik et al., 2014), *etc*.

On account of their importance, enzymes and their catalysis principles, have been the theme of scientific studies for over several years, in which time, different enzyme mechanisms of actions have been investigated in great detail (Bartlett et al., 2002; Hedstrom, 2002; Perona and Craik, 1997). One aspect of understanding how enzymes exercise their functional roles is to examine how they use the limited set of residue side chains that form their "catalytic toolkit". These catalytic units are basically combinations of different residues that are frequently found in diverse unrelated enzymes (Gutteridge and Thornton, 2005). Over years, both experimental and computational efforts have been made to assign various attributes of evolution, diversity in families, mechanism of action and biochemical function for enzymes. They have also been classified to organise perspectives for relevant studies and their application (Holliday et al., 2007; Nagano, 2005).

In view of the fact that the computational approaches offer time and resource utilisation advantages, they have gained impetus through the years, broadly including similarity-transfer and *ab initio* or *de novo* techniques, reviewed in a previous study (Zhang et al., 2009). The similarity-transfer based methods spot supposed catalytic residues in uncharacterised sequences based on their homology with sequences whose catalytic residues are known. Thus, they depend on templates, alignment and pattern matching for catalytic residue mapping. The *ab initio* methods, on the other hand, foreshow catalytic residues by capitalising on several general properties of enzyme catalytic residues which distinguish them from non-catalytic residues. These methods are of assistance especially when the catalytic residues of query enzymes are largely dissimilar from the characterised enzyme catalytic residues. By means of sequence and structure information, various computational approaches have been reported over years for developing knowledge bases (Fleischmann et al., 2004; Furnham et al., 2014; Holliday et al., 2005; Nagano, 2005; Pegg et al., 2006; Schomburg et al., 2004), analysing important biological properties (Bartlett et al., 2002; Bate and Warwicker, 2004; Ben-Shimon and Eisenstein, 2005; del Sol et al., 2006; Lichtarge et al., 1996; Meroz and Horn, 2008; Youn et al., 2007) and catalytic residue prediction (Chien and Huang, 2012; Choi and Kim, 2011; Chou and Cai, 2004; Dou et al., 2012; Fajardo and Fiser, 2013; Gao et al., 2013; Izidoro et al., 2014; Lichtarge et al., 1996; Lu et al., 2014; Sankararaman et al., 2010; Zhang et al., 2008, 2009).

There are many complexities associated with the computational identification of catalytic residues, such as available definition of catalytic residue in literature, their reaction involvement, their sequence-structure-function relationships and available knowledge or resources to ascertain discriminative properties. Among the many biological aspects, enzymes have been studied in the context of catalysis for residue type, location in secondary structure, residue-residue separation, solvent accessibility, intra-protein electrostatic interactions, mobility as assessed based on crystallographic temperature factors, environment polarity and the sequence conservation between homologous enzymes in terms of residues that were in the catalytic residue neighbourhood (Zvelebil and Sternberg, 1988). On account of the inherent data imbalance and biochemical diversity, the prediction scenario is often mired with a significant number of false positives (Zhang et al., 2009). More recent approaches that have been developed in the last five years include L1pred (Dou et al., 2012), neural networks with gravitational center of mass based distance (Fajardo and Fiser, 2013), Random forests with minimum redundancy maximum relevance features (Gao et al., 2013), CLIPS-4D (Janda et al., 2013), EFPrf and rf-SDRs (Nagao et al., 2014), CMASA extension (Flores et al., 2014), EXIA2 (Lu et al., 2014), GASS (Izidoro et al., 2014), *etc.*

Despite the challenges associated, many a milestones have been achieved in unraveling enzyme function and its catalytic architecture. With the help of this fundamental information, important pointers in host-pathogen interactions leading to various diseases, some even life threatening like cancer, have been unveiled (Goldberg, 1992). Rapid advances in high-throughput technologies have yielded large numbers of uncharacterised protein sequences, presenting a pressing need for knowledge acquisition in this context. This chapter presents an approach for obtaining catalytic residue predictions with improved performance by using selected physicochemical properties and evolutionary information from enzyme sequences in a supervised machine learning based prediction architecture and post-processing.

2.2 Materials and Methods

In this chapter, prediction of catalytic residue was delineated in a classification based construct comprising of two classes: (i) the catalytic residue (positive class) and (ii) the non-catalytic residue (negative class). As shown in the Figure 2.1, supervised machine learning is employed for the classification purposes. This process comprises of training and testing phases broadly using non-redundant datasets. Based on the available information a model (or discriminative function) is determined that can describe and distinguish between the two classes. The discrimination is usually based on inherent characteristic traits of the two classes. In this study, evolutionary and biochemical information of enzymes are presented for a given residue along with its sequence neighbours. After the model development, an assessment of the performance is done using cross-validation for determining how well the model is capable of performing on independent data set. Development of a supervised machine learning approach requires careful dataset construction, feature extraction (and selection, if any) and classification using an appropriate discriminative function. In this chapter, an additional step is described to facilitate highly precise prediction, *i.e*, the use of post-processing based on domain knowledge of enzymes. These are described in the following:

2.2.1 Datasets

For the construction of suitable benchmark training and independent test datasets, updated list of enzymes with catalytic residue information was collected for the predictor development from

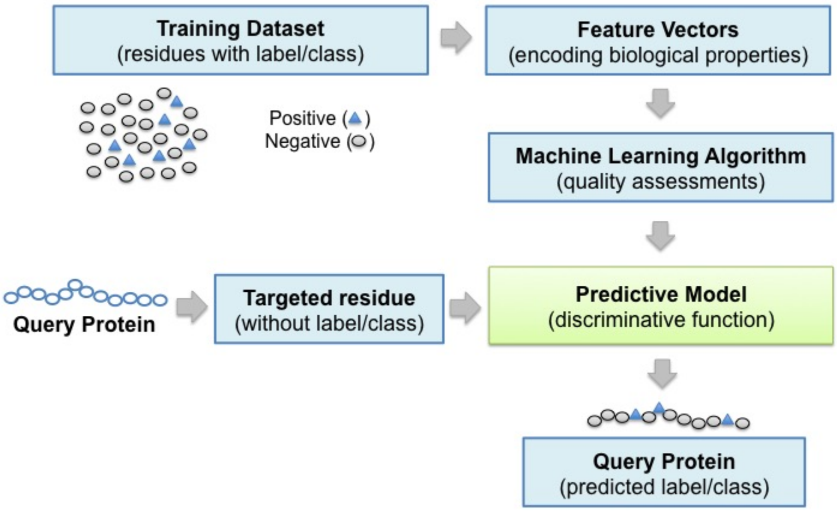


Figure 2.1: Schematic workflow of a typical supervised machine learning approach.

the study L1-pred (Dou et al., 2012) and the Catalytic Site Atlas (CSA) 2.0 (Furnham et al., 2014). The sequence information was generated for the collected Protein Data Bank (PDB) code using the ATOM record of the enzyme structures available in the PDB (Berman et al., 2000). From the pool of enzyme sequences, the sequence fragments with lengths less than 60 amino acids were filtered out. Additionally, to limit the possible scope of overestimation and for the inclusion of diversity, remaining enzyme sequences were clustered. Clustering was performed using BLASTClust (Altschul et al., 1997) into groups with $\geq$ 30% intra-cluster pairwise sequence identity over a 60% overlap on both sequences. A total of 850 clusters were returned with 2819 catalytic residues and 312222 non-catalytic residues. From this parent non-redundant dataset, 650 enzymes were randomly allocated into the training and 200 enzymes in the independent test dataset. The training dataset was named as Dset650 and independent test dataset as Dtestset200 and employed in this study for predictor development.

2.2.2 Encoding of biological properties as features

After making the benchmark dataset, a set of biologically relevant informative features was created and employed for predictor development in this study. This step involved representing the protein P as a discrete model (Chou, 2011) of residues in its sequence of length L as follows:

$$R_1R_2R_3R_4R_5R_6R_7R_8R_9R_{10}...R_i.....R_{L-3}R_{L-2}R_{L-1}R_L \tag{2.1}$$

For a given residue R_i the window size reflects the flanking region, presented by w number of residues in each side as shown below.

$$R_{i-w}..R_{i-2}R_{i-1}R_iR_{i+1}R_{i+2}..R_{i+w} \tag{2.2}$$

In this discrete model, each of the residues were represented using their sequence-based biological properties and their local neighbourhood information was used using various window sizes.

Polarity index

The rich variation in physicochemical properties of the twenty naturally occurring amino acids in protein sequences such as their polar nature, can guide the functional and architectural specificity of proteins. A major hurdle in performing rigorous statistical analyses of biological sequence data is the so-called "sequence metric problem", *i.e.*, which arises because sequences are essentially represented as alphabets rather than arrays of numerical values. In order to overcome this problem, a multivariate statistical analysis was performed in earlier studies, on almost 500 amino acid attributes to yield a small set of distinctly interpretable numeric patterns of amino acid variability (Atchley et al., 2005). These high-dimensional attribute data are summarised by five multidimensional patterns of covariation in attributes that reflect polarity, secondary structure, molecular volume, codon diversity, and electrostatic charge. Factor I is bipolar (substantially different positive and negative factor coefficients) and displays simultaneous covariation in portion of exposed residues versus buried residues, non-bonded energy versus free energy, number of hydrogen bond donors, polarity versus non-polarity, and hydrophobicity versus hydrophilicity; factor II is a secondary structure factor; factor III relates to molecular size or volume with high factor coefficients for bulkiness, residue volume, average volume of a buried residue, side chain volume, and molecular weight; factor IV reflects relative amino acid composition in various proteins, number of codons coding for an amino acid, and amino acid composition. Factor V refers to electrostatic charge with high coefficients on isoelectric point and net charge. Since the catalytic architecture generally tends to have polar residues (Bartlett et al., 2002), it could be used for discriminating them from non-catalytic residues. After initial screening, in this study, one of these five factors mentioned above, factor I, called the Polarity Index, representing various aspects of catalytic residues was chosen to present the physicochemical properties. The values used for each of the 20 naturally occurring amino acids are: (A: -0.591, C: -1.343, D: 1.050, E: 1.357, F: -1.006, G: -0.384, H: 0.336, I: -1.239, K: 1.831, L: -1.019, M: -0.663, N: 0.945, P: 0.189, Q: 0.931, R: 1.538, S: -0.228, T: -0.032,V: -1.337, W: -0.595, Y: 0.260).

PSSM and conservation score

The evolution of enzymes have been widely studied to gain insights into their function (Anderson et al., 2016; Buljan and Bateman, 2009; Choi and Kim, 2006; Fischer et al., 2016; Kinch and Grishin, 2002; Soskine and Tawfik, 2010; Studer et al., 2013; Todd et al., 1999; Vogel et al., 2004; Yuen and Liu, 2007) . Findings over years have revealed that enzymes, in order to facilitate their diverse functional roles, may undergo different types of evolutionary changes (Alcalde, 2017; Galperin and Koonin, 2012; Galperin et al., 1998; Harayama et al., 1992; Mannervik et al., 2009; Rost, 2002; Tabita et al., 2007; Whisstock and Lesk, 2003). However, even when they alter their function over the course of evolution, there are several different properties which it might in principle conserve as others change (Gutteridge and Thornton, 2005). The catalytic mechanism might remain the same, the substrate specificity might remain the same, or the catalytic architecture might remain the same. Therefore, it may be understood that residues that are important for the structure and function of a protein are conserved through evolution (George et al., 2005); such as catalytic residues, as stated by another study (Petrova and Wu, 2006). Based on all these previous remarks made about catalytic residues, it was inferred that evolutionary information could be used to discriminate among conserved and non-conserved residues, of which an important class is the catalytic residue, as also reviewed in the context of computational efforts (Zhang et al., 2009). Disentangling evolutionary signals have shown that it is possible to improve the prediction of catalytic residues by using sequence evolutionary information and sequence conservation (Teppa et al., 2012).Thus, evolutionary patterns were presented for the catalytic residue prediction in this study using PSSM and entropy. PSI-BLAST(Altschul et al., 1997) was used to create PSSM features, *i.e.*, where for each of the residues, 20-dimensional Weighted Observed Percentages (WOP) vectors were obtained. Each of these vector for a residue depicts the log-likelihood of the substitution of 20 amino acids at that sequence position. PSSM values (x) for each residue is normalised by $1/(1+e^{-x})$.

$$P_L^{PSSM} = \begin{bmatrix} S_{1\to 1} & S_{1\to 2} & \cdots & S_{1\to 20} \\ S_{2\to 1} & S_{2\to 2} & \cdots & S_{2\to 20} \\ . & . & S_{i\to j} & . \\ . & . & \cdots & . \\ S_{L\to 1} & S_{L\to 2} & \cdots & S_{L\to 20} \end{bmatrix} \tag{2.3}$$

Since the frequency distribution of 20 amino acids for a given residue position is given by WOP, the entropy (EntWOP) (Zhang et al., 2008) was computed using the equation 2.4:

$$\sum_{i=1}^{20} -p_i log(p_i), where\ \ p_i = n_i / \sum_{j=1}^{20} n_i \tag{2.4}$$

The conservation values depicted by EntWOP lies in between between 0 (most conserved) and 2.996 (least conserved).

2.2.3 Feature selection

Selective presentation of biological properties can also provide for clarity during discrimination in the prediction process as also described in a previous study (Gao et al., 2013). The central thought process behind using a feature selection technique is that the data may contain many redundant or irrelevant features, which can be removed without incurring much loss of information (Bermingham et al., 2015). In order to select properties which could appropriately discriminate between the catalytic and non-catalytic nature of the residues using evolutionary and physicochemical perspectives, feature selection using Fischer-score (F-score) technique was performed in this study. F-score estimates the discrimination of two sets of real numbers (Chen YW, 2006). Given training vectors $x_k, k = 1, ..., m$ if the numbers of positive and negative instances are n^+ and n^-, respectively, then the F-score of the i^{th} feature is defined as:

$$F_i = \frac{(\bar{x}_i^{(+)} - \bar{x}_i)^2 + (\bar{x}_i^{(-)} - \bar{x}_i)^2}{\frac{1}{n^+ +1}\sum_{k=1}^{n^+}(x_{k,i}^{(+)} - \bar{x}_i^{(-)})^2 + \frac{1}{n^- -1}\sum_{k=1}^{n^-}(x_{k,i}^{(-)} - \bar{x}_i^{(-)})^2} \tag{2.5}$$

where $\bar{x}_i$, $\bar{x}_i^{(+)}$, $\bar{x}_i^{(-)}$ are the average of the i^{th} feature of the whole, positive, and negative datasets, respectively; $\bar{x}_{k,i}^{(+)}$ is the i^{th} feature of the k^{th} positive instance, and $\bar{x}_{k,i}^{(-)}$ is the i^{th} feature of the k^{th} negative instance. The difference between the positive and negative sets is depicted by the numerator and the denominator indicates discrimination within each of the two sets. Larger F-scores have been reported to have more discriminative potential (Chen YW, 2006) and was thus, employed in this study.

2.2.4 Supervised machine learning based discriminative functions

Based on the inherent advantages and reported applications in a previous review (Zhang et al., 2009), three classifiers were chosen for relevant studies, *i.e.* support vector machines (SVM), *l*1- regularised logistic regression (LLR) and radial basis function networks (RBFN). These are described in the following:

Support vector machines

SVM is a method for the classification of both linear and non-linear data based on the structural risk minimisation principle of statistics learning theory (Vapnik, 2013). Basically it is an algorithm that uses a non-linear mapping to transform the original training data into a higher dimension. Within this new high dimension, the linear optimal hyperplane is searched for that separates one class from the other. With an appropriate non-linear mapping to a sufficiently high dimension, a hyperplane that separates data from two classes can be achieved. The SVM finds the said hyperplane by means of support vectors and margins defined by them. Several studies reported their applications, especially in biological datasets (Bradford and Westhead, 2005; Cai et al., 2002, 2003; Chou and Cai, 2002; Ding and Dubchak, 2001; Kumar Kandaswamy et al., 2010; Mohabatkar et al., 2011; Nugent and Jones, 2009; Petrova and Wu, 2006; Zhang et al., 2008; Zhou et al., 2007). Mathematically, a training vector $x_i \in R_n$, and class values $y_i \in \{-1, 1\}, i = 1, ..., N$ as depicted in the following equation, are used:

$$Minimise\ \frac{1}{2}w^T\dot{w} + C\sum_{j=1}^{N}\xi_i \tag{2.6}$$

$$Subject\ to\ y_i(w^T\dot{x}_i + b) \geq 1 - \xi_i and \xi \geq 0 \tag{2.7}$$

where w is the normal vector perpendicular to the hyperplane and ξ_i are slake variables for permitting misclassifications. A penalty parameter $C(> 0)$ is employed for balancing the trade-off between the margin and the training error (Scholkopf and Burges, 1999). Number of parameters and kernels (e.g. linear, polynomial, radial basis function and sigmoidal) are optimisable and further, the kernel may be defined by the user. In this study, radial basis function kernel was selected and models were generated using SVMlight Version 6.02 package which is available at `http://svmlight.joachims.org/`.

L1-regularized logistic regression

In statistics, logistic regression, or logit regression, or logit model is a regression model where the dependent variable is categorical. It may be binary, such as in this problem, catalytic or non-catalytic. L1-logreg classifier (Koh et al., 2007) has innate feature ranking capacity, which is beneficial for optimal selection of information from features encoded as shown in a previous study (Dou et al., 2012). This classifier is basically an implementation of an interior-point method for large-scale solver for problems based on $l1$- regularised logistic regression. The logistic model measures the conditional probability of $b \in \{-1, 1\}$ given $x \;\; \in \;\; R_n$,

$$P(b|x) = \frac{exp(b(w^T x + v)}{1 + exp(b(w^T x + v)} \tag{2.8}$$

where x denotes a vector of feature variables and b denotes the associated binary outcome (class). The model has parameters $w \in R_n$ (the weight vector) and $v \in R$ (the intercept); $w^T x + v = 0$ defines the neutral hyper-plane in the data vector space. The classifier locates the optimal model by maximising the estimation of likelihood from the observed examples, *i.e.*, minimising the average logistic loss:

$$Minimise \frac{1}{2} \sum_{i=1}^{m} \log(1 + exp(-b_i(x_i^T w + v))) + \lambda \sum_{i=1}^{n} |w - i| \tag{2.9}$$

where $\lambda > 0$ is the regularisation parameter that can balance the average logistic loss and the size of the weight vector. The software package of L1-logreg classifier available at `http://www.stanford.edu/~boyd/l1_logreg/`, was used for studies presented in this chapter.

Radial basis function networks

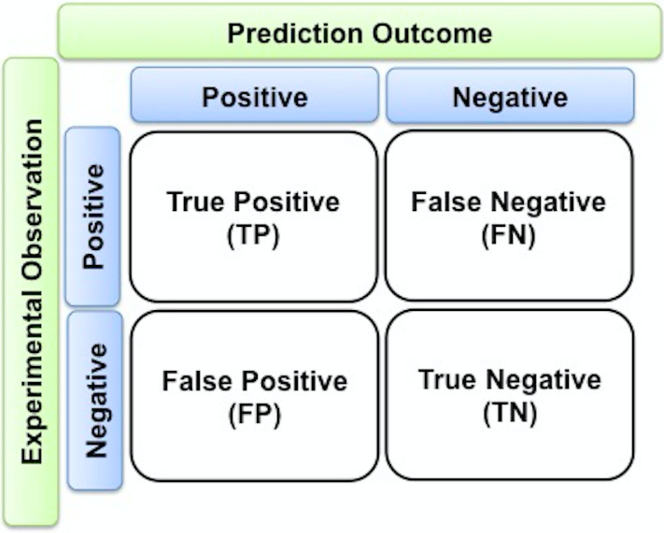
A radial basis function network is an artificial neural network that uses radial basis functions as activation functions (Broomhead and Lowe, 1988). They typically have three layers: an input layer, a hidden layer with a non-linear RBF activation function and a linear output layer. The input can be modeled as a vector of real numbers $x_i \in R_n$ The input nodes transfer the information to the hidden nodes directly and the first layer connections are not weighted. The general mathematical form of the output nodes in an RBF network is as follows:

$$g_j(x) = \sum_{i=1}^{k} w_{ji}\phi((||x - \mu_i||; \sigma_i)) \tag{2.10}$$

where $g_j(x)$ is the function corresponding to the j^{th} output node and is a linear combination of k radial basis functions $\phi()$ with center μ_i and bandwidth r_i; The value of r can be estimated with data-driven methods. Also, w_{ji} is the weight associated with the link between the j^{th} output node and the i^{th} hidden node. Functions that depend only on the distance from a center vector are radially symmetric about that vector, hence the name radial basis function. In this chapter, the QuickRBF package (Ou, 2005) which has been reported to have successfully contributed in an earlier protein related studies (Chen et al., 2010; Ou, 2012; Ou and Chen, 2009), was used to construct RBFN classifiers with all training data as centers. RBF networks have the same properties as back-propagation networks such as generalisation ability and robustness, and further, they present an additional advantage of quick learning and outlier detection. This package was used with a bandwidth = 5, available at `http://www.csie.ntu.edu.tw/~yien/quickrbf/`.

2.2.5 Prediction performance assessment

Performance measure is the way a solution to a given problem can be evaluated. For this study and others in this thesis, in order to evaluate how well experimental observations have been predicted by the developed model, counts of correctly identified and incorrectly identified residues were generated as shown in the Figure 2.2. Using the counts of true positives (TP; residues correctly predicted as catalytic), false positives (FP; residues incorrectly predicted as catalytic), true negatives (TN; residues correctly predicted as non-catalytic) and false negatives (FN; residues incorrectly predicted as non-catalytic), assessment parameters providing various class-wise and general insights such as sensitivity (SN) or recall (RC), precision (PR), specificity (SP), accuracy (AC), Matthews correlation coefficient (MCC), F-measure (FM), *etc*., were calculated



		Prediction Outcome	
		Positive	Negative
Experimental Observation	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Figure 2.2: An illustration of performance assessment using confusion matrix.

during cross-validation, which is described below. ***RC or SN*** is the relative frequency of the correctly classified positive examples. ***SP*** is the relative frequency of the correctly classified negative examples. ***PR*** measures the proportion of the correctly identified residues among examples predicted as positive. ***AC*** is the proportion of the known residues that are correctly predicted in all predictions. ***MCC*** indicates the degree of the correlation between the actual and predicted classes of the residues. MCC values range between *zero* and *one* (*one* where all the predictions are correct, and *zero* where none are correct. ***FM*** combines precision and recall into their harmonic mean. Mathematically,

$$SN \;\; or \;\; RC = \frac{TP}{(TP+FN)} * 100 \tag{2.11}$$

$$PR = \frac{TP}{(TP+FP)} * 100 \tag{2.12}$$

$$SP = \frac{TN}{(TN+FN)} * 100 \tag{2.13}$$

$$AC = \frac{(TP+TN)}{(TP+FN+FP+FN)} * 100 \tag{2.14}$$

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{((TP + FP)(TP + FN)(TN + FP)(TN + FN))}} \tag{2.15}$$

$$FM = \frac{2\ (PR * RC)}{PR + RC} * 100 \tag{2.16}$$

In this study, performance evaluation of models trained on Dset650 was done using tenfold cross-validation (10CV). The enzymes of training dataset were divided into ten sets. One enzyme set was taken out of the ten sets and used as test dataset, and the remaining were used as datasets for training. This process was repeated ten times and the results arising out of each attempt were averaged over all the test results and best models were selected based on best F-measure.

2.3 Results and Discussion

Biochemically diverse benchmarked enzymes were used for supervised machine learning including steps of training and independent testing. The results obtained were analysed and applied in real scenario and efforts towards using domain knowledge for exploring further scope of improvement were made.

2.3.1 Biochemical diversity of enzymes

Representing the diversity in enzymes can help in generalising the prediction approach. For dealing with the challenges of accurate catalytic residue prediction, firstly, an analysis of variations observed in enzymes was done. This included attempts to understand aspects such as in the number of catalytic residues per chain, type of constituting amino acids, sequence length and non-catalytic residues to catalytic residues per chain (NnCS) were analysed. The count of catalytic residues occurring in an enzyme chain ranged from 1 to 23 in the Dset650 (total 2136) and 1 to 10 in the Dtestset200 (total 683). Most chains comprised of less than 10 catalytic residues as can be seen in Figure 2.3.

Further examination into the amino acid composition of these residues in the datasets showed that there were representations of catalytic residues of different naturally occurring

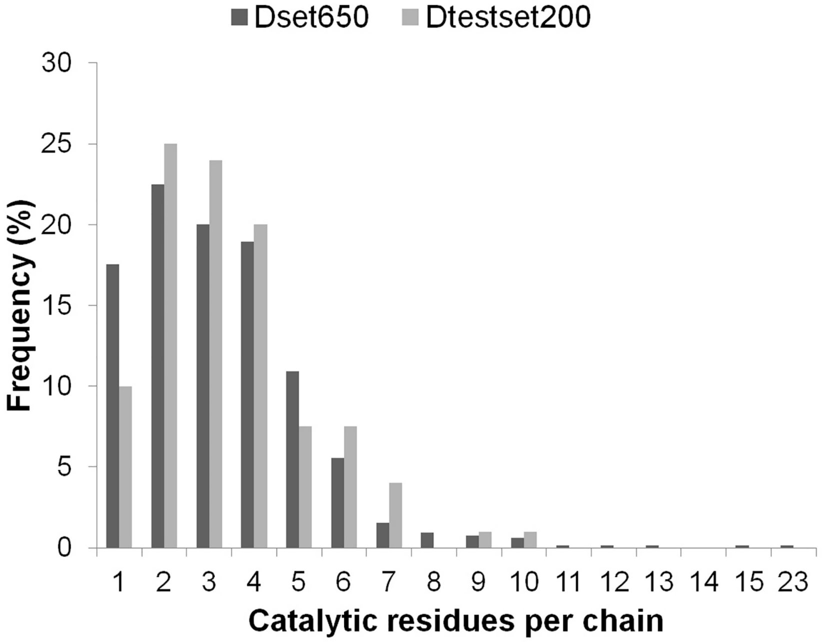


Figure 2.3: Number of catalytic residues per enzyme (chain) in the datasets.

amino acid types. The amino acid distribution in groups of charged (HERKD), polar (QTSNCYW) and hydrophobic (GFLMAIPV) was observed to be 61.6%, 28.6%, and 9.8% for Dataset650 and 62.1%, 26.21%, and 11.7% for Dtestset200. The overall trend is shown in Figure 2.4, which is similar to amino acid distribution in catalytic residues reported in previous study (Bartlett et al., 2002). An examination of the enzyme sequence lengths in the datasets (shown in Figure 2.5) indicates that lengths of Dset650 enzymes ranges from as small as 67 amino acids (aa) to as large as 1520 aa and that of Dtestset200 ranges from 62 aa to1023 aa. Therefore, the study was inclusive of variety in enzyme sequence length as can be seen in Figure 2.5. These diverse sequence lengths have vivid NnCS in datasets with Dset650 having NnCS ranging from over 14 to 1024 residues, with most enzymes having NnCS in between 80 and 100. In Dtestset200, NnCS ranged from 16 to 848, with a most enzymes having NnCS in between 80 and 100. This observation highlights the fact that the distribution of catalytic residues in enzymes is highly skewed, as also reported in earlier studies (Figure 2.6). With this preliminary idea on the included enzymes' diversity, their biological properties were encoded from sequences. These

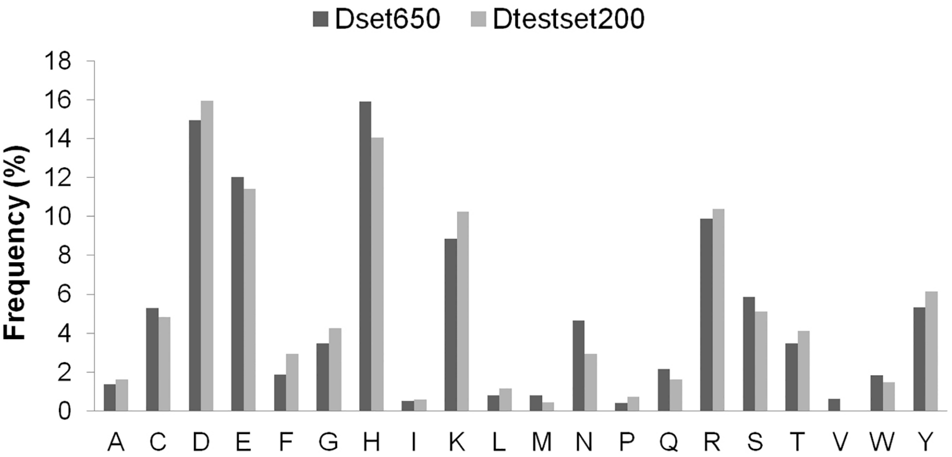


Figure 2.4: Type of amino acids constituting the catalytic residues of enzymes.

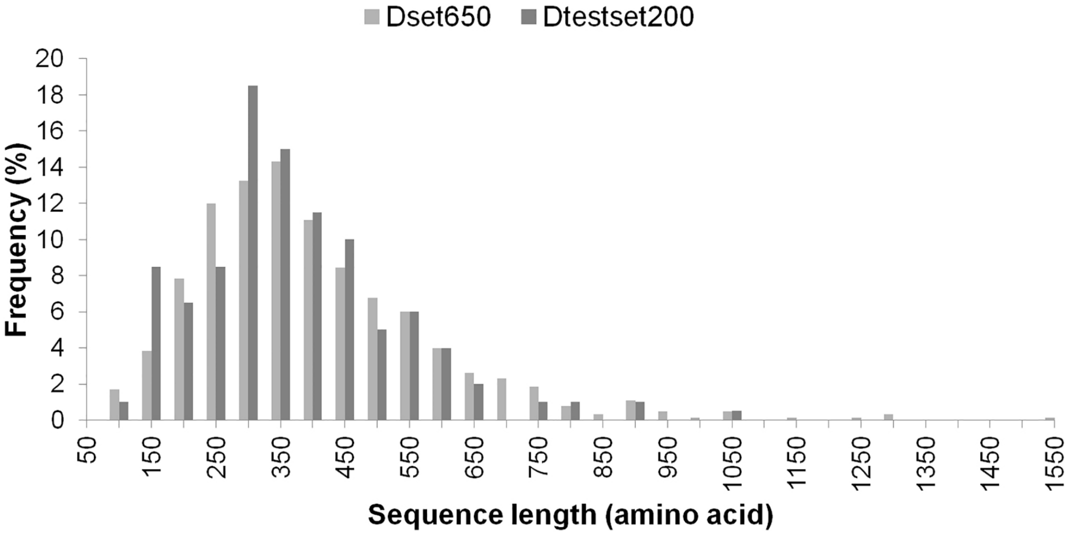


Figure 2.5: Sequence lengths (number of amino acids in a chain) of enzymes.

were then used as an input into various classification models (described below) for obtaining the best possible catalytic residue prediction.

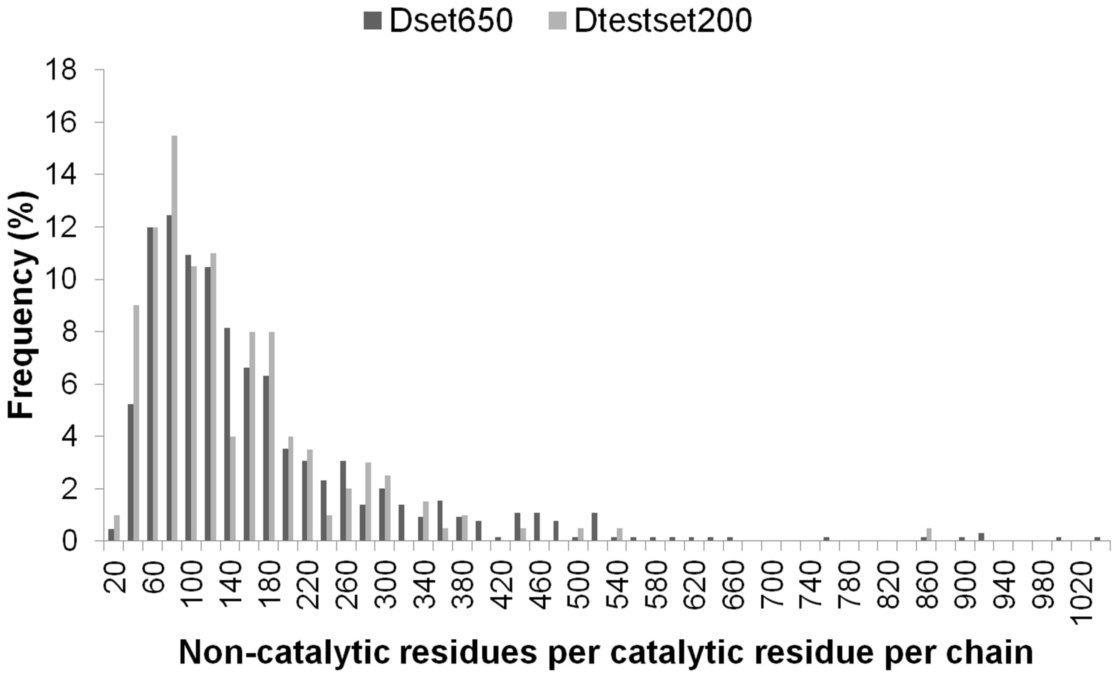


Figure 2.6: Non-catalytic residues per catalytic residue in enzymes.

2.3.2 Choice of classification models

Catalytic residues have prominent physicochemical properties and evolutionary information that differentiate them from non-catalytic residues. These properties were extracted from the enzyme sequence as features of polarity index of amino acids, position specific scoring matrix and entropy information. They were then used for discrimination purpose with the help of three classifiers SVM, LLR and RBF as described in methodology. The prediction performance obtained upon 10CV shown in Table 2.1. The impact of using imbalanced training on prediction is also depicted. As shown in the table, the models trained with balanced number of catalytic residues and non-catalytic residues show a MCC of 0.652 and FM of 83.1% (SVM). This is better than the performance obtained using LLR and RBF. When the number of non-catalytic residues per

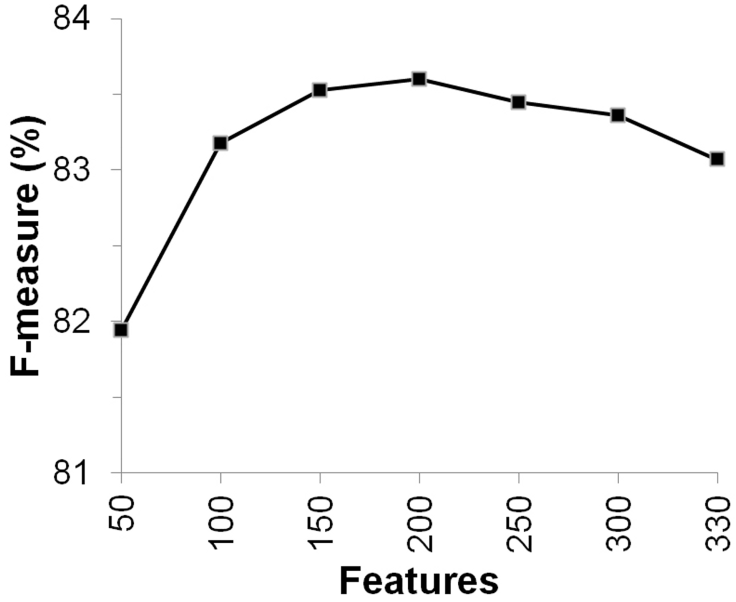


Figure 2.7: Selected number of features showing the best training performance on Dset650.

catalytic residue was greater than one (NnCS > 1), the training performance showed a dip in sensitivity with greater NnCS values. Because catalytic residues have very scanty occurrence in the enzyme sequence, it is essential to have a model than can predict as many of them. Although there may be many false positives using this model, in order to obtain better identification of both the classes, a compromise at this stage was explored for facilitating end-user eventually. The issue of false prediction was addressed separately. Thus, the best SVM models (training NnCS = 1 in a window size = 15) were used for prediction on enzymes in the independent test-dataset. However, before independent testing, the scope of using selective features was explored using F-score as described in methodology. Study findings suggested that 200 out of 330 features were sufficient to reach the best prediction performance under the considerations, in this study as shown in Figure 2.7. Upon examining the 200 optimal features, details shown in Table 2.2, 10 features were that of Polarity index, 175 of PSSM and 15 from EntWOP. The model developed using selected features yielded a promising performance, with an FM of 83.6% and MCC of 0.665, and this was used for independent testing as discussed in the next section.

Table 2.1: Ten-fold cross-validation results on Dset650

NnCS	Classifier	SN	PR	SP	AC	MCC	FM
1	LLR	84.7	79.4	80.0	81.4	0.629	81.9
	RBF	85.6	79.8	78.3	81.9	0.641	82.6
	SVM	85.6	80.8	79.5	82.5	0.652	83.1
2	LLR	73.2	76.0	88.4	83.3	0.622	74.6
	RBF	71.9	76.6	89.0	83.3	0.619	74.2
	SVM	75.4	76.7	88.4	84.1	0.642	76.0
3	LLR	58.6	71.7	94.2	87.1	0.571	64.4
	RBF	58.3	75.4	95.4	87.6	0.580	64.4
	SVM	62.0	72.4	94.0	87.6	0.595	66.7
4	LLR	48.6	70.1	96.5	89.7	0.528	57.3
	RBF	45.1	74.2	97.4	90.0	0.528	56.1
	SVM	54.3	71.0	96.3	90.3	0.567	61.5

Table 2.2: Summary of selected features for independent on Dtestset200

Position in window	Feature composition and frequency of occurrence			
	Polarity index	PSSM	EntWOP	Total
-7	0	5	1	6
-6	1	7	1	9
-5	1	13	1	15
-4	1	13	1	15
-3	1	14	1	16
-2	1	17	1	19
-1	1	14	1	16
0	1	20	1	22
+1	1	15	1	17
+2	0	16	1	17
+3	1	11	1	13
+4	0	8	1	9
+5	0	8	1	9
+6	0	7	1	8
+7	1	7	1	9

2.3.3 Incrementally challenging predefined settings for independent testing

Prediction in a balanced fashion (among one catalytic residue and one non-catalytic residue) in independent test dataset showed an MCC of 0.629 and an FM of 81.5%. Encouraged by the discriminative potential, these models were posed with incrementally challenging prediction scenarios. This was done by increasing NnCS in the setting for identification. Results indicated that the developed models were able to identify most of the catalytic residues despite their scanty occurrence in the entire sequence of amino acids. However, in addition to these true catalytic residues, some non-catalytic residues were mistakenly identified as catalytic. This, could also be noted in Table 2.3 as the decrease in precision from 81.2% (NnCS = 1) to 13.4 % (NnCS = 30). However, as described earlier, our interest in accurately identifying as many catalytic residues is achieved with an overall sensitivity (81.9%) throughout the increments in NnCS. The specificity was also high ($\geq$ 80%) throughout, implying most non-catalytic residues were also correctly identified. FM and MCC over the varied NnCS are shown in Table 2.3. Based on the obtained promising results, the performance of these models was explored in the real scenario (where all the catalytic residues and non- catalytic residues of a chain were included) and is described next.

Table 2.3: PINGU prediction on Independent test-dataset Dtestset200.

NnCS	SN	PR	SP	AC	MCC	FM
1	81.9	81.2	81.0	81.5	0.629	81.5
6	81.9	40.1	79.6	79.9	0.473	53.8
12	81.9	26.3	80.9	81.0	0.392	39.8
18	81.9	20.1	81.9	81.9	0.347	32.3
24	81.9	15.9	81.9	81.9	0.309	26.6
30	81.9	13.4	82.1	82.1	0.285	23.0

2.3.4 Predictions in real scenario

Upon independent testing, the best model obtained during the independent testing with (c = 50.0 and g = 0.08) was named **PINGU**: **P**red**I**ction of e**N**zyme catalytic residues usin**G** seq**U**ence information. From the independent test 60 enzymes were chosen and the discrimination power of PINGU was tested on them. These enzymes were diverse with respect to enzyme classes, sequence lengths and NnCS. Performance evaluation was done and results are shown in Figure 2.8. It was observed that the performance of PINGU on an average is 86.4% sensitive, 80.9% specific with an MCC of 0.203 and FM of 12.6%. The premise of predicting most catalytic residues is preserved (detailed performance results are shown in Table 2.4). Further, the prediction scenario was explored to seek for peculiarities if any. Upon analysis, it was marked that some of these residues were present in the enzyme sub-unit interface in homo-dimeric proteins with PDB code *1bd0*, *1dqr* and *1q6l* (based on CSA database (Furnham et al., 2014) records). For predictions at residue level, one chain per enzyme was included in this study and mapping of the residues that were present on the subunit interface of the other chain was done. Predictions for the mentioned enzymes showed that the following residues occurring on the subunit interface were missed. *1dqr*A position 388 (Histidine) and *1q6l*B positions 68 (Alanine) and 139 (Arginine), with numbering based on PDB (Berman et al., 2000). Further, whether the residues missed were solvent exposed (surface) or buried could provide an interesting understanding of

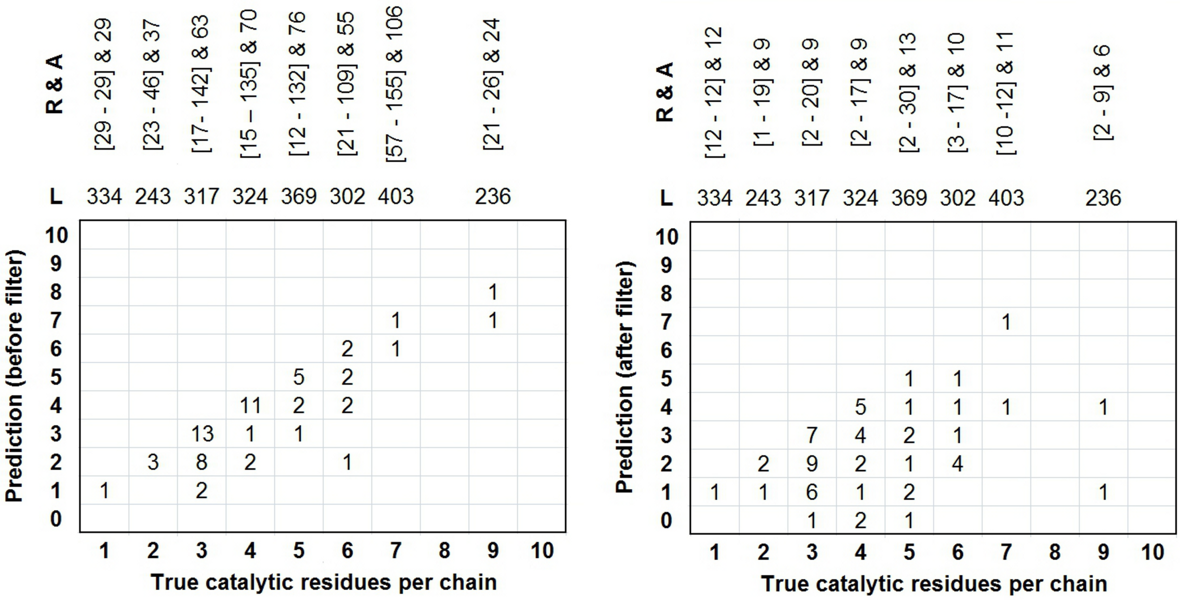


Figure 2.8: PINGU prediction performance of in real scenario on 60 diverse enzymes.

the trend in prediction by PINGU. This was also explored using GETAREA web server available at `http://curie.utmb.edu/getarea.html`. It was found that nine of the 60 protein chains (*1aopA*, *1bd3A*, *1bibA*, *1djlB*, *1dqrA*, *1f6dA*, *1g99A*, *1kezB*, *2tdtA*) have one of their catalytic residues exposed and five others (*1cvrA*,*1jhfA*, *1nsfA*, *1sesB*, *2a86B*) have two. PINGU was able to predict 15 out of 19 solvent exposed catalytic residues from above-mentioned 14 protein chains. Details of the performance of PINGU for these proteins are given in Table 2.4. However, no specific bias or prediction trend was observed with regard to residues present on the subunit interface or those occurring on the surface of the enzymes.The false positives causing dip in predictor precision (Figure 2.9) is addressed separately by application of predicted ligand-binding information. This is described in the following.

Table 2.4: Prediction performance of PINGU on 60 diverse enzymes in real scenario.

PDB code	TP	FN	FP	TN	SN	PR	SP	AC	MCC	FM
*12as*A	3	0	33	280	100.0	8.3	89.5	89.6	0.273	15.4
*1a0i*A	1	0	29	304	100.0	3.3	91.3	91.3	0.174	6.5
*1ald*A	3	0	81	265	100.0	3.6	76.6	76.8	0.165	6.9
*1aop*A	5	0	118	360	100.0	4.1	75.3	75.6	0.175	7.8
*1b6b*B	3	2	12	125	60.0	20.0	91.2	90.1	0.307	30.0
*1b6t*A	2	1	36	106	66.7	5.3	74.7	74.5	0.134	9.8
*1bd0*A	4	0	85	285	100.0	4.5	77.0	77.3	0.186	8.6
*1bd3*A	2	1	32	175	66.7	5.9	84.5	84.3	0.165	10.8
*1bib*A	2	0	46	259	100.0	4.2	84.9	85.0	0.188	8.0
*1ca2*A	3	0	36	203	100.0	7.7	84.9	85.1	0.256	14.3
*1cgk*A	4	0	135	236	100.0	2.9	63.6	64.0	0.135	5.6
*1chk*A	2	0	23	199	100.0	8.0	89.6	89.7	0.268	14.8
*1cvr*A	4	0	43	374	100.0	8.5	89.7	89.8	0.276	15.7
*1czf*A	4	0	49	295	100.0	7.6	85.8	85.9	0.254	14.0
*1db3*A	4	0	111	243	100.0	3.5	68.6	69.0	0.155	6.7
*1de6*A	3	0	52	347	100.0	5.5	87.0	87.1	0.218	10.3
*1dil*A	1	2	26	257	33.3	3.7	90.8	90.2	0.084	6.7
*1djl*B	3	0	41	124	100.0	6.8	75.2	75.6	0.226	12.8
*1dli*A	6	0	109	273	100.0	5.2	71.5	71.9	0.193	9.9
*1dnk*A	4	0	21	221	100.0	16.0	91.3	91.5	0.382	27.6
*1dnp*B	3	0	108	344	100.0	2.7	76.1	76.3	0.143	5.3
*1do8*A	3	0	109	413	100.0	2.7	79.1	79.2	0.146	5.2
*1dqr*A	6	1	155	381	85.7	3.7	71.0	71.3	0.140	7.2

*1eq2*A	3	0	59	234	100.0	4.8	79.9	80.1	0.197	9.2
*1f6d*A	4	0	113	245	100.0	3.4	68.4	68.8	0.153	6.6
*1f8x*A	2	1	17	123	66.7	10.5	87.9	87.4	0.230	18.2
*1fcb*A	5	0	80	412	100.0	5.9	83.7	83.9	0.222	11.1
*1fq0*A	2	0	42	155	100.0	4.6	78.7	78.9	0.189	8.9
*1g99*A	5	0	102	287	100.0	4.7	73.8	74.1	0.186	8.9
*1gcu*A	2	4	21	254	33.3	8.7	92.4	91.1	0.136	13.8
*1gim*A	3	0	142	272	100.0	2.1	65.7	66.0	0.117	4.1
*1gsa*A	3	0	65	234	100.0	4.4	78.3	78.5	0.186	8.5
*1h7a*A	5	0	132	412	100.0	3.7	75.7	76.0	0.166	7.0
*1hrk*A	6	0	53	286	100.0	10.2	84.4	84.6	0.293	18.5
*1hto*X	3	0	131	329	100.0	2.2	71.5	71.7	0.127	4.4
*1jhf*A	4	1	22	161	80.0	15.4	88.0	87.8	0.317	25.8
*1kas*A	4	0	114	280	100.0	3.4	71.1	71.4	0.155	6.6
*1kez*B	4	0	19	230	100.0	17.4	92.4	92.5	0.401	29.6
*1l7n*B	5	1	23	165	83.3	17.9	87.8	87.6	0.350	29.4
*1m9c*A	4	2	63	82	66.7	6.0	56.6	57.0	0.091	11.0
*1n2c*C	4	2	69	389	66.7	5.5	84.9	84.7	0.160	10.1
*1nn4*A	4	0	36	105	100.0	10.0	74.5	75.2	0.273	18.2
*1nsf*A	2	1	21	209	66.7	8.7	90.8	90.6	0.217	15.4
*1o98*A	3	0	141	353	100.0	2.1	71.5	71.6	0.122	4.1
*1ok4*J	2	1	25	208	66.7	7.4	89.3	89.0	0.197	13.3
*1p7m*A	2	1	33	137	66.7	5.7	80.6	80.4	0.154	10.5
*1pym*B	3	1	39	225	75.0	7.1	85.2	85.1	0.201	13.0
*1q6l*B	7	2	21	171	77.8	25.0	89.1	88.6	0.399	37.8
*1qfe*B	2	1	23	212	66.7	8.0	90.2	89.9	0.207	14.3
*1ses*B	4	1	73	329	80.0	5.2	81.8	81.8	0.174	9.8
*1w1o*A	1	2	63	454	33.3	1.6	87.8	87.5	0.049	3.0
*1ytw*A	5	1	44	242	83.3	10.2	84.6	84.6	0.258	18.2
*1z9h*A	2	2	15	241	50.0	11.8	94.1	93.5	0.220	19.0
*2a86*B	7	0	57	199	100.0	10.9	77.7	78.3	0.292	19.7
*2dln*A	5	0	72	215	100.0	6.5	74.9	75.3	0.221	12.2
*2f9r*B	8	1	26	236	88.9	23.5	90.1	90.0	0.427	37.2
*2oat*A	3	0	96	326	100.0	3.0	77.3	77.4	0.153	5.9
*2pgd*A	4	0	121	343	100.0	3.2	73.9	74.2	0.154	6.2
*2tdt*A	2	1	74	183	66.7	2.6	71.2	71.2	0.089	5.1
*2ts1*A	2	2	85	316	50.0	2.3	78.8	78.5	0.069	4.4

2.3.5 Application of enzyme domain knowledge in post-processing

Biological functions at deeper levels are inclusive of specific biochemical activities such as catalysis. Looking for catalytic residues in a reduced search space may improve the prediction scenario (Zhang et al., 2009). As catalytic residues are also ligand-binding residues in a broader perspective, it was hypothesised that upon searching for catalytic residues within ligand binding residue pools may reduce the search space, thus sparing the scenario from false predictions. To explore this idea, predicted ligand information was used as a post-processing filter (Bruha, 2001). For this, predictions of a template-recognition based ligand binding site predictor, S-SITE (Yang et al., 2013b), was combined with PINGU predictions hoping to improve its precision. The class labels of each of the test protein residues were re-labeled. Those predictions of PINGU that were also present in the set of S-SITE based predictions were considered as catalytic, the remaining residues were filtered out and regarded as non-catalytic.

Table 2.5: Prediction performance upon application of post-processing filter in real scenario on 60 diverse enzymes.

PDB code	TP	FN	FP	TN	SN	PR	SP	AC	MCC	FM
*12as*A	3	0	2	311	100.0	60.0	99.4	99.4	0.772	75.0
*1a0i*A	1	0	12	321	100.0	7.7	96.4	96.4	0.272	14.3
*1ald*A	3	0	17	329	100.0	15.0	95.1	95.1	0.378	26.1
*1aop*A	5	0	30	448	100.0	14.3	93.7	93.8	0.366	25.0
*1b6b*B	2	3	2	135	40.0	50.0	98.5	96.5	0.429	44.4
*1b6t*A	2	1	14	128	66.7	12.5	90.1	89.7	0.258	21.1
*1bd0*A	2	2	13	357	50.0	13.3	96.5	96.0	0.244	28.6
*1bd3*A	2	1	9	198	66.7	18.2	95.7	95.2	0.322	33.3
*1bib*A	2	0	19	286	100.0	9.5	93.8	93.8	0.299	50.0
*1ca2*A	2	1	9	230	66.7	18.2	96.2	95.9	0.334	66.7
*1cgk*A	4	0	16	355	100.0	20.0	95.7	95.7	0.436	NA
*1chk*A	1	1	1	221	100.0	8.0	89.6	89.7	0.268	27.3
*1cvr*A	4	0	4	413	100.0	8.5	89.7	89.8	0.276	37.5
*1czf*A	0	4	2	342	100.0	7.6	85.8	85.9	0.254	18.2
*1db3*A	3	1	15	339	100.0	3.5	68.6	69.0	0.155	27.3
*1de6*A	3	0	10	389	100.0	5.5	87.0	87.1	0.218	17.4
*1dil*A	1	2	7	276	33.3	3.7	90.8	90.2	0.084	36.4
*1djl*B	3	0	16	149	100.0	6.8	75.2	75.6	0.226	8.0
*1dli*A	2	4	15	367	100.0	5.2	71.5	71.9	0.193	17.4

*1dnk*A	2	2	5	237	100.0	16.0	91.3	91.5	0.382	36.4
*1dnp*B	1	2	20	432	100.0	2.7	76.1	76.3	0.143	8.4
*1do8*A	2	1	9	513	100.0	2.7	79.1	79.2	0.146	28.6
*1dqr*A	4	3	10	526	85.7	3.7	71.0	71.3	0.140	38.1
*1eq2*A	3	0	14	279	100.0	4.8	79.9	80.1	0.197	29.9
*1f6d*A	4	0	15	343	100.0	3.4	68.4	68.8	0.153	34.8
*1f8x*A	2	1	4	136	66.7	10.5	87.9	87.4	0.230	44.4
*1fcb*A	4	1	16	476	100.0	5.9	83.7	83.9	0.222	32.0
*1fq0*A	2	0	7	190	100.0	4.6	78.7	78.9	0.189	36.3
*1g99*A	3	2	16	373	100.0	4.7	73.8	74.1	0.186	25.0
*1gcu*A	2	4	9	266	33.3	8.7	92.4	91.1	0.136	23.5
*1gim*A	3	0	19	395	100.0	2.1	65.7	66.0	0.117	23.9
*1gsa*A	2	1	6	293	100.0	4.4	78.3	78.5	0.186	36.4
*1h7a*A	1	4	9	535	100.0	3.7	75.7	76.0	0.166	13.3
*1hrk*A	2	4	17	322	100.0	10.2	84.4	84.6	0.293	16.0
*1hto*X	1	2	12	448	100.0	2.2	71.5	71.7	0.127	12.5
*1jhf*A	0	5	2	181	80.0	15.4	88.0	87.8	0.317	NA
*1kas*A	4	0	10	384	100.0	3.4	71.1	71.4	0.155	44.5
*1kez*B	3	1	2	247	100.0	17.4	92.4	92.5	0.401	66.7
*1l7n*B	5	1	3	185	83.3	17.9	87.8	87.6	0.350	71.4
*1m9c*A	4	2	6	139	66.7	6.0	56.6	57.0	0.091	50.0
*1n2c*C	2	4	9	449	66.7	5.5	84.9	84.7	0.160	23.5
*1nn4*A	3	1	8	133	100.0	10.0	74.5	75.2	0.273	40.0
*1nsf*A	1	2	6	224	66.7	8.7	90.8	90.6	0.217	20.0
*1o98*A	1	2	3	491	100.0	2.1	71.5	71.6	0.122	28.6
*1ok4*J	2	1	7	226	66.7	7.4	89.3	89.0	0.197	33.3
*1p7m*A	2	1	3	167	66.7	5.7	80.6	80.4	0.154	50.0
*1pym*B	3	1	12	252	75.0	7.1	85.2	85.1	0.201	31.6
*1q6l*B	1	8	9	183	77.8	25.0	89.1	88.6	0.399	10.5
*1qfe*B	2	1	8	227	66.7	8.0	90.2	89.9	0.207	30.8
*1ses*B	3	2	18	384	80.0	5.2	81.8	81.8	0.174	23.1
*1w1o*A	0	3	5	512	33.3	1.6	87.8	87.5	0.049	NA
*1ytw*A	3	3	10	276	83.3	10.2	84.6	84.6	0.258	31.6
*1z9h*A	1	3	2	254	50.0	11.8	94.1	93.5	0.220	28.6
*2a86*B	7	0	12	244	100.0	10.9	77.7	78.3	0.292	53.8
*2dln*A	1	4	12	275	100.0	6.5	74.9	75.3	0.221	11.1
*2f9r*B	4	5	2	260	88.9	23.5	90.1	90.0	0.427	53.3
*2oat*A	3	0	13	409	100.0	3.0	77.3	77.4	0.153	31.6
*2pgd*A	4	0	11	453	100.0	3.2	73.9	74.2	0.154	42.1
*2tdt*A	1	2	3	254	66.7	2.6	71.2	71.2	0.089	28.6
*2ts1*A	0	4	17	384	50.0	2.3	78.8	78.5	0.069	NA

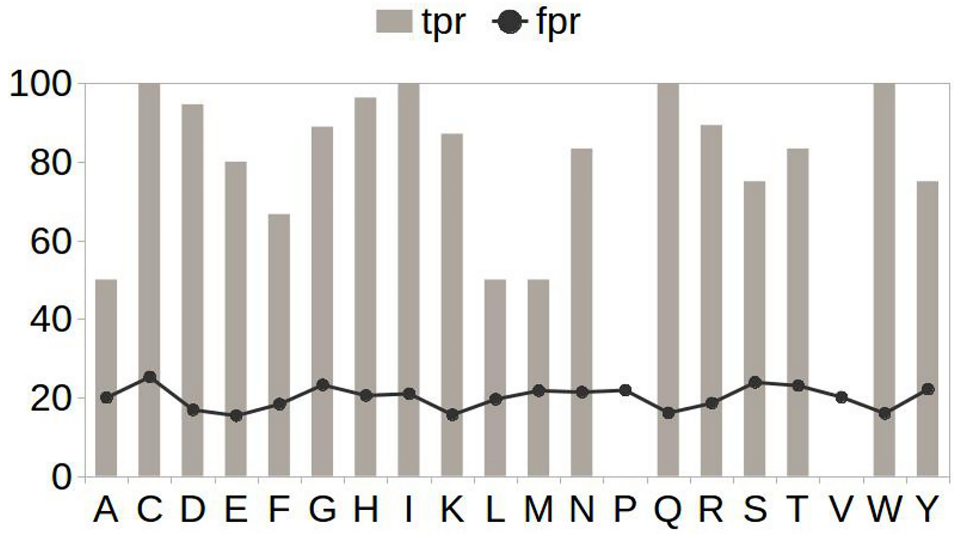


Figure 2.9: Exploring preference of amino acids during PINGU predictions in real scenario. Abbreviations tpr (Sensitivity): true positive rate; fpr (1- Specificity): false positive rate.

Figure 2.8 shows the predictor performance before and after application of this post-processing filter. On an average, without much compromise in sensitivity, specificity or accuracy of the predictor, an overall improvement of 16% was observed in precision. This reflected in the assessment parameters also, where an achievement of 20% rise in FM and 0.138 in MCC was marked (details in Table 2.4, 2.5). The results of this attempt suggest that application of post-processing filter such as the one used in this study can be useful in obtaining accurate prediction of catalytic residues in real scenarios, with minimal false positives, which has been a challenge in biology.

2.3.6 Case study

PINGU predictions were so far generalised for a pool of enzymes. An insight into its working on an individual enzyme, 4-hydroxyproline betaine 2-epimerase (Zhao et al., 2013b), is provided here. The enzyme reportedly takes part in multiple biochemical reactions resulting in different biologically relevant functions in the catabolic pathway depending upon osmotic stress. The residues that directly take part in the catalytic activity, *i.e.*, the catalytic residues, occur at position 163 (Lysine) and 265 (Lysine) in a sequence length of 367 residues. Prediction performance of PINGU is shown in Figure 2.10. Notably among 353 residues (residues analysed excluding termini), the two catalytic residues were correctly predicted. However, along with the true positives, 52 non-catalytic residues (false positives) were also predicted. To reduce the false positive rate, predicted ligand binding residue data for this enzyme was used as a filter. Consequently, it was observed that, of the 52 falsely predicted residues, 40 false positives could be reassigned correctly as non-catalytic. This implied in a pool of predictions, if there had to be experimental validation to be done, it would require scanning of only 6 other residues per catalytic residue. These findings clearly indicate that PINGU with post-processing filtering can boost enzyme applications further.

```
mkiaeiqlfq hdlPvvngpy riAsGdvwsl tttivkiiae dgtiGwGEtc PvgptYAEAh aggalaalev 70
lasglagaea lplplhtrmd sllcgHnyAK salDiAvhDl wgkrlgvpvh elLggaltds vssyySLgvm 140
epdeaarqal ekqregysrl QvKLGarpie idieairkvw eavrgtgial aaDgNrgwtt rdalrfsrec 210
pdipfvmEQP cnsfedleai rplchhalym DEdgtslntv itaaatslvd gfgmKvsRiG Glqhmrafrd 280
gcaarNlpht cDdAWGgdIv saActHIasT vlprlmegaW LAQPYvaehy daengvrieg grirvpqgPG 350
lGltidperf gpplfsa 367
```

Figure 2.10: Prediction performance of PINGU on 4-hydroxyproline betaine 2-epimerase. The alphabets in upper case indicate PINGU predictions; alphabets underlined are predicted ligand binding residues; and those highlighted are catalytic residues.

2.3.7 Software availability

The software **PINGU** along with the user manual and associated data is available at `http://dx.doi.org/10.6084/m9.figshare.1492931`.

2.4 Conclusion

Based on the findings obtained, it can be understood that despite several efforts through many years, there are still many issues associated with the prediction of catalytic residues and efforts towards this direction have been continuing as shown in some recent studies (Sun et al., 2016;

Xiao et al., 2015). However, using domain knowledge and supervised machine learning techniques in combination can certainly improve the prediction scenario, such as demonstrated in chapter[1] by selecting robust features and application of predicted ligand binding residues as a post-processing filter. The findings of this study are hoped to boost the enzyme function annotation eventually for use in various biotechnological applications.

[1] Relevant findings: **Pai, P. P.**, Ranjani, S. S. S., and Mondal, S. (2015). PINGU: PredIction of eNzyme catalytic residues usinG seqUence information. *PLoS ONE*, 10(8): e0135122. `http://doi.org/10.1371/journal.pone.0135122`

Chapter 3

MOWGLI: prediction of protein-MannOse interacting residues With ensemble classifiers usinG evoLutionary Information

This chapter has details of studies performed for sequence-based identification of mannose-interacting residues in proteins showcasing efforts for achieving enhanced precision. Application of consensus information using varied type of neighbourhood information in an ensemble architecture is demonstrated here for improving the prediction scenario.

3.1 Introduction

Among the many protein-ligands interactions, the ones involving carbohydrates are essential for cellular processes such as signaling, structural support, inter-cell interactions, cell-matrix adhesion, growth, and immune response (De Schutter and Van Damme, 2015). Since proteins interacting with carbohydrate ligands are present almost ubiquitously in different tissues as cell surface conjugates, over evolutionary time, they have been utilised as receptors for attachment and invasion, by several disease causing microorganisms (Vliegenthart, 2007). For example, infections involving Ebola virus (Lin et al., 2003), malaria, dengue, African sleeping sickness, tick-borne fevers, and human immunodeficiency virus (HIV) (Dinglasan and Jacobs-Lorena, 2005), involve protein interactions with mannose and its variants, a special class of carbohydrates. In order to survive, the host requires to respond adequately to these infectious agents (Dinglasan and Jacobs-Lorena, 2005), which may be externally supplemented by global control efforts including strategies for diagnosing diseases, infection control and treatment. A need for progress in the novel interventions has become more prominent.

Computational efforts in aiding experiment driven therapeutic strategies can be directed towards analysis and identification of these interactions at greater depth. For example, a protein interacting with mannose, also known as mannose-binding lectin, is a calcium-dependent serum protein, that participates in the innate immune response. Essentially, it binds to carbohydrates on the pathogen surfaces, where it can elicit complement system activation or directly act as an opsonin (Koch et al., 2001). Another protein, antibody 2G12, uniquely counterpoises a wide range of HIV-1 isolates. It does so by binding the high-mannose glycans on the HIV-1 surface glycoprotein gp120 (Sanders et al., 2002). Based on their functional implications, it can be clearly understood that the potential of understanding these residues, at greater depth, could have applications in HIV-1 vaccine development. With increasing availability of protein-sequence and structure information, materialising computational contribution to the biomedicine community has become feasible.

General and specific carbohydrate interacting predictors are available for use. Methods based on structure, for example, use properties such as those discriminating sugar-binding surface patches (Taroni et al., 2000) or three-dimensional probability density distribution of interacting atoms (Tsai et al., 2012) or binding energy (Gromiha et al., 2014) for general prediction of carbohydrate binding sites. Specific prediction approaches have also been reported such as for galactose binding sites (Sujatha et al., 2004), glucose binding sites (Nassif et al., 2009), and mannose binding sites (Khare et al., 2012). Sequence-based methods have also progressed over years since the first prediction approach, from general protein-carbohydrate interacting sites predictors (Malik and Ahmad, 2007), to predictors which are more specific, such as, for mannose binding sites (Agarwal et al., 2011).

Though this field has been advancing, accurate sequence or structure based prediction of specific protein-carbohydrate binding residues such as (mannose-interacting residues) MIR is still challenging. This is because of their non-uniform distribution (Agarwal et al., 2011). This may be because such interactions are not highly selective and they exhibit multiple specificities. Any subtle changes in the size and nature of constituting amino acids can affect ligand affinity of binding proteins. Studies suggest that within a family there seems to be a general pattern in the nature of protein-carbohydrate interaction but gaining insights into distinguishing common pattern between the families and specific carbohydrate binding has ample scope for research, for example, lectin families and their interaction with mannose (Srinivasan et al., 1999). There are reports of structural features with discriminative potential but at sequence level, needless to say, the identification process gets very challenging. This presents an exciting scope of exploration for the development of novel medical interventions (Raz and Nakahara, 2008), biological findings such as in characterisation of enzyme dynamics (Virgens et al., 2014), understanding of biomolecular recognition mechanisms (Mamidi and Surolia, 2015), and unraveling molecular etiology of diseases (Fernandes et al., 2014).

The approach proposed here for identification of mannose interacting residues is called **MOWGLI** (prediction of protein-**M**ann**O**se interacting residues **W**ith ensemble classifiers usin**G** evo**L**utionary **I**nformation). Its development and implementation is described in the next section.

3.2 Materials and Methods

For developing a mannose-interacting residue predictor, as described in the previous study, first a benchmark dataset comprising of protein-mannose interacting complex information has to be created for learning and testing. Biological properties of mannose-interacting and non-interacting residues have to be encoded for discrimination using a function. In this study, many such functions have been generated to draw a consensus for exploring the scope of achieving enhanced predictions. This is described in details below:

3.2.1 Datasets

For this study, PDB codes of proteins which are bound to mannose and its derivatives, along with information of their interacting residues were collected from BioLip database (Yang et al., 2013a). This database available at `http://zhanglab.ccmb.med.umich.edu/BioLiP/`, has semi-manually curated high-quality, biologically relevant ligand-protein binding interactions information, was used to prepare the non-redundant datasets. Basically, a list of protein chains (PDB (Berman et al., 2000) code and chain identifier) interacting with mannose and its variants (listed in Table 3.1) were collected. Mannose interacting residues were noted. Following this, the sequence information was generated using the ATOM record of their structures (solved by X-ray crystallography with a resolution of 3.0 Å) and filtered for presence of fragments and non-naturally occurring amino acids. Further, sequences which were more than 25% similar to any other sequence in the collection were removed using BlastClust (Altschul et al., 1997). Altogether, a non-redundant benchmark data set consisting of 157 protein chains with 1311 MIR was obtained. Of the total, 917 MIR (128 protein chains) were randomly reserved for training and 394 MIR (29 protein chains) for testing. The training and testing datasets were named as Dset128 and Dtestset29, respectively.

Table 3.1: List of mannose and its variants used in the development of MOWGLI. The three letter representation are as per the PDB code.

Ligand ID	Name
DRI	4-O-methyl-2,6-dideoxy-beta-D-glucose
FVQ	3-pyridin-3-ylprop-2-yn-1-yl alpha-D-mannopyranoside
G1P	Alpha-D-glucose-1-phosphate
GDD	Guanosine-5’-diphosphate-alpha-D-mannose
GL1	1-O-phosphono-alpha-D-galatopyranose
HNV	3-(4-methoxyphenyl)prop-2-yn-1-yl alpha-D-mannopyranoside
HNW	D-mannose alpha1O P-hydroxylpropynyl-phenyl
LRY	N-acetylmannosamine-6-phosphate
M13	Methyl 3-O-alpha-D-mannopyranosyl-beta-D-altropyranoside
M6D	6-O-phosphono-beta-D-mannopyranose
M6P	Alpha-D-mannose-6-phosphate
MA2	4-S-methyl-4-thio-alpha-D-glucopyranose
MAF	2-deoxy-2-fluoro-alpha-D-mannose
MAN	Alpha-D-mannose
MBF	2-deoxy-2-fluoro-beta-D-mannose
MDM	Methyl-O3-(alpha-D-mannose)-alpha-D-mannose
MMA	O1-methyl-mannose
MN9	2-acetylamino-2-deoxy-D-mannose
OPM	O1-pentyl-mannose
RNS	L-rhamnose
SHG	2-deoxy-2-fluoro-beta-D-glycopyranose
UFM	Uridine-5’-diphosphate-mannose
XGP	1-O-phosphono-beta-D-glucopyranose
XMM	(2R,3S,4S,5S,6R)-2-(5-bromo-4-chloro-1H-indol-3-yloxy)-tetrahydro-6-(hydroxymethyl)-2H-pyran-3,4,5-triol

3.2.2 Amino acid composition

Biological functions of proteins depend upon the amino acids that constitute its architecture. To know if certain types of amino acids are favored over the others in mannose interaction, composition of the mannose-interacting residues (MIR), and non-mannose-interacting residues (Non-MIR) were examined. Essentially, the amino acid composition (AAC) of a given protein data set which represents 20 natural amino acids is computed using the following formula for MIR:

$$AAC_i^{MIR} = \frac{f_i^{MIR}}{N^{MIR}} \tag{3.1}$$

And, for Non-MIR:

$$AAC_i^{Non-MIR} = \frac{f_i^{Non-MIR}}{N^{Non-MIR}} \tag{3.2}$$

where AAC^{MIR} and $AAC^{Non-MIR}$ are the AACs of MIR and Non-MIR; $i \in \{I, V, L, F, C, M, A, G, T, W, S, Y, P, H, E, Q, D, N, K, R\}$ f_i^{MIR}, $f_i^{Non-MIR}$ are the frequency of occurrence in MIRs and Non-MIRs in the data set; and N^{MIR} , $N^{Non-MIR}$ are the total numbers of MIRs and Non-MIRs, respectively, in the data set.

3.2.3 Features

In this study, sequence based protein evolutionary information was obtained in the form of PSSM using the PSI-BLAST program (Altschul et al., 1997), searching the National Center for Biotechnology Information Non-redundant database available at `ftp://ftp.ncbi.nlm.nih.gov/blast/db/` with three iterations and E-value cut-off = .001 for multiple sequence alignment. The final PSSM profile is a matrix comprising of 20-dimensional weighted observed percentages (WOP) for each residue, and each of these values (x) is normalised to the range of [0, 1] using $1/(1+e^{-x})$. For a protein P of length L, $S_{i \to j}$ represents the normalised occurrence probability of amino acid at position i of the protein sequence, when it is mutated by j during the evolutionary process; $j \in \{A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V\}$. After obtaining the normalised PSSM values for a protein, multi-dimension features for the target residues were created by encoding the PSSM values into a vector. The neighbourhood information along with the evolutionary information for a residue, has been previously reported (Agarwal et al., 2011) to contribute to the distinct properties of the mannose-interacting region. Therefore, in a window of $(2w + 1)$ residues, details of w residues in the flanking region from each side of the target residue was included, where $w \in \{4, 5, 6, 7, 8, 9, 10\}$ giving rise to window sizes of 9, 11, 13, 15, 17, 19, and 21. Upon exploring in this window sizes, the features vectors of $20 * (2w + 1)$ dimension were made.

3.2.4 Prediction engine and performance assessment

A variety of their ensemble architectures in this study, as can be seen in Figure 3.1, were created based on the optimised base classifiers trained on Dset128 using evolutionary information. These architectures comprised of varied G, where $G \in \{10, 15, 20, 25, 30\}$ denoting the number of base classifiers used in the ensemble. The idea behind exploring a variety of G is to understand, for a given set of positive examples, with varied negative examples, how best the prediction performance can be improved. Two classifiers, *i.e.*, RF and SVM were explored and optimised for developing the base-classification algorithm. The performance assessed using 10-fold cross-validation (10CV) using Dset128. Since MCC is considered to be an assessor of how well predictions correlate with observed class labels (Baldi et al., 2000), the best models were selected based on best MCC. Thus, for a set of 10 training folds ($k \in \{1, 2, ..., 10\}$) 10 best models were obtained for a base-classifier. To create the ensemble architecture, for each G, G number of base-classifiers were created using Dset128. With an objective of enhancing the true class prediction and diminishing false prediction based on randomness, the constituting base-classifiers were created in the following manner. Basically, in all base-classifiers, k^{th} fold where $k \in \{1, 2, ..., 10\}$ contained same positive examples and negative examples were subsampled randomly from the pool of negative examples.With the created ensembles, for each G, predictions were obtained for protein chains in Dtestset29. For a protein chain, k^{th} fold, $k \in \{1, 2, ..., 10\}$ prediction from each base-classifier was considered for every residue position and a consensus drawn to assign the new class. In this way, consensus-based predictions

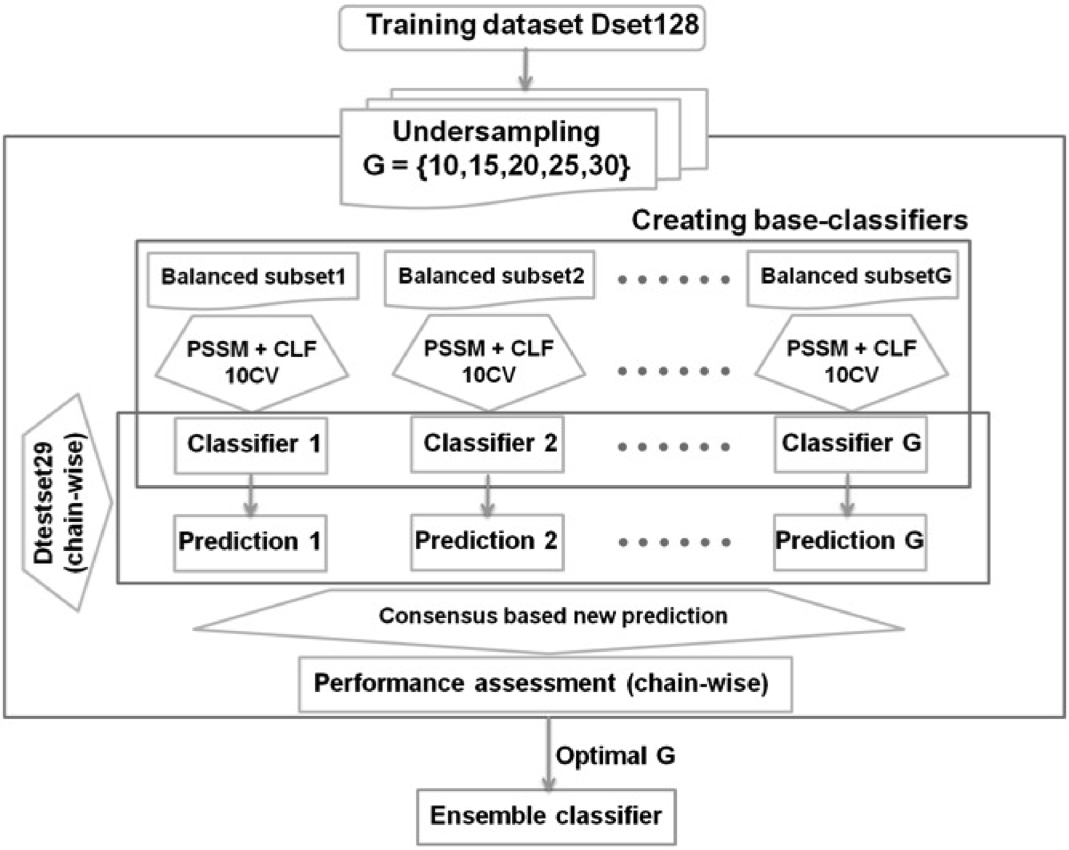


Figure 3.1: The overall study workflow where an ensemble architecture is built from created training dataset and optimised for prediction purposes.

were obtained for all protein chains and the performance assessed fold-wise. That is, prediction performance of all the 29 chains was averaged for a given fold. The best-performing fold (best MCC) was considered for comparison across various G ensembles.

3.3 Results and discussion

3.3.1 Dataset analysis: insights into residues interacting with mannose and its variants

Proteins that interact with mannose and its variants were studied to gain insights into their characteristics traits, if any. Among the chains in Dset128 and Dtestset29, the sequence length and distribution of interacting residues were explored. The sequence lengths of chains varied from 74 (aa) to 971 (aa). They belonged to multiple-classes of proteins including enzymes, lectins, periplasmic protein, and signaling protein from a variety of organisms including lower microorganisms to higher plants and animals. Figure 3.2 shows the frequency of occurrence of amino acids constituting interacting and non-interacting residues. After the dataset analysis, based on the interacting residue particulars, the classification algorithm was chosen for creation of ensemble architecture.

3.3.2 Base-classifer and ensemble performance

Considering the imbalanced data as shown above, if the entire data are used, the classifier may be prone to ignoring the minority class, *i.e.*, the class of interest consisting of MIR. Therefore, multiple comparatively small and balanced subsets (where representatives from the larger pattern are to be selected randomly) were employed in the study. First, an optimisation of individual classification was done and then they were combined into ensemble architecture. For base-classifier creation, biologically relevant properties such as PSSM and local AAC (AAC of the fragment containing interacting residues along with their neighbourhood) were extracted, for the interacting residues present in Dset128 chains. The process was optimised using 10CV. The so obtained best results are shown in Table 3.2.

Clearly, all assessment parameters suggest PSSM to have relatively more discriminatory

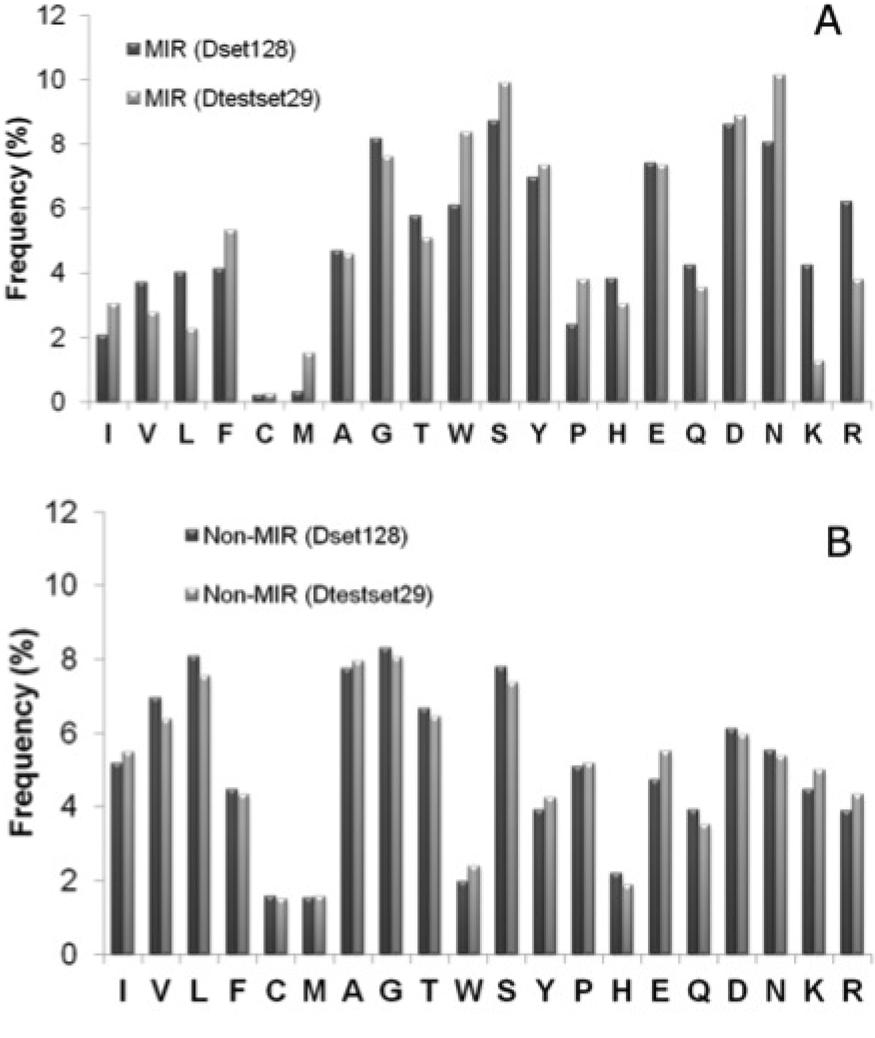


Figure 3.2: Frequency of occurrence of the mannose (A) interacting residues (MIR) and (B) non-interacting (Non-MIR) in the datasets.

Table 3.2: Training prediction performance using different feature types.

Window size	Classifier	Features	MCC	FM
9	RF	PSSM	0.368	66.6
		LAAC	0.197	54.2
	SVM	PSSM	0.390	60.1
		LAAC	0.251	52.1
11	RF	PSSM	0.368	67.0
		LAAC	0.145	50.2
	SVM	PSSM	0.374	65.4
		LAAC	0.215	58.1
13	RF	PSSM	0.385	67.6
		LAAC	0.165	51.0
	SVM	PSSM	0.398	65.2
		LAAC	0.183	56.9
15	RF	PSSM	0.373	67.0
		LAAC	0.189	50.9
	SVM	PSSM	0.389	58.6
		LAAC	0.204	57.0
17	RF	PSSM	0.364	66.1
		LAAC	0.217	50.4
	SVM	PSSM	0.384	57.5
		LAAC	0.222	57.4
19	RF	PSSM	0.368	66.4
		LAAC	0.187	48.6
	SVM	PSSM	0.374	62.1
		LAAC	0.242	59.8
21	RF	PSSM	0.357	66.0
		LAAC	0.201	49.3
	SVM	PSSM	0.374	65.4
		LAAC	0.237	58.6

potential than LAAC with both RF and SVM with maximum MCC in the window size of 13. The RF showed an MCC of 0.385 and FM of 67.6%. SVM showed an MCC of 0.398 and FM of 65.2%. Therefore, in this study, PSSM has been used as a discriminating feature of MIR. Both RF and SVM using a window size 13 were employed further in the classification ensemble architecture. This was developed as described in Materials and Methods section 2.6 with G sets of 10, 15, 20, 25, and 30 base-classifiers on Dset128. The performance evaluation parameters of the training phase are shown in Table 3.3.

Table 3.3: Prediction performance of base-classifiers on an average. Each base-classifier obtained using 10-fold cross-validation on Dset128.

G	Classifier	SN	PR	SP	AC	MCC	FM
10	RF	62.1	69.6	72.7	67.4	0.352	65.3
	SVM	54.8	78.8	79.5	67.2	0.382	59.3
15	RF	62.2	69.7	72.8	67.5	0.355	65.4
	SVM	54.6	78.8	79.5	67.0	0.380	59.0
20	RF	62.0	69.6	72.7	67.4	0.352	65.2
	SVM	54.3	78.7	79.4	66.8	0.376	58.6
25	RF	62.0	69.5	72.6	67.3	0.351	65.2
	SVM	53.4	79.2	79.8	66.6	0.374	57.8
30	RF	61.8	69.4	72.6	67.2	0.349	65.0
	SVM	52.6	79.6	80.2	66.4	0.372	57.2

Consequently, by means of the best trained models of each G set, predictions were obtained for protein chains in Dtestset29 which had considerable diversity. The predictions were further analysed and this process repeated for all the sets of ensembles. The number of votes needed was studied and optimised based on best averaged MCC as shown in Figure 3.3.

The ensemble with a set of 25 base-classifiers, with a consensus from all 25 for both approaches, showed an MCC of 0.370 for RF and 0.333 for SVM. This was analysed further to understand the prediction scenario obtained using ensemble and non- ensemble approach. Looking into Figure 3.4, it can be clearly understood that there is an enhancement in prediction using the ensemble approach as compared to the base-classifiers, with an increase in MCC from 0.208 to 0.370 upon using RF and an increase from 0.202 to 0.333 with SVM.

Based on the above, the 25 RF base-classifier based ensemble architecture is selected for

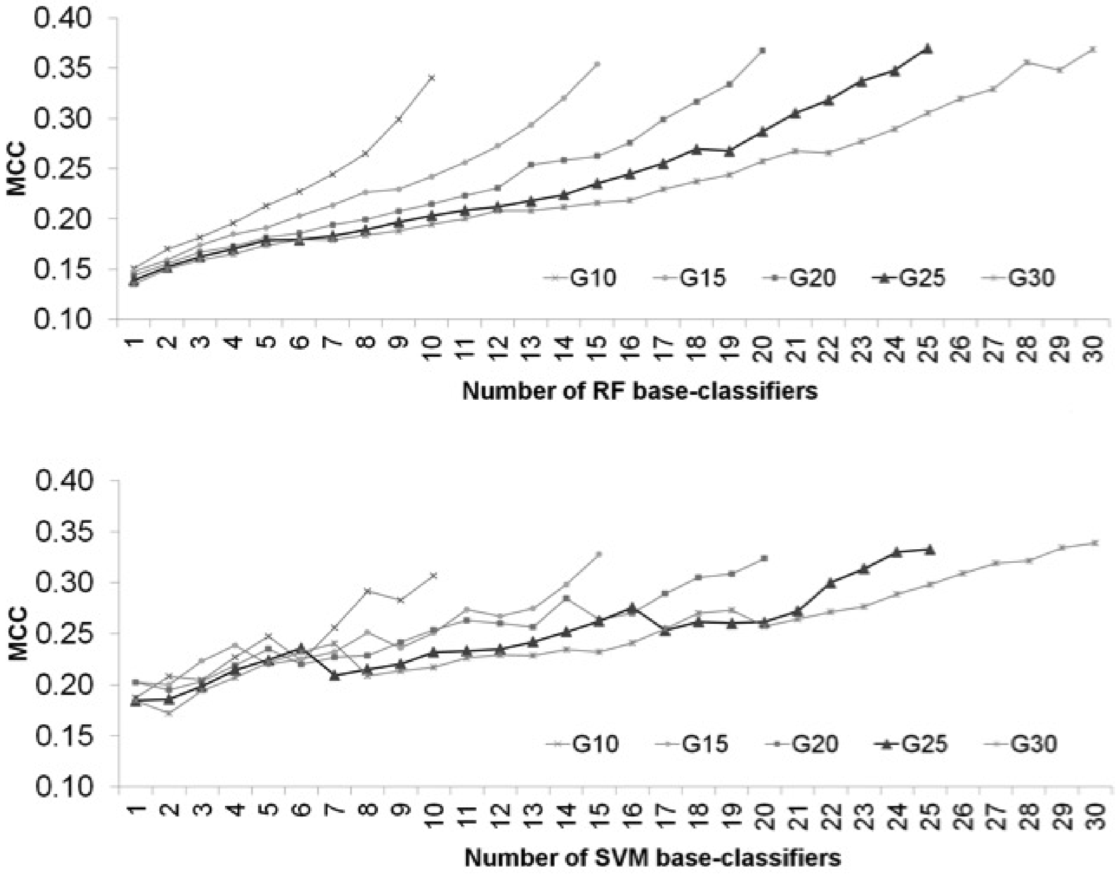


Figure 3.3: Prediction performance of RF (top panel) and SVM (bottom panel) based ensemble classifiers.

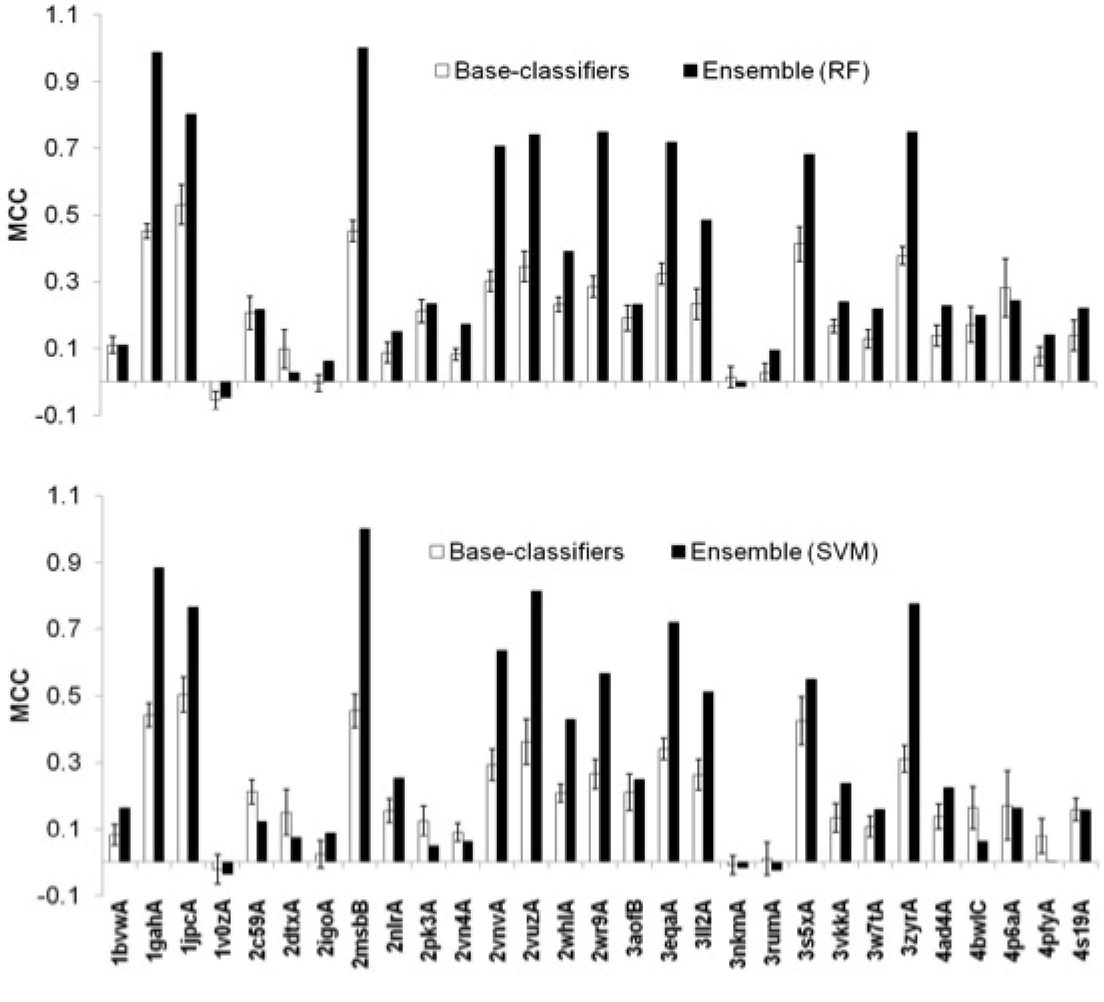


Figure 3.4: Comparative analysis of ensemble and non-ensemble classification approaches using RF (top panel) and SVM (bottom panel).

further analysis and named **MOWGLI** (prediction of protein-**M**ann**O**se interacting residues **W**ith ensemble classifiers usin**G** evo**L**utionary **I**nformation). In order to understand its discriminatory potential for proposing this work as a new prediction approach, it is important that MOWGLI 's performance be comparable with the state-of-art mannose-specific and general carbohydrate binding site predictors. The comparative prediction performance obtained for MOWGLI is described below.

3.3.3 Comparison with state-of-art

Comparison of a newly put forth method with previously reported solutions is an important step towards development of an effective computational approach. Owing to the differences in data sets, definitions of problems, and approaches, a direct comparison with the performance published in the literature is next to impossible (Murakami and Mizuguchi, 2010). However, upon careful considerations with the available state-of-art and relative performance analysis of MOWGLI was performed with sequence-based mannose-specific predictor PreMieR (Agarwal et al., 2011), and general carbohydrate binding site predictor, CBS-PSSM (Malik and Ahmad, 2007).

For comparison with mannose-specific predictor, the 29 chains in Dtestset29 were submitted to PreMieR web server available at `http://www.imtech.res.in/raghava/premier/` and analysed. Detailed protein chain-wise prediction performances obtained using both the predictors are shown in Tables 3.4,3.5 and 3.6.

There is an improvement of 22.7% in sensitivity, 26.6% in precision, and 9.2% on an average in specificity, suggesting that in the predictions obtained by MOWGLI, improved discrimination of MIRs from non-MIRs can be achieved. Figure 3.5 shows the diminishing false positive rate that can be achieved by MOWGLI indicating that ensemble approach can help address the imbalance in data better than individual classification algorithms. The overall performance showed an increase of 0.286 in MCC and 19.0% in FM which is indeed reassuring. The performance obtained by MOWGLI in the test cases presents its enhanced prediction power, but whether MOWGLI's performance is comparable with general carbohydrate binding site predictor still required further insights. So, the same test cases were submitted to the CBS-PSSM (Malik and Ahmad, 2007) web-server and the predictions evaluated. CBS-PSSM uses evolutionary information in neural networks for prediction of carbohydrate binding sites including mannose as one of the ligands. Table 3.6 shows the detailed predictions obtained by CBS-PSSM for each of the 29 test cases.

Table 3.4: Prediction performance of MOWGLI as compared to state-of-art. The parameters which could not be assessed have been labelled as 'NA'

Protein ID	SN	PR	SP	AC	MCC	FM
1bvwA	50.0	4.7	82.8	82.2	0.110	8.6
1gahA	100.0	97.8	99.8	99.8	0.988	98.9
1jpcA	78.9	88.2	97.8	94.4	0.802	83.3
1v0zA	0.0	0.0	91.1	88.9	-0.048	NA
2c59A	20.0	33.3	97.1	91.8	0.217	25.0
2dtxA	12.5	4.8	91.9	89.4	0.028	6.9
2igoA	22.2	4.1	91.7	90.6	0.062	6.9
2msbB	100.0	100.0	100.0	100.0	1.000	100.0
2nlrA	50.0	9.3	81.8	80.6	0.150	15.7
2pk3A	13.6	50.0	99.0	92.9	0.235	21.4
2vn4A	54.5	7.9	88.1	87.5	0.172	13.8
2vnvA	90.0	60.0	94.5	94.2	0.707	72.0
2vuzA	66.7	85.7	99.2	96.9	0.740	75.0
2whlA	88.9	20.0	88.8	88.8	0.390	32.7
2wr9A	88.9	66.7	96.4	95.8	0.749	76.2
3aofB	50.0	15.2	90.6	89.3	0.233	23.3
3eqaA	83.9	65.0	96.7	95.9	0.717	73.3
3ll2A	66.7	44.4	91.0	88.6	0.484	53.3
3nkmA	0.0	0.0	96.4	95.8	-0.015	NA
3rumA	20.0	6.7	96.2	95.2	0.095	10.0
3s5xA	50.0	100.0	100.0	93.2	0.681	66.7
3vkkA	55.6	12.8	92.6	91.8	0.239	20.8
3w7tA	35.3	16.7	96.0	94.6	0.218	22.7
3zyrA	84.6	68.8	97.8	97.1	0.748	75.9
4ad4A	55.6	12.8	90.0	89.1	0.229	20.8
4bwlC	20.0	25.0	97.9	95.3	0.199	22.2
4p6aA	18.8	50.0	96.9	86.0	0.244	27.3
4pfyA	25.9	14.9	92.2	88.9	0.140	18.9
4s19A	50.0	12.0	93.8	93.1	0.221	19.4

Table 3.5: Prediction performance of PreMieR (Agarwal et al., 2011)

Protein ID	SN	PR	SP	AC	MCC	FM
1bvwA	0.0	0.0	90.7	89.2	-0.04	NA
1gahA	97.8	31.2	77.2	79.2	0.481	47.3
1jpcA	15.8	13.6	78.7	67.6	-0.053	14.6
1v0zA	22.2	16.7	97.4	95.6	0.170	19.1
2c59A	12.0	50.0	99.1	93.1	0.221	19.4
2dtxA	0.0	0.0	85.4	82.8	-0.073	NA
2igoA	0.0	0.0	88.9	87.5	-0.044	NA
2msbB	100.0	22.6	77.4	78.8	0.418	36.8
2nlrA	25.0	3.9	77.1	75.2	0.009	6.8
2pk3A	13.6	8.8	89.2	83.8	0.023	10.7
2vn4A	36.4	2.9	77.6	76.8	0.045	5.4
2vnvA	40.0	22.2	87.3	83.3	0.211	28.6
2vuzA	22.2	13.3	89.2	84.5	0.091	16.7
2whlA	22.2	2.6	73.3	71.8	-0.017	4.6
2wr9A	44.4	22.2	82.3	84.0	0.234	29.6
3aofB	10.0	4.8	93.3	90.6	0.023	6.5
3eqaA	93.6	20.9	74.2	75.6	0.370	34.1
3ll2A	100.0	20.3	57.7	61.8	0.342	33.8
3nkmA	0.0	0.0	87.3	86.7	-0.03	NA
3rumA	0.0	0.0	87.3	86.2	-0.044	NA
3s5xA	11.1	9.1	82.5	72.7	-0.059	10.0
3vkkA	11.1	1.6	86.7	85.2	-0.009	2.8
3w7tA	17.7	2.1	80.8	79.3	-0.006	3.7
3zyrA	92.3	20.0	78.9	79.6	0.372	32.9
4ad4A	0.0	0.0	95.9	93.4	-0.033	NA
4bwlC	0.0	0.0	95.1	91.9	-0.042	NA
4p6aA	6.3	14.3	93.9	81.6	0.002	8.7
4pfyA	0.0	0.0	89.3	84.8	-0.007	NA
4s19A	0.0	0.0	85.1	83.7	-0.054	NA

Table 3.6: Prediction performance of CBS-PSSM (Malik and Ahmad, 2007)

Protein ID	SN	PR	SP	AC	MCC	FM
1bvwA	0.0	0.0	98.0	96.4	-0.018	NA
1gahA	26.3	0.0	99.1	89.6	-0.03	NA
1jpcA	26.3	62.5	96.6	84.3	0.334	37.0
1v0zA	0.0	NA	100.0	97.7	NA	NA
2c59A	4.0	14.3	98.2	91.8	0.041	6.3
2dtxA	12.5	7.1	94.7	92.2	0.055	9.1
2igoA	0.0	0.0	99.1	97.6	-0.012	NA
2msbB	14.3	12.5	93.4	88.5	0.072	13.3
2nlrA	25.0	7.4	88.3	86.0	0.076	11.4
2pk3A	0.0	0.0	97.2	90.3	-0.045	NA
2vn4A	0.0	0.0	99.3	97.5	-0.011	NA
2vnvA	0.0	NA	100.0	91.7	NA	NA
2vuzA	22.2	15.4	90.8	86.0	0.110	18.2
2whlA	44.4	14.8	91.9	90.5	0.217	22.2
2wr9A	0.0	NA	100.0	92.4	NA	NA
3aofB	20.0	13.3	95.6	93.2	0.129	16.0
3eqaA	0.0	0.0	98.4	91.7	-0.034	NA
3ll2A	16.7	18.2	91.9	84.6	0.089	17.4
3nkmA	0.0	0.0	99.9	99.3	-0.003	NA
3rumA	20.0	12.5	98.1	97.1	0.144	15.4
3s5xA	5.6	14.3	94.7	82.6	0.004	8.0
3vkkA	11.1	6.7	96.9	95.3	0.063	8.4
3w7tA	0.0	0.0	99.9	97.6	-0.005	NA
3zyrA	23.1	33.3	97.4	93.3	0.243	27.3
4ad4A	0.0	0.0	98.2	95.7	-0.022	NA
4bwlC	10.0	20.0	98.6	95.6	0.121	13.3
4p6aA	12.5	22.2	92.9	81.6	0.069	16.0
4pfyA	25.9	43.8	98.2	94.6	0.311	0.326
4s19A	33.3	20.0	97.7	96.7	0.242	25.0

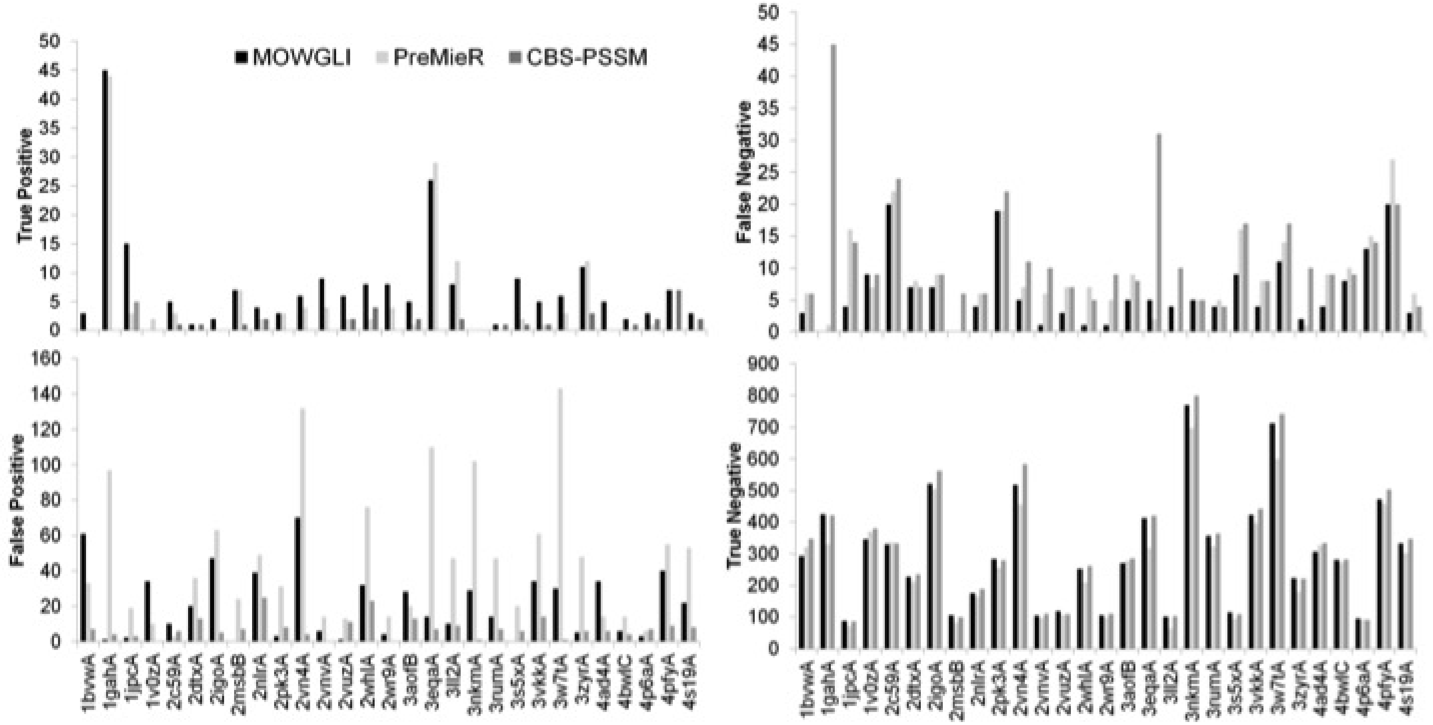


Figure 3.5: Prediction performance of MOWGLI (this work) in comparison with state-of-art mannose-specific predictor PreMieR (Agarwal et al., 2011) and general carbohydrate binding site predictor CBS-PSSM (Malik and Ahmad, 2007).

From the table, it can be understood that, though the prediction accuracy is similar, predictions obtained by MOWGLI are 38.8% more sensitive and 24.1% precise. Figure 3.5 showed high false negative rate in top right panel marked in CBS-PSSM predictions as compared to MOWGLI. The specificity that can be obtained using MOWGLI is 94.1% and CBS-PSSM is 96.7%. Clearly, it can be marked that MOWGLI is able to identify MIRs from non-MIRs more discretely than the state-of-art general sequence-based carbohydrate binding site predictor. The results based on the analysed 29 sequences are optimistic and it would be encouraging to see a similar trend as when more test cases can be included in the study in future. In the following, examples of the enhanced prediction scenario that can be achieved using MOWGLI are shown using case studies.

3.3.4 Case study

In this section, example predictions obtained for proteins using MOWGLI and state-of-art approaches above-mentioned is explained using examples from Dtestset29. Figure 3.6 (A) and (B) show details of the prediction obtained residue wise along the sequence suggesting how MOWGLI can be used for eventual applications in industry and vaccine development.

Mannose-specific agglutinin: This (PDB ID: *1jpcA*) is a protein of the lectin family from Snowdrop bulbs (*Galanthus nivalis*) and it has been determined that residues that interact with mannose and its variants, in a length of 108 amino acids. Upon submitting this sequence for prediction by MOWGLI, 15 of the MIRs were predicted accurately, with a precision of 88.2% obtained upon performance assessment. Prediction for the same chain with PreMieR (Agarwal et al., 2011) suggested that only three MIRs could be identified correctly with a precision of 13.6%. The general carbohydrate binding site predictor CBS-PSSM, though better performing than PreMieR in this case, was also able to identify only five MIRs with predictions showing a precision of 62.5%. After exploring the scenario at sequence level, structural information of this protein chain was used for the prediction. Using COACH (Yang et al., 2013b), a template-based meta-server for ligand binding sites of which mannose and its variants are a part, for the protein under consideration, predictions were obtained. The predictions were then screened considering those MIR derived from a template other than the query. Analysis of COACH (Yang et al., 2013b) predictions revealed that it was able to identify more residues than PreMieR and CBS-PSSM, *i.e.*, 11 residues with a precision of 22.9%, but not as many as that could be obtained with MOWGLI.

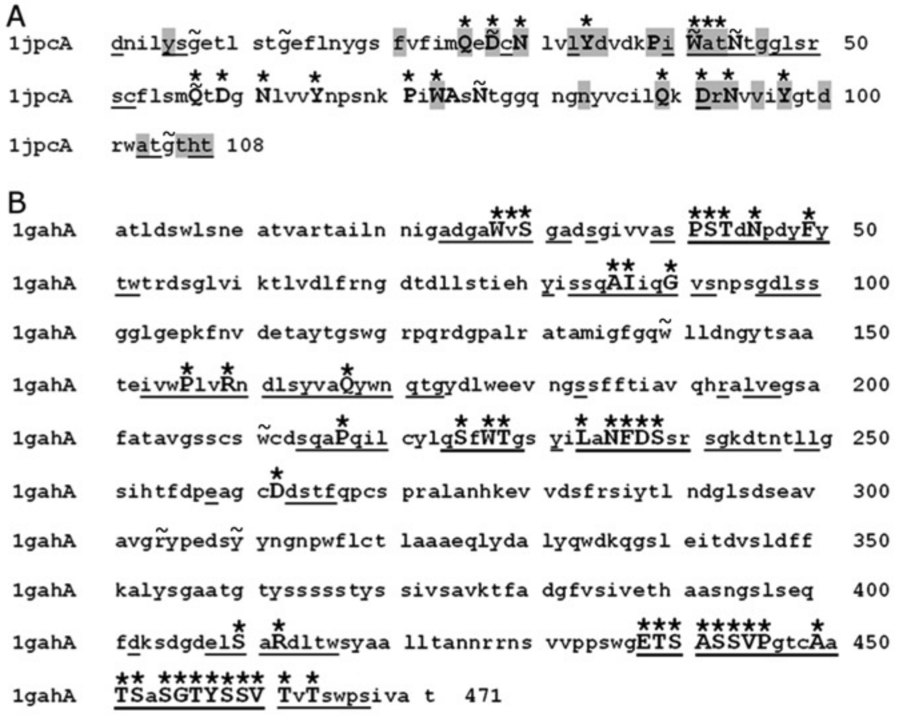

A
1jpcA dnilysgetl stgeflnygs fvfimQeDcN lvlYdvdkPi WatNtgglsr 50
1jpcA scflsmQtDg NlvvYnpsnk PiWAsNtggq ngnyvcilQk DrNvviYgtd 100
1jpcA rwatgtht 108
B
1gahA atldswlsne atvartailn nigadgaWvS gadsgivvas PSTdNpdyFy 50
1gahA twtrdsglvi ktlvdlfrng dtdllstieh yissqAIiqG vsnpsgdlss 100
1gahA gglgepkfnv detaytgswg rpqrdgpalr atamigfgqw lldngytsaa 150
1gahA teivwPlvRn dlsyvaQywn qtgydlweev ngssfftiav qhralvegsa 200
1gahA fatavgsscs wcdsqaPqil cylqSfWTgs yiLaNFDSsr sgkdtntllg 250
1gahA sihtfdpeag cDdstfqpcs pralanhkev vdsfrsiytl ndglsdseav 300
1gahA avgrypedsy yngnpwflct laaaeqlyda lyqwdkqgsl eitdvsldff 350
1gahA kalysgaatg tyssssstys sivsavktfa dgfvsiveth aasngslseq 400
1gahA fdksdgdelS aRdltwsyaa lltannrrns vvppswgETS ASSVPgtcAa 450
1gahA TSaSGTYSSV TvTswpsiva t 471

Figure 3.6: Case study. Prediction scenario for protein (A) Mannose-specific agglutinin and (B) Glucoamylase. The interacting residues are denoted by uppercase. The prediction marked by star, underline, tilde, and highlight are obtained by MOWGLI (this work), PreMieR (Agarwal et al., 2011), CBS-PSSM (Malik and Ahmad, 2007) and COACH (Yang et al., 2013b) respectively.

Glucoamylase: This (PDB ID: *1gahA*) is an enzyme of the hydrolase class from a fungus called *Aspergillus awamori* of length 471 amino acids containing 45 MIR (Aleshin et al., 1996). Of the 45 MIR, using MOWGLI, all residues were accurately identified with a precision of 97.8%. Using PreMieR (Agarwal et al., 2011), 44 residues were identified but 99 residues falsely predicted as mannose-interacting with an overall precision of 31.2%. CBS-PSSM was unable to identify any of the MIRs. Further, COACH (Yang et al., 2013b) predictions also could not be obtained in this case, as templates other than the query were not available.

3.3.5 Software availability

The package for **MOWGLI** along with the user manual is available for the users at `http://sites.google.com/site/sukantamondal/software`. Prediction scenario is shown in Figure 3.7

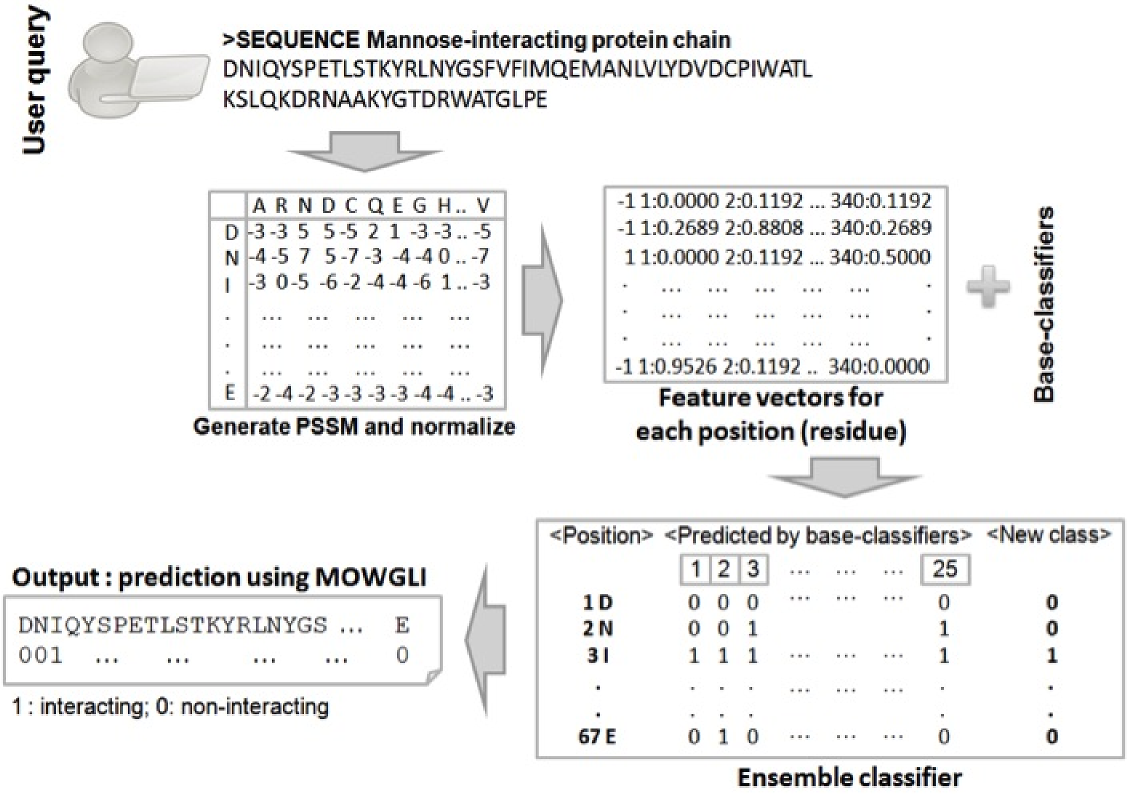

User query
>SEQUENCE Mannose-interacting protein chain
DNIQYSPETLSTKYRLNYGSFVFIMQEMANLVLYDVDCPIWATL
KSLQKDRNAAKYGTDRWATGLPE
A R N D C Q E G H.. V
D -3 -3 5 5 -5 2 1 -3 -3 .. -5
N -4 -5 7 5 -7 -3 -4 -4 0 .. -7
I -3 0 -5 -6 -2 -4 -4 -6 1 .. -3
E -2 -4 -2 -3 -3 -3 -3 -4 -4 .. -3
Generate PSSM and normalize
-1 1:0.0000 2:0.1192 ... 340:0.1192
-1 1:0.2689 2:0.8808 ... 340:0.2689
1 1:0.0000 2:0.1192 ... 340:0.5000
-1 1:0.9526 2:0.1192 .. 340:0.0000
Feature vectors for each position (residue)
Base-classifiers
<Position> <Predicted by base-classifiers> <New class>
1 2 3 25
1 D 0 0 0 ... ... 0 0
2 N 0 0 1 1 0
3 I 1 1 1 ... ... 1 1
67 E 0 1 0 ... ... 0 0
Ensemble classifier
Output : prediction using MOWGLI
DNIQYSPETLSTKYRLNYGS ... E
001 0
1 : interacting; 0: non-interacting

Figure 3.7: Schematic prediction scenario for a user provided query mannose-interacting protein

3.4 Conclusion

Attempts to enhance the prediction scenario of protein-mannose interacting residues can help unveil the underlying mechanisms of host-pathogen interactions and aid in development of infection management strategies. This chapter[1] clearly shows the positive impact of using evolutionary information in the binding region in an random forest based ensemble setting, outperforming the state-of-art. The vast pool of negative examples in the non-interacting region can be used advantageously for discriminative learning purposes. With as minimal as sequence information, enhanced performances can be achieved in the prediction scenario.

[1] Relevant findings: **Pai, P. P.** and Mondal, S. (2016). MOWGLI: prediction of protein-MannOse interacting residues With ensemble classifiers usinG evoLutionary Information. *Journal of Biomolecular Structure and Dynamics*, 34(10): 2069-2083. `http://doi.org/10.1080/07391102.2015.1106978.`

Chapter 4

ROBBY: pRediction Of Biologically relevant small molecule Binding residues on enzYmes

This chapter has details of studies performed for sequence-based identification of diverse ligand binding residues in enzymes showcasing efforts for achieving enhanced precision. Application of consensus information, as well as, template-based and structural insights are shown to positively influence the prediction scenario.

4.1 Introduction

The biological functions of proteins are intertwined to facilitate various processes in the cell. This requires proteins to be evolutionarily designed in such a way that depending upon the biochemical needs, the core architecture can interact with one or multiple ligands at same or different sites in the various stages of the involved biochemical pathway. Over many decades, with increasing number of resolved and available ligand-enzyme complexes, attempts to understand these interactions using sequence and structure information have been made (Roche et al., 2015). Some of these are general (Agostino et al., 2013; Brylinski and Feinstein, 2013; Chen et al., 2016, 2014; Heo et al., 2014; Hu et al., 2016b; Qiu and Wang, 2011; Roche et al., 2013; Singh et al., 2016; Tsujikawa et al., 2016; Yang et al., 2013b) whereas some others are more specific such as DNA (Hu et al., 2016a; Ma et al., 2016) , RNA (Pai et al., 2017; Yasser et al., 2016), Heme (Liu and Hu, 2011), zinc (Liu et al., 2014), vitamin (Yu et al., 2014), mannose (Agarwal et al., 2011). Modelling interactions using general or specific methods use template-based or alternative techniques, contribute to open questions in the field of drug discovery (Konc et al., 2015).

Despite multiple prediction efforts, determining the properties that can be used for discriminative purposes in the identification of protein-ligand binding regions is challenging because of their varied functional demands. Common assumptions such as those related to conservation in evolution, solvent accessibility, presence on surface or pockets, *etc.*, which have been used to discriminate between ligand interacting and non-interacting residues, may require additional considerations. For example, some of the proteins may have a cavity for longer and more specific interactions with ligands, often found in enzyme catalysis (Bartlett et al., 2002). On the contrary, some other proteins such as those involved in molecular recognition or adhesion processes may require more surface-based interactions, where proteins provide the ligands rather exposed shallow clefts for binding or temporary influence (Konc et al., 2015; Krivak and Hoksza, 2015).At higher resolution, the arrangement of residues in the sequence provides for a certain module in structure that can in turn facilitate a specific type of interaction (Boraston et al., 2004; Gutteridge and Thornton, 2005). This arrangement is often conserved in proteins through evolution because of its functional implications in the life processes (Capra et al., 2009). However, the degree of conservation varies depending upon the nature of function and overall biochemical requirement (Rost, 2002). Mutations or modifications in these residues may play an important role in functional regulation (Fu et al., 2000) of these protein-ligand interactions and implicating in disorders or diseases (Gonzalez and Kann, 2012). Scientists have been attempting to understand the evolution and design of the protein-interaction architecture focussing on their innate structure, conformational modifications, target and off-target binding, modes of interaction and function (Konc et al., 2015).

From a computational perspective, the potential in various biological properties may not be discretely discriminative, which puts forth the need to look for alternate means of achieving enhanced identification of these varyingly populated ligand interacting residues. Such problems have been increasingly approached in bioinformatics using ensemble learning as already mentioned in the previous chapter. Since binding site evolution affects the specificity and selectivity of interactions in the protein architecture (Najmanovich, 2017), the framework for prediction is designed using consensus information available in the sequence neighbourhood. In this chapter, studies for developing a *de novo* approach for identifying all the regions which might bind to small molecule ligands are described. For this, the ligand binding regions in biochemically diverse enzymes have been considered. The focus was to achieve enhanced precision without compromising much on sensitivity by exploring supervised machine learning techniques in an ensemble and applying domain knowledge.

4.2 Materials and Methods

The methodology used to create ensembles for the development of ROBBY is shown in Figure 4.1.

4.2.1 Datasets

For this study, enzymes which had at least one bound and unbound protein or complex information were required. So, a list of protein chains, *i.e.*, PDB code and chain identifier, interacting with ligands were collected from the LigASite database. (version 9.7 released April 2012). LigASite consists exclusively of biologically relevant protein-ligand binding sites for which at least one apo- and one holo-structure are available. In defining the protein-ligand binding sites, information from all holo-structures is combined, considering in each case the quaternary structure defined by the PQS server (Dessailly et al., 2008). Their sequence information for all the PDB codes were extracted from the ATOM record of their experimentally solved structures. Fragments and those sequences containing non-naturally occurring amino acids or other ambiguity were removed from the study. The collection was then filtered for enzymes using a software called SIFTS: Structure Integration with Function, Taxonomy and Sequence available at `http://www.ebi.ac.uk/pdbe/docs/sifts`. Then sequences which were more than 30%

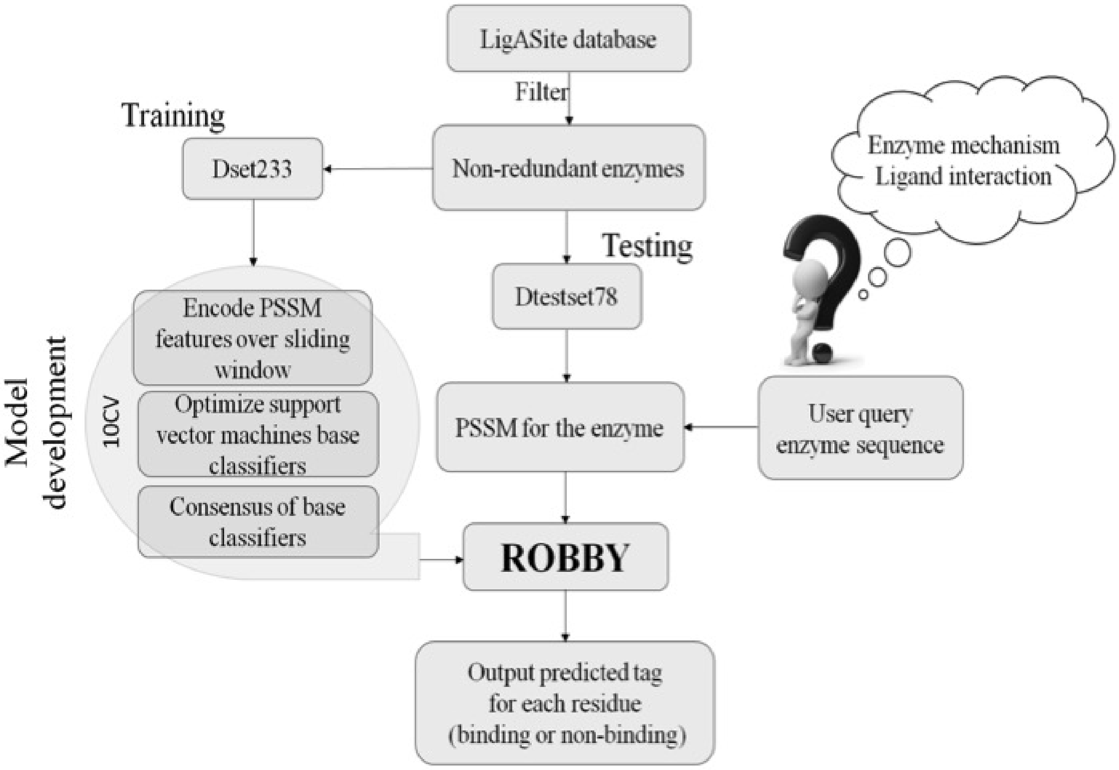

LigASite database
Filter
Training
Dset233
Non-redundant enzymes
Enzyme mechanism
Ligand interaction
Testing
Dtestset78
Model development
10CV
Encode PSSM features over sliding window
Optimize support vector machines base classifiers
Consensus of base classifiers
PSSM for the enzyme
User query enzyme sequence
ROBBY
Output predicted tag for each residue (binding or non-binding)

Figure 4.1: An overview of ROBBY development.

similar to any other sequence in the collection were removed using BlastClust (Altschul et al., 1997). Altogether, a non-redundant benchmark dataset comprising of 311 protein chains with 8682 interacting residues and 87430 non-interacting residues were obtained. Of the total, 6512 interacting residues (233 enzymes) were randomly allocated for training and 2170 interacting residues (78 enzymes) for testing. The training and testing dataset were named as Dset233 and Dtestset78 respectively. To make sure the study holds true for newly identified enzymes, not only learning from gold standard dataset is crucial, but also, testing has to be performed on recent information too. With this objective, using the advanced search interface of the PDB, proteins with at least one *apo-* (Note: these had Enzyme Classification Search: EC=1 to 6 and Ligand Search: Has free ligands=no) and *holo*-structures released during May 2012 to December 2016 were collected and filtered for more than 30% similarity (with each other, as well as, other previously created datasets) as mentioned above using BlastClust (Altschul et al., 1997). Fragments and proteins not matching with the requisites were excluded from the study. After processing for necessary attributes based on LigASite, a set of 17 enzymes was eventually obtained(with 587 interacting residues and 4343 non-interacting residues) and named Dtestset17 for further use in independent testing.

4.2.2 Features

Evolutionary information has been widely used for protein-ligand interaction related annotation such as in works (Capra et al., 2009; Nagl et al., 1999; Pai and Mondal, 2016; Pai et al., 2015; Panwar et al., 2013). For the development of a general ligand binding predictor, position specific evolutionary information in the form of position-specific scoring matrix has been used. It was generated using PSI-BLAST program by searching the UniRef50 (Suzek et al., 2007) database with three iterations and e-value cut-off 0.001 for multiple sequence alignment. The final PSSM profile is a matrix comprising of 20-dimensional weighted observed percentages for each residue. Each of the PSSM scores (x) that is generally depicted as integers was normalised to the range of [0,1] using $1/(1 + e^{-x})$ for this study. After obtaining the normalised PSSM values for a protein, features for the target residues were created by encoding the PSSM values into a multi-dimension vector along with varied neighbourhood information. Firstly, features were extracted for a given set of interacting residues in the training dataset along with randomly chosen balanced set of non-interacting residues, for a given window of $(2w + 1)$ residues. Here, the information of w residues flanking on each side of the interacting residue were included, where $w \in \{4, 5, 6, 7, 8, 9, 10\}$ giving rise to window sizes of 9, 11, 13, 15, 17, 19 and 21. Then two techniques were applied:

(i) Smoothing and condensing: This was explored to reduce noise while encoding features as described to be advantageous in a relevant study (Fang et al., 2013). This was done by representing each row of the vector for a residue C_i according to the following equation 4.1 for smoothing:

$$Smoothing\_C_i = \frac{1}{2m+1} \sum_{j=i-m}^{j=i+m} PSSM\_C_j, where\ \ i = (1, 2, ..., N) \tag{4.1}$$

where $PSSM\_C_j$ represents the score in the original PSSM, $smoothing\_C_i$ represents the score in the smoothed PSSM, N is the length of the sequence, $2m + 1$ is the smoothing-window size. After smoothing, for condensing, the Kidera factors (Kidera et al., 1985) were used. The smoothed PSSMs are then divided into sliding windows of size m. Each window is a matrix $E_{ij} i =, ..., m, j = 1, ...20$, where j represents each of the standard 20 amino acids. Each feature is calculated according to the following equation:

$$F_{i,p} = \sum E_{i,j} f_{j,p} (i = 1, ..., m, p, = 1, ...10) \tag{4.2}$$

where $f_{j,p}$ means the p^{th} Kidera factor of j (each j has 10 Kidera factors). Finally, each value in the condensed and smoothed PSSM matrix is scaled to the range of [-1, 1].

(ii) Ensemble learning: Further for creating ensembles, the basic feature extraction procedure was repeated three times, where the type of non-interacting residues varied for the same set of interacting residues. And, thus, for each of the 21 features for a given interacting residue, $20 * (2w + 1)$ dimension vectors were constructed and applied for identification purposes.

4.2.3 Prediction engine and performance assessment

The basic concepts and architecture employing SVM and RF that have been used here are described in details in the previous chapter. Models were trained on Dset233 using different window sizes and negative instances while learning with one of SVM and RF at a time. All the selected SVM models (21 models: for a given set of positive examples, three different sets of negative examples were collected across various windows 9 to 21) were combined for obtaining the consensus prediction. The same was repeated for RF followed by performance assessment. The performances of individual base-classifiers and ensemble models were analysed using Dset233 by means of 10CV. This technique along with the assessment parameters such as SN, PR, SP, AC, MCC and FM (details described in chapter 2) have been used. Model selection was based on best MCC. The consensus approach was tested on Dtestset78 and Dtestset17.

4.3 Results and discussion

4.3.1 Prediction insights with sequence information alone

Small-molecule or ligand interacting residue identification has been attempted over many decades now. Many studies have brought to fore, the importance of evolutionary information in the prediction of protein-functional residues. Since enzyme interactions are complex, delineating the involved residues with minimal assumption and information was explored in this study and has been discussed under three headings: (i) Choice of prediction architecture, (ii) Including different types and numbers of neighbours in an ensemble, and (iii) Analysis of prediction performance.

(i) Choice of prediction architecture:

Evolutionary information in the form of PSSM has been extensively explored for application in identification of proteins and their functional aspects. Studies have also brought to fore, the impact of smoothing and condensing PSSM over biochemical factors such as reported for identification of Flavin Adenine Dinucleotide (FAD), Nicotinamide Adenine Dinucleotide (NAD), Adenosine triphosphate (ATP) (Fang et al., 2013) and Heme (Xiong et al., 2012).

Impact of smoothing and condensing: In this study, the scope of using smoothened and condensed PSSM has been explored using Kidera factors (Kidera et al., 1985). Kidera factors carry information about physical properties of all the 20 naturally occurring amino acids. Figure 4.2 shows the prediction performance (10CV) upon condensing PSSM over varied number of neighbors, *i.e.*, in different window sizes, using these factors. The best MCC 0.407 was shown in window size 21 by SVM. This model was investigated further for the impact of smoothing before condensing PSSM (Fang et al., 2013; Xiong et al., 2012). For this, varying smoothing windows were explored. Figure 4.2 clearly suggests that there was no improvement in performance upon smoothing. In order to gain further insights into whether the impact of condensing evolutionary information showed an overall improvement, the contribution of PSSM and Kidera Factors alone (for the chosen window size using SVM) was checked. Results suggested that upon using PSSM alone an MCC of 0.456 was obtained and that with Kidera Factors alone is just 0.242. This indicated the advantage of using PSSM without smoothing and condensing for this study.

Role of position-specific information: Additionally, for the same window, we also explored the role of position-specific and non-specific information, by using BLOSUM62 matrix for the latter. The BLOSUM (BLOcks SUbstitution Matrix) matrix is a substitution matrix used for scoring sequence alignments evolutionarily divergent protein sequences (Henikoff and Henikoff, 1992). BLOSUM62 showed an MCC of 0.252. This suggests that position-specific information is important and encoding it without any smoothing and condensing is more discriminative for prediction in this study.

(ii) Including different types and numbers of neighbours in an ensemble:

It is well-known that the neighbourhood of interacting residues is important for conferring upon them distinct functionally relevant properties. So, the prediction power of models using different types and numbers of neighbours were explored. Prediction performances for various windows using randomly chosen set of negative examples for three sets on Dset233 are summarized in Table 4.1. The best performance obtained for RF was 0.449 MCC and that for SVM is 0.482.

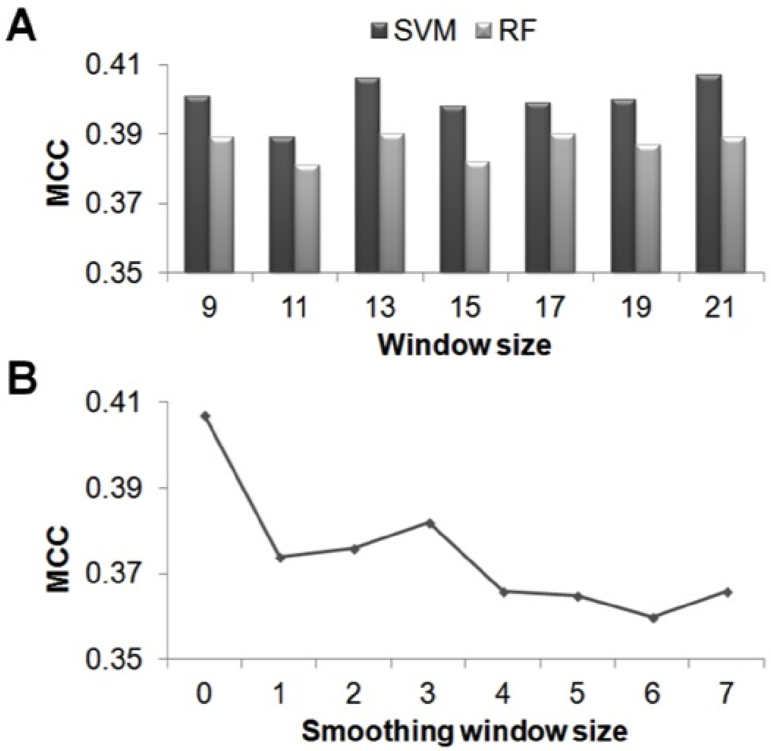


Figure 4.2: Assessment of prediction performance using (A) condensed PSSM with two different classifiers on Dset233 and (B) smoothened neighborhood information with best model.

Table 4.1: Training performance on Dset233

Window	Approach	SN	PR	SP	AC	MCC	FM
9	RF	71.2	72.9	73.6	72.4	0.449	72.0
	SVM	73.3	73.7	73.8	73.6	0.472	73.5
11	RF	70.7	73.0	73.8	72.3	0.446	71.7
	SVM	73.1	74.0	74.2	73.7	0.474	73.5
13	RF	70.8	72.9	73.6	72.2	0.445	71.8
	SVM	73.6	74.2	74.4	74.0	0.480	73.9
15	RF	71.2	73.0	73.6	72.4	0.449	72.0
	SVM	73.5	74.0	74.2	73.8	0.477	73.7
17	RF	70.7	73.2	74.0	72.4	0.448	71.9
	SVM	73.4	74.4	74.7	74.1	0.482	73.9
19	RF	71.0	72.8	73.4	72.2	0.445	71.8
	SVM	73.6	74.2	74.4	74.0	0.480	73.8
21	RF	70.5	73.2	74.3	72.4	0.448	71.8
	SVM	73.3	74.3	74.6	73.9	0.479	73.7

Upon testing these models on Dtestset78, the best performance for SVM showed an MCC of 0.297 and 35.0% F-measure as shown in Table 4.2. With an objective of exploring the scope of using the influence of type and number of neighbours to improve upon precision, consensus of the trained models was drawn. For this, three models trained with same set of positive examples and randomly chosen different negative examples, for each of the different windows, altogether 21 models (base- classifiers) were used. Study findings as summarised in Table 4.3 suggest that a combination of neighbourhood information in an SVM ensemble architecture helps achieve enhanced precision over others. This is named ROBBY: pRediction Of Biologically relevant small molecule Binding residues on enzYmes and analysed for robustness and further applicability.

Table 4.2: Testing performance of various models on Dtestset78

Approach	SN	PR	SP	AC	MCC	FM
RF base classifier: bclRF	69.3	23.9	73.5	72.7	0.276	33.3
RF ensemble classifier: ensRF	54.4	32.7	86.3	82.5	0.316	37.5
SVM base classifier: bclSVM	71.8	25.2	74.1	73.3	0.297	35.0
SVM ensemble classifier: ensSVM	53.2	36.2	88.2	83.9	0.337	39.0

(iii) Analysis of prediction performance

Enzymes are made of different amino acids varying in sequence lengths, number of required interacting sites and biological units in macromolecular assembly (Nelson et al., 2008). They also vary in the number and type of ligands that they interact with for performing functions. This brings into picture the fact that identification of ligand binding sites in enzymes is not a straight forward process. Many perspectives and approaches in a comprehensive setting can perhaps amount to more precise predictions. For example in Dtestset78, it was found that for protein Uridylate (2'-deoxyuridine 5'-monophosphate, UMP) Kinase from a Gram-negative phytopathogen *Xanthomonas campestris* (*3ek6A*) which catalyses the reversible phosphorylation of UMP to UDP (Uridine-5'-diphosphate), Figure 4.3 (A). ROBBY predictions did not have an overlap with experimentally identified residues as per LigASite record, depicted in Figure

4.3 (B). Since in LigASite records, the interacting residues are mapped based on experimentally solved structures, nascent information during annotation may have resulted in such a gap. Figure 4.3 (C) has the sequence based scenario where residues are mapped with experimental observations. One of the directions in which the above goal can be achieved for ligand binding residue identification is by bringing into picture template based methods such as HOMCOS (Fukuhara and Kawabata, 2008). HOMCOS (HOMology modeling of COmplex Structure) is a server for modeling complex 3D structures using 3D molecular similarities based on template complex 3D structures in PDB. For a given amino acid sequence or a chemical structure, the server provides list of contacting molecules in PDB, predicted complex 3D structure based on the template PDB structures. Based on the mapping, it can be inferred from templates that ligands such as UTP (Uridine 5'-triphosphate) (PDB code: *2bnfB*), GTP (Guanosine-5'-triphosphate) (PDB code: *2v4yC*), 4TC (P1-(5'-Adenosine)P4-(5'-Uridine)-beta,gamma-methylene tetraphosphate)(PDB code: *2j4jF*), ATP (Adenosine-5'-triphosphate)(PDB code: *2jjxC*) and ANP (Phosphoaminophosphonic acid-adenylate ester) (*2bmuB*) are preferable by the protein in regions some of which overlap with ROBBY predictions. The above-mentioned ligands are represented by three letter PDB ligand code. As summarised in Table 4.3, for the remaining 77 enzymes in Dtestset78 and Dtestset17, it could be inferred that a stable prediction with above 80% prediction accuracy can be achieved using ROBBY. And further, findings also suggest that as the availability of experimentally observed information improves, some of what constitutes falsely predicted interacting residues may be addressed more appropriately.

4.3.2 Enhancing success rates, validity and applicability using structure information

Protein structural insights have been reported to aid in enhancing prediction scenarios. In this study, we have explored the scope of using predicted pocket information to filter out false predictions if possible. For this purpose, a meta-approach that used the consensus of a variety of computational algorithms and tools developed in the recent decade, MetaPocket 2.0 (Zhang et al., 2011) has been used. Table 4.3 shows the average performance for enzymes in Dtestset78

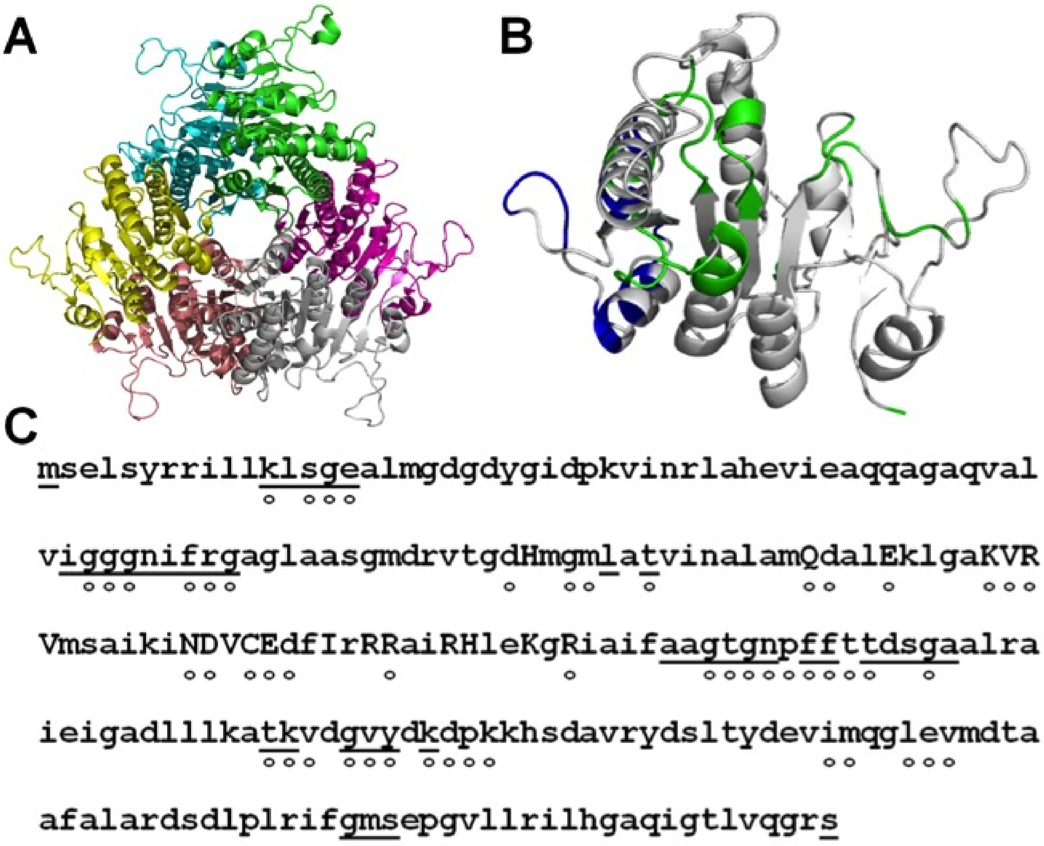

A
B
C
mselsyrrillklsgealmgdgdygidpkvinrlahevieaqqagaqval
vigggnifrgaglaasgmdrvtgdHmgmlatvinalamQdalEklgaKVR
VmsaikiNDVCEdfIrRRaiRHleKgRiaifaagtgnpffttdsgaalra
ieigadlllkatkvdgvydkdpkkhsdavrydsltydevimqglevmdta
afalardsdlplrifgmsepgvllrilhgaqigtlvqgrs

Figure 4.3: Interaction residues of Uridylate kinase (PDB code: 3ek6A) from the Gram-negative plant pathogen *Xanthomonas campestris*. (A) Hexameric biological unit (B) LigASite mapped on one protein chain in blue and ROBBY in green (C) Sequence information: upper-case denoting interacting residues, underline representing ROBBY predictions, open circles denoting putative ligand binding residues using HOMCOS.

(*3ek6A* excluded) where around 50% of interacting residues were identified with around 52% precision upon combination of both the methods. On a general level, upon using a combination of sequence-structure information, the overall precision is enhanced and this is summarised in Table 4.3.

Table 4.3: Testing performance of various models on Dtestset78

Identification using	**SN**	**PR**	**SP**	**AC**	**MCC**	**FM**
Dtestset78						
ROBBY	53.9	36.7	88.3	84.0	0.343	39.0
Pocket information alone	88.2	22.6	65.5	67.9	0.322	34.9
ROBBY with pocket information	50.0	52.3	97.8	89.5	0.436	46.6
Dtestset17						
ROBBY	51.8	37.1	87.1	80.7	0.309	37.2
Pocket information alone	88.2	26.0	62.7	64.8	0.313	36.7
ROBBY with pocket information	49.4	52.8	93.9	86.2	0.410	44.1

4.3.3 Case study

Alcohol dehydrogenase: Belonging to the oxidoreductase class of enzymes, Drosophila alcohol dehydrogenase (DADH; EC 1.1.1.1, organism: *Scaptodrosophila lebanonensis*, length: 254 amino acids) is a NAD(H)-dependent oxido-reductase belonging to the short-chain dehydrogenases/reductases (SDR) family (Benach et al., 1998). It is a homo-dimeric enzyme that catalyses the dehydrogenation of alcohols to their respective ketones or aldehydes in the fruit-fly Drosophila, both for metabolic assimilation and detoxification purposes. In Figure 4.4 (A) on the structure of this protein, the experimentally observed ligand.binding region is presented comprising of 45 interacting residues altogether, of which 31 residues were identified by ROBBY with a precision of 81.6 % indicated in Figure 4.4 (B). MetaPocket 2.0 was able to identify all the interacting residues indicated by stars, but the precision of this prediction obtained was only 37.2 %. Upon combining the two, the overall precision could be enhanced to 91.2 %. Further upon adding insights obtained using template based approaches such as COACH (Yang et al., 2013b), 3DLigandSite (Wass et al., 2010) and FunFOLD2 (Roche et al., 2013) server, it was suggested that predicted regions had the potential to interact with ATP. Using information in combination can help make the prediction precise and informative as shown in this example.

Sensor kinase: Sensor kinases in the bacterial two-component system share a unique ATP-binding Bergerat fold with the GHL (gyrase, Hsp90, and MutL) family of proteins. Salmonella Sensor Kinase PhoQ (Guarnieri et al., 2008) catalytic domain (PhoQcat; EC 2.7.13.3, organism: *Salmonella enterica*, length: 143 amino acids) regulates the expression of genes involved in virulence, adaptation to acidic and low Mg^{2+} environments and resistance to host defense antimicrobial peptides. Figure 4.4 shows individual and combination predictions of ROBBY and MetaPocket 2.0. For this protein, LigASite database lists out only one holo-structure which is monomeric as shown in Figure 4.4 (A), housing 18 experimentally observed interacting residues. Of these, ROBBY was able to identify 10 residues with a precision of 55.6 % and MetaPocket 2.0 identified 9 residues with a precision of 12.0 %. Upon combination of the two, a precision of 55.6 % was obtained which is same as shown by ROBBY but 43.6 % more precise than MetaPocket 2.0. In this scenario, the sequence based method can stand alone. Moreover, the region identified as false positives at 97-100 positions (Gly-Gln-Gly) by ROBBY are putative substrate binding residues (UniProtKB AC: P0DM80). This suggests that, with more experimentally solved structures of ligand bound complexes, the prediction scenario will hopefully enhance. Nevertheless, even as this proceeds, the sequence based methods such as ROBBY, along with structure or template based methods can help understand and validate prediction for experimental studies better.

4.3.4 Software availability

The approach described in this study, ROBBY, is available as a standalone package along with the user-manual at http://doi.org/10.6084/m9.figshare.3472346.

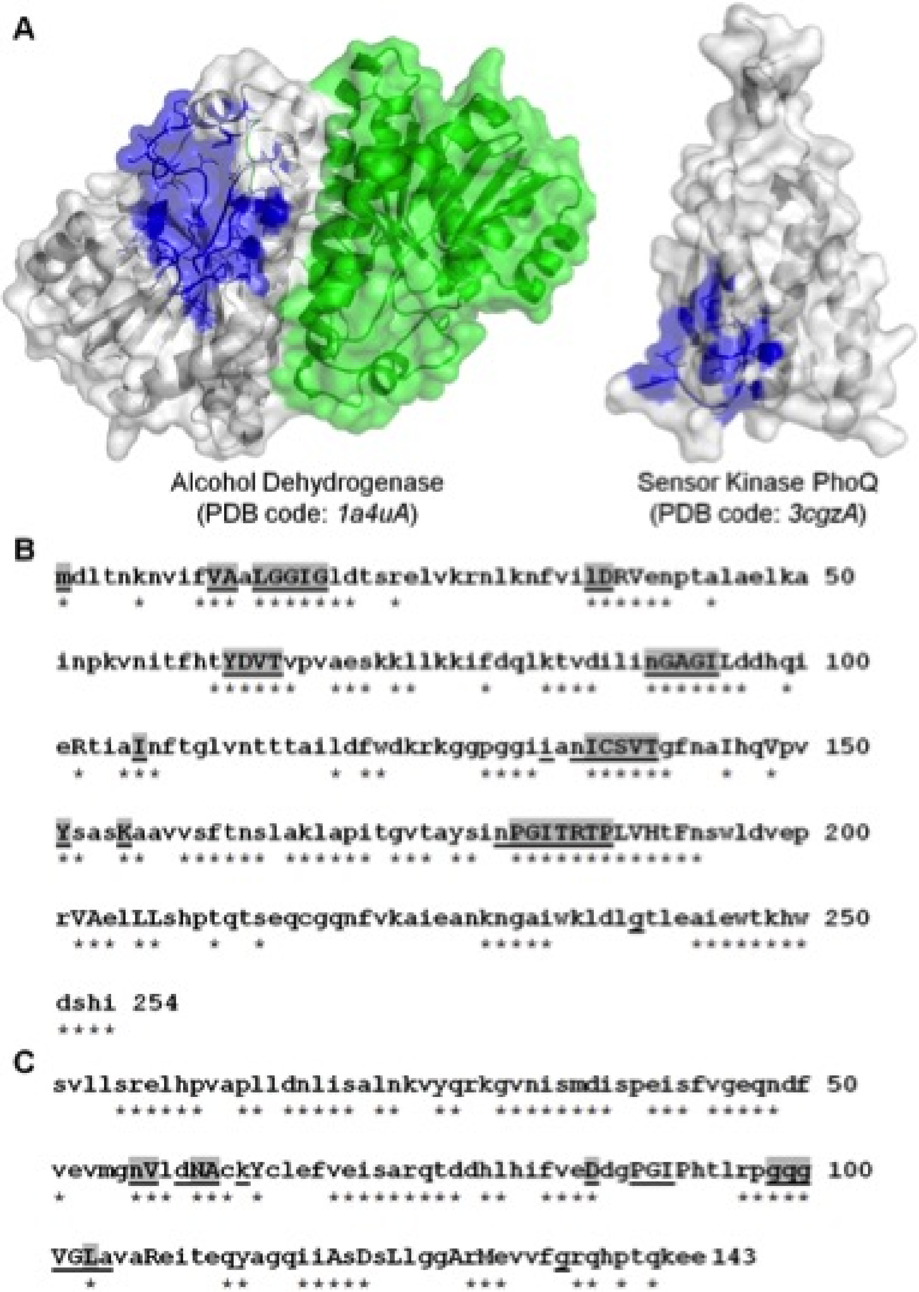

A
Alcohol Dehydrogenase
(PDB code: 1a4uA)
Sensor Kinase PhoQ
(PDB code: 3cgzA)
B
mdltnknvifVAaLGGIGldtsrelvkrnlknfviIDRVenptalaelka 50
inpkvnitfhtYDVTvpvaeskkllkkifdqlktvdilinGAGILddhqi 100
eRtiaInftglvnttttaildfwdkrkggpggiianICSVTgfnaIhqVpv 150
YsasKaavvsftnslaklapitgvtaysinPGITRTPLVHtFnswldvep 200
rVAelLLshptqtseqcgqnfvkaieankngaiwkldlgtleaiewtkhw 250
dshi 254
C
svllsrelhpvaplldnlisalnkvyqrkgvnismdispeisfvgeqndf 50
vevmgnVldNAckYclefveisarqtddhlhifveDdgPGIPhtlrpggg 100
VGLavaReiteqyagqiiAsDsLlggArMevvfgrqhptqkee 143

Figure 4.4: Case study. (A) biological assembly with highlighted ligand binding region in blue. Predictions obtained using ROBBY and MetaPocket 2.0 along with the experimental observation for (B) Alcohol Dehydrogenase (PDB code: *1a4uA*) and (C) Sensor Kinase PhoQ (PDB code: *3cgzA*). Sequence information provided with interacting residues in uppercase, ROBBY predictions underlined, MetaPocket 2.0 predictions denoted by star and their overlap is highlighted.

4.4 Conclusion

Rapidly evolving knowledge of enzyme complexes with small molecules and their potential biotechnological applications have accelerated computational characterisation of their interaction mechanism using similarity based and *de novo* methods. When as minimal information as the sequence is available, there are various challenges and concerns associated with prediction of enzyme properties or interactions. This could be because the some basic assumptions which might not hold true for all the enzymes as described and reviewed in relevant studies (Konc et al., 2015; Rost, 2002). To facilitate prediction in such a scenario, an approach called ROBBY has been presented[1]. It is based on evolutionary information in support vector machines ensemble architecture and tested for robustness. Adding domain knowledge has proven to be useful, as shown in this chapter and the previous one, for achieving enhanced results in the prediction scenario.

[1] Relevant findings: **Pai, P. P.**, Dattatreya, R.K., and Mondal, S. (2017). Ensemble Architecture for Prediction of Enzyme-Ligand Binding Residues using Evolutionary Information.*Molecular Informatics*. Doi:1002/minf.201700021.[Epub ahead of print]

Chapter 5

DORAEMON: conDitiOnal pRobabilistic Approach for idEntification of aMino acids interacting with ribONucleic acids

This chapter presents a novel perspective for sequence-based identification of residues interacting with ribonucleic acids in a non-numeric feature space. A simple-yet-efficient conditional probabilistic approach based on the local occurrence of amino acids in the interacting region is proposed for discrimination purposes.

5.1 Introduction

RNA interactions with proteins are essential for regulation of various cellular processes, such as protein synthesis, sequence encoding, RNA transfer, and gene regulation at the transcriptional and post-transcriptional levels. A variety of proteins including metabolic enzymes such as vertebrate cytoplasmic aconitase, glyceraldehyde-3-phosphate dehydrogenase, aldolase, lactate dehydrogenase *etc.*, interact with RNA (Alberts et al., 2008; Ciesla, 2006) performing vital molecular functions. Disruptions in protein-RNA interactions have known to have implications in several diseases of the central nervous system, including fronto-temporal lobar degeneration, amyotrophic lateral sclerosis and fragile X syndrome (Modic et al., 2013). In order to gain a deeper understanding of the functioning of RNA-binding proteins, their mechanisms of interaction and eventual role in development of various diseases, various experimental and computational approaches have been used in protein-centric and RNA-centric perspectives. The experimental methods that have been developed for the determination of protein-RNA interactions include techniques such as immunoprecipitation (Modic et al., 2013), RNA affinity capture methods (McHugh et al., 2014), mass spectrometry (McHugh et al., 2014), NMR (Theimer et al., 2012) and X-ray crystallography (Jones et al., 2001). The computational approaches on the other hand have applied various experimental observations determining the nature of protein-RNA interactions for prediction of RNA-binding proteins, the type of protein-RNA interactions and further, have shed light on important sites involved (Si et al., 2015a). Protein-RNA interaction information has been stored and is available in databases such as PDB (Berman et al., 2000), PRIDB (Lewis et al., 2011), RBPDB (Cook et al., 2011), NPIDB (Kirsanov et al., 2012), DOMMINO 2.0 (Kuang et al., 2012), RNAcentral (Consortium et al., 2014), RAID (Zhang et al., 2014), ATtRACT (Giudice et al., 2016), URS database (Baulin et al., 2016), RAIN (Junge et al., 2017), *etc*.

Computational approaches make use of sequence and structure information of protein-RNA complexes for prediction from low to high resolution. A low-resolution prediction is a simple two-state prediction of whether a protein is RNA binding or non-RNA binding. A medium-resolution prediction locates the interacting region of an RBP that binds to RNA (RNA binding site/residue/motif prediction). At higher resolutions prediction indicates the types of RNA binding to an RBP and other aspects at 3D structure levels. Most computational methods developed so far have focused on low to medium resolution prediction (Zhao et al., 2013a).

Several methods have been developed over decades using similarity and machine-learning methods as mentioned in a detailed review (Si et al., 2015a) for protein-RNA interaction identification. Some of them are general predictors such as RBRIdent (Xiong et al., 2015), RBRDetector (Yang et al., 2014), PRIdictor (Tuvshinjargal et al., 2016), FastRNABindR (Yasser et al., 2016), RBscore & NBench (Miao and Westhof, 2016), DRNApred (Yang et al., 2014), *etc*. Properties such as amino acid composition, sequence similarity, evolutionary information, secondary structure, accessible surface area, hydrophobicity, electrostatic patches and cleft size have been used for computational identification in different approaches so far (Miao and Westhof, 2016; Si et al., 2015a; Yasser et al., 2016). These approaches display great diversity in definition of binding sites, the data used for method development, algorithms used for matching, mapping or pattern recognition and availability of prediction programs. Owing to the importance of understanding protein-RNA interactions, despite the inherent challenges in the characteristic properties of protein-RNA interacting complexes, their identification has been widely attempted. However, recent comparative studies suggest that despite these developments, many problems are faced with respect to the usability, prerequisites, and accessibility of various tools, thereby calling for an alternative approach and perspective supplementation in the prediction scenario (Yasser et al., 2016). State-of-art sequence-based approaches use various evolutionary and biochemical properties, among them the best performing ones use PSSM and are comparable with structure-based methods (Yasser et al., 2016). Though PSSM has shown significant contributive influence in the prediction processes prediction of protein functional sites, even in the prediction of protein-nucleic acid binding residues, it may also cause the approach to suffer when there is lack of homology or limitations in terms of resources (Butenko et al., 2009). This presents the need for alternative approaches and perspective supplementation in order to be able to achieve an enhanced prediction of RNA binding regions.

Inspired by the need for exploring alternative strategies for identification of protein-RNA binding regions, ideas which were inherently competitive, widely applicable and value adding, were sought after. Based on available information, it could be inferred that since RNAs have diverse functional roles, they vary in lengths, modes of action and requirements of interaction. The local sequence occurrence plays a crucial role in determining structural requirements for the complex formation and interaction. In this chapter, a novel probabilistic approach based on this local occurrence is proposed and it is named as DORAEMON.

5.2 Materials and Methods

Unlike previous chapters in this thesis which were based on supervised machine learning techniques, in this chapter, a non-parametric approach has been used. Non-redundant datasets are created with the information of interacting residues (at various distance cut-offs defining the interaction). They are then used for optimisation and testing purposes.

5.2.1 Using local occurrence in a conditional probability-based perspective

In this section, first, the preliminary technical aspects of the probabilistic approach are described, followed by the implementation details in the protein-RNA interaction datasets used in this study.

Essentials of developing probabilistic models

An ***experiment*** is a procedure that is carried out in anticipation of the results. For example, in this work, the goal is to predict whether a residue of an RNA-binding protein is interacting, or non-interacting with a RNA. So, the testing of all possible residues in the protein is regarded as an experiment. In such a case, testing one of the residues is called a ***trial***. Finding out whether the residue subjected to trial is interacting or non-interacting (*i.e.*, generally the outcome of a trial) is called an ***event***. In a similar sense, finding the occurrence of a particular amino acid at some specific position in the protein sequence is another event. The measure of the likelihood of an event in an experiment comprising of several trials is referred to as ***probability***. For example, estimation of the probability of interaction of a residue of a protein with RNA. Probability estimation, denoted as $P(\cdot)$, gives a value in the range [0,1] where zero(0) indicates impossibility of occurrence, and one(1) indicates certainty. Two events e_i and e_j are said to be independent if occurrence of e_i does not influence occurrence of e_j . The probability of occurrence of an event e_i, given the occurrence of another event e_j is called conditional probability. Conditional probability is written as $P(e_i|e_j)$. Mathematically, it is computed using a well-known formula called Bayes' theorem (or rule):

$$P(e_i|e_j) = \frac{P(e_i, e_j)}{P(e_j)} \tag{5.1}$$

where $P(e_i, e_j)$, also written as $P(e_i \; and \; e_j)$ or $P(e_i \cap e_j)$, and is estimated as $P(e_i, e_j) = P(e_j|e_i)P(e_i)$, is the probability of both the events e_i and e_j occurring together and is called the joint probability of the events e_i and ej. It may be noted that the occurrence of an event could also be influenced by the occurrence of another set of events. In equation 5.1, the term $P(e_j|e_i)$ is called the likelihood, $P(e_i)$ is called the prior, and $P(e_i|e_j)$ is called the posterior. The Bayes' theorem has been redefined in statistical machine learning term for developing the proposed approach. Here, a data sample is represented as a vector $\mathtt{x}$. To include local occurrence of amino acids in the interacting region, a fragment is created for every residue that is to be subjected to learning and prediction. This means, a fragment is interacting if its central residue is interacting and likewise non-interacting, if its central residue is non-interacting. Equal number of residues neighbouring the central or target residue are selected in the fragment. The total length of the fragment reflects the window size used in the learning process. The posterior probability (that is, the probability which is to be observed) of a hypothesis $\mathtt{h}$ (here, a model which tells whether $\mathtt{x}$ is interacting or non-interacting), denoted as $P(\mathtt{h}|\mathtt{x})$ can be written as:

$$P(\mathtt{h}|\mathtt{x}) = \frac{P(\mathtt{x}|h)P(\mathtt{h})}{P(\mathtt{x})} \tag{5.2}$$

where $P(\mathtt{x}|h)$ (*i.e.*, the likelihood) is the probability of occurrence of a set of $x_i \in \mathtt{x}$ $(i = 1, 2, ...)$, given a hypothesis $\mathtt{h}$. The term $P(\mathtt{h})$ (*i.e.*, the prior) gives the information of occurrence of $\mathtt{h}$, and is called the prior (*i.e.*, previously observed). The denominator term $P(\mathtt{x})$ is the total probability of occurrence of $\mathtt{x}$, rather all $x_i \in \mathtt{x}$. In Bayesian inference, the occurrence of the event $\mathtt{x}$ is fixed and becomes a constant, and the remaining thing varying is only the numerator term based on the hypothesis $\mathtt{h}$. Therefore, in this case, the posterior is directly proportional to the product of likelihood and the prior. Mathematically,

$$P(\mathtt{h}|\mathtt{x}) \propto P(\mathtt{x}|\mathtt{h})P(\mathtt{h}) \tag{5.3}$$

For the studies, x implies the occurrence of a set of events with an event being the occurrence of an amino acid in a specific position of the protein sequence (specifically, the protein fragment or region under consideration). So, the fragment of the protein has length w, also referred to as window size, as mentioned briefly before. The vector term x can be represented as $\mathtt{x} = x_1, x_2, ..., x_r, ..., x_w$, where x_r is the residue of interest (*i.e.*, $r = (w+1)/2$) and each $x_i(i = 1, 2, ..., r, ..., w)$ is a residue in the protein fragment at position i. For convenience, the terms x_i and $i = 1, 2, ..., w$ are used in most of the following descriptions. Each x_i can be one(1) out of the twenty(20) possible amino acids (naturally occurring that are {'A', 'C', 'D', 'E', 'F', 'G', 'H', 'I', 'K', 'L', 'M', 'N', 'P', 'Q', 'R', 'S', 'T', 'V', 'W', 'Y'}). The probability estimate of $P(\mathtt{x}|\mathtt{h})$ can be converted to $P(x_1, x_2, ..., x_w|\mathtt{h})$ or simply $P(x_1, x_2, ..., x_w|\mathtt{h})$. This can be interpreted as the probability of occurrence of each (rather all) of these x_is given a hypothesis h. As mentioned earlier, the hypothesis h is nothing but a model which tells that the fragment is interacting (we can write this as h = true) or non-interacting (*i.e.*, h = false). So, in simple words, one could regard the term $P(x_1, x_2, ..., x_w|\mathtt{h})$ as the probability of occurrence of amino acids at each i^{th} position (*i.e.*, x_i is one of the twenty amino acid) given h is true or h is false. So, this problem reduces to the problem of estimating the probability of occurrence of amino acids at individual position given h is true or h is false which incorporates an assumption that occurrence of amino acid at some position j does not depend on the occurrence of an amino acid at present position $i(i \neq j)$. Moreover, the overall sequence of occurrences of the amino acids are independent of each other, and the occurrences that are conditional upon the hypothesis h to be true or false are therefore independent. We know that when two events e_i and e_j are independent we can write the joint probability $P(e_i, e_j) \;\; as \;\; P(e_i, e_j) = P(e_i)P(e_j)$. This implies that it also holds, given a condition. This can be devised as:

$$P(e_i, e_j|\mathtt{h}) = P(e_i|\mathtt{h})P(e_j|\mathtt{h}) \tag{5.4}$$

Upon applying the above concept in this work, the following is obtained:

$$P(x_1, x_2, ..., x_w|\mathtt{h}) = P(x_1|\mathtt{h})P(x_2|\mathtt{h})...P(x_w|\mathtt{h}) \tag{5.5}$$

or,

$$P(x_1, x_2, ..., x_w|\mathtt{h}) = \prod_{i=1}^{w} P(x_i|\mathtt{h}) \tag{5.6}$$

In this work, basically a model is being built to estimate a probability that a fragment $\mathtt{x}$ is interacting or non-interacting. Mathematically, this implies interest in the estimation of $P(\mathtt{h}|\mathtt{x})$. So, a combination of equations 5.3 and 5.6 gives us:

$$P(\mathtt{h}|\mathtt{x}) \propto P(\mathtt{h}) \prod_{i=1}^{w} P(x_i|\mathtt{h}) \tag{5.7}$$

Since this study deals with discriminating whether a given fragment is interacting or non-interacting based on the available knowledge base (facts) about previously observed characteristics, for the purpose of bias-free discrimination between these two classes, an equal cost of classification(discrimination) is considered. Hence, the proportionality constant (c) which comes into the picture in equation 5.7 is the same. Since the constant is same, the final estimate is scaled up by c. For simplicity, a unit cost is considered, *i.e.*, $c = 1$, which does not affect the final estimates for decision making purposes. Therefore, the equation 5.7 can now be written as:

$$P(\mathtt{h}|\mathtt{x}) = P(\mathtt{h}) \prod_{i=1}^{w} P(x_i|\mathtt{h}) \tag{5.8}$$

Here, the equation 5.8 essentially estimates the posterior of whether an unknown fragment, say x_unknown, is interacting or non-interacting, given the likelihood over the past fragments *i.e.*,

$P(x_{known}|\mathtt{h})$ and the prior $P(\mathtt{h})$. Thus, equation 5.8 estimates two different posterior probabilities which are $P(\mathtt{h} = true|x_{unknown})$ and $P(\mathtt{h} = false|x_{unknown})$. These two estimations are:

$$P(\mathtt{h} = true|x_{unknown}) = P(\mathtt{h} = true)\prod_{i=1}^{w} P(x_{known}i|\mathtt{h} = true) \tag{5.9}$$

and

$$P(\mathtt{h} = false|x_{unknown}) = P(\mathtt{h} = false)\prod_{i=1}^{w} P(x_{known}i|\mathtt{h} = false) \tag{5.10}$$

Based on the equations, the proposed model is implemented in the following manner.

Model implementation

A major part of the implementation of this model lies in estimation of $P(x_i|\mathtt{h})$, where $i = 1, 2, ..., w$; where h could be either true or false, for a given dataset of known samples or known protein fragments. These have contained within them the information on whether they are interacting or not. Computing the probabilities from the known dataset of such samples is quite easy as the probability is the fraction of the frequency of occurrence of an event over the total number of experiments. Here, an experiment is the set of all the known samples and event is the occurrence of an amino acid (that takes 1 out of the 20 amino acids) given whether the considered fragment is interacting (*i.e.*, true) or non- interacting (*i.e.*, false). This implementation could be clearly visualised as a directed acyclic graph (DAG), called Bayesian network (Jensen, 1996) which is depicted in Figure 5.1

5.2.2 Implementation on benchmark datasets

RNA-interacting proteins were collected from assorted datasets used for the development of various approaches available at NBench (Miao and Westhof, 2016) and another latest dataset used for the development of FastRNABindR (Yasser et al., 2016). From the collection, 122 non-redundant protein sequences were reserved for independent testing (RB122) and 3213 were reserved for cross-estimation (CE3123) which is explained in details in the forthcoming pages. Protein sequences in CE3213 were not more than 40% similar to any of the test

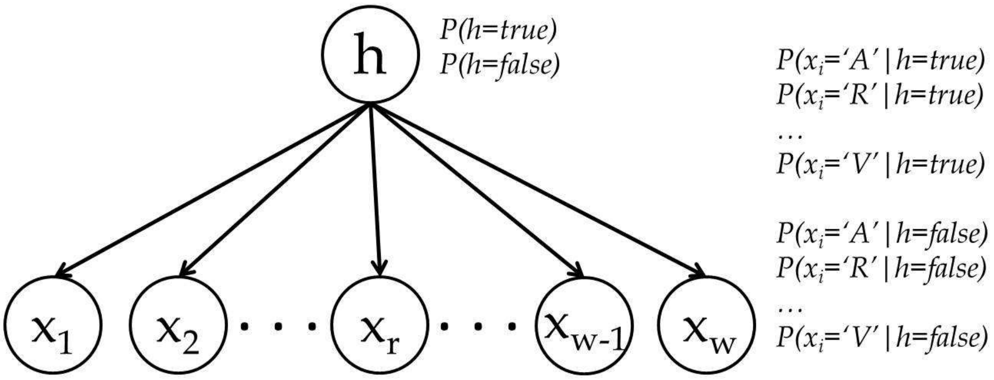

h
P(h=true)
P(h=false)
$P(x_i=\text{'A'} \mid h=true)$
$P(x_i=\text{'R'} \mid h=true)$
...
$P(x_i=\text{'V'} \mid h=true)$
$P(x_i=\text{'A'} \mid h=false)$
$P(x_i=\text{'R'} \mid h=false)$
...
$P(x_i=\text{'V'} \mid h=false)$
x_1
x_2
x_r
x_{w-1}
x_w

Figure 5.1: Bayesian network depicted for a protein fragment of length w where h is the hypothesis tested to know if the fragment is interacting or not given the probability of local occurrence of amino acids in that region.

proteins; redundancy was removed using the CD-HIT (Fu et al., 2012) software suite available at `http://weizhongli-lab.org/cdhitsuite/cgi-bin/index.cgi`. Fragments and sequences containing non-naturally occurring amino acids or other ambiguity (filters) were excluded from the study to ensure the reliability of generated information and study findings. The above implementation is known as Naïve Bayes model. For efficient implementation and access of the probability values, the conditional probabilities as shown in the Figure 5.1 are stored as matrices for each of the positions $x_i s$. Each matrix is called conditional probability matrix or conditional probability table (CPT). Each matrix in this study is of the order 20×2 and there are total w matrices. However, it just a mere representation of the probabilities in the storage space.

Besides the underlying motivation behind the work, there is also a computational aspect of observing the efficiency of the task. The computational merit of using this model over the available machine learning models is that it does not require any training whereas the machine learning models need sufficient training to generalise the situation.

5.2.3 Performance assessment

The prediction performance was assessed from various angles in this study to ensure robustness across different definitions of interacting residues and neighbourhood information. A cross-estimation procedure was used where the protein sequences reserved for exploring various ideas

(CE3123) were evenly distributed into five parts based on the interacting residue information. The number of interacting samples and non-interacting samples are denoted as n^+ and n^- respectively. For a given fold of n^+ interacting residues, $n^- = kn^+$ non-interacting residues were randomly selected without replacement from the pool of non-interacting residues, where $k \in \{1.0, 1.5, 2.0\}$. The final performance results were averaged over all the folds to gain insights in the study. Performance measures such as sensitivity (SN), precision (PR), specificity (SP), accuracy (AC), Matthew's correlation coefficient (MCC) and F-measure (FM) were calculated as described in Chapter 2.

5.3 Results and discussion

In nature, proteins are designed in such a way that they perform very specific interactions using highly specialised architecture. Residues involved in the interaction may range from few to almost all of the protein depending upon their biological roles and multitude in function. So, developing an approach that is suitable for identification of RNA-interacting residues despite the variation in their natural occurrence would make it widely applicable. Motivated by this thought, estimation of position-specific probabilities was done using balanced and imbalanced examples from the pool of RNA-interacting and non-interacting residues. These were subsequently applied for discrimination purposes.

5.3.1 Position-specific probability estimation and discrimination

Influence of balanced and imbalanced number of examples:

Results showed that upon using probabilities derived from balanced examples, the discrimination was more meaningful with an overall sensitivity and specificity of greater than 65%. On the other hand, probabilities derived from imbalanced examples led to an undesirable strong bias in the prediction scenario. This is evident because of the following fact. In a dataset of known examples containing both interacting and non-interacting types, if the number of a type of examples is increased, it means we are increasing the prior for that type *i.e.*, $P(\mathtt{h})$. So, if $P(\mathtt{h} = true) < P(\mathtt{h} = false)$, then the posterior probability estimate for the unknown residue, $P(\mathtt{h} = false|\mathtt{x}_{unknown})$, will be higher than $P(\mathtt{h} = true|\mathtt{x}_{unknown})$; and as we are considering unit cost of discrimination, each of the unknown residues shall be discriminated as non-interacting. This could be observed from the results obtained for the imbalanced combination of the examples (*i.e.*, $k = \{1.5, 2\}$). The details of the performance assessment achieved with balanced examples for a window size (w) 11, with interacting residue definition based on 5.0 Ådistance cut-off and 100% similarity cut-off for fragments are shown in Table 5.1. The results of studies performed using imbalanced examples were extremely biased towards the majority class and are hence not depicted in the table. These were used for further analysis across different definition of residues.

Table 5.1: Performance assessment measures for discrimination based on probability estimation from CE3213

$n^+ : n^-$	**SN**	**PR**	**SP**	**AC**	**MCC**	**FM**
1:1	66.4	67.3	67.7	67.0	0.341	66.8

Influence of interacting residue definition:

Residues involved in RNA-interaction have been defined using various distance cut-offs in as per earlier work summarised in a recent study (Miao and Westhof, 2016). In order to study the robustness of discrimination across these definitions (ranging from 3.5 Åto 6.0 Å), probability estimations for positions in the target residue neighbourhood were performed as shown in Table 5.2. This was done using balanced examples for a window size (w) 11 and 100% similarity cut-off for fragments. These, when used for discrimination suggested in the range of 0.334 to 0.394, for 2145 to 3378 interacting residues (n^+) respectively. The sensitivity and specificity achieved throughout was 66.0 % or more. This shows that the prediction is not biased and in characterisation of proteins, the chances that the interacting residues are correctly identified is similar to that of non-interacting residues.

Table 5.2: Performance assessment measures across various distance cut-offs (definition) for dataset CE3213

Definition	n^+	SN	PR	SP	AC	MCC	FM
3.5	2145	69.4	69.8	70.0	69.7	0.394	69.6
4.0	2565	67.7	68.2	68.5	68.1	0.362	67.9
4.5	2838	66.9	67.7	68.2	67.5	0.351	67.3
5.0	3180	66.4	67.3	67.7	67.0	0.341	66.8
5.5	3263	66.0	67.2	67.7	66.9	0.338	66.6
6.0	3378	66.0	66.9	67.4	66.7	0.334	66.4

Influence of different types and numbers of neighbours:

It is well-known that the neighbourhood of interacting residues is important for conferring upon them distinct functionally relevant properties. So, the prediction power of this probabilistic approach was explored using different numbers and types of flanking residues for target residue prediction, window ranging in between 11 and 21, keeping the interacting residue definition based on 5.0 Åcut-off and 100% similarity cut-off for fragments. Such a distance cut-off is chosen to facilitate comparative analysis with state-of-art methods. The Table 5.2 nevertheless shows that even if RNA-binding residues are defined based on other distance cut-offs (definition) this approach is still stable. Table 5.3 has a summary of study findings which show balanced performance across different numbers and types of neighbours. To ensure minimal loss of information, while applying this approach for the uncharacterised cases, a window of 11 residues was considered for further analysis. To check if similar or dissimilar neighbourhood showed better discrimination, we removed a redundancy at fragment level to about 60%. Study findings are detailed in Table 5.4, indicating that considering similarity in the neighbourhood while estimating probabilities for discrimination was more contributive. Based on these inferences, the testing was performed and analysed.

Analysis of prediction performance and large-scale applicability:

As mentioned in methodology of this chapter, a set of 122 RNA-interacting proteins were used for testing the prediction power of this approach. For each of the proteins, the effective length

Table 5.3: Performance assessment measures across various window sizes (w) for dataset CE3213

w	**SN**	**PR**	**SP**	**AC**	**MCC**	**FM**
11	66.4	67.3	67.7	67.0	0.341	66.8
13	66.3	67.6	68.2	67.3	0.345	66.9
15	66.3	67.8	68.4	67.4	0.348	67.0
17	66.1	67.2	67.8	66.9	0.339	66.6
19	65.9	67.9	68.7	67.3	0.346	66.9
21	65.9	68.0	68.8	67.4	0.348	66.9

Table 5.4: Performance assessment measures across various fragment similarities for dataset CE3213

%Cut-off	**SN**	**PR**	**SP**	**AC**	**MCC**	**FM**
100	66.4	67.3	67.7	67.0	0.341	66.8
90	65.5	67.1	67.9	66.7	0.334	66.3
80	66.0	67.1	67.6	66.8	0.336	66.5
70	65.6	66.4	66.8	67.2	0.325	66.0
60	65.3	66.6	67.3	66.3	0.326	65.9

evaluated excludes the terminal residues for window size 11, because the fragments generated were of inadequate length. The results are summarised in Table 5.5 and suggest that this approach can be contributive to the prediction scenario from perspectives alternate to the existing approaches. Meanwhile, analysis of this approach for bias towards any specific amino acid type was performed to see if certain residues which occur more in the interacting region were predicted better. Of the interacting residues, about 1259 were charged (HERKD), 947 were polar (QTSNCYW) and 1180 were hydrophobic (GFLIMAIPV) in nature. Among them, this approach was found to be relatively more precisely sensitive towards charged residues, followed by polar and hydrophobic residues. The coverage, however, was slightly better with respect to hydrophobic residues, considering they were more frequently and correctly discriminated in the non-interacting regions. However, there was no specific bias or undesirable influence towards any of these groups. The prediction was more or less balanced with around 60 ± 8 % sensitivity and specificity. Thus, this approach is hoped to add complementarily in a comprehensive understanding of RNA-interaction based on templates, machine learning, and probabilities. An example prediction is shown as follows.

Table 5.5: Independent testing of DORAEMON on RB122

Approach	**SN**	**PR**	**SP**	**AC**	**MCC**	**FM**
DORAEMON	55.4	22.4	67.1	66.7	0.158	29.5

5.3.2 Comparison with state-of-art RNA interacting residue predictors

Though a direct comparison between various approaches is not possible, in order to demonstrate the discrimination power of DORAEMON, we performed a comparative analysis on New_R15 dataset (Miao and Westhof, 2016) which is summarised in Table 5.6. The assessment measures for FastRNABindR are calculated based on predictions obtained using the web server (Yasser et al., 2016) and for other approaches, the values are based on previous study (Miao and Westhof, 2016).

Table 5.6: Comparison with state-of-art on New_R15 dataset

Approach	**SN**	**SP**	**AC**	**MCC**	**FM**
DORAEMON	60.4	65.2	65.5	0.213	37.8
FastRNABindR (Yasser et al., 2016)	53.2	78.0	74.9	0.268	42.6
RNABindR (**?**)	66.1	68.1	67.8	0.258	39.3
RBScore SVM (Miao and Westhof, 2016)	4.6	98.7	83.9	0.090	8.3
RBRIdent (Xiong et al., 2015)	16.9	95.2	86.3	0.160	21.9
PPRInt (**?**)	35.6	81.4	74.5	0.150	29.7
BindN^{+}_RNA (Wang and Brown, 2006)	37.8	83.5	76.3	0.194	33.4

Based on the performance assessment measures, it can be seen that DORAEMON has a discrimination power comparable to FastRNABindR (Yasser et al., 2016) which is a PSSM-based approach and has been proven to perform at par with relevant state-of-art structure-based methods.

5.3.3 Case study

The protein-RNA complex shown in Figure 5.2. is essentially a rescue factor YaeJ (Gagnon et al., 2012) bound to the *Thermus thermophilus* 70S ribosome in complex with the initiator tRNAfMet and a short mRNA. The chain used here is Y and it is approximately 132 amino acids in length. Excluding five residues on either side, DORAEMON showed a sensitivity of 92.6%. More specifically, DORAEMON achieved a performance of $TP = 25, FN = 2, FP = 28, and \ \ TN = 67$; at the same time, the FastRNABindR (Yasser et al., 2016) achieved a performance of $TP = 25, FN = 2, FP = 30, and \ \ TN = 65$. This suggests an encouraging capability of DORAEMON that may be further subjected to post-processing using knowledge-based filters or inclusion of structural insights.

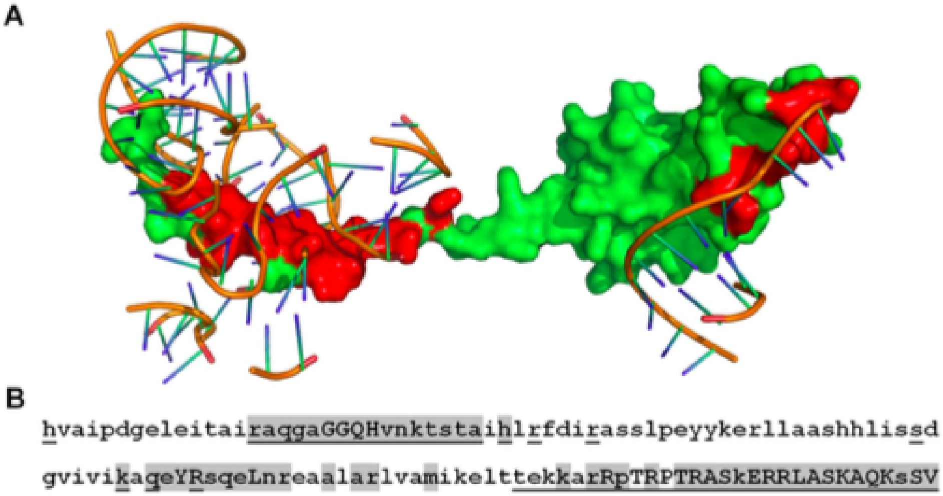

A
B
hvaipdgeleitairaqgaGGQHvnktstaihlrfdirasslpeyykerllaashhlissd
gvivikaqeYRsqeLnreaalarlvamikelttekkarRpTRPTRASkERRLASKAQKsSV

Figure 5.2: Case study. (A) The structure of YaeJ (PDB code 4dh9Y) with Y chain shown in surface and interacting RNA is shown in cartoon. (B) The sequence of YaeJ. (A and B) The interacting regions are highlighted in red in structure and upper case in sequence, whereas the non-interacting regions are marked in green in structure and lower case in sequence. DORAEMON prediction is underlined, and FastRNABindR is highlighted.

5.3.4 Software availability

The software DORAEMON along with the user manual and associated data is available at `https://github.com/ABCgrp/DORAEMON`.

5.4 Conclusion

Despite the enormous progress done in the last few decades, it is clear that both experimental and computational approaches might often need supplementation for eventually gaining a deeper understanding of protein-RNA interactions, and the research in this area is still progressing. The inherent challenges associated with the identification of protein-RNA interactions including the important residues, render the prediction scenario mired with false predictions and burden on resources. This has led to a pressing for alternate perspectives such as the one that has been proposed in this study, based on conditional probability of local amino acids occurrence in the protein interaction architecture[1]. As suggested in recent studies, scope of diminishing cross-predictions with DNA interacting residues, as well as, using combined nucleic-acid interaction data for identification purposes has also been explored. This is presented in the next chapter. Meanwhile, the findings of this study are hoped to add to a comprehensive understanding of low to high-resolution prediction scenario in protein-RNA interaction biology for further applications in therapeutics and industry.

[1] Relevant findings: **Pai, P. P.**, Dash, T., and Mondal, S. (2017). Sequence-based discrimination of protein-RNA interacting residues using a probabilistic approach. *Journal of Theoretical Biology*, 418: 77-83. `http://dx.doi.org/10.1016/j.jtbi.2017.01.040`

Chapter 6

DORAMI: conDitiOnal pRobAbility based prediction Model of protein-deoxyribonucleic acid Interacting residues

This chapter presents findings pertaining to prediction of protein-DNA interacting residues using a probabilistic approach (concept reported in the previous chapter). It addresses concerning issues related to cross-predictions and explores the scope of using combined datasets containing both protein-DNA and protein-RNA interacting residues .

6.1 Introduction

DNA, the genetic material of many living organisms, stores information for the working of almost all processes in the cell. It not only codes for RNA leading to formation of proteins, but also interacts with them, in order to facilitate various molecular functions such as replication, transcription and repair (Alberts et al., 2008). Sequence-specific DNA-binding proteins regulate gene-expression and also serve many structural and catalytic roles. Insights into how the flow of genetic information occurs can boost targeted biotechnological manipulations (Konc et al., 2015). This is particularly useful for protein regulated gene-based therapeutic strategies. Studies suggest that various experimental techniques including biochemical, genetic and crystallographic insights (Pabo and Sauer, 1984), have brought to fore, remarkable observations on transcription factor binding sites which have been deposited and maintained in repositories such as TFBSbank (Chen et al., 2017), CollecTF (Kilic et al., 2013), TFinDit (Turner et al., 2012), *etc*. Apart from the transcription factors, protein-DNA interactions have been studied in general and such information is also stored in databases, for examples in DOMMINO 2.0 (), NPIDB updated (Zanegina et al., 2016), DBBP (Park et al., 2014). As mentioned in previous chapters, correctly locating DNA-binding residues solely from protein sequences is also an important but challenging task for protein function annotations and drug discovery, especially in the post-genomic era where large volumes of protein sequences have quickly accumulated (Ofran et al., 2007). Using experimental information, computational efforts have been put forth to predict various aspects of protein-DNA interactions such as the specificity of proteins which bind to single stranded DNA or double stranded DNA (Wang et al., 2014), in elucidating mutational landscape in transcription factors (Sneha and Doss, 2017), in understanding mechanisms of their action (Dutta et al., 2016), in redesigning interfacial amino acids (Havranek et al., 2004), as well as, most widely in identifying DNA-binding candidate residues in a protein sequence (Ahmad and Sarai, 2005). Several algorithms have been designed for this purpose over the years (reviewed in detail in a recent study (Si et al., 2015b)), some of these even combined to achieve enhanced prediction scenario, such as MetaDBSite (Si et al., 2011), which integrates the prediction results from six available online web servers: DISIS (Ofran et al., 2007), DNABindR (**?**), BindN (Wang and Brown, 2006), BindN-rf (Wang et al., 2009), DP-Bind (Hwang et al., 2007) and DBS-PRED(Ahmad et al., 2004). More recent approaches such as DNABind (Liu and Hu, 2013), TargetDNA (Hu et al., 2016a), PDNA (Zhou et al., 2016), *etc* are available using sequence, structure or combination information. Conventional template- and machine learning based approaches have been rigorously investigated in protein-DNA interactions at various resolutions and has been extensively reviewed in recent studies (Kauffman and Karypis, 2012; Si et al., 2015b; Yan et al., 2016) still presenting ample for scope of improvement. These studies show that prediction approaches for DNA-interacting (RNA-interacting) residues offer comparably strong predictive performance but they are unable to properly discriminate DNA- from

RNA-binding residues. That is, the prediction scenario is often challenged in terms of false predictions. Further, they propose that development of a new generation of predictors would profit from using training data sets that combine both RNA- and DNA-binding proteins, designing new input data that specifically target either DNA- or RNA-binding residues and seeking combined prediction of DNA- and RNA-binding residues. In this chapter, studies have been aimed at exploring the scope of extending the conditional probabilistic approach for nucleic acid prediction, and in the process perform investigations for protein-DNA interactions. Issues related to cross-predictions have been addressed and the possibilities of using combined datasets for enhanced prediction is also presented.

6.2 Materials and Methods

For this study, the probabilistic approach described in the previous chapter has been used. In the following the implementation is explained.

6.2.1 Datasets

DNA binding protein benchmark datasets were created using data collected from diverse studies performed on DNA-binding proteins with interacting residue information from 3.5 to 6.0 Å. Conditional probability matrices were developed using 821 proteins (Dset821). Another set of 31 proteins (New D31) which were utilised by a recent comparative study (Miao and Westhof, 2016) for understanding the performance of various nucleic acid predictors. These proteins were reserved for testing and comparative analysis in this study. It was ensured that the Dset821 did not have any sequence more than 40% similar to the sequences in New D31 using CD-HIT Suite available at `http://weizhongli-lab.org/cdhitsuite/cgi-bin/index.cgi`.

6.2.2 Implementation of probabilistic approach and performance assessment

Basically, the Naïve Bayes model, estimates the posterior probability $P(\mathtt{h}|\mathtt{x})$ from the known prior probability $P(\mathtt{h})$ and the likelihood $P(\mathtt{x}|h)$ by considering that $P(\mathtt{h}|\mathtt{x}) \propto P(\mathtt{x}|\mathtt{h})P(\mathtt{h})$.

The proportionality constant is basically the total probability of occurrence of a protein fragment $\mathtt{x}$, $P(\mathtt{x})$ and is independent of h. So, removing it from the equation does not affect the decision. Now, since the protein fragment $\mathtt{x}$, is a vector of amino acids, that is, $\mathtt{x} = x_1, x_2, ..., x_r, ..., x_w$,the equation becomes

$$P(\mathtt{h}|\mathtt{x}) \propto P(\mathtt{h}) \prod_{i=1}^{w} P(x_i|\mathtt{h}) \tag{6.1}$$

Here, r is the residue for which a decision is desired, as to whether it is an interacting with DNA or not; w is the size of the fragment and is called the 'window size'. The symbol h signifies the hypothesis that whether a residue is interacting *i.e.*, $h = TRUE$, or non-interacting *i.e.*, $h = FALSE$. Each x_i may have one(1) out of the twenty(20) possible amino acids that are $\{'A','C','D','E','F','G','H','I','K','L','M','N','P','Q','R','S','T','V','W','Y'\}$. The equation 6.1 essentially estimates the posterior that whether an unknown fragment $x_{unknown}$ is interacting or non-interacting, given the likelihood over the known fragments $P(\mathtt{x}_{unknown}|h)$ and the prior $P(\mathtt{h})$. Therefore, equation 6.1 can be written as two different posterior probabilities which are $P(\mathtt{h} = true|\mathtt{x}_{unknown})$; $P(\mathtt{h} = false|\mathtt{x}_{unknown})$;

$$P(\mathtt{h} = true|x_{unknown}) \propto P(\mathtt{h} = true) \prod_{i=1}^{w} P(x_{known}i|\mathtt{h} = true) \tag{6.2}$$

and

$$P(\mathtt{h} = false|x_{unknown}) \propto P(\mathtt{h} = false) \prod_{i=1}^{w} P(x_{known}i|\mathtt{h} = false) \tag{6.3}$$

Based on the above implementation on datasets Dset831, various interacting residue definitions were explored. Further, studies for cross-predictions were performed using datasets used in the development of DORAEMON. Also, the scope of using combined datasets were explore. Performance assessment was done using parameters described in our previous studies.

6.3 Results and Discussion

6.3.1 DNA-binding residue prediction

Based on various available and optimisable parameters, the conditional probability matrix was generated on the earlier mentioned benchmark dataset of 821 proteins. Upon testing the developed model, now referred to as **DORAMI** (con**D**iti**O**nal p**R**ob**A**bility based prediction **M**odel of protein-deoxyribonucleic acid **I**nteracting residues), an encouraging performance was observed and is shown as follows. An example case study is shown in Figure. The state-of-art prediction approaches are diverse in many aspects and direct comparison is not possible. Testing is done on 31 proteins benchmarked for comparative analysis purpose in a latest study (Miao and Westhof, 2015) is summarised in Table 6.1 and 6.2 using various definition (Def). For certain studies, even if DNA binding residues are defined based on other distance cut-offs ranging from 3.5 Å to 5.0 Å this approach is still stable and can be seen below.

Table 6.1: Comparative analysis on New_D31 dataset

Approach	**Cut-off**	**SN**	**SP**	**AC**	**MCC**	**FM**
DORAMI	6.0	69.1	63.9	65.1	0.240	35.6
BindN^{+}_DNA (Wang et al., 2010)	6.0	42.1	90.8	85.6	0.307	38.7
DORAMI	5.5	68.5	64.6	65.6	0.241	35.2
DBS_PSSM (Ahmad and Sarai, 2005)	5.5	54.1	82.5	79.5	0.273	35.6

Table 6.2: Performance with various definition

Def (Å)	**SN**	**SP**	**AC**	**MCC**	**FM**
5.0	69.2	64.9	66.1	0.245	34.3
4.5	71.7	64.6	65.6	0.244	31.8
4.0	72.3	65.7	66.7	0.237	30.1
3.5	74.4	66.6	67.3	0.223	26.2

6.3.2 Cross-predictions

Cross predictions have been a point of concern in nucleic acid interacting residue prediction from times immemorial and it has been extensively analysed in a recent study (Yan et al., 2016).

For exploring this issues, the approach developed for identification of protein-DNA interacting residues (DORAMI) was tested upon RNA-binding protein sequences of New_R15. Similarly, the approach developed for identification of protein-RNA interacting residues (DORAEMON) was tested upon DNA-binding protein sequences of New_D31.

Table 6.3: Prediction on New_D31 by means of CPT developed using RNA binding dataset (Pai et al., 2017)

Def (Å)	SN	SP	AC	MCC	FM
5.0	67.8	65.2	66.0	0.230	35.4
	69.2	64.9	66.1	0.245	34.3
6.0	68.9	62.4	64.1	0.221	34.9
	69.1	63.9	65.1	0.240	35.6

Table 6.4: Comparative analysis of prediction on New_R15 by means of CPT developed using DNA binding dataset Dset831 and DORAEMON for the optimised definition 5.0 Å.

Approach	SN	SP	AC	MCC	FM
DORAMI	62.5	62.9	63.8	0.207	37.1
DORAEMON	60.5	65.2	65.5	0.213	37.8

Results show that DORAMI was able to predict RNA binding residues with more sensitivity. But this came at the cost of reduced specificity, reducing the overall coverage to a slight disadvantage as reflected in MCC, which reduced from 0.213 to 0.207 as shown in Table 6.3 and 6.4. Further studies on using combined datasets were performed to seek benefits from combination learning if possible.

6.3.3 Comparative analysis of prediction using combined datasets

In order to understand if a combination of DNA and RNA binding proteins based CPT would produce better predictions, as proposed for other parametric sequence based approaches (Yan et al., 2016), an exercise with the selected definition of nucleic acid binding residues to a distance cut-off of 5 Åwas performed. Results shown in Table 6.5 reflect an improvement in sensitivity of the predictors upon using a combined dataset. The MCC was similar, *i.e.* 0.233 and 0.232,

Table 6.5: Comparative analysis of prediction on New_R15 and New_D31 by means of CPT using combined datasets (selected definition 5.0 Å).

Dataset	SN	SP	AC	MCC	FM
New_R15	67.3	60.6	63.0	0.233	39.1
New_D31	70.9	62.4	64.2	0.232	33.5

for the predictors, implying similar achievement in terms of coverage. There was a difference of 5.6 % in FM in the prediction, which might have manifested on account of difference in false positives. In a nutshell, findings suggest that upon using combined datasets better sensitivity and coverage can be obtained. Nevertheless, the cost at which this comes must be also determined, and this could much depend on the study under consideration.

6.3.4 Case study

Visualising the prediction scenario just for the DNA-interacting residue prediction by DORAMI, the following case study is presented on a global regulatory protein, NoIR (Lee et al., 2014). This protein is basically a component of the transcriptional machinery for formation of nodules and symbiosis across a range of *Rhizobium*, involved in the symbiosis between rhizobial microbes and host plants, which leads to nodule formation and nitrogen fixation. Despite the false positives, DORAMI predictions suggest that the number of interacting residues predicted correctly is high, and those missed is low, as indicated in green in Figure 6.1, leading to high sensitivity. Large parts of the non-interacting residues, in grey, are also predicted correctly, showing the scope of discrimination and coverage.

6.3.5 Software availability

The software **DORAMI** along with the user manual and associated data is available at `https://github.com/ABCgrp/DORAMI`.

6.4 Conclusion

There has been a considerable rise in protein sequence and structure information leading to a pressing need for rapid functional annotation to boost protein-based therapeutic and industrial

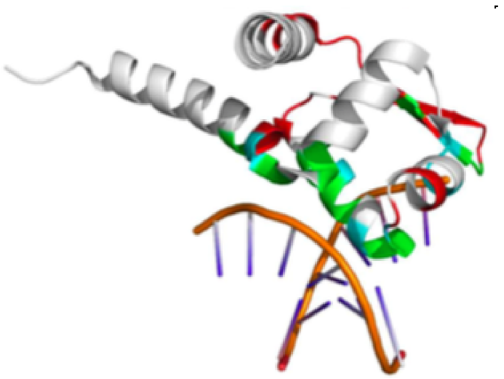


Figure 6.1: Case study: Crystal Structure of regulatory protein NolR from *Sinorhizobium fredii* in complex with DNA (Lee et al., 2014) Color code: TP=Green, FN=Cyan, FP=Red, TN=Grey

applications. Despite several developments in computational approaches, ensuring widely applicable predictors that do not depend heavily on protein homology, or existing templates has ample scope of research. The findings of this study[1] promisingly show that sequence-based predictions can be obtained in a non-numeric feature space, yet showing comparable performance to the state-of-art sequence based methods. This is value-adding to the comprehensive scenario for better understanding of the protein-DNA interaction mechanisms.

[1] Relevant findings: **Pai, P. P.**, Dash, T., and Mondal, S. (Manuscript in preparation 2017). DORAMI: conDitiOnal pRobAbility based prediction Model of protein-deoxyribonucleic acid Interacting residues.

Chapter 7

Conclusion

The human mind has been limitlessly fascinated by how things came into existence, what are they made of, how they evolve, how they work, what happens if some of their aspects are changed and how they can be manipulated advantageously. This has led to a huge range of discoveries, right from understanding the details of the essential molecules of life to its complex manifestations in nature. This thesis describes a journey of learning from, through and about one of the fundamental aspects orchestrating various processes leading to life sustenance. If we look at it, life sustenance is one of the most simply obvious, macro and ubiquitous phenomenon. As much ironically, it is also, a complexly-intertwined and balanced interplay of interactions among various biological molecules, *or* biomolecules, at much microscopic levels. Needless to say, gaining insights in this direction is indeed - a challenging affair. Innately attracted towards exploring the unexplored, researchers have achieved several milestones over years, in understanding these biomolecular interactions and their potential implications in health and diseases. Glimpses of this journey can be seen here in the thesis chapters, each being an elucidation of some of the widely known challenges in this regard, with important biomedical implications and industrial applications, with a focus on protein interactions with ligands .

Protein interactions with ligands are crucial for orchestrating various biochemical processes in living organisms. Disruptions in these interactions can lead to undesirable loss or gain in protein functions, which could eventually manifest as diseases (Goldberg, 1992). Therefore, gaining insights into how these interactions occur may aid in developing a deeper understanding of protein activities and their biomedical implications. Several experimental and computational attempts have been made to understand protein-ligand interactions over the years at various resolutions, from identifying interacting partners to learning about the involved atomic contacts in the structures of complexes. With the advent of high-throughput technologies, there has been an increasing availability of protein sequence and structure information. Consequently, the demand of rapid protein characterisation for functional annotation has also risen, leading to a pressing need for widespread computational efforts.

Computational identification relies broadly on template-based similarity transfer or pattern recognition-based *de novo* methods. Each of these strategies have their own advantages and limitations in the identification process (Roche et al., 2015). The inherent biological diversities based on functional requirements make the scenario very complex and challenging, especially at sequence levels (Gutteridge and Thornton, 2005; Rost, 2002). Therefore, multidirectional approaches are required for obtaining a comprehensive understanding of protein interaction biology. Despite several attempts to achieve accurate predictions (Du et al., 2016; Si et al., 2015a,b; Yugandhar and Gromiha, 2017; Zhang et al., 2009), there is still ample scope of improvement for real-time applications (Pegg et al., 2006; Yan et al., 2016; Zhao et al., 2013a), for which multiple novel perspectives are presented here.

Essentially, this thesis has attempts to understand and address various challenges associated with sequence-based computational identification of protein-ligand interacting residues, which is an important step towards understanding protein function and mechanism of action as summarised in ***Chapter one***. It focuses on the key issues associated with the diversity in the nature and frequency of interacting residues which often adversely manifest in the prediction scenario, through various studies. These include exploring the scope of using novel perspectives in supervised machine learning and non-parametric approaches through feature selection, application of post-processing, ensemble architecture, prediction combination with multiple computational insights.

The study findings in ***Chapter two*** suggest the use of support vector machines based architecture using selected evolutionary and biochemical information for obtaining *de novo* prediction with high sensitivity, which is required for scantily occurring interacting residues such as those involved in catalysis. Although mired with false positives, this helps picking of many catalytic residues in the prediction pool, which can be then filtered using structural inputs or sequence-based predicted insights. Since catalytic residues are subsets of ligand binding residues in proteins, by reducing the search space, the overall prediction performance can be improved (Pai et al., 2015). Another strategy reported in this thesis for enhancing prediction performance concerns the use of ensemble or consensus in prediction. Findings of ***Chapter three*** and ***Chapter four*** demonstrate how neighbourhood information can be used to obtain consensus information for improved learning in populations of interacting residues, where the contrast between interacting and non-interacting regions is not very discrete. In proteins which allow multiple types of ligand binding in same or different sites, use of different type and number of non-interacting instances for a set of interacting instances shows considerable improvement in method precision. Further, domain knowledge or more specific information about the protein can be applied to achieve residues of more relevance. An illustration of how structural insights such as pocket information and template-based information can be used for obtaining a comprehensive understanding of the protein interactions is shown (Pai and Mondal, 2016). This method is suitable for identification of ligand interacting residues where their occurrence is not very scanty or limited. From here, explorations for diverse populations of interacting residues which occur in case of ligands such as nucleic acids reveal the promising potential of non-parametric approaches in the prediction scenario using local occurrence of amino acids in a conditional probability perspective, have been made. Study findings in ***Chapter five*** and ***Chapter six*** suggest the developed probabilistic approach to be as competitive as PSSM-based approaches which show comparable performance with structure-based approaches in this context (Pai et al., 2017) and further, reinstates the fact that next generation approaches for protein-nucleic acid interactions should have commonalities and a combined predictor may be a good address to the prediction scenario.

Study findings altogether present multi-directional efforts towards gaining a comprehensive understanding of how efficiently available information can be used to achieve enhanced predictions of protein-ligand interacting residues. It gives examples of staying focussed on objectives despite deterring challenges to achieve a larger goal, by seeking help of knowledge guided filters, utilising negative examples for improving outcomes, relying on history and evolution but not so much and to go beyond convention and explore possibilities as much as possible. Though the scope of research in this field has been explored widely (Pai and Mondal, 2017), providing the scientific community many slants on how proteins perform their function, it is not surprising that the field is still fast-growing and would continue to do so, considering the immense importance of relevant knowledge[1]. This thesis is an effort towards this direction and is hoped to present alternate perspectives for achieving enhanced results, in almost the same circumstances despite the limitations in available resources, for a common goal, with a hope to be a part of studies targeted at protein design applications for therapeutics and industrial production.

[1] Relevant examples and further details: **Pai, P. P.** and Mondal, S. (2017). Applying Knowledge of Enzyme Biochemistry in Prediction of Functional Sites for Aiding Drug Discovery.*Current Topics in Medicinal Chemistry*, 17: Epub Ahead of Print. `doi:10.2174/1568026617666170329153858`

Future scope and directions

Proteins have been under the focus of scientific studies since many years. Because of their crucial roles in cellular processes and consequently, biomedical implications such as in diseases, plenty of efforts have been put forth for their functional annotation. This thesis presents perspectives for identification of protein-ligand interacting residues using feature selection, domain knowledge guided post-processing, ensemble architecture and non-parametric methods. However, research has been growing in this field. And, following future directions may arise based on current findings presented:

- *Use of modelled structure* The current results of PINGU are promising as far as sequence-based approaches are concerned. However, it might be useful to explore the scope of using modelled structure information for a prediction improvement and better understanding of the catalytic architecture.

- *Extension to other proteins and identification of sub-sites* MOWGLI identifies mannose and its variants, whereas, ROBBY identifies different possible ligand interacting residues in the proteins. This could be extended to non-enzymes and scope of allocating specific class of ligands to the overall scenario may be investigated further. The chances of enhancing the prediction scenario with template-based insights may be explored.

- *Inclusion of global information* Currently, DORAEMON and DORAMI utilise the local occurrence of the amino acids in the interacting region. It may be useful to investigate the impact of adding global features.

References

Agarwal, S., Mishra, N. K., Singh, H., and Raghava, G. P. (2011). Identification of mannose interacting residues using local composition. *PLoS One*, 6(9):e24039.

Agostino, M., Mancera, R. L., Ramsland, P. A., and Yuriev, E. (2013). Automap: A tool for analyzing protein–ligand recognition using multiple ligand binding modes. *Journal of Molecular Graphics and Modelling*, 40:80–90.

Ahmad, S., Gromiha, M. M., and Sarai, A. (2004). Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. *Bioinformatics*, 20(4):477–486.

Ahmad, S. and Sarai, A. (2005). Pssm-based prediction of dna binding sites in proteins. *BMC bioinformatics*, 6(1):33.

Alberts, B. J., Lewis, A., Raff, J., et al. (2008). *Molecular biology of the cell.* Number 574.87 M6/2008. Garland Science.

Alcalde, M. (2017). *Directed Enzyme Evolution: Advances and Applications.* Springer.

Aleshin, A. E., Stoffer, B., Firsov, L. M., Svensson, B., and Honzatko, R. B. (1996). Crystallographic complexes of glucoamylase with maltooligosaccharide analogs: Relationship of stereochemical distortions at the nonreducing end to the catalytic mechanism? *Biochemistry*, 35(25):8319–8328.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

Anderson, D. P., Whitney, D. S., Hanson-Smith, V., Woznica, A., Campodonico-Burnett, W., Volkman, B. F., King, N., Thornton, J. W., and Prehoda, K. E. (2016). Evolution of an ancient protein function involved in organized multicellularity in animals. *Elife*, 5:e10147.

Ardenkjaer-Larsen, J.-H., Boebinger, G. S., Comment, A., Duckett, S., Edison, A. S., Engelke, F., Griesinger, C., Griffin, R. G., Hilty, C., Maeda, H., et al. (2015). Facing and overcoming sensitivity challenges in biomolecular nmr spectroscopy. *Angewandte Chemie International Edition*, 54(32):9162–9185.

Arenas-Salinas, M., Ortega-Salazar, S., Gonzales-Nilo, F., Pohl, E., Holmes, D. S., and Quatrini, R. (2014). Afal: a web service for profiling amino acids surrounding ligands in proteins. *Journal of computer-aided molecular design*, 28(11):1069–1076.

Atchley, W. R., Zhao, J., Fernandes, A. D., and Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the United States of America*, 102(18):6395–6400.

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424.

Bartlett, G. J., Porter, C. T., Borkakoti, N., and Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *Journal of molecular biology*, 324(1):105–121.

Bate, P. and Warwicker, J. (2004). Enzyme/non-enzyme discrimination and prediction of enzyme active site location using charge-based methods. *Journal of molecular biology*, 340(2):263–276.

Baulin, E., Yacovlev, V., Khachko, D., Spirin, S., and Roytberg, M. (2016). Urs database: universe of rna structures and their motifs. *Database*, 2016:baw085.

Ben-Shimon, A. and Eisenstein, M. (2005). Looking at enzymes from the inside out: The proximity of catalytic residues to the molecular centroid can be used for detection of active sites and enzyme–ligand interfaces. *Journal of molecular biology*, 351(2):309–326.

Benach, J., Atrian, S., Gonzalez-Duarte, R., and Ladenstein, R. (1998). The refined crystal structure of drosophila lebanonensis alcohol dehydrogenase at 1.9 å resolution. *Journal of molecular biology*, 282(2):383–399.

Berg, J. M., Stryer, L., and Tymoczko, J. L. (2015). *Stryer Biochemie*. Springer-Verlag.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

Bermingham, M. L., Pong-Wong, R., Spiliopoulou, A., Hayward, C., Rudan, I., Campbell, H., Wright, A. F., Wilson, J. F., Agakov, F., Navarro, P., et al. (2015). Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific reports*, 5.

Blikstad, C. and Ivarsson, Y. (2015). High-throughput methods for identification of protein-protein interactions involving short linear motifs. *Cell Communication and Signaling*, 13(1):38.

Boraston, A. B., Bolam, D. N., Gilbert, H. J., and Davies, G. J. (2004). Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *Biochemical Journal*, 382(3):769–781.

Bradford, J. R. and Westhead, D. R. (2005). Improved prediction of protein–protein binding sites using a support vector machines approach. *Bioinformatics*, 21(8):1487–1494.

Broomhead, D. S. and Lowe, D. (1988). Radial basis functions, multi-variable functional interpolation and adaptive networks. Technical report, DTIC Document.

Bruha, I. (2001). Pre-and post-processing in machine learning and data mining. In *Machine Learning and Its Applications*, pages 258–266. Springer.

Brylinski, M. and Feinstein, W. P. (2013). efindsite: improved prediction of ligand binding sites in protein models using meta-threading, machine learning and auxiliary ligands. *Journal of computer-aided molecular design*, 27(6):551–567.

Buljan, M. and Bateman, A. (2009). The evolution of protein domain families.

Butenko, S., Chaovalitwongse, W. A., and Pardalos, P. M. (2009). *Clustering challenges in biological networks*. World Scientific.

Cai, Y.-D., Liu, X.-J., Xu, X.-b., and Chou, K.-C. (2002). Prediction of protein structural classes by support vector machines. *Computers & chemistry*, 26(3):293–296.

Cai, Y.-D., Zhou, G.-P., and Chou, K.-C. (2003). Support vector machines for predicting membrane protein types by using functional domain composition. *Biophysical journal*, 84(5):3257–3263.

Capra, J. A., Laskowski, R. A., Thornton, J. M., Singh, M., and Funkhouser, T. A. (2009). Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3d structure. *PLoS Comput Biol*, 5(12):e1000585.

Chai, H., Zhang, J., Yang, G., and Ma, Z. (2016). An evolution-based dna-binding residue predictor using a dynamic query-driven learning scheme. *Molecular BioSystems*, 12(12):3643–3650.

Chakrabarti, S. and Lanczycki, C. J. (2007). Analysis and prediction of functionally important sites in proteins. *Protein Science*, 16(1):4–13.

Chen, D., Jiang, S., Ma, X., and Li, F. (2017). Tfbsbank: a platform to dissect the big data of protein–dna interaction in human and model species. *Nucleic Acids Research*, 45(D1):D151–D157.

Chen, P., Hu, S., Zhang, J., Gao, X., Li, J., Xia, J., and Wang, B. (2016). A sequence-based dynamic ensemble learning system for protein ligand-binding site prediction. *IEEE/ACM transactions on computational biology and bioinformatics*, 13(5):901–912.

Chen, P., Huang, J. Z., and Gao, X. (2014). Ligandrfs: random forest ensemble to identify ligand-binding residues from sequence information alone. *BMC bioinformatics*, 15(15):S4.

Chen, S.-A., Ou, Y.-Y., and Gromiha, M. M. (2010). Topology prediction of α-helical and β-barrel transmembrane proteins using rbf networks. In *International Conference on Intelligent Computing*, pages 642–649. Springer.

Chen YW, L. C. (2006). *Feature Extraction*, volume 207. Springer, Berlin Heidelberg.

Chien, Y.-T. and Huang, S.-W. (2012). Accurate prediction of protein catalytic residues by side chain orientation and residue contact density. *PLoS One*, 7(10):e47951.

Choi, I.-G. and Kim, S.-H. (2006). Evolution of protein structural classes and protein sequence families. *Proceedings of the National Academy of Sciences*, 103(38):14056–14061.

Choi, K. and Kim, S. (2011). Sequence-based enzyme catalytic domain prediction using clustering and aggregated mutual information content. *Journal of bioinformatics and computational biology*, 9(05):597–611.

Chou, K.-C. (2011). Some remarks on protein attribute prediction and pseudo amino acid composition. *Journal of theoretical biology*, 273(1):236–247.

Chou, K.-C. and Cai, Y.-D. (2002). Using functional domain composition and support vector machines for prediction of protein subcellular location. *Journal of Biological Chemistry*, 277(48):45765–45769.

Chou, K.-C. and Cai, Y.-d. (2004). A novel approach to predict active sites of enzyme molecules. *PROTEINS: Structure, Function, and Bioinformatics*, 55(1):77–82.

Ciesla, J. (2006). Metabolic enzymes that bind rna: yet another level of cellular regulatory network? *Acta Biochimica Polonica*, 53(1):11–32.

Consortium, R. et al. (2014). Rnacentral: an international database of ncrna sequences. *Nucleic acids research*, page gku991.

Cook, K. B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T. R. (2011). Rbpdb: a database of rna-binding specificities. *Nucleic acids research*, 39(suppl 1):D301–D308.

Darnell, J. E., Lodish, H., Baltimore, D., et al. (1990). *Molecular cell biology*, volume 2. Scientific American Books New York.

De Schutter, K. and Van Damme, E. J. (2015). Protein-carbohydrate interactions, and beyond?

del Sol, A., Fujihashi, H., Amoros, D., and Nussinov, R. (2006). Residue centrality, functionally important residues, and active site shape: Analysis of enzyme and non-enzyme families. *Protein Science*, 15(9):2120–2128.

Dessailly, B. H., Lensink, M. F., Orengo, C. A., and Wodak, S. J. (2008). Ligasite–a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic acids research*, 36(suppl 1):D667–D673.

Ding, C. H. and Dubchak, I. (2001). Multi-class protein fold recognition using support vector machines and neural networks. *Bioinformatics*, 17(4):349–358.

Dinglasan, R. R. and Jacobs-Lorena, M. (2005). Insight into a conserved lifestyle: protein-carbohydrate adhesion strategies of vector-borne pathogens. *Infection and immunity*, 73(12):7797–7807.

Dou, Y., Wang, J., Yang, J., and Zhang, C. (2012). L1pred: a sequence-based prediction tool for catalytic residues in enzymes with the l1-logreg classifier. *PloS one*, 7(4):e35666.

Du, X., Li, Y., Xia, Y.-L., Ai, S.-M., Liang, J., Sang, P., Ji, X.-L., and Liu, S.-Q. (2016). Insights into protein–ligand interactions: mechanisms, models, and methods. *International journal of molecular sciences*, 17(2):144.

Dutta, S., Madan, S., and Sundar, D. (2016). Exploiting the recognition code for elucidating the mechanism of zinc finger protein-dna interactions. *BMC Genomics*, 17(13):109.

El Dib, R. P., Nascimento, P., and Pastores, G. M. (2013). Enzyme replacement therapy for anderson-fabry disease. *The Cochrane Library*.

Fajardo, J. E. and Fiser, A. (2013). Protein structure based prediction of catalytic residues. *BMC bioinformatics*, 14(1):63.

Fang, C., Noguchi, T., and Yamana, H. (2013). Scpssmpred: A general sequence-based method for ligand-binding site prediction. *IPSJ Transactions on Bioinformatics*, 6:35–42.

Fernandes, C. L., Escouto, G. B., and Verli, H. (2014). Structural glycobiology of heparinase ii from pedobacter heparinus. *Journal of Biomolecular Structure and Dynamics*, 32(7):1092–1102.

Fischer, M., Kang, M., and Brindle, N. P. (2016). Using experimental evolution to probe molecular mechanisms of protein function. *Protein Science*, 25(2):352–359.

Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K. B., Bairoch, A., Schomburg, D., Tipton, K. F., and Apweiler, R. (2004). Intenz, the integrated relational enzyme database. *Nucleic acids research*, 32(suppl 1):D434–D437.

Flores, D. I., Sotelo-Mundo, R. R., and Brizuela, C. A. (2014). A simple extension to the cmasa method for the prediction of catalytic residues in the presence of single point mutations. *PloS* *one*, 9(9):e108513.

Fu, H., Subramanian, R. R., and Masters, S. C. (2000). 14-3-3 proteins: structure, function, and regulation. *Annual review of pharmacology and toxicology*, 40(1):617–647.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152.

Fukuhara, N. and Kawabata, T. (2008). Homcos: a server to predict interacting protein pairs and interacting sites by homology modeling of complex structures. *Nucleic acids research*, 36(suppl 2):W185–W189.

Furnham, N., Holliday, G. L., de Beer, T. A., Jacobsen, J. O., Pearson, W. R., and Thornton, J. M. (2014). The catalytic site atlas 2.0: cataloging catalytic sites and residues identified in enzymes. *Nucleic acids research*, 42(D1):D485–D489.

Gagnon, M. G., Seetharaman, S. V., Bulkley, D., and Steitz, T. A. (2012). Structural basis for the rescue of stalled ribosomes: structure of yaej bound to the ribosome. *Science*, 335(6074):1370–1372.

Galperin, M. Y. and Koonin, E. V. (2012). Divergence and convergence in enzyme evolution. *Journal of Biological Chemistry*, 287(1):21–28.

Galperin, M. Y., Walker, D. R., and Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Research*, 8(8):779–790.

Gao, Y.-F., Li, B.-Q., Cai, Y.-D., Feng, K.-Y., Li, Z.-D., and Jiang, Y. (2013). Prediction of active sites of enzymes by maximum relevance minimum redundancy (mrmr) feature selection. *Molecular BioSystems*, 9(1):61–69.

George, R. A., Spriggs, R. V., Bartlett, G. J., Gutteridge, A., MacArthur, M. W., Porter, C. T., Al-Lazikani, B., Thornton, J. M., and Swindells, M. B. (2005). Effective function annotation through catalytic residue conservation. *Proceedings of the National Academy of Sciences of the United States of America*, 102(35):12299–12304.

Giudice, G., Sanchez-Cabo, F., Torroja, C., and Lara-Pezzi, E. (2016). Attract–a database of rna-binding proteins and associated motifs. *Database*, 2016:baw035.

Goldberg, D. M. (1992). Enzymes as agents for the treatment of disease. *Clinica chimica acta*, 206(1-2):45–76.

Gonzalez, A. J., Liao, L., and Wu, C. H. (2012). Predicting ligand binding residues and functional sites using multipositional correlations with graph theoretic clustering and kernel cca. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 9(4):992–1001.

Gonzalez, M. W. and Kann, M. G. (2012). Protein interactions and disease. *PLoS Comput Biol*, 8(12):e1002819.

Greener, J. G. and Sternberg, M. J. (2015). Allopred: prediction of allosteric pockets on proteins using normal mode perturbation analysis. *BMC bioinformatics*, 16(1):335.

Gromiha, M. M., Veluraja, K., and Fukui, K. (2014). Identification and analysis of binding site residues in proteincarbohydrate complexes using energy based approach. *Protein and peptide letters*, 21(8):799–807.

Guarnieri, M. T., Zhang, L., Shen, J., and Zhao, R. (2008). The hsp90 inhibitor radicicol interacts with the atp-binding pocket of bacterial sensor kinase phoq. *Journal of molecular biology*, 379(1):82–93.

Gutteridge, A. and Thornton, J. M. (2005). Understanding nature's catalytic toolkit. *Trends in biochemical sciences*, 30(11):622–629.

Harayama, S., Kok, M., and Neidle, E. (1992). Functional and evolutionary relationships among diverse oxygenases. *Annual Reviews in Microbiology*, 46(1):565–601.

Havranek, J. J., Duarte, C. M., and Baker, D. (2004). A simple physical model for the prediction and design of protein–dna interactions. *Journal of molecular biology*, 344(1):59–70.

Hedstrom, L. (2002). Serine protease mechanism and specificity. *Chemical reviews*, 102(12):4501–4524.

Henikoff, S. and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919.

Heo, L., Lee, H., Baek, M., and Seok, C. (2016). Binding site prediction of proteins with organic compounds or peptides using galaxy web servers. *Computational Design of Ligand Binding Proteins*, pages 33–45.

Heo, L., Shin, W.-H., Lee, M. S., and Seok, C. (2014). Galaxysite: ligand-binding-site prediction by using molecular docking. *Nucleic acids research*, 42(W1):W210–W214.

Hoffer, L. and Horvath, D. (2012). S4mple–sampler for multiple protein–ligand entities: Simultaneous docking of several entities. *Journal of chemical information and modeling*, 53(1):88–102.

Holliday, G. L., Almonacid, D. E., Bartlett, G. J., O'boyle, N. M., Torrance, J. W., Murray-Rust, P., Mitchell, J. B., and Thornton, J. M. (2007). Macie (mechanism, annotation and classification in enzymes): novel tools for searching catalytic mechanisms. *Nucleic acids research*, 35(suppl 1):D515–D520.

Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O'boyle, N. M., Murray-Rust, P., Thornton, J. M., and Mitchell, J. B. (2005). Macie: a database of enzyme reaction mechanisms. *Bioinformatics*, 21(23):4315–4316.

Hu, B., Zhu, X., Monroe, L., Bures, M. G., and Kihara, D. (2014). Pl-patchsurfer: a novel molecular local surface-based method for exploring protein-ligand interactions. *International journal of molecular sciences*, 15(9):15122–15145.

Hu, J., Li, Y., Zhang, M., Yang, X., Shen, H.-B., and Yu, D.-J. (2016a). Predicting protein-dna binding residues by weightedly combining sequence-based features and boosting multiple svms. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*.

Hu, X., Wang, K., and Dong, Q. (2016b). Protein ligand-specific binding residue predictions by an ensemble classifier. *BMC bioinformatics*, 17(1):470.

Hwang, S., Gou, Z., and Kuznetsov, I. B. (2007). Dp-bind: a web server for sequence-based prediction of dna-binding residues in dna-binding proteins. *Bioinformatics*, 23(5):634–636.

Izidoro, S. C., de Melo-Minardi, R. C., and Pappa, G. L. (2014). Gass: identifying enzyme active sites with genetic algorithms. *Bioinformatics*, page btu746.

Jacobson, M. P., Kalyanaraman, C., Zhao, S., and Tian, B. (2014). Leveraging structure for enzyme function prediction: methods, opportunities, and challenges. *Trends in biochemical sciences*, 39(8):363–371.

Janda, J.-O., Meier, A., and Merkl, R. (2013). Clips-4d: a classifier that distinguishes structurally and functionally important residue-positions based on sequence and 3d data. *Bioinformatics*, 29(23):3029–3035.

Jensen, F. V. (1996). *An introduction to Bayesian networks*, volume 210. UCL press London.

Jensen, R. A. (1976). Enzyme recruitment in evolution of new function. *Annual Reviews in Microbiology*, 30(1):409–425.

Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M., and Thornton, J. M. (2001). Protein–rna interactions: a structural analysis. *Nucleic acids research*, 29(4):943–954.

Junge, A., Refsgaard, J. C., Garde, C., Pan, X., Santos, A., Alkan, F., Anthon, C., von Mering, C., Workman, C. T., Jensen, L. J., et al. (2017). Rain: Rna–protein association and interaction networks. *Database*, 2017(1).

Kantardjiev, A. A. (2012). Quantum. ligand. dock: protein–ligand docking with quantum entanglement refinement on a gpu system. *Nucleic acids research*, 40(W1):W415–W422.

Kastritis, P. L. and Bonvin, A. M. (2013). On the binding affinity of macromolecular interactions: daring to ask why proteins interact. *Journal of The Royal Society Interface*, 10(79):20120835.

Kauffman, C. and Karypis, G. (2012). Computational tools for protein–dna interactions. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(1):14–28.

Kern, C., Gonzalez, A. J., Liao, L., and Vijay-Shanker, K. (2013). Predicting interacting residues using long-distance information and novel decoding in hidden markov models. *IEEE transactions on nanobioscience*, 12(3):158–164.

Khare, H., Ratnaparkhi, V., Chavan, S., and Jayraman, V. (2012). Prediction of protein-mannose binding sites using random forest. *Bioinformation*, 8(24):1202.

Khazanov, N. A. and Carlson, H. A. (2013). Exploring the composition of protein-ligand binding sites on a large scale. *PLoS Comput Biol*, 9(11):e1003321.

Kidera, A., Konishi, Y., Oka, M., Ooi, T., and Scheraga, H. A. (1985). Statistical analysis of the physical properties of the 20 naturally occurring amino acids. *Journal of Protein Chemistry*, 4(1):23–55.

Kilic, S., White, E. R., Sagitova, D. M., Cornish, J. P., and Erill, I. (2013). Collectf: a database of experimentally validated transcription factor-binding sites in bacteria. *Nucleic acids research*, page gkt1123.

Kinch, L. N. and Grishin, N. V. (2002). Evolution of protein structures and functions. *Current opinion in structural biology*, 12(3):400–408.

Kirsanov, D. D., Zanegina, O. N., Aksianov, E. A., Spirin, S. A., Karyagina, A. S., and Alexeevski, A. V. (2012). Npidb: nucleic acid-protein interaction database. *Nucleic acids research*, page gks1199.

Koch, A., Melbye, M., Sørensen, P., Homøe, P., Madsen, H. O., Mølbak, K., Hansen, C. H., Andersen, L. H., Hahn, G. W., and Garred, P. (2001). Acute respiratory tract infections and mannose-binding lectin insufficiency during early childhood. *Jama*, 285(10):1316–1321.

Koh, K., Kim, S.-J., and Boyd, S. (2007). An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine learning research*, 8(Jul):1519–1555.

Konc, J. and Janezic, D. (2007). Protein- protein binding-sites prediction by protein surface structure conservation. *Journal of chemical information and modeling*, 47(3):940–944.

Konc, J., Lesnik, S., and Janezic, D. (2015). Modeling enzyme-ligand binding in drug discovery. *Journal of cheminformatics*, 7(1):48.

Krivak, R. and Hoksza, D. (2015). Improving protein-ligand binding site prediction accuracy by classification of inner pocket points using local features. *Journal of cheminformatics*, 7(1):12.

Kuang, X., Han, J. G., Zhao, N., Pang, B., Shyu, C.-R., and Korkin, D. (2012). Dommino: a database of macromolecular interactions. *Nucleic acids research*, 40(D1):D501–D506.

Kumar Kandaswamy, K., Pugalenthi, G., Moller, S., Hartmann, E., Uwe Kalies, K., N Suganthan, P., and Martinetz, T. (2010). Prediction of apoptosis protein locations with genetic algorithms and support vector machines through a new mode of pseudo amino acid composition. *Protein and Peptide Letters*, 17(12):1473–1479.

Leahy, J. L. (2005). Pathogenesis of type 2 diabetes mellitus. *Archives of medical research*, 36(3):197–209.

Lee, S. G., Krishnan, H. B., and Jez, J. M. (2014). Structural basis for regulation of rhizobial nodulation and symbiosis gene expression by the regulatory protein nolr. *Proceedings of the National Academy of Sciences*, 111(17):6509–6514.

Lewis, B. A., Walia, R. R., Terribilini, M., Ferguson, J., Zheng, C., Honavar, V., and Dobbs, D. (2011). Pridb: a protein–rna interface database. *Nucleic acids research*, 39(suppl 1):D277–D282.

Lichtarge, O., Bourne, H. R., and Cohen, F. E. (1996). An evolutionary trace method defines binding surfaces common to protein families. *Journal of molecular biology*, 257(2):342–358.

Lin, C.-T., Lin, K.-L., Yang, C.-H., Chung, I.-F., Huang, C.-D., and Yang, Y.-S. (2005). Protein metal binding residue prediction based on neural networks. *International journal of neural systems*, 15(01n02):71–84.

Lin, G., Simmons, G., Pöhlmann, S., Baribaud, F., Ni, H., Leslie, G. J., Haggarty, B. S., Bates, P., Weissman, D., Hoxie, J. A., et al. (2003). Differential n-linked glycosylation of human immunodeficiency virus and ebola virus envelope glycoproteins modulates interactions with dc-sign and dc-signr. *Journal of virology*, 77(2):1337–1346.

Liu, D., Tang, Y., Fan, C., Chen, Z., and Deng, L. (2016). Predrbr: Accurate prediction of rna-binding residues in proteins using gradient tree boosting. In *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*, pages 47–52. IEEE.

Liu, R. and Hu, J. (2011). Computational prediction of heme-binding residues by exploiting residue interaction network. *PLoS One*, 6(10):e25560.

Liu, R. and Hu, J. (2013). Dnabind: A hybrid algorithm for structure-based prediction of dna-binding residues by combining machine learning-and template-based approaches. *Proteins: Structure, Function, and Bioinformatics*, 81(11):1885–1899.

Liu, Z., Wang, Y., Zhou, C., Xue, Y., Zhao, W., and Liu, H. (2014). Computationally characterizing and comprehensive analysis of zinc-binding sites in proteins. *Biochimica et Biophysica*

Acta (BBA)-Proteins and Proteomics, 1844(1):171–180.

Lo, Y.-T., Wang, H.-W., Pai, T.-W., Tzou, W.-S., Hsu, H.-H., and Chang, H.-T. (2013). Protein-ligand binding region prediction (plb-save) based on geometric features and cuda acceleration. *BMC bioinformatics*, 14(4):S4.

Lu, C.-H., Lin, Y.-F., Lin, J.-J., and Yu, C.-S. (2012). Prediction of metal ion–binding sites in proteins using the fragment transformation method. *PloS one*, 7(6):e39252.

Lu, C.-H., Yu, C.-S., Chien, Y.-T., and Huang, S.-W. (2014). Exia2: web server of accurate and rapid protein catalytic residue prediction. *BioMed research international*, 2014.

Lu, C.-H., Yu, C.-S., Lin, Y.-F., and Chen, J.-Y. (2015). Predicting flavin and nicotinamide adenine dinucleotide-binding sites in proteins using the fragment transformation method. *BioMed research international*, 2015.

Ma, X., Guo, J., and Sun, X. (2016). Dnabp: Identification of dna-binding proteins based on feature selection using a random forest and predicting binding residues. *PloS one*, 11(12):e0167345.

Malik, A. and Ahmad, S. (2007). Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. *BMC Structural Biology*, 7(1):1.

Mamidi, A. S. and Surolia, A. (2015). Hierarchical sampling for metastable conformers determines biomolecular recognition: the case of malectin and diglucosylated n-glycan interactions. *Journal of Biomolecular Structure and Dynamics*, 33(6):1363–1384.

Mannervik, B., Runarsdottir, A., and Kurtovic, S. (2009). Multi-substrate–activity space and quasi-species in enzyme evolution: Ohno's dilemma, promiscuity and functional orthogonality.

McHugh, C. A., Russell, P., and Guttman, M. (2014). Methods for comprehensive experimental identification of rna-protein interactions. *Genome biology*, 15(1):203.

Meroz, Y. and Horn, D. (2008). Biological roles of specific peptides in enzymes. *Proteins: Structure, Function, and Bioinformatics*, 72(2):606–612.

Miao, Z. and Westhof, E. (2015). A large-scale assessment of nucleic acids binding site prediction programs. *PloS Comput Biol*, 11(12):e1004639.

Miao, Z. and Westhof, E. (2016). Rbscore&nbench: a high-level web server for nucleic acid binding residues prediction with a large-scale benchmarking database. *Nucleic acids research*, 44(W1):W562–W567.

Ming, D., Han, M., and An, X. (2017). Sequence-based prediction of function site and protein-ligand interaction by a functionally annotated domain profile database. *arXiv preprint arXiv:1701.08086*.

Modic, M., Ule, J., and Sibley, C. R. (2013). Cliping the brain: studies of protein–rna interactions important for neurodegenerative disorders. *Molecular and Cellular Neuroscience*, 56:429–435.

Mohabatkar, H., Beigi, M. M., and Esmaeili, A. (2011). Prediction of gaba a receptor proteins using the concept of chou's pseudo-amino acid composition and support vector machine. *Journal of Theoretical Biology*, 281(1):18–23.

Morris, J. H., Knudsen, G. M., Verschueren, E., Johnson, J. R., Cimermancic, P., Greninger, A. L., and Pico, A. R. (2014). Affinity purification–mass spectrometry and network analysis to understand protein-protein interactions. *Nature protocols*, 9(11):2539–2554.

Murakami, Y. and Mizuguchi, K. (2010). Applying the naive bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. *Bioinformatics*, 26(15):1841–1848.

Nagano, N. (2005). Ezcatdb: the enzyme catalytic-mechanism database. *Nucleic acids research*, 33(suppl 1):D407–D412.

Nagao, C., Nagano, N., and Mizuguchi, K. (2014). Prediction of detailed enzyme functions and identification of specificity determining residues by random forests. *PloS one*, 9(1):e84623.

Nagl, S. B., Freeman, J., and Smith, T. F. (1999). Evolutionary constraint networks in ligand-binding domains: an information-theoretic approach. In *Pacific Symposium on Biocomputing*, volume 4, pages 90–101.

Najmanovich, R. J. (2017). Evolutionary studies of ligand binding sites in proteins. *Current Opinion in Structural Biology*, 45:85–90.

Nassif, H., Al-Ali, H., Khuri, S., and Keirouz, W. (2009). Prediction of protein-glucose binding sites using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 77(1):121–132.

Nelson, D. L., Lehninger, A. L., and Cox, M. M. (2008). *Lehninger principles of biochemistry*. Macmillan.

Nugent, T. and Jones, D. T. (2009). Transmembrane protein topology prediction using support vector machines. *BMC bioinformatics*, 10(1):159.

Ofran, Y., Mysore, V., and Rost, B. (2007). Prediction of dna-binding residues from sequence. *Bioinformatics*, 23(13):i347–i353.

Ou, Y. (2005). Quickrbf: a package for efficient radial basis function networks. *Software available at http://csie. org/˜yien/quickrbf.*

Ou, Y.-Y. (2012). Predicting protein metal binding sites with rbf networks based on pssm profiles and additional properties. *Current Bioinformatics*, 7(2):180–186.

Ou, Y.-Y. and Chen, S.-A. (2009). Using efficient rbf networks to classify transport proteins based on pssm profiles and biochemical properties. In *International Work-Conference on Artificial Neural Networks*, pages 869–876. Springer.

Ou, Y.-Y. et al. (2016). Incorporating efficient radial basis function networks and significant amino acid pairs for predicting gtp binding sites in transport proteins. *BMC Bioinformatics*, 17(19):183.

Pabo, C. O. and Sauer, R. T. (1984). Protein-dna recognition. *Annual review of biochemistry*, 53(1):293–321.

Pai, P. P., Dash, T., and Mondal, S. (2017). Sequence-based discrimination of protein-rna interacting residues using a probabilistic approach. *Journal of Theoretical Biology*, 418:77–83.

Pai, P. P. and Mondal, S. (2016). Mowgli: prediction of protein–mannose interacting residues with ensemble classifiers using evolutionary information. *Journal of Biomolecular Structure* *and Dynamics*, 34(10):2069–2083.

Pai, P. P. and Mondal, S. (2017). Applying knowledge of enzyme biochemistry in prediction of functional sites for aiding drug discovery. *Current Topics in Medicinal Chemistry*, 17:Epub Ahead of Print.

Pai, P. P., Ranjani, S. S., and Mondal, S. (2015). Pingu: Prediction of enzyme catalytic residues using sequence information. *PloS one*, 10(8):e0135122.

Panjkovich, A. and Daura, X. (2014). Pars: a web server for the prediction of protein allosteric and regulatory sites. *Bioinformatics*, 30(9):1314–1315.

Panwar, B., Gupta, S., and Raghava, G. P. (2013). Prediction of vitamin interacting residues in a vitamin binding protein using evolutionary information. *BMC bioinformatics*, 14(1):44.

Park, B., Kim, H., and Han, K. (2014). Dbbp: database of binding pairs in protein-nucleic acid interactions. *BMC bioinformatics*, 15(15):S5.

Pegg, S. C.-H., Brown, S. D., Ojha, S., Seffernick, J., Meng, E. C., Morris, J. H., Chang, P. J., Huang, C. C., Ferrin, T. E., and Babbitt, P. C. (2006). Leveraging enzyme structure- function relationships for functional inference and experimental design: the structure- function linkage database. *Biochemistry*, 45(8):2545–2555.

Perona, J. J. and Craik, C. S. (1997). Evolutionary divergence of substrate specificity within the chymotrypsin-like serine protease fold. *Journal of Biological Chemistry*, 272(48):29987–29990.

Perozzo, R., Folkers, G., and Scapozza, L. (2004). Thermodynamics of protein–ligand interactions: history, presence, and future aspects. *Journal of Receptors and Signal Transduction*, 24(1-2):1–52.

Petrova, N. V. and Wu, C. H. (2006). Prediction of catalytic residues using support vector machine with selected protein sequence and structural properties. *BMC bioinformatics*, 7(1):312.

Qiu, Z. and Wang, X. (2011). Improved prediction of protein ligand-binding sites using random forests. *Protein and peptide letters*, 18(12):1212–1218.

Raz, A. and Nakahara, S. (2008). Biological modulation by lectins and their ligands in tumor progression and metastasis. *Anti-Cancer Agents in Medicinal Chemistry (Formerly Current Medicinal Chemistry-Anti-Cancer Agents)*, 8(1):22–36.

Rennie, W., Liu, C., Carmack, C. S., Wolenc, A., Kanoria, S., Lu, J., Long, D., and Ding, Y. (2014). Starmir: a web server for prediction of microrna binding sites. *Nucleic acids research*, 42(W1):W114–W118.

Roche, D. B., Brackenridge, D. A., and McGuffin, L. J. (2015). Proteins and their interacting partners: an introduction to protein–ligand binding site prediction methods. *International journal of molecular sciences*, 16(12):29829–29842.

Roche, D. B., Buenavista, M. T., and McGuffin, L. J. (2013). The funfold2 server for the prediction of protein–ligand interactions. *Nucleic acids research*, 41(W1):W303–W307.

Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., et al. (2014). A proteome-scale map of the human interactome network. *Cell*, 159(5):1212–1226.

Ross, C. A. and Tabrizi, S. J. (2011). Huntington's disease: from molecular pathogenesis to clinical treatment. *The Lancet Neurology*, 10(1):83–98.

Rost, B. (2002). Enzyme function less conserved than anticipated. *Journal of molecular biology*, 318(2):595–608.

Roy, A. and Zhang, Y. (2012). Recognizing protein-ligand binding sites by global structural alignment and local geometry refinement. *Structure*, 20(6):987–997.

Sanders, R. W., Venturi, M., Schiffner, L., Kalyanaraman, R., Katinger, H., Lloyd, K. O., Kwong, P. D., and Moore, J. P. (2002). The mannose-dependent epitope for neutralizing antibody 2g12 on human immunodeficiency virus type 1 glycoprotein gp120. *Journal of virology*, 76(14):7293–7305.

Sankararaman, S., Sha, F., Kirsch, J. F., Jordan, M. I., and Sjölander, K. (2010). Active site prediction using evolutionary and structural information. *Bioinformatics*, 26(5):617–624.

Schiavo, V. L., Robert, P., Limozin, L., and Bongrand, P. (2012). Quantitative modeling assesses the contribution of bond strengthening, rebinding and force sharing to the avidity of biomolecule interactions. *PLoS One*, 7(9):e44070.

Scholkopf, B. and Burges, C. J. (1999). *Advances in kernel methods: support vector learning*. MIT press.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004). Brenda, the enzyme database: updates and major new developments. *Nucleic acids research*, 32(suppl 1):D431–D433.

Si, J., Cui, J., Cheng, J., and Wu, R. (2015a). Computational prediction of rna-binding proteins and binding sites. *International journal of molecular sciences*, 16(11):26303–26317.

Si, J., Zhang, Z., Lin, B., Schroeder, M., and Huang, B. (2011). Metadbsite: a meta approach to improve protein dna-binding sites prediction. *BMC systems biology*, 5(1):S7.

Si, J., Zhao, R., and Wu, R. (2015b). An overview of the prediction of protein dna-binding sites. *International journal of molecular sciences*, 16(3):5194–5215.

Singh, H., Srivastava, H. K., and Raghava, G. P. (2016). A web server for analysis, comparison and prediction of protein ligand binding sites. *Biology direct*, 11(1):14.

Sneha, P. and Doss, C. (2017). Elucidating the mutational landscape in hepatocyte nuclear factor 1β (hnf1b) by computational approach. *Advances in Protein Chemistry and Structural Biology*.

Soskine, M. and Tawfik, D. S. (2010). Mutational effects and the evolution of new protein functions. *Nature Reviews Genetics*, 11(8):572–582.

Srinivasan, N., Rao, V., Vijayan, M., Yathindra, N., and Kolaskar, A. (1999). Structural features of protein–carbohydrate interactions in galactose and mannose binding proteins complexes. *Perspectives in structural biology*, pages 355–366.

Studer, R. A., Dessailly, B. H., and Orengo, C. A. (2013). Residue mutations and their impact on protein structure and function: detecting beneficial and pathogenic changes. *Biochemical Journal*, 449(3):581–594.

Sujatha, M. S., Sasidhar, Y. U., and Balaji, P. V. (2004). Energetics of galactose–and glucose–aromatic amino acid interactions: Implications for binding in galactose-specific proteins. *Protein Science*, 13(9):2502–2514.

Sun, J., Wang, J., Xiong, D., Hu, J., and Liu, R. (2016). Crhunter: integrating multifaceted information to predict catalytic residues in enzymes. *Scientific Reports*, 6.

Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288.

Tabita, F. R., Hanson, T. E., Li, H., Satagopan, S., Singh, J., and Chan, S. (2007). Function, structure, and evolution of the rubisco-like proteins and their rubisco homologs. *Microbiology and Molecular Biology Reviews*, 71(4):576–599.

Taroni, C., Jones, S., and Thornton, J. M. (2000). Analysis and prediction of carbohydrate binding sites. *Protein Engineering*, 13(2):89–98.

Tawfik, D. S. (2010). Messy biology and the origins of evolutionary innovations. *Nature chemical biology*, 6(10):692.

Teppa, E., Wilkins, A. D., Nielsen, M., and Buslje, C. M. (2012). Disentangling evolutionary signals: conservation, specificity determining positions and coevolution. implication for catalytic residue prediction. *BMC bioinformatics*, 13(1):235.

Theimer, C. A., Smith, N. L., and Khanna, M. (2012). Nmr studies of protein–rna interactions. *Protein NMR Techniques*, pages 197–218.

Todd, A. E., Orengo, C. A., and Thornton, J. M. (1999). Evolution of protein function, from a structural perspective. *Current opinion in chemical biology*, 3(5):548–556.

Tsai, K.-C., Jian, J.-W., Yang, E.-W., Hsu, P.-C., Peng, H.-P., Chen, C.-T., Chen, J.-B., Chang, J.-Y., Hsu, W.-L., and Yang, A.-S. (2012). Prediction of carbohydrate binding sites on protein surfaces with 3-dimensional probability density distributions of interacting atoms. *PloS one*, 7(7):e40846.

Tsujikawa, H., Sato, K., Wei, C., Saad, G., Sumikoshi, K., Nakamura, S., Terada, T., and Shimizu, K. (2016). Development of a protein–ligand-binding site prediction method based on interaction energy and sequence conservation. *Journal of Structural and Functional Genomics*, 17(2-3):39–49.

Turner, D., Kim, R., and Guo, J.-t. (2012). Tfindit: transcription factor-dna interaction data depository. *BMC bioinformatics*, 13(1):220.

Tuvshinjargal, N., Lee, W., Park, B., and Han, K. (2016). Pridictor: Protein–rna interaction predictor. *BioSystems*, 139:17–22.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Vinik, A., Feliberti, E., and Perry, R. R. (2014). Carcinoid tumors.

Virgens, M. Y. F., Pol-Fachin, L., Verli, H., and Saraiva-Pereira, M. L. (2014). Effects of glycosylation and ph conditions in the dynamics of human arylsulfatase a. *Journal of Biomolecular Structure and Dynamics*, 32(4):567–579.

Visser, T. J. (1988). Metabolism of thyroid hormone. *New Comprehensive Biochemistry*, 18:81–103.

Vliegenthart, J. (2007). Protein–carbohydrate interactions in infectious diseases. edited by carole a. bewley.

Vogel, C., Bashton, M., Kerrison, N. D., Chothia, C., and Teichmann, S. A. (2004). Structure, function and evolution of multidomain proteins. *Current opinion in structural biology*, 14(2):208–216.

Wang, K., Gao, J., Shen, S., Tuszynski, J. A., Ruan, J., and Hu, G. (2013). An accurate method for prediction of protein-ligand binding site on protein surface using svm and statistical depth function. *BioMed research international*, 2013.

Wang, L. and Brown, S. J. (2006). Bindn: a web-based tool for efficient prediction of dna and rna binding sites in amino acid sequences. *Nucleic acids research*, 34(suppl 2):W243–W248.

Wang, L., Huang, C., Yang, M. Q., and Yang, J. Y. (2010). Bindn+ for accurate prediction of dna and rna-binding residues from protein sequence features. *BMC Systems Biology*, 4(1):S3.

Wang, L., Yang, M. Q., and Yang, J. Y. (2009). Prediction of dna-binding residues from protein sequence information using random forests. *Bmc Genomics*, 10(1):S1.

Wang, W., Liu, J., Xiong, Y., Zhu, L., et al. (2014). Analysis and classification of dna-binding sites in single-stranded and double-stranded dna-binding proteins using protein information. *IET systems biology*, 8(4):176–183.

Wass, M. N., Kelley, L. A., and Sternberg, M. J. (2010). 3dligandsite: predicting ligand-binding sites using similar structures. *Nucleic acids research*, page gkq406.

Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, 36(03):307–340.

Wu, M.-Y., Dai, D.-Q., and Yan, H. (2012). Prl-dock: Protein-ligand docking based on hydrogen bond matching and probabilistic relaxation labeling. *Proteins: Structure, Function, and Bioinformatics*, 80(9):2137–2153.

Xiao, X., Hui, M.-J., Liu, Z., and Qiu, W.-R. (2015). icataly-pseaac: identification of enzymes catalytic sites using sequence evolution information with grey model gm (2, 1). *The Journal of membrane biology*, 248(6):1033–1041.

Xie, Z.-R., Liu, C.-K., Hsiao, F.-C., Yao, A., and Hwang, M.-J. (2013). Lise: a server using ligand-interacting and site-enriched protein triangles for prediction of ligand-binding sites. *Nucleic acids research*, 41(W1):W292–W296.

Xiong, D., Zeng, J., and Gong, H. (2015). Rbrident: An algorithm for improved identification of rna-binding residues in proteins from primary sequences. *Proteins: Structure, Function, and Bioinformatics*, 83(6):1068–1077.

Xiong, Y., Liu, J., Zhang, W., and Zeng, T. (2012). Prediction of heme binding residues from protein sequences with integrative sequence profiles. *Proteome science*, 10(1):S20.

Yan, J., Friedrich, S., and Kurgan, L. (2016). A comprehensive comparative review of sequence-based predictors of dna-and rna-binding residues. *Briefings in bioinformatics*, 17(1):88–105.

Yan, J. and Kurgan, L. (2017). Drnapred, fast sequence-based method that accurately predicts and discriminates dna-and rna-binding residues. *Nucleic Acids Research*, page 1.

Yang, J., Roy, A., and Zhang, Y. (2013a). Biolip: a semi-manually curated database for biologically relevant ligand–protein interactions. *Nucleic acids research*, 41(D1):D1096–D1103.

Yang, J., Roy, A., and Zhang, Y. (2013b). Protein–ligand binding site recognition using complementary binding-specific substructure comparison and sequence profile alignment. *Bioinformatics*, 29(20):2588–2595.

Yang, X., Wang, J., Sun, J., and Liu, R. (2015). Snbrfinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. *PloS one*, 10(7):e0133260.

Yang, X.-X., Deng, Z.-L., and Liu, R. (2014). Rbrdetector: Improved prediction of binding residues on rna-binding protein structures using complementary feature-and template-based strategies. *Proteins: Structure, Function, and Bioinformatics*, 82(10):2455–2471.

Yasser, E.-M., Abbas, M., Malluhi, Q., and Honavar, V. (2016). Fastrnabindr: Fast and accurate prediction of protein-rna interface residues. *PloS one*, 11(7):e0158445.

Youn, E., Peters, B., Radivojac, P., and Mooney, S. D. (2007). Evaluation of features for catalytic residue prediction in novel folds. *Protein Science*, 16(2):216–226.

Yu, D.-J., Hu, J., Huang, Y., Shen, H.-B., Qi, Y., Tang, Z.-M., and Yang, J.-Y. (2013). Targetatpsite: A template-free method for atp-binding sites prediction with residue evolution image sparse representation and classifier ensemble. *Journal of computational chemistry*, 34(11):974–985.

Yu, D.-J., Hu, J., Yan, H., Yang, X.-B., Yang, J.-Y., and Shen, H.-B. (2014). Enhancing protein-vitamin binding residues prediction by multiple heterogeneous subspace svms ensemble. *BMC bioinformatics*, 15(1):297.

Yuen, C. M. and Liu, D. R. (2007). Dissecting protein structure and function using directed evolution. *Nature methods*, 4(12):995–997.

Yugandhar, K. and Gromiha, M. M. (2017). Computational approaches for predicting binding partners, interface residues, and binding affinity of protein–protein complexes. *Prediction of Protein Secondary Structure*, pages 237–253.

Zanegina, O., Kirsanov, D., Baulin, E., Karyagina, A., Alexeevski, A., and Spirin, S. (2016). An updated version of npidb includes new classifications of dna–protein complexes and their families. *Nucleic acids research*, 44(D1):D144–D153.

Zhang, J., Chai, H., Gao, B., Yang, G., and Ma, Z. (2016). Hemespred: Structure-based ligand-specific heme binding residues prediction by using fast-adaptive ensemble learning scheme. *IEEE/ACM transactions on computational biology and bioinformatics*.

Zhang, T., Zhang, H., Chen, K., Shen, S., Ruan, J., and Kurgan, L. (2008). Accurate sequence-based prediction of catalytic residues. *Bioinformatics*, 24(20):2329–2338.

Zhang, X., Wu, D., Chen, L., Li, X., Yang, J., Fan, D., Dong, T., Liu, M., Tan, P., Xu, J., et al. (2014). Raid: a comprehensive resource for human rna-associated (rna–rna/rna–protein) interaction. *rna*, 20(7):989–993.

Zhang, Z., Li, Y., Lin, B., Schroeder, M., and Huang, B. (2011). Identification of cavities on protein surface using multiple computational approaches for drug binding site prediction. *Bioinformatics*, 27(15):2083–2088.

Zhang, Z., Tang, Y.-R., Sheng, Z.-Y., and Zhao, D. (2009). An overview of the de novo prediction of enzyme catalytic residues. *Current Bioinformatics*, 4(3):197–206.

Zhao, H., Yang, Y., and Zhou, Y. (2013a). Prediction of rna binding proteins comes of age from low resolution to high resolution. *Molecular bioSystems*, 9(10):2417–2425.

Zhao, M., Chang, H.-T., Zhou, Q., Zeng, T., Shih, C.-S., Liu, Z.-P., Chen, L., and Wei, D.-Q. (2014). Predicting protein-ligand interactions based on chemical preference features with its application to new d-amino acid oxidase inhibitor discovery. *Current pharmaceutical design*, 20(32):5202–5211.

Zhao, S., Kumar, R., Sakai, A., Vetting, M. W., Wood, B. M., Brown, S., Bonanno, J. B., Hillerich, B. S., Seidel, R. D., Babbitt, P. C., et al. (2013b). Discovery of new enzymes and metabolic pathways by using structure and genome context. *Nature*, 502(7473):698–702.

Zheng, H., Handing, K. B., Zimmerman, M. D., Shabalin, I. G., Almo, S. C., and Minor, W. (2015). X-ray crystallography over the past decade for novel drug discovery–where are we heading next? *Expert opinion on drug discovery*, 10(9):975–989.

Zhou, J., Xu, R., He, Y., Lu, Q., Wang, H., and Kong, B. (2016). Pdnasite: identification of dna-binding site from protein sequence by incorporating spatial and sequence context. *Scientific* *reports*, 6.

Zhou, X.-B., Chen, C., Li, Z.-C., and Zou, X.-Y. (2007). Using chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes. *Journal of theoretical biology*, 248(3):546–551.

Zhu, X., Ericksen, S. S., and Mitchell, J. C. (2013). Dbsi: Dna-binding site identifier. *Nucleic acids research*, page gkt617.

Zvelebil, M. J. and Sternberg, M. J. (1988). Analysis and prediction of the location of catalytic residues in enzymes. *Protein Engineering*, 2(2):127–138.

Appendix I: Amino acid representations

Amino acid	Three letter code	Single letter code
Alanine	Ala	A
Cysteine	Cys	C
Aspartic acid	Asp	D
Glutamic acid	Glu	E
Phenylalanine	Phe	F
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Lysine	Lys	K
Leucine	Leu	L
Methionine	Met	M
Asparagine	Asn	N
Proline	Pro	P
Glutamine	Gln	Q
Arginine	Arg	R
Serine	Ser	S
Threonine	Thr	T
Valine	Val	V
Tryptophan	Trp	W
Tyrosine	Tyr	Y

Appendix II: List of Publications

Related to work described in this thesis:

Pai, P. P. and Mondal, S*. (2017). Applying Knowledge of Enzyme Biochemistry in Prediction of Functional Sites for Aiding Drug Discovery. *Current Topics in Medicinal Chemistry*, 17: (In press) `doi:10.2174/1568026617666170329153858`.

Pai, P. P., Dattatreya, R.K., Mondal, S.(2017)Ensemble Architecture for Prediction of Enzyme-Ligand Binding Residues using Evolutionary Information. *Molecular Informatics*.DOI:10.1002/minf.201700021.[Epub ahead of print]

Pai, P. P., Dash, T. and Mondal, S*. (2017). Sequence-based discrimination of protein-RNA interacting residues using a probabilistic approach. *Journal of Theoretical Biology*, 418: 77-83.

Pai, P. P. and Mondal, S*. (2016). MOWGLI: prediction of protein-MannOse interacting residues With ensemble classifiers usinG evoLutionary Information. *Journal of Biomolecular Structure & Dynamics*, 34(10): 2069-2083.

Pai, P. P., Ranjani, S.S.S., Mondal S*. (2015). PINGU: PredIction of eNzyme catalytic residues usinG seqUence information. *PLoS ONE*, 10: e0135122.

Other publications:

Mondal, S* and **Pai, P. P.** (2014). Chou's pseudo amino acid composition improves sequence- based antifreeze protein prediction. *Journal of Theoretical Biology*, 356: 30-35.

Dhole, K., Singh, G., **Pai, P. P.**, Mondal S*. (2014). Sequence-based prediction of protein- protein interaction sites with L1-logreg classifier. *Journal of Theoretical Biology*, 348: 47-54.

Book chapters:

Pai, P. P. and Mondal, S*. (2016). Intriguing Cystine Knot Miniproteins in Drug Design and Therapeutics. Toxinology. Toxins and Drug Discovery, Editor-in-Chief: Gopalakrishnakone P, Publisher: Springer, `doi:10.1007/978-94-007-6726-3_25-1`.

Pai, P. P. and Mondal, S*. (2015). Computational approaches for animal toxins to aid drug discovery. Toxinology. Toxins and Drug Discovery, Editor-in-Chief: Gopalakrishnakone P, Publisher: Springer, `doi:10.1007/978-94-007-6726-3_20-1`.

* corresponding author

Appendix III: List of conferences and workshops

- **Pai, P. P.** and Mondal, S. (2016). Learning from, through and about Biomolecular Interactions in Nature. Grace Hopper India Celebrations, Bangalore, India.
- **Pai, P. P.** [Workshop] (2015). Biomolecular Interactions. National Center for Biological Sciences, Bangalore, India.
- **Pai, P. P.** and Mondal, S. (2015). Prediction of RNA-interacting proteins and sites for therapeutics and drug discovery. International Conference on Trends in Cell and Molecular Biology, Goa, India.
- **Pai, P. P.** and Mondal, S. (2015). Identification of catalytic residues in enzymes using support vector machines. 2015 NextGen Genomics, Biology, Bioinformatics and Technologies (NGBT) Conference, Hyderabad, India.
- **Pai, P. P.** and Mondal, S*. (2015). Advances in Computational Identification of Protein Interaction Sites. BIT's 6th World Gene Convention, Qingdao, China.*presenting author
- **Pai, P. P.** and Mondal, S. (2014). Computational predictability of biochemically diverse cofactors. International Conference on Biotechnology and Bioinformatics (ICBB-2014), Pune, India.

Biodata of the Candidate

Name: Priyadarshini P. Pai

ID: 2012PHXF0007G

Education: M.Sc. in Medical Biotechnology, Manipal University, 2009

Priyadarshini P. Pai is a doctoral research fellow in the Department of Biological Sciences, Birla Institute of Technology & Science-Pilani, K. K. Birla Goa Campus, working in the area of protein interaction biology under the supervision of Dr. Sukanta Mondal. She has been actively involved with studies related to protein interactions at residue levels for the development of various computational approaches using statistics and supervised machine learning techniques, in the Annotate Biomolecules Computationally group. Her study findings presented in this thesis are hoped to eventually boost protein engineering for drug-design and industrial production.

ORCID ID: orcid.org/0000-0001-8450-0085

LinkedIn: `https://in.linkedin.com/in/priyadarshini-pai-8ab15926`

ResearchGate: `https://www.researchgate.net/profile/Priyadarshini_Pai`

Professional Memberships:

- Association for Computing Machinery (ACM), December 2016-17.

- Bioinformatics Experts and Troubleshooters (BET), RSG India - ISCB Student Council, 2016 onward.

Awards/Honors/Accomplishments:

- The FEBS Journal Poster Prize, 2015 NextGen Genomics, Biology, Bioinformatics and Technologies (NGBT) Conference, Hyderabad, India, October 2015. `http://febs.onlinelibrary.wiley.com/hub/journal/10.1111/(ISSN)1742-4658/features/the-febs-journal-poster-prize.html`
- Key scientific article by Global Medical Discovery (GMD) contributing to excellence in biomedical research, October 2014. `https://globalmedicaldiscovery.com/key-scientific-articles/`

Biodata of the Supervisor

Sukanta Mondal, Ph.D.

Assistant Professor, Department of Biological Sciences

Birla Institute of Technology & Science - Pilani

K. K. Birla Goa Campus

Profile: `http://universe.bits-pilani.ac.in/goa/suku/profile`

Dr. Mondal works in the field of computational biology and bioinformatics to address various challenging questions in bio-molecular science. He obtained his Ph.D. degree from Indian Institute of Science in 2007, under the supervision of Prof. Ramakumar S, for thesis work titled "Contributions to venominformatics: sequence-structure-function studies of toxins from marine cone snails. Application of order-statistics filters for detecting membrane-spanning helices". After his doctoral studies, Dr. Mondal moved to National Institutes of Biomedical Innovation, Health and Nutrition (NIBIOHN), Japan, for pursuing post-doctoral research on "Development of an international pharmaceutical innovation value chain for *in silico* drug discovery" under the supervision of Prof. Kenji Mizuguchi. Gathering rich experience at both national and international levels, he initiated his Annotate Biomolecules Computationally (ABC) group in the year 2012, which currently has various doctoral, graduate and undergraduate students working on protein functional annotation.